System Design Strategies

26th Edition

An ESRI ® Technical Reference Document • August 2009

Prepared by:

Dave Peters
Systems Integration

Environmental Systems Research Institute, Inc.
380 New York Street
Redlands, California 92373-8100

Document History

Revision History

Date	Description	Author
June 1993	Technology Update	Dave Peters
March 1994	Technology Update	Dave Peters
April 1995	Technology Update	Dave Peters
May 1996	Technology Update	Dave Peters
June 1997	Technology Update	Dave Peters
June 1998	Technology Update	Dave Peters
July 1999	Technology Update	Dave Peters
February 2000	Technology Update	Dave Peters
June 2000	Technology Update	Dave Peters
February 2001	Technology Update	Dave Peters
June 2001	Technology Update	Dave Peters
May 2002	Technology Update	Dave Peters
July 2002	Technology Update	Dave Peters

Date	Description	Author
February 2003	Technology Update	Dave Peters
June 2003	Technology Update	Dave Peters
March 2004	Technology Update	Dave Peters
July 2004	Technology Update	Dave Peters
April 2005	Technology Update	Dave Peters
August 2005	Technology Update	Dave Peters
March 2006	Technology Update	Dave Peters
October 2006	Technology Update	Dave Peters
January 2007	Technology Update	Dave Peters
July 2007	Technology Update	Dave Peters
August 2008	Technology Update	Dave Peters
August 2009	Technology Update	Dave Peters

August 2009 Change Highlights

1. Editorial text and graphic updates throughout the document
2. Figure 1-3, update GIS Deployment Stages with more current slide
3. Figure 1-6, Included Capacity Planning Tool and ESRI Press *Building a GIS* Release
4. Figure 2-1, expanded history to include 2009 highlights
5. Figure 2-8, update to address new ArcGIS Desktop software options
6. Figure 2-9, update to address new ArcGIS Server software options
7. Figure 2-10, update to include new online resource center options
8. Figure 3-2, update to include wireless communication options
9. Figure 3-9, update to expand wait time discussion
10. Figure 3-12, update to include 10 Gbps LAN network
11. Figure 4-2, update to include Imagery and map cache data sources
12. Figure 4-4, modified to remove ArcIMS
13. Figure 4-5, update to include access to ArcGIS Online
14. Figure 4-8, update to remove ArcIMS and include ArcGIS Online services
15. Chapter text modified to remove ArcIMS architecture discussion
16. Figure 4-9, update server component tier names and include REST and OGC APIs
17. Section 4.3.3, update to reflect Image Server integration with ArcGIS Server software
18. Figure 7-1, update to demonstrate both light and medium performance targets
19. Figure 7-8, new chart showing sensitivity of transaction queue time to number of platform core
20. Add new section 7.2.5 on Performance Processing Delays
21. Figure 7-12, update to show total of 10 platform tier in the current Capacity Planning Tool
22. Figure 7-13, update User Productivity chart and discuss minimum think times
23. Figure 8-3, move forward closer to the display complexity discussion
24. Figure 8-7, update the ArcGIS Standard ESRI Workflow sample with current Capacity Planning tool
25. Figure 8-8, provide more conceptual discussion on reducing map display layers
26. Figure 8-9, introduce MXDperfstat performance measurement tool
27. Figure 8-10, introduce ArcGIS 9.3.1 ArcMap display optimization tools
28. Figure 8-11, introduce Capacity Planning Tool use of MXDperfstat performance measurements
29. Figure 9-1, update to include ArcInfo 2009 baseline platform
30. Figure 9-2, update to include Arc09 baseline
31. Figure 9-3, update to include 2009 map display performance
32. Figure 9-9, update chart to show performance history of Intel platform technology
33. Figure 9-10, add chart to show performance history of AMD platform technology
34. Figure 9-11, update chart to show performance history of UNIX platform technology

35. Figure 9-12, update chart to show current hardware selection challenges
36. Figure 9-13, update chart to show hardware selection performance variations
37. Figure 9-14, update chart to show current recommended GIS Workstation performance considerations
38. Figure 9-16, update chart to show current ArcGIS Desktop workflow service times
39. Figure 9-17, update chart to address current Windows Terminal Server platform sizing
40. Figure 9-18, update chart to show current ArcSDE DBMS workflow service times
41. Figure 9-19, update chart to address current ArcSDE Geodatabase windows server sizing (up to 8 core)
42. Figure 9-20, update chart to address current ArcSDE Geodatabase UNIX server sizing (up to 8 core)
43. Figure 9-21, update chart to address current ArcSDE Geodatabase large capacity platforms
44. Figure 9-30, update chart to show current ArcGIS Desktop workflow performance summary
45. Figure 9-31, update chart to show current ArcGIS Server two-tier workflow service times
46. Figure 9-32, update chart to address current ArcGIS Server two-tier platform sizing
47. Figure 9-33, update chart to show current ArcGIS Server three-tier workflow service times
48. Figure 9-34, update chart to address current ArcGIS Server three-tier platform sizing
49. Figure 9-35, update chart to address current Web Server platform sizing
50. Figure 9-36, update chart to address current ArcGIS Server Image Extension platform sizing
51. Figure 9-37, update chart to show current ArcGIS Server workflow performance summary
52. Chapter 10, updated to reflect the August 2009 Capacity Planning Tool.
53. Chapter 11, updated to reflect the August 2009 Capacity Planning Tool and the 2009 platform baseline
54. Section 12.3.1, new section on Virtual Desktop and Server technology
55. Figure 12-11, new graphics on common performance monitoring tools
56. Figure 12-12, update to introduce new CPT Test tab tools

ESRI released a new GIS Capacity Planning Forums for the purpose of building a community of users interested in capacity planning and modeling system performance and scalability. ESRI user forums are available on the ESRI Support Center (http://support.esri.com/index.cfm?fa=forums.gateway).

ESRI Press released *Building a GIS, System Architecture Design Strategies for Managers* in August this year. This book represents a considerable amount of effort to document and share our experience helping ESRI customers design, implement, and manage successful GIS operations. The *System Design Strategies TRD* provides an overview for reference by our consulting services, while *Building a GIS* shares the rest of the story. *Building a GIS* also includes a copy of the Capacity Planning Tool introduced by ESRI in 2006 as a framework for performance and capacity planning and a tool for managing implementation success. *Building a GIS* is available through ESRI Press at www.esri.com/esripress/buildingagis or through Amazon.com. Updates to the Capacity Planning Tool are published on the *Building a GIS* Online Resource Center.

Table of Contents

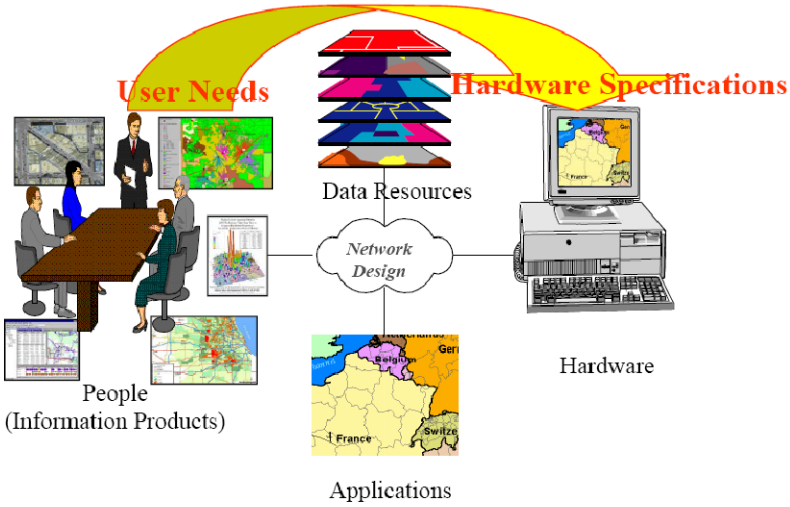
1 System Design Process

The purpose of this document is to share a system design methodology that promotes successful deployment of geographic information system (GIS) technology. Guidelines include appropriate rationale and logic to deploy and support a system that will satisfy initial performance needs for most of ESRI's customers. Once the initial implementation is operational, the system environment can be further tuned and adjusted to fit specific customer requirements.

1.1 What Is System Architecture Design?

System architecture design is a process developed by ESRI to promote successful GIS implementations. This process supports existing infrastructure requirements and provides specific recommendations for hardware and network solutions based on existing and projected user needs. Application requirements, data resources, and people within an organization are all important in determining the optimum hardware solution as shown in figure 1-1. The ESRI system architecture design process provides specific deployment strategies and associated hardware specifications based on identified operational workflow requirements.

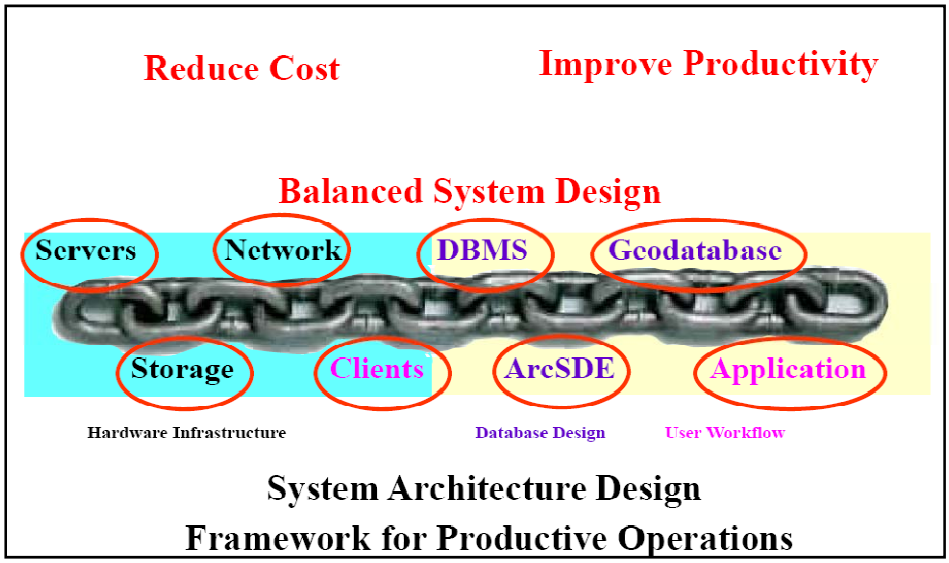
Figure 1-1
System Architecture Design Process



1.2 Why Is System Architecture Design Important?

A distributed computer environment must be designed properly to support user performance requirements. The weakest "link" in the system will limit performance. The system architecture design process develops specifications for a balanced hardware solution. Investment in hardware and network components based on a balanced system load model provides the highest possible system performance at the lowest overall cost as shown in figure 1-2.

Figure 1-2
Why Is System Architecture Design Important?

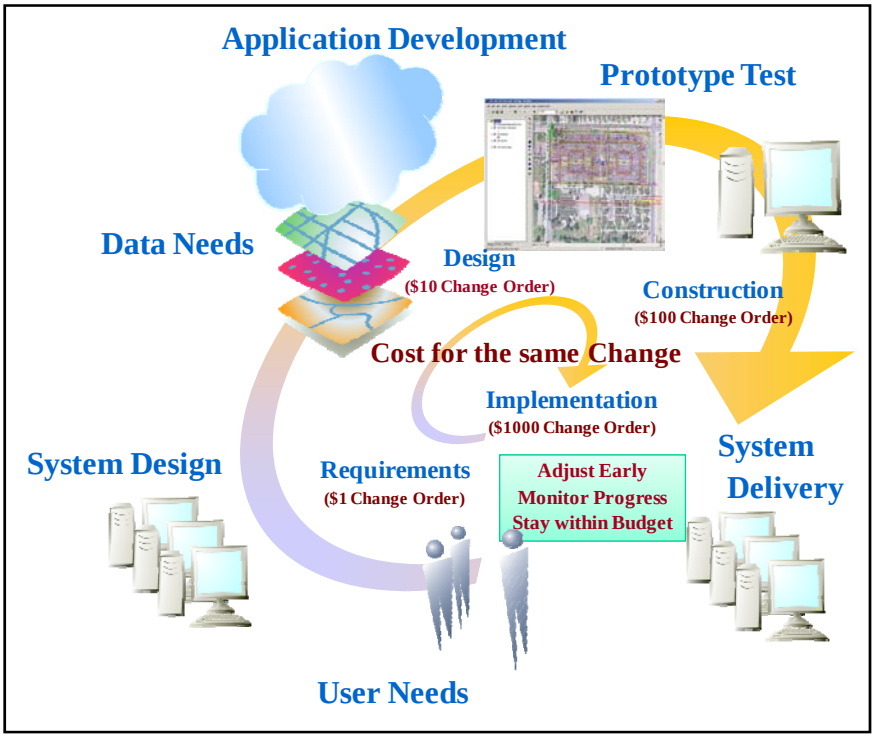


System architecture design provides a framework for supporting the implementation of a successful enterprise GIS. User workflows must be designed to optimize interactive client productivity and efficiently manage heavier geoprocessing loads. The geodatabase design and database selection should be optimized to support performance requirements. The selected system platform components (servers, client workstations, storage systems) must perform adequately and have the capacity to support peak user workflow requirements. The system architecture design strategy must address performance needs and bandwidth constraints over distributed communication networks—technology and configuration must be selected to conserve shared infrastructure resources. System architecture design provides a solid foundation for building a productive operational environment.

1.3 What Does It Take to Support Successful GIS Operations?

There are several critical deployment stages that support a successful implementation. Understanding the importance of each stage and the key objectives that support success leads to more effective enterprise implementations. Figure 1-3 shows the different stages of GIS deployment.

Figure 1-3
GIS Deployment Stages



Requirements Stage: Understanding GIS technology alternatives, quantifying GIS user requirements, and establishing an appropriate system architecture design deployment strategy are critical during the requirements stage. Capacity planning during this phase can establish preliminary software performance specifications. This is a planning stage where "getting it right" can save considerable effort and money throughout the implementation.

Design Stage: System development and prototype functional testing build the components and confidence to support the follow-on deployment. This is where time and money are invested to build and test the selected environment. Initial prototype testing demonstrates functionality and reduces implementation risk. Preliminary software performance testing can validate initial capacity planning assumptions.

Construction Stage: A successful initial system deployment can set the stage for success. This is where the solution comes together and final operational support needs are validated. This is an important time to demonstrate performance and the capacity of the deployed system and validate that the selected hardware and infrastructure will support full production deployment.

Implementation Stage: Final system procurement and deployment demonstrate operational success. Capacity planning metrics can be used to monitor and maintain system performance objectives. Good planning, development, and testing will support a smooth deployment, productive operations, and satisfied users.

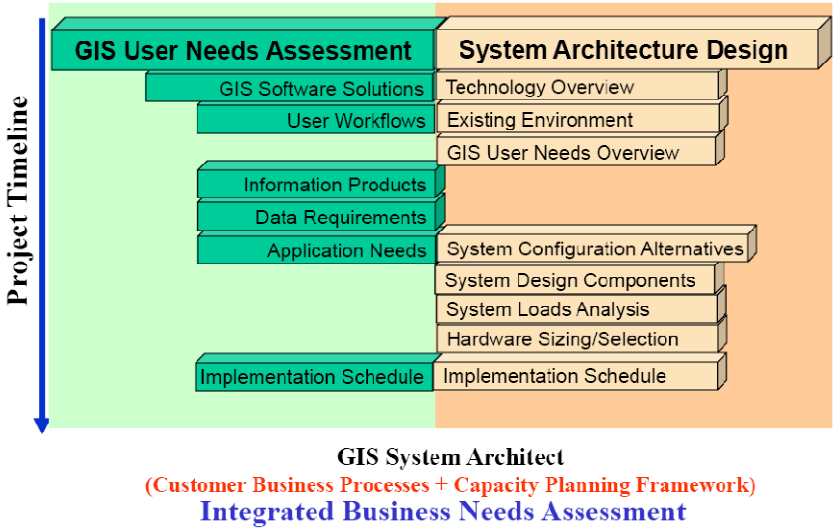
Getting it right from the start is best done by taking the time to understand the technology, quantify user requirements, select the right software technology, and deploy the right hardware. Not getting it right from the start will cost money to make it right later. The cost of change increases exponentially as the project implementation proceeds.

Establishing workflow performance target milestones and managing software performance to achieve performance goals throughout deployment is a positive recipe for success. Building systems without regard to performance and scalability can lead to disappointing results and costly recovery.

1.4 System Design Process

The traditional system design process includes a GIS needs assessment and a system architecture design. The system architecture design is based on user workflow requirements identified in the GIS needs assessment. The most effective system design approach considers user needs and system architecture constraints throughout the design process. Figure 1-4 provides an overview of the system design process.

Figure 1-4
System Design Process



GIS Needs Assessment: The GIS needs assessment includes a review of user workflow requirements and identifies where GIS applications can improve user productivity. This assessment identifies GIS application and data requirements and an implementation strategy for supporting GIS user needs. The user requirements analysis is a process that must be accomplished by the user organization. A GIS professional consultant familiar with current GIS solutions and customer business practices can help facilitate this planning effort.

System Architecture Design: The system architecture design is based on user requirements identified by the GIS needs assessment. Customers must have a clear understanding of their GIS application and data requirements before they are ready to develop system design specifications. System implementation strategies should identify hardware purchase requirements "just in time" to support user deployment needs.

The system design begins with technology exchange. Technology exchange provides the foundation for client support throughout the design process. Client participation is a key ingredient in the design process. The design process includes a review of the existing computer environment, GIS user requirements, and system design alternatives. The system design capacity planning tools provided by ESRI translate projected peak user workflow requirements to specific platform specifications. An integrated implementation strategy is developed to support GIS deployment milestones.

Traditionally, the user needs assessment and the system architecture design were two separate efforts. There are some key advantages in completing these efforts together. GIS Software solutions should include a discussion of architecture options and system deployment strategies for each technology option. The existing hardware environment and information on peak user workflows and user locations can be identified during the user workflow interviews. Technology selection should consider configuration options, required platforms, peak system loads for each technology option, and overall system design costs. And finally the system implementation schedule must consider delivery milestones. A primary goal of developing the new capacity

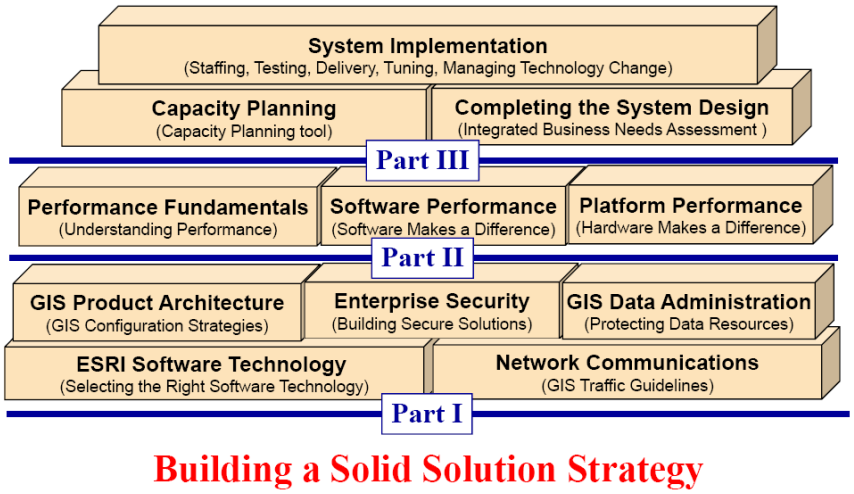
planning tools presented later in this document is to automate the system architecture design analysis in such a way that GIS professional consultants will be able to use the capacity planning tools to support an integrated business needs assessment.

1.5 Supporting Technology

Distributed GIS solutions include integration of a variety of vendor products. Each vendor technology is a part of the total enterprise environment that must be integrated together as a whole. Integration of this multivendor environment is made possible through voluntary compliance with generally accepted industry interface standards.

A general understanding of the fundamental supporting technologies associated with a GIS enterprise solution provides a foundation for supporting the system architecture design process. Figure 1-5 identifies the key technology building blocks supporting a distributed GIS environment.

Figure 1-5
Supporting Technology



GIS Software Technology: A variety of GIS software was developed by ESRI over the past 38 years. Each product solution was developed to support specific user requirements. Section 2 describes how these solutions fit together to support the growing needs of the GIS user community.

Network Communications: Network communications provide the connectivity for powerful GIS enterprise solutions. A fundamental understanding of network communications helps developers build better applications and establish efficient user access to shared data resources. Section 3 provides an overview of network communication concepts and identifies GIS design standards for successful distributed GIS environments.

GIS Product Architecture: An enterprise GIS solution includes a variety of ESRI and third-party vendor products. These products must interface with one another to support an integrated GIS solution. Section 4 provides an overview of the system architecture components required to support distributed GIS operations, Section 5 addresses enterprise security, and Section 6 addresses geodatabase administration.

System Design Sizing Models: The system design models translate peak capacity hardware loads developed during the user workflow analysis to platform specifications to support operational requirements. Section 7 discusses the technical assumptions supporting the ESRI sizing models, identifying fundamental performance relationships between the distributed computing platform components supporting the GIS hardware solution. Capacity planning tools are introduced to record target performance metrics and model system performance relationships. Section 8 identifies how to apply these sizing models and use the capacity planning tools in the real world of rapidly changing hardware technologies. The capacity planning tools incorporate vendor-published relative performance benchmarks with the performance sizing models developed in section 7 to identify required hardware specifications.

User Needs and Configuration Strategy: The system architecture design process starts with a customer user needs assessment. The results of the GIS user needs assessment establish a foundation for the system architecture design analysis. The design process provides a methodology to translate peak user workflow requirements to an appropriate GIS configuration architecture. Peak capacity needs are translated to peak hardware and network loads for design analysis. Section 6 data administration provides an overview of the user needs assessment and analysis required to develop an appropriate distributed GIS architecture design strategy.

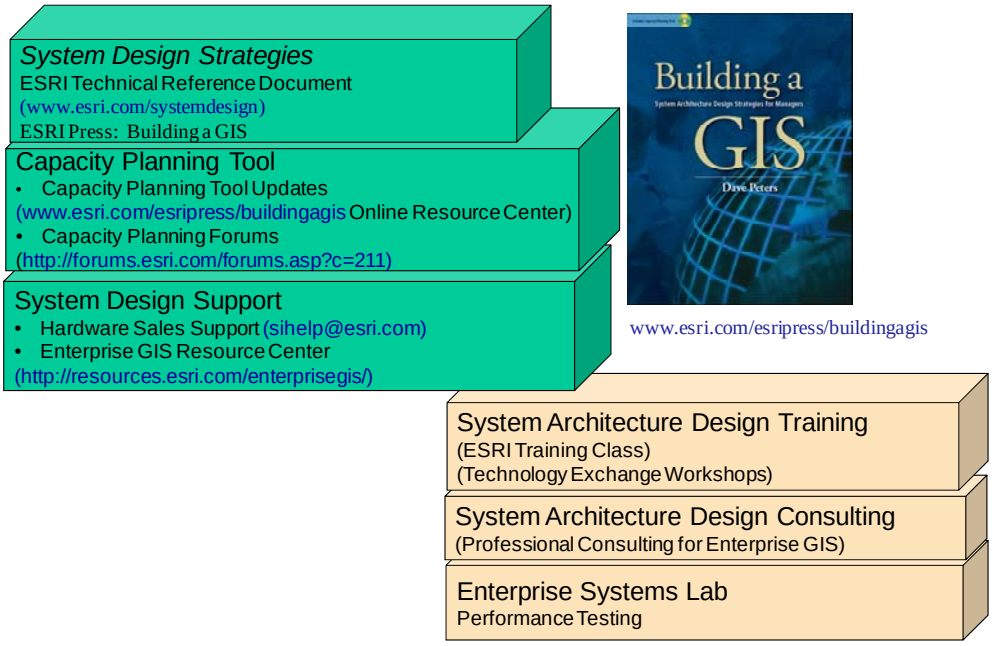
Implementation Strategies. An appropriate implementation strategy sets the framework for successful deployment. Proper staffing and management sponsorship are critical for developing long-term support for enterprise GIS operations. Project management and well-defined deployment strategy establish achievable expectations and set critical milestones to ensure success. An effective deployment strategy supports early problem identification and reduces implementation risk. Knowing what and when to test can save money and effectively promote successful deployment. System deployment schedules and implementation planning can facilitate timely delivery. System tuning and selected performance validation testing can ensure productive operations during peak performance loads.

Technology is changing very rapidly. Most enterprise GIS deployments evolve over years of commitment, planning, and hard work. It is essential in today's world to plan for technology change and update these plans on an annual basis. GIS projects should be planned to complete within an annual business cycle. Enterprise GIS is an evolving program that changes each year to support business objectives and keep pace with technology.

1.6 System Design Support Efforts

The ESRI systems integration team was established in November 1990 with the charter to promote successful GIS implementations. Several initiatives were developed over the years to simplify the design of successful GIS environments and reduce implementation risk. Figure 1-6 shows services maintained to support the needs of the ESRI user community and promote effective enterprise GIS solutions.

Figure 1-6
System Design Support Efforts



System Design Strategies Technical Reference Document: The System Design Strategies technical reference document defines the technical issues associated with the design of an effective enterprise GIS solution and includes specific sizing guidelines to support hardware selection. ESRI customers, hardware consultants, and GIS technical staff can use this document as a guide for designing distributed GIS solutions. The guidelines in

this document are also useful in understanding existing hardware-related performance issues and identifying proper solutions. The document is updated twice each year and is accessible over the Internet at http://www.esri.com/systemdesign.

Capacity Planning Tool: The Capacity Planning Tool is an application developed in Microsoft Excel providing an easy to use framework to collect user requirements and complete the system design. The Capacity Planning Tool is included with the ESRI Press *Building a GIS* book and through System Architecture Design Strategies training and consulting services. Updates to the Capacity Planning Tool are published on the Building a GIS online resource center.

Capacity Planning Forums: A series of GIS Capacity Planning forums are provided on the ESRI User Forums Web site. User forums provide a special opportunity to establish a user community that can share new ideas and best practices. New updates to the capacity planning tool are announced on the forum pages.

System Design Support: The ESRI Implementation Services Department provides online hardware marketing support through the sihelp@esri.com e-mail alias. Simple questions related to hardware selection and sizing requirements can be addressed through this site. A new Enterprise GIS Resource Center is maintained on http://resources.esri.com/gateway/index.cfm to share system design best practices, test tools, and test benchmarks demonstrating performance of a variety of ESRI workflows.

System Architecture Design Training: A three-day System Architecture Design Strategies training class is available through the ESRI Learning Center. This class provides an in-depth review of the material presented in the System Design Strategies technical reference document. A class brochure is included as Attachment B. A one-day System Architecture Design workshop is held the weekend prior to the annual ESRI User Conference providing a technical overview of the material presented in the training class.

System Architecture Design Consulting: Professional system architecture design consulting services are available for ESRI customers. These services can be used to resolve performance problems with existing environments and develop infrastructure support strategies for future successful GIS implementations. These services provide IT professionals with the information they need to support GIS users within their organizations.

Enterprise Systems Lab: The Enterprise Systems Lab supports the testing and evaluation of ESRI and third-party products with a primary focus on system architecture design performance and capacity planning for enterprise GIS environments. The test lab completes functional and performance testing for each software release along with a variety of customer implementation projects. Test results are used to validate performance and scalability of ESRI software technology and provide relative performance information used for development and maintenance of the ESRI capacity planning tools.

ESRI Press *Building a GIS*: Building a GIS is an ESRI Press book release that documents and shared experience collected from over 18 years of system design consulting services. Building a GIS includes all the material shared through the System Design Strategies TRD plus a CD copy of the Capacity Planning Tool, an application developed in Excel to capture and simplify the system design analysis design process. www.esri.com/esripress/buildingagis

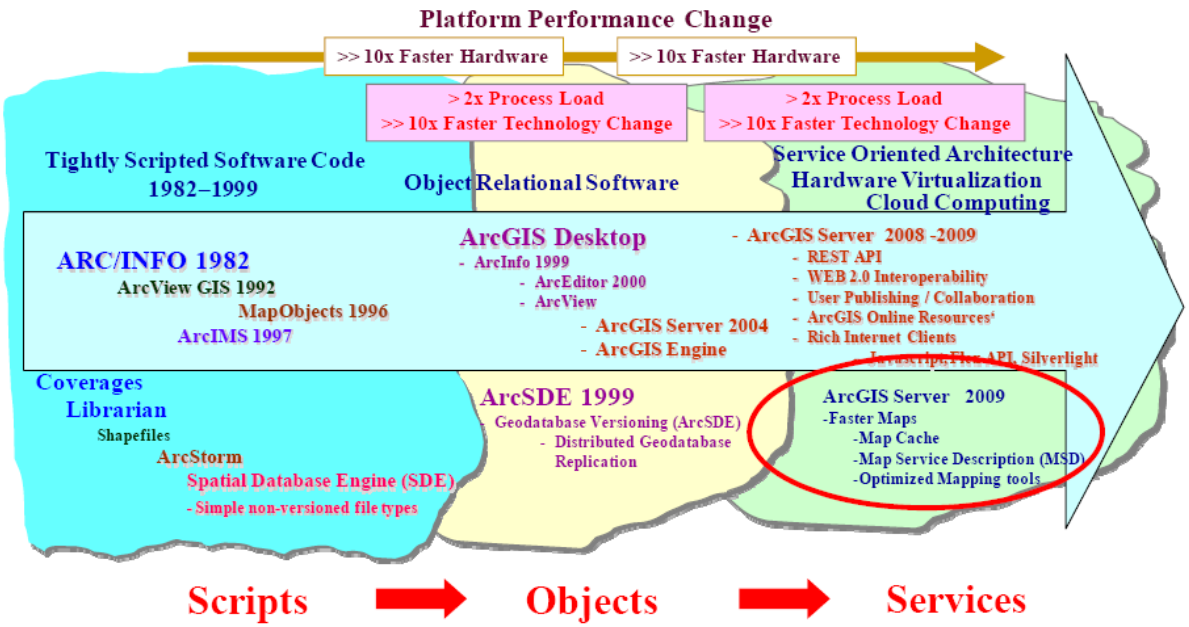
Several system architecture design activities are included in the annual ESRI International User Conference in San Diego, California, each summer. System architecture design technical workshops provide a system design strategy refresher. System design architects are available at the Professional Services island in the ESRI Showcase area to provide customer design support during the conference. A System Implementation papers track provides an opportunity for ESRI users to present their GIS implementation experiences

2 ESRI Software Evolution

For the past 38 years, ESRI has continued to develop evolving GIS software technology supporting functional requirements identified by the GIS user community. Sensitivity to software development trends and enterprise architecture strategies provide guidelines for development investment. ESRI software developers leverage the latest computer hardware and software technology to maintain ESRI's leadership in the GIS marketplace. ESRI aligns its resources to provide the best software and services based on GIS customer needs.

This section provides an overview of the ESRI software and associated product technologies. Understanding the primary role of each member of the ESRI software family will help users identify technology needs and develop a road map for migration to successful enterprise GIS operations. Figure 2-1 provides an overview of the ESRI software history and the associated third-party technologies supporting operational GIS enterprise environments.

Figure 2-1
GIS Enterprise Evolution



The early ARC/INFO software provided developers and professional GIS users with a rich toolkit to support geospatial query and analysis and demonstrate the value of GIS technology. ArcView introduced easy-to-use commercial off-the-shelf (COTS) software that could be used directly by GIS operational users. Map Objects empowered developers with a simple way to integrate GIS in common standard application environments. Terminal servers provide remote user access to centrally managed GIS desktop applications. ArcIMS Web services introduced a framework for publishing GIS information products to Web browser clients. ArcStorm and ArcSDE introduced better ways to manage and support GIS data resources.

Hardware performance improvements led to more efficient programming techniques deployed in the late 1990s. ArcGIS Desktop software provides users with a simple and powerful GIS application interface to support standard operations. ArcGIS Server and ArcGIS Engine provide GIS developers with rich processing tools and full GIS functionality for custom application development and deployment. Distributed geodatabase management tools and replication services support remote user access to centrally managed spatial resources and provide better protection and sharing of geospatial data.

Web technology introduced more ways to share data and services, enabling a new services oriented component architecture along with interoperability standards that provide open and adaptive solutions developed from a multi-vendor component architecture.

Software technology migration from scripts to objects to services accelerated technology change while

increasing demands on hardware performance and network capacity. The change in technology impacted business process in an evolutionary way opening new opportunities for GIS to support enterprise and community operations, helping customers better understand their world and make more informed decisions.

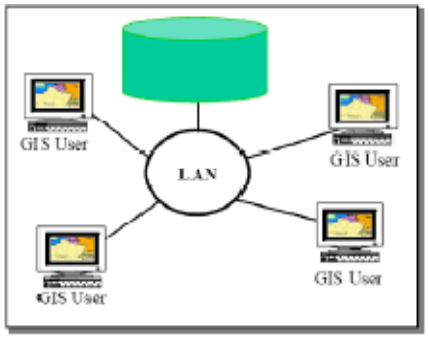
Many ESRI customers developed effective enterprise solutions with the Workstation ARC/INFO and ArcView GIS software provided in the 1990s. The current ArcGIS software provides operational capabilities that were not available with the older technology. Most ESRI legacy customers have migrated their data and applications to the current ArcGIS object-based geodatabase technology. New customers support enterprise GIS solutions directly with ArcGIS desktop and server software.

This past year saw some remarkable gains in performance and scalability of ArcGIS Server software. ArcGIS 9.3.1 introduced a new map service description (MSD) document with a new graphics rendering engine, new optimize mapping tools, and worldwide access to high quality cached image base maps. ArcGIS Online, new Resource Centers, and a growing number of ESRI User Forums expand and connect the GIS user community on a global scale. This was a remarkable year for GIS performance and scalability, with hardware performance gains of over 70 percent and software reducing processing loads by over 50 percent, expanding entry level GIS software capacity to over four times what was available just one year ago.

2.1 Organizational GIS Evolution

GIS implementations grew in size and complexity throughout the 1990s. GIS started on the user desktops and evolved to support GIS workgroup operations with department-level file servers. A majority of the GIS community is currently supported by department-level GIS operations. Figure 2-2 provides an overview of department-level GIS architecture.

Figure 2-2
Departmental GIS

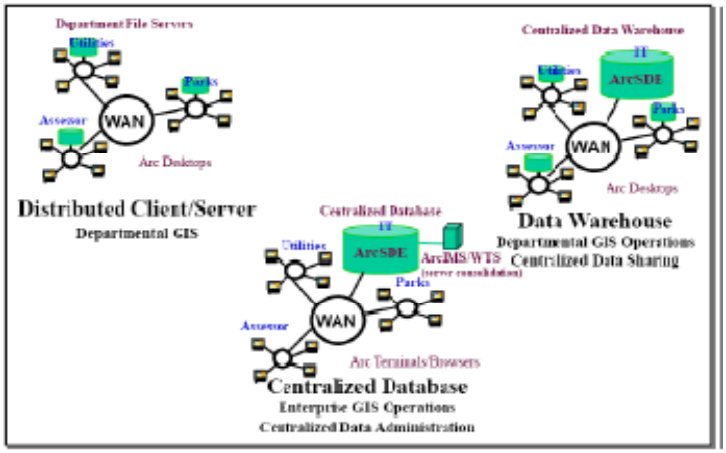


As GIS evolved within a typical organization, several departments had local GIS operations that required data resources from other departments. The organization's wide area network (WAN) became a way of sharing data between department servers. Data standardization and integrity issues surfaced within the organization since data was developed and managed from different department-level sources.

The initial Spatial Database Engine (ArcSDE) release in the mid 1990s supported the first GIS enterprise data warehouse operations, and many organizations shared GIS data resources from a central ArcSDE data warehouse. GIS responsibilities were assigned within IT departments to consolidate enterprise-wide GIS data and establish common data standards across the organization. The enterprise data warehouse provided a reliable shared GIS data source for departments throughout the organization. This was a very common migration path taken by most local government GIS operations.

Figure 2-3 provides an overview of the organizational architecture alternatives.

Figure 2-3
Organizational GIS



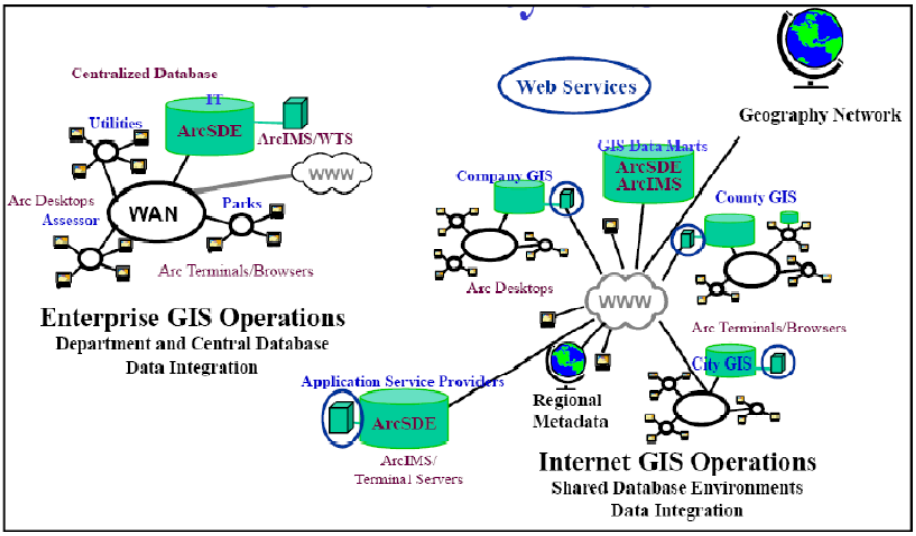
The electric and gas utilities started using GIS in the early 1990s to support management of their power distribution facilities. Most implementations were supported by a central database. Remote users were supported with terminal access to application compute servers (terminal servers) located in the central computer facility with the GIS database.

Many organizations today are moving their geospatial data from department file-based GIS database environments to a common central enterprise geodatabase and providing terminal client access to centrally managed server environments. Departments retain responsibility for their data resources, updating and maintaining data through terminal access to central ArcGIS Desktop applications. The central IT computer center supports the general administration tasks of data backups, operating system upgrades, platform administration, and so forth. Users throughout the organization enjoy browser access to published Web services over the intranet. The complexity and sophistication of the geodatabase for central administration and support make centralized servers the most productive alternative for most organizations.

2.2 Community GIS Evolution

Year 2000 introduced a growing Internet awareness, demonstrating the tremendous value of sharing information between organizations and nations. Internet access was extended from the workplace to the home, rapidly expanding the user community. Figure 2-4 shows how communities and companies developed and deployed services to customers over the Internet. The Internet provided opportunities for organizations to share data and services between organizations. Users had access to data and services from a multitude of organizations through the Internet.

Figure 2-4
Community GIS



ESRI introduced the Geography Network, providing a metadata search engine that published information about GIS data services and provided direct Internet links between the ArcGIS Desktop application and the data or service provider. ArcIMS introduced a way for organizations throughout the world to share GIS data and services. The Geography Network established a framework to bring GIS data and services together, supporting a rapidly expanding infrastructure of worldwide communities sharing information about the world we all live in. Promotion of data standards and improved data collection technologies unlock enormous possibilities for sharing geospatial information and help us better understand and improve our world.

GIS data resources are expanding exponentially. In the 1990s, GIS data servers seldom required a database that was more than 25 to 50 gigabytes in size. Today it is common for organizations to operate geodatabase servers supporting several terabytes to petabytes of GIS data (one petabyte is equal to eight quadrillion bits of data).

State-level agencies are consolidating data to support municipalities and commercial activities throughout their states. National agencies are consolidating data to support their user requirements and sharing data between state and national communities. Community-level data marts are being established to consolidate GIS data resources and support Internet data sharing to organizations throughout county and state regional areas.

Many organizations are outsourcing their IT operations to commercial Internet service providers (ISPs). Application service providers (ASPs) support organizations with IT administration, providing opportunities for smaller organizations to take advantage of high-end GIS database and application solutions to support their business needs. State governments are hosting applications and data for smaller municipalities throughout their states so the smaller communities can take advantage of GIS technology to support their local operations.

Regional geography network sites support sharing data throughout the region and within large state and federal agencies. ArcIMS software provides a metadata search engine that can be used by organizations to share their data and support their community operations. Cities can establish metadata sites to promote local commercial and public interests. States can consolidate metadata search engines for sharing data and services with municipalities throughout the state. Law enforcement can establish search engines to support national datasets. Businesses can establish metadata search engines to support distributed operational environments. Web services support community data sharing and integrated workflows.

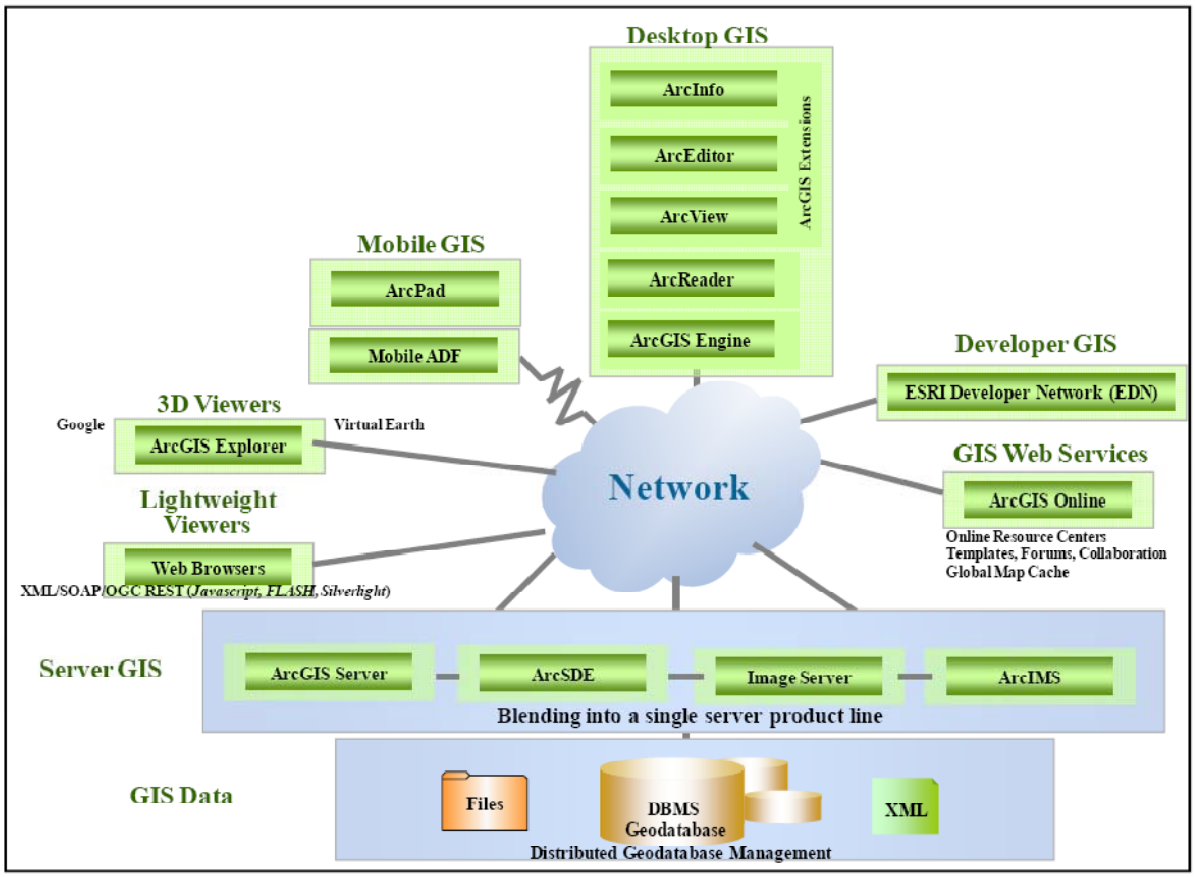
Deployment of ArcGIS Server with ArcGIS 9 expanded Web services technology to include geoprocessing and a broad range of service-oriented Web operations. GIS technology, in conjunction with Web standards and open systems architecture, has opened new opportunities for more improved business operations. As GIS data and services are shared among a growing number of organizations and technology provides real-time access to geographic information products, GIS becomes an integral part of community operations. GIS is rapidly expanding as a primary technology for understanding our world and related geospatial business opportunities.

ArcGIS 9.3 opens GIS to enable an Internet technology revolution, new opportunities for business and communities alike through GIS services deployed on ESRI, Google, and Microsoft geospatial Web 2.0 platform environments.

2.3 ESRI Product Family

The ESRI product family, illustrated in figure 2-5, includes a mix of software developed to support a full range of GIS user requirements. GIS software is provided to support desktop, server, and mobile user operations. Data management solutions are provided to support data file, geodatabase, and Extensible Markup Language (XML) based formats.

Figure 2-5
ESRI Product Family



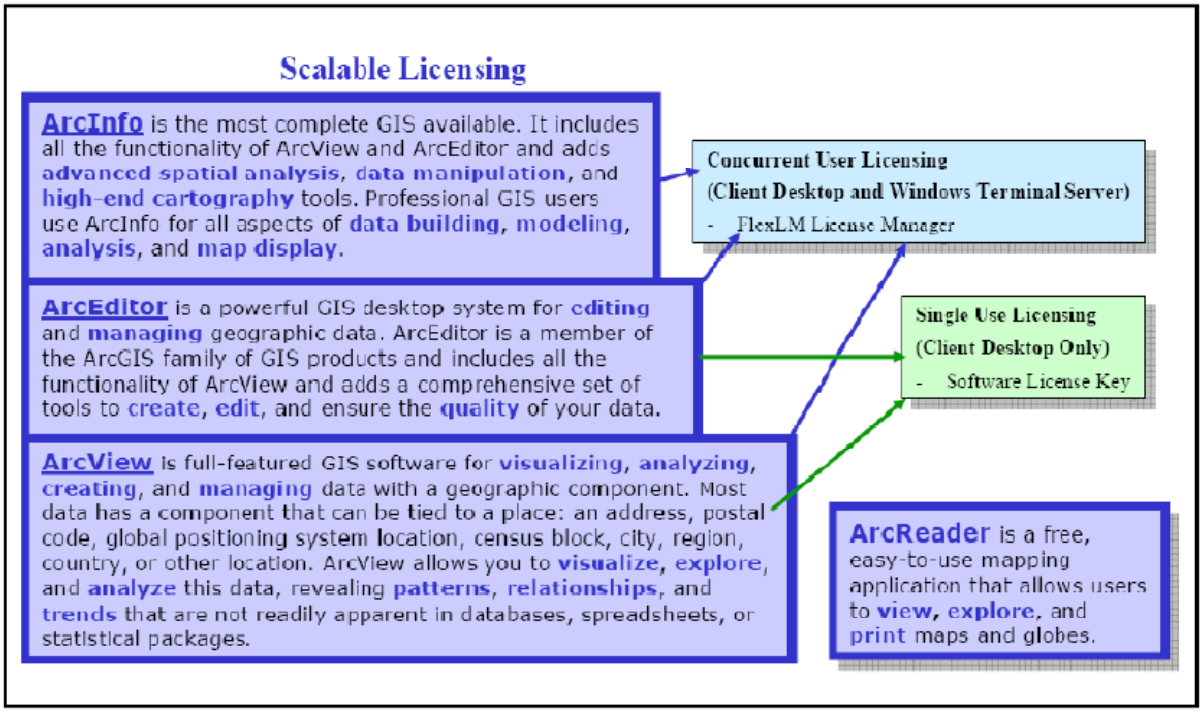
GIS Web services are provided by ESRI to support a variety of managed, hosted, and shared GIS Internet services. Online resource centers, templates, forums, Web map hosting, collaboration, and global map cache all make ArcGIS Online a growing part of enterprise GIS core technology. ArcGIS Server provides technology for publishing GIS services that can be consumed by ArcGIS Desktop, mobile GIS, and standard Web browsers. HTML JavaScript and new Internet Rich Clients (Adobe Flex and Microsoft Silverlight) provide users with high map quality and improved user display performance over the Web. ESRI Developer Network (EDN) provides a range of technical services to the ESRI developer community through a bundled low-cost developer software license.

Desktop GIS: Desktop GIS is divided into four licensed solutions based on user functionality needs including ArcGIS Engine, a desktop development environment that provides a complete set of ArcGIS ArcObjects components for custom desktop application development.

Figure 2-6 provides an overview of the ArcGIS Desktop software. ArcGIS Desktop is licensed as scalable software options based on user functional needs. ArcReader is free desktop software for viewing and sharing a variety of dynamic geographic data. ArcView includes all the functionality of ArcReader and adds geographic data visualization, query, analysis, and integration capabilities. ArcEditor includes all the functionality of ArcView and adds the power to create and edit data in a geodatabase. ArcInfo is the complete GIS data creation, update, query, mapping, and analysis system. Internet access to worldwide Microsoft Virtual Earth (Bing) high resolution imagery is included with current ArcGIS Desktop maintenance licensing starting with the ArcGIS 9.3.1 release.

A range of desktop extension licenses are available that provide enhanced functionality for supporting more focused GIS operations. Desktop extensions operate with the foundation ArcGIS Desktop license. ArcGIS Desktop extensions are listed on the Web at the following URL:
http://www.esri.com/software/arcgis/about/desktop_extensions.html

Figure 2-6
ArcGIS Desktop

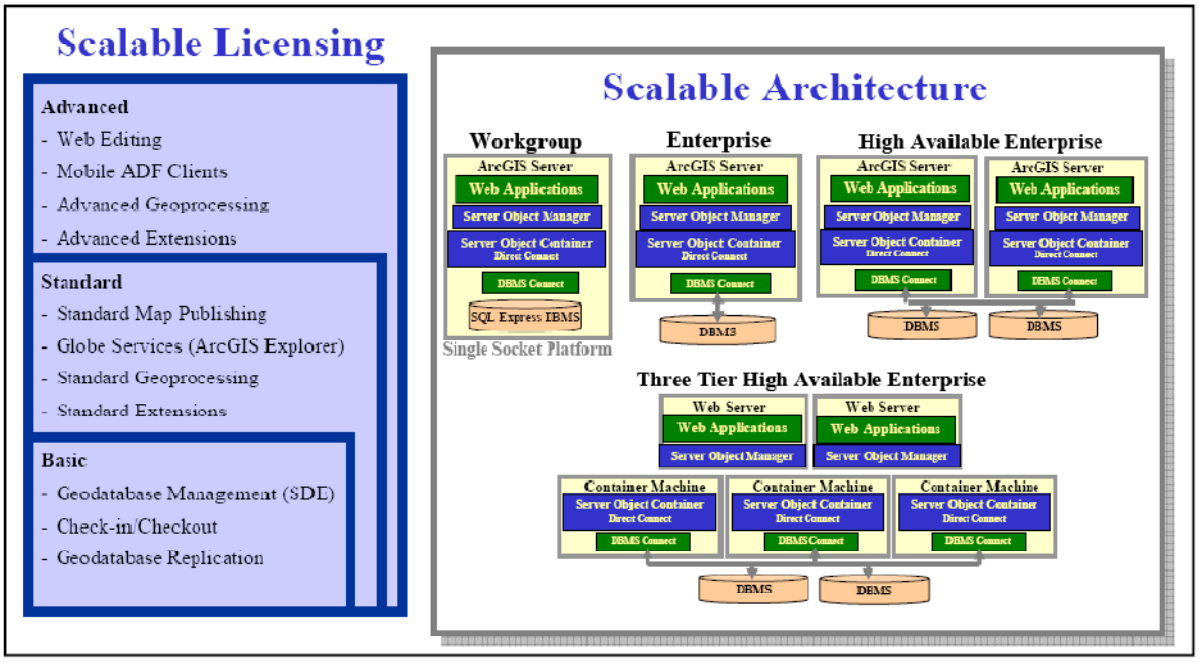


Single use licensing is provided for ArcView and ArcEditor client desktop workstations and is managed by a software license key. Concurrent use licensing is provided for ArcView, ArcEditor, and ArcInfo user sessions and is managed by a networked FLEXlm (Flexible License Manager) hardware key.

Server GIS: Server GIS is used for a variety of centrally hosted GIS services. The server-based GIS technology trend is growing. GIS software can be centralized in application servers to deliver GIS capabilities to large numbers of users over networks. Enterprise GIS users connect to central GIS servers using traditional desktop GIS as well as Web browsers, mobile computing devices, and digital appliances.

Figure 2-7 provides an overview of the ArcGIS Server software. ArcGIS Server is divided into three licensed solutions with pricing based on available functionality and system capacity. ArcGIS Server Basic includes geodatabase management (ArcSDE), geodatabase check-in/checkout, and geodatabase replication services. ArcGIS Server Standard includes all the functionality of ArcGIS Server Basic plus standard map publishing, globe services (ArcGIS Explorer), and standard geoprocessing. ArcGIS Explorer is a free, lightweight ArcGIS Server desktop client. It can be used to access, integrate, and utilize GIS services, geographic content, and other Web services. ArcGIS Server Advanced includes all the functionality of ArcGIS Server Standard plus Web editing, mobile client application development framework (mobile ADF), advanced geoprocessing, and support for ArcGIS Server extensions. ArcGIS Server extensions are listed on the Web at the following URL: http://www.esri.com/software/arcgis/arcgisserver/extensions.html

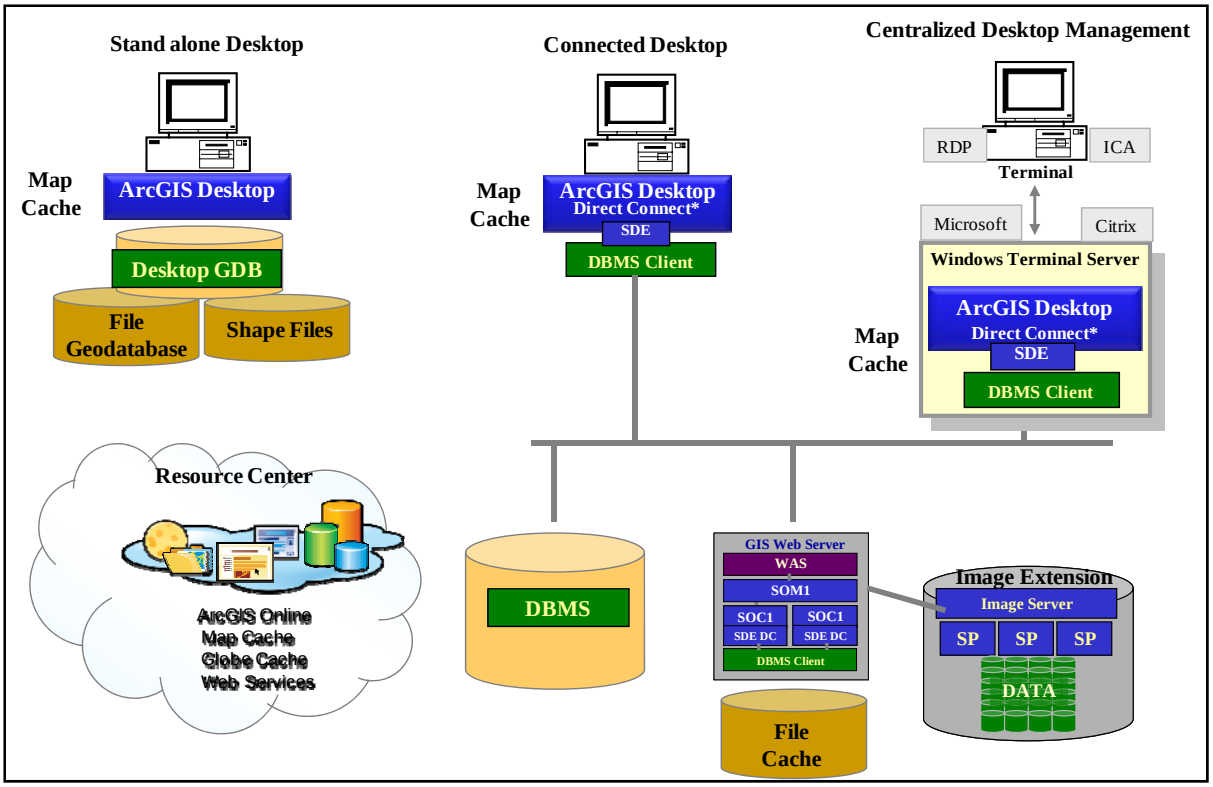
Figure 2-7
ArcGIS Server



ArcGIS Server supports a scalable architecture, and license pricing is based on the number of core processors supporting the deployed server configuration. Workgroup pricing is provided for an entry-level server configuration (all software must be supported by a single server with a Microsoft SQL Express database). Enterprise pricing is provided for an entry-level four core server configuration, with the option for expanding capacity on per core increments to support peak capacity needs. All core included in server platforms supporting ArcGIS Server software components are included in determining capacity-based software licensing requirements (ArcGIS Server Web ADF run time omponents are no longer included in determining licensed server core starting with the ArcGIS Server 9.3.1 release).

ArcGIS Desktop Software Solutions. Figure 2-8 provides an overview of the primary ArcGIS Desktop client operations. Potential candidate workflows support standalone Desktop, connected Desktop, and centralized Desktop configurations.

Figure 2-8
ArcGIS Desktop Operations



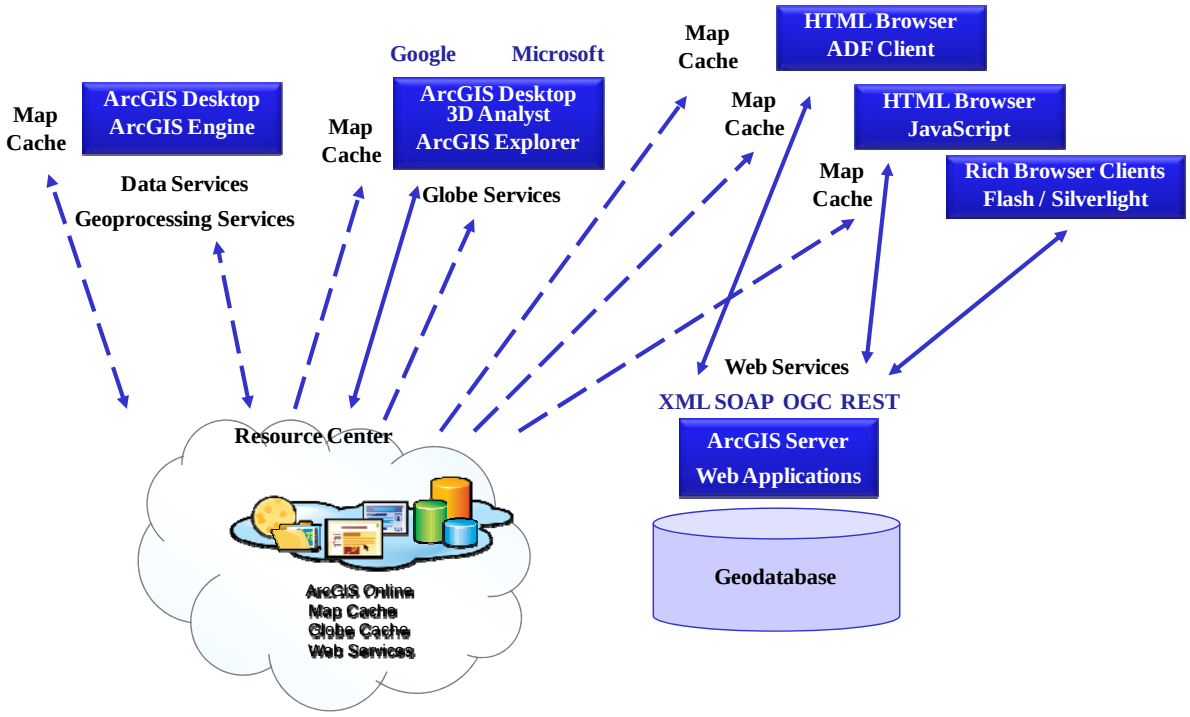
Standalone ArcGIS Desktop workstation. ArcGIS Desktop workflows can operate as a single standalone workstation using a variety of local data sources. The personal geodatabase is now supported by a Microsoft SQL Server Express database (name was changed this year to Desktop Geodatabase), which can operate as a replica version of a central ArcSDE geodatabase. The SQL Server Express database will support up to 4 GB or geospatial vector data. The ArcGIS Desktop workstation can also support a File Geodatabase, which provides up to 1 TB of data per data file. The File Geodatabase can provide a replica of reference data layers that can be incrementally updated from a central ArcSDE geodatabase.

Connected ArcGIS Desktop workstation. ArcGIS Desktop user workflows can operate in a connected local area network (LAN). Standard architectures include ArcGIS Desktop workstations connected over a LAN to a central ArcSDE geodatabase, Web services, and Image data sources (ArcGIS Server Image Extension was introduced this year as the new name for Image Server as it is integrated into the ArcGIS Server core software.

Centralized ArcGIS Desktop server. ArcGIS Desktop user workflows can use terminal clients to access centrally managed ArcGIS Desktop applications. ArcGIS Desktop can be deployed on Windows Terminal Server using Microsoft or Citrix terminal clients. Most ESRI clients use Citrix XenApp terminal clients for optimum compute and display performance.

ArcGIS Server Web Operations. Figure 2-9 provides an overview of the primary ArcGIS Server Web client operations. Potential candidate workflows include ArcGIS Desktop, ArcGIS Engine, ArcGIS Explorer, and standard Web browser applications including ADF clients, JavaScript, Adobe Flash and Microsoft Silverlight rich internet clients.

Figure 2-9
Web Operations



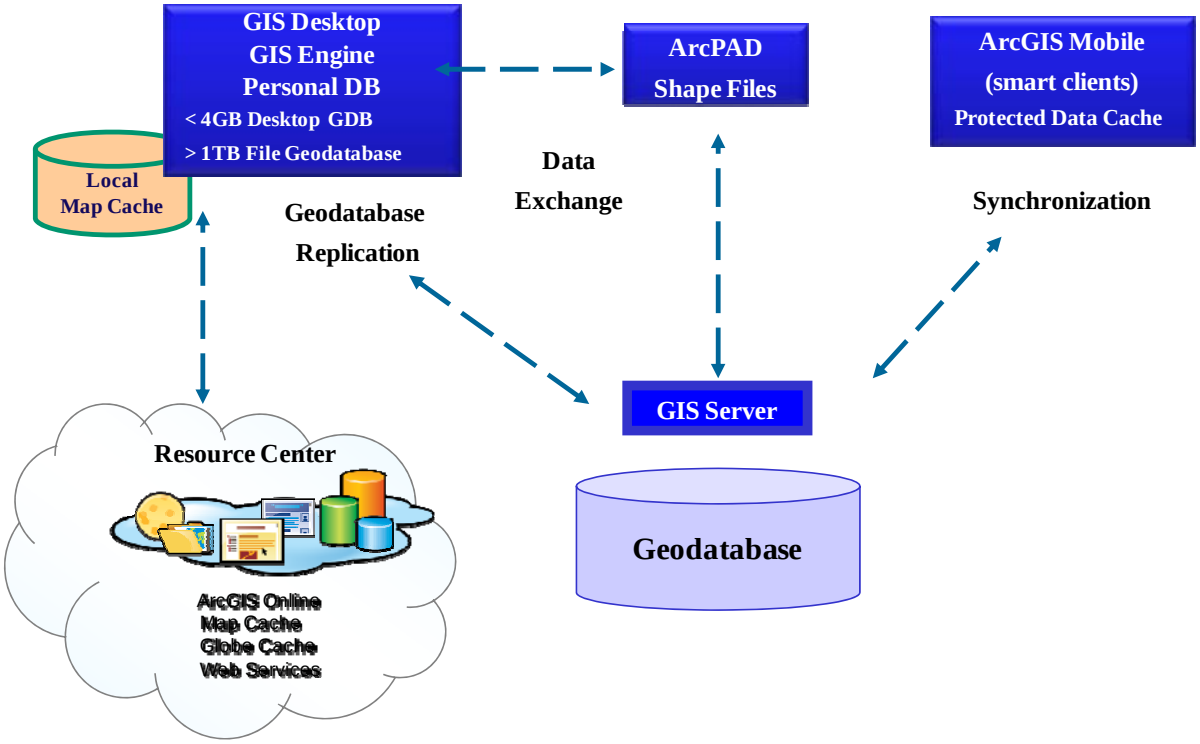
ArcGIS Server can provide Simple Object Access Protocol (SOAP)/XML-based data services (published reference images) and geoprocessing services to ArcGIS Desktop and ArcGIS Engine client applications, provide a 3D globe cached file data source for ArcGIS 3D Analyst and ArcGIS Explorer clients, and host a full range of map view and edit applications for Web HTML browser clients supported by out-of-the box .NET and Java Web map and editor server development kit components. ArcGIS 9.3 supports a REST and JavaScript API providing a variety of new Web client options, including KML services for use with Google and Microsoft geospatial client environments.

ArcIMS is a popular solution for delivering dynamic maps and GIS data and services via the Web. It provides a highly scalable framework for GIS Web publishing that meets the needs of corporate intranets and demands of worldwide Internet access. ArcIMS customers are rapidly moving to ArcGIS Server software to leverage the rich functionality available with the new ArcGIS Server software release. ArcGIS 9.3.1 provides a new optimized dynamic map service that outperforms equivalent map services deployed using the ArcIMS Image service. ArcGIS Server with cached map services provide high quality and high performance well beyond what was available with the legacy ArcIMS technology.

ArcGIS Image Server (now called the ArcGIS Server Image Extension) changes how imagery is managed, processed, and distributed. The Image Extension provides fast access and visualization of large quantities of file-based imagery, processed on the fly and on demand. It provides rapid display of imagery for a number of users working simultaneously, without the need to preprocess the data and load it into a database management system (DBMS). ArcGIS Image Extension can be used as a data source for ArcGIS Desktop, ArcGIS Server, and ArcIMS Web mapping services. The ArcGIS Server image service extends Image Extension access to include the full range of Web Internet clients. Additional support is provided for AutoCAD and MicroStation CAD clients. On-the-fly processing can include image enhancement, orthorectification, pan sharpening, and complex image mosaicking.

Mobile GIS: Mobile GIS supports a range of mobile systems from lightweight devices to PDAs, laptops, and Tablet PCs. ArcPad is software for mobile GIS and field-mapping applications. All ArcGIS Desktop products—ArcReader, ArcView, ArcEditor, and ArcInfo—and custom applications can be used on high-end mobile systems such as laptops and Tablet PCs. Figure 2-10 provides an overview of the primary connected mobile workflow alternatives.

Figure 2-10
Mobile Operations



The ArcGIS Server Basic license supports distributed geodatabase replication. Geodatabase replication provides loosely connected synchronization services for distributed geodatabase versions maintained in supported database platforms. Web-based disconnected check-in and checkout services are also provided. Distributed geodatabase replication is discussed later in section 6 Data Administration.

ArcGIS also provides geodatabase support in a Microsoft SQL Server Express personal database. Microsoft SQL Server Express is bundled with each ArcGIS Desktop software license. ArcGIS Desktop clients (including custom ArcGIS Engine runtime deployments) can support a distributed geodatabase client replica and synchronize changes with the central parent geodatabase. The SQL Server Express database has a data capacity of 4 gigabyte (GB).

ArcGIS also supports a file based geodatabase. The ArcGIS Server 9.3 data checkout/check-in and one-way incremental replication Web interface supports distributed file geodatabase clients. A file geodatabase can support up to 1 TB of data with impressive performance (single use performance comparable to an ArcSDE Geodatabase experience).

ArcGIS Server Advanced license supports ArcGIS mobile software development kit. ArcGIS Mobile lets developers create centrally managed, high-performance, GIS-focused applications for mobile clients. Mobile applications powered by ArcGIS Server contribute to increased field productivity and more informed personnel.

Developer GIS: EDN is an annual subscription-based program designed to provide developers with comprehensive tools that increase productivity and reduce the cost of GIS development. EDN provides a comprehensive library of developer software, a documentation library, and a collaborative online Web site that

offers an easy way to share information.

GIS Web Services: GIS Web services offer a cost-effective way to access up-to-date GIS content and capabilities on demand. With ArcGIS Web Services, data storage, maintenance, and updates are handled by ESRI, eliminating the need for users to purchase and maintain the data. Users can access data and GIS capabilities directly using ArcGIS Desktop or use ArcWeb Services to build unique Web-based applications. ArcGIS Online Services provide instant and reliable access to terabytes of data including street maps, live weather and traffic information, extensive demographic data, topographic maps, and high-resolution imagery from an extensive list of world-class data providers.

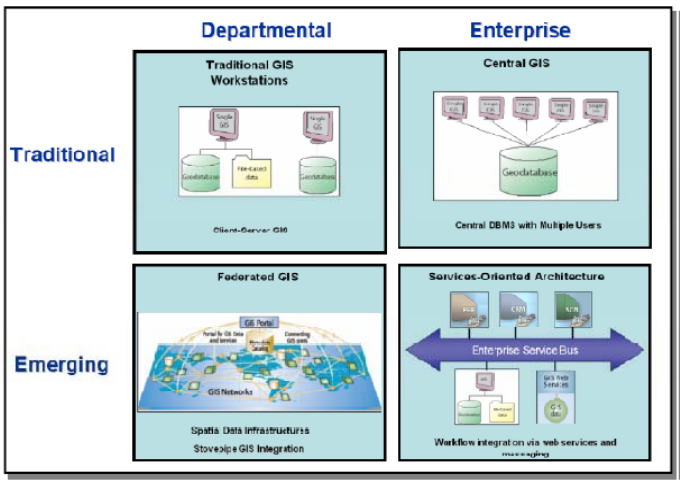
2.4 Expanding GIS Technology Trends

GIS software and computer infrastructure technology continue to expand capabilities and introduce new business opportunities. Organizations are expanding operations to incorporate mobile users as an integral part of their enterprise workflow. Improved availability and capacity of wireless technology support mobile communication connectivity for a growing number of GIS users as shown in figure 2-11.

Figure 2-11
Mobile and Wireless GIS Technology

Figure 2-12 provides a simple overview of common ArcGIS deployment alternatives. Traditional department-level GIS client/server operations are looking for ways to improve access and data sharing with other organizations and introducing new emerging federated GIS architecture strategies. Traditional enterprise-level operations are looking for ways to integrate GIS with other centrally managed business operations and introducing new emerging integrated business solutions based on service-oriented architecture strategies.

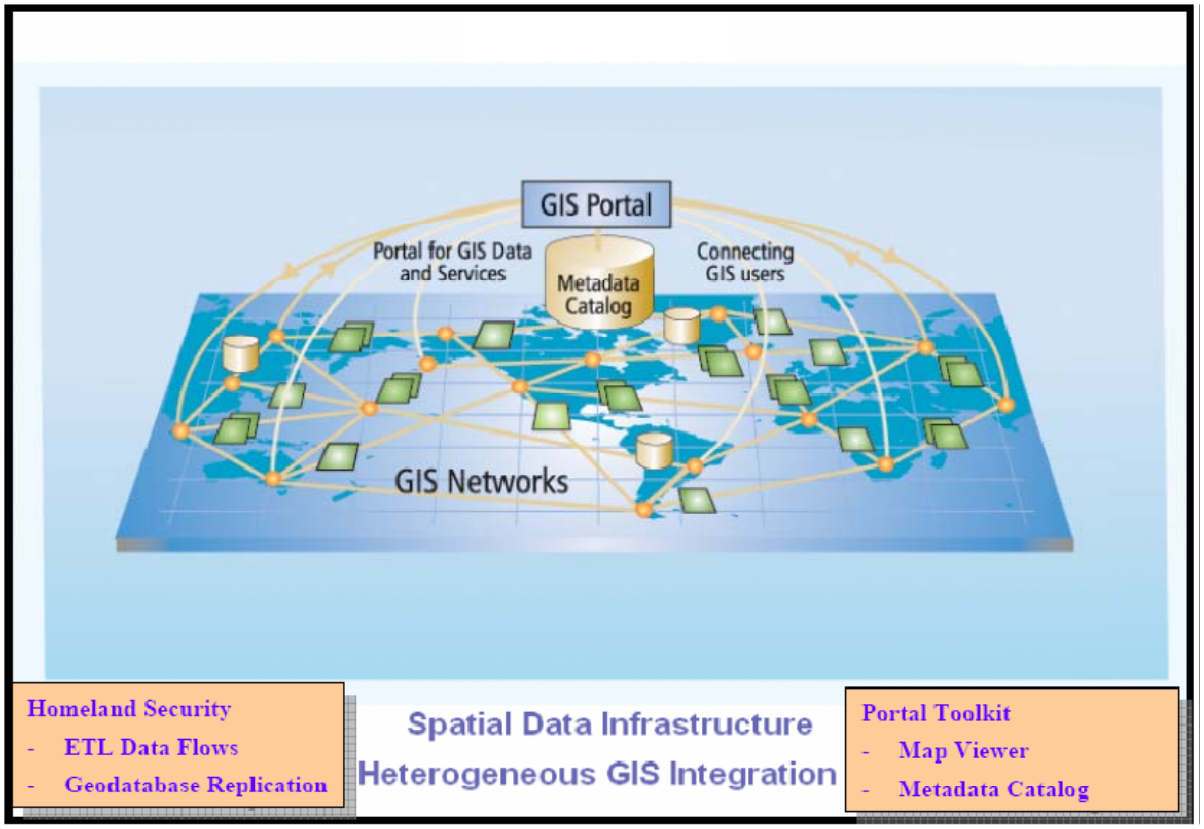
Figure 2-12
GIS Is Deployed in Many Ways



2.4.1 Federated GIS Technology

Database and Web technology standards provide new opportunities to better manage and support user access to a rapidly growing volume of geospatial data resources. Web services and rich XML communication protocols support efficient data migration between distributed databases and storage locations. Web search engines and standard Web mapping services support integrated geospatial information products published from a common portal environment with data provided from a variety of distributed service locations. Federated architectures identified in figure 2-13 promote better data management, integrating community and national GIS operations. Geodatabase replication services and managed extract transform, and load (ETL) processes support loosely coupled distributed geodatabase environments.

Figure 2-13
Federated GIS Technology



2.4.2

2.4.3 Service-Oriented Architecture

Technology is changing faster each year, and organizations are searching for more effective ways to manage technology change. During the 1990s, there was a shift in programming methods promoted by commercial software acceptance of component architecture standards. Software development migrated from compiled, scripted legacy languages to object-based programming environments. ArcGIS technology is based on common ArcObjects components used to support a broad range of desktop and server software. Developing new applications and functionality in an object-based programming environment is much more powerful than developing in the traditional scripted software languages.

Technology change is again being influenced by general acceptance of standard Web communication protocols and more stable and available network bandwidth connectivity. Software development is taking advantage of Internet communication standards and network connectivity with a new service-oriented enterprise architecture strategy.

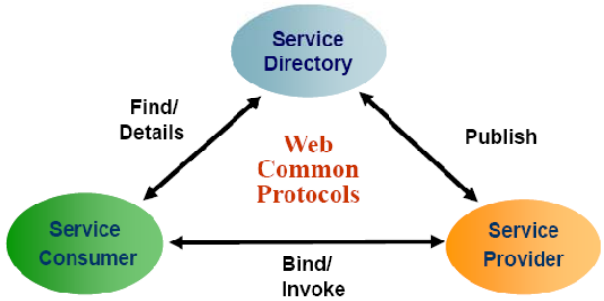
Business environments are influenced by the rate of technology change. Change introduces risk contributing to business success or failure. Selecting the right technology investment strategies is critical. Service-oriented architecture deployment strategies reduce business risk through diversification and reduced vendor dependence. Open standards reduce the time and effort involved in developing integrated business systems, providing integrated information products (common operating picture) that support more informed business decisions. Advantages of a service-oriented architecture are highlighted in figure 2-14.

Figure 2-14
Advantages of a Service-Oriented Architecture

- Technology Change → Component Architecture
- Business Continuance → Reduce Vendor Dependence
- Leverage Investments → Reusable Components
- Customer Flexibility → More Vendor Choices
- Business Integration → Open System Communications

The core components supporting a service-oriented architecture (SOA) are presented in figure 2-15. These components include service providers, service consumers, and implementation of a service directory.

Figure 2-15
Service-Oriented Architecture Technology

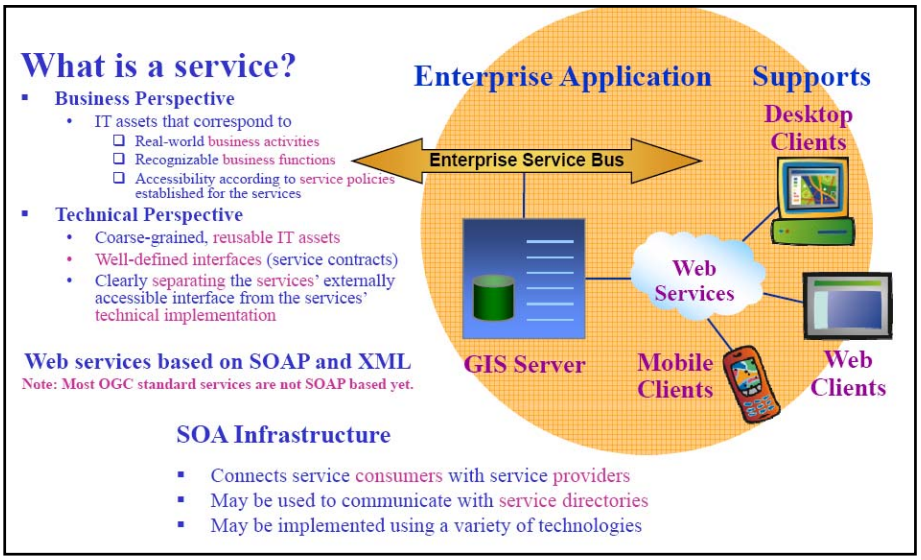


An approach for building distributed computing systems based on encapsulating business functions as services that can be easily accessed in a loosely coupled fashion

Common Web protocols and network connectivity are essential to support this type of architecture. Business functions are encapsulated as Web services that can be consumed by Web clients and desktop applications.

The basic language of an SOA is introduced in figure 2-16. New business functions are provided as Web services, which are IT assets that correspond to real-world business activities or recognizable business functions, in which accessibility is based on service policies established to support enterprise operations (loosely coupled to the business applications).

Figure 2-16
SOA Infrastructure

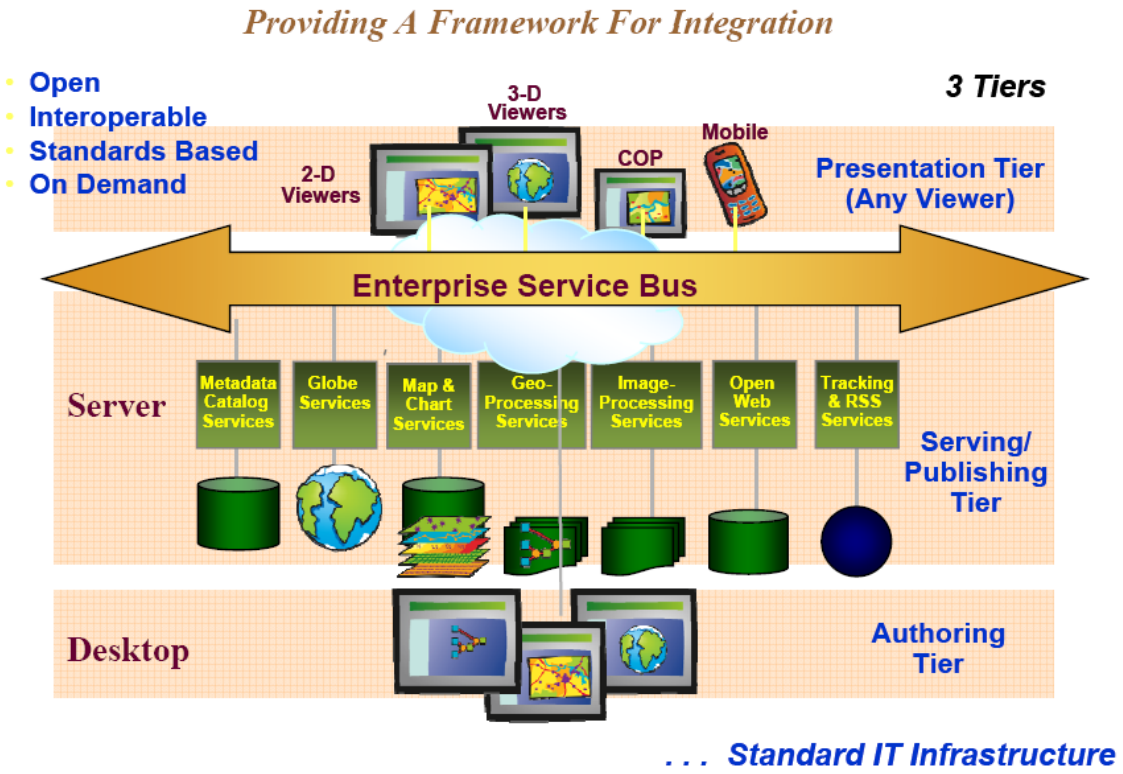


From a technical perspective, services are coarse-grained, reusable IT assets. They have well-defined interfaces (service contracts) with the software technology providing the service abstracted from the external accessible service interface. The trend is to support SOA through Web services based on SOAP and XML.

The SOA infrastructure connects service consumers with service providers, may be used to communicate with service directories, and may be implemented using a variety of technologies.

ESRI embraced open standards during the 1990s and has actively participated in the Open GIS Consortium and a variety of other standards bodies in an effort to promote open GIS technology. The initial ArcIMS Web services, Geography Network metadata search engines, Geospatial One-Stop, and the ESRI Portal Toolkit technology are all examples of service-oriented solutions supporting ESRI's current customer implementations. Figure 2-17 provides a view of how current ESRI software supports the evolving SOA enterprise infrastructure.

Figure 2-17
ESRI Fits into SOA

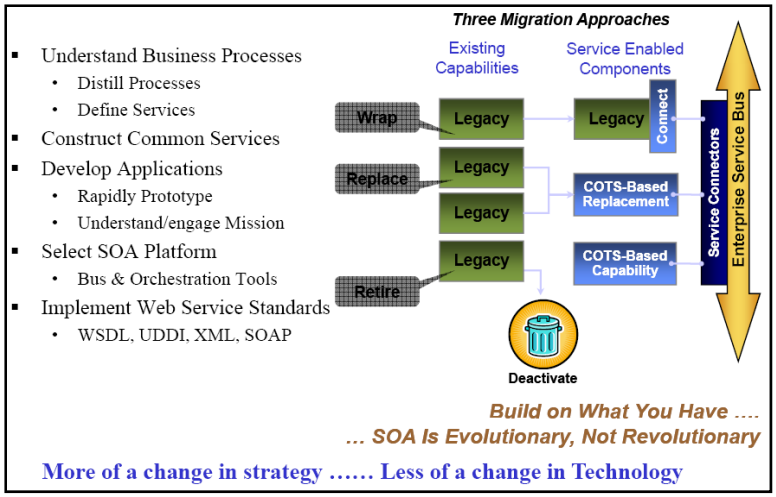


The SOA framework includes multiple access layers connecting producers and consumers, based on current client/software technology and incorporating Web application and service communication tiers. Consumers connect to producers through a variety of communication paths. This framework supports a presentation tier of viewers with access to available published services, a serving/publishing tier of services, and an authoring tier of professional ArcGIS Desktop users. This framework supports current client/server connections (client applications), Web applications, and Web services—all available today with current technology. Future vendor compliance and maturity of Web interface standards are expected to gradually migrate business applications from tightly coupled proprietary client/server environments to a more loosely coupled service-oriented architecture. The ideal environment would decouple business services and workflows from the underlying software technology providing an adaptive business environment that can effectively manage and take advantage of rapid technology change.

GIS is by nature a service-oriented technology with inherent fundamental characteristics that bring diverse information systems together to support real-world decisions. GIS technology flourishes in a data-rich environment, and ArcGIS technology can help ease the transition from existing "stovepipe" GIS environments. The geodatabase technology provides a spatial framework for establishing and managing integrated business operations. Many spatial data resources are available to support organizations as they migrate their operations to take advantage of GIS technology.

Migration toward a service-oriented architecture is more a change in attitude than a change in technology. Moving a business from high-risk, tightly coupled, monolithic stovepipe operations to a more integrated responsive service-oriented architecture will take time. Figure 2-18 provides some basic guidelines for moving existing systems to a more dynamic and supportable SOA environment.

Figure 2-18
Migrating to a Service-Oriented Architecture

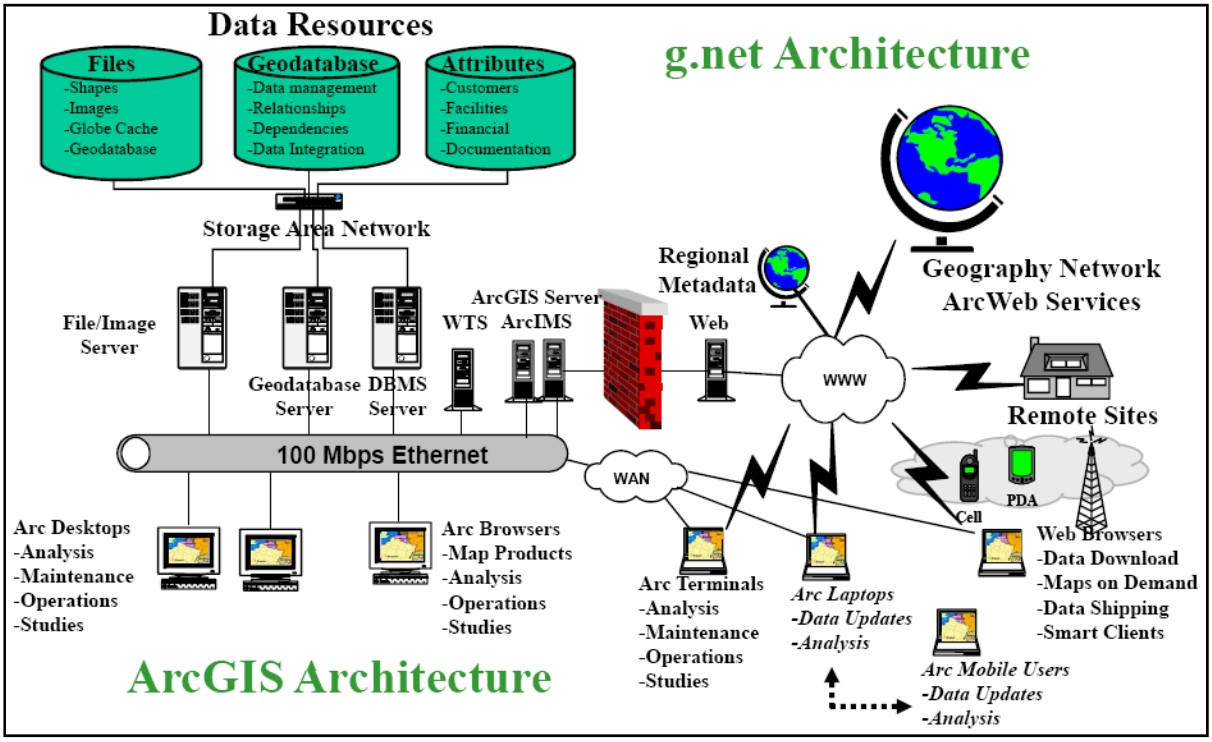


Understanding SOA and how it enables business process integration and helps control and manage technology change is important. Organizations must build an infrastructure that can effectively take advantage of new technology to stay competitive and productive in today's rapidly changing environment.

2.5 GIS Technology Alternatives

Current GIS technology is available to support a rapidly expanding spectrum of GIS user needs as depicted in figure 2-19. Solutions are supported by ESRI products integrated with a variety of vendor technologies.

Figure 2-19
ESRI Core GIS Technology



Data storage and data management technologies are growing in importance as organizations continue to develop and maintain larger volumes of GIS data. Individual server storage solutions are being replaced by more adaptive storage area networks (SANs), enhancing the IT's ability to respond to changing data storage needs and providing options for efficiently managing large volumes of data.

GIS data sources include file servers, geodatabase servers, and a variety of business database solutions. Desktop ArcGIS applications can be supported on local workstation clients or on centrally managed Windows Terminal Server farms.

Web services are supported by ArcGIS Server and legacy ArcIMS mapping services to Web browser clients throughout the organization and the community. ArcGIS clients are able to connect to ArcGIS Server Web products as intelligent browser clients, enabling connection to unlimited data resources through the ESRI Geography Network as well as organization resources served through a variety of ESRI customer portals. Users can access applications from the Internet or through intranet communication channels. Mobile ArcGIS users can be integrated into central workflow environments to support seamless integrated operations over wireless or remote connected communication. ArcGIS Desktop applications can include Web services as data sources integrated with local geodatabase or file data sources, expanding desktop operations to include available Internet data sources. GIS enterprise architecture is typically supported by a combination of ArcGIS Desktop, ArcGIS Server, and geodatabase software technology. Selecting the right combination of technology will make a big difference in the level of support for user operational needs and business productivity.

2.6 GIS Configuration Alternatives

GIS environments commonly begin with single-user workstations at a department level within the organization. Many organizations start with one GIS manager and evolve from a department level to an enterprise operation. This was common through the early 1990s, as many organizations worked to establish digital representation of their spatial data. Once this data is available, organizations expand their GIS operations to support enterprise business needs.

GIS is a very compute-intensive and data-rich technology. A typical GIS workflow can generate a remote user desktop display every 6–10 seconds with a client application requiring hundreds of sequential data requests to a shared central data server to support each desktop display. GIS workflows can place high processing demands on central servers and generate a relatively high volume of network traffic. Selecting the right configuration strategy can make a significant impact on user productivity.

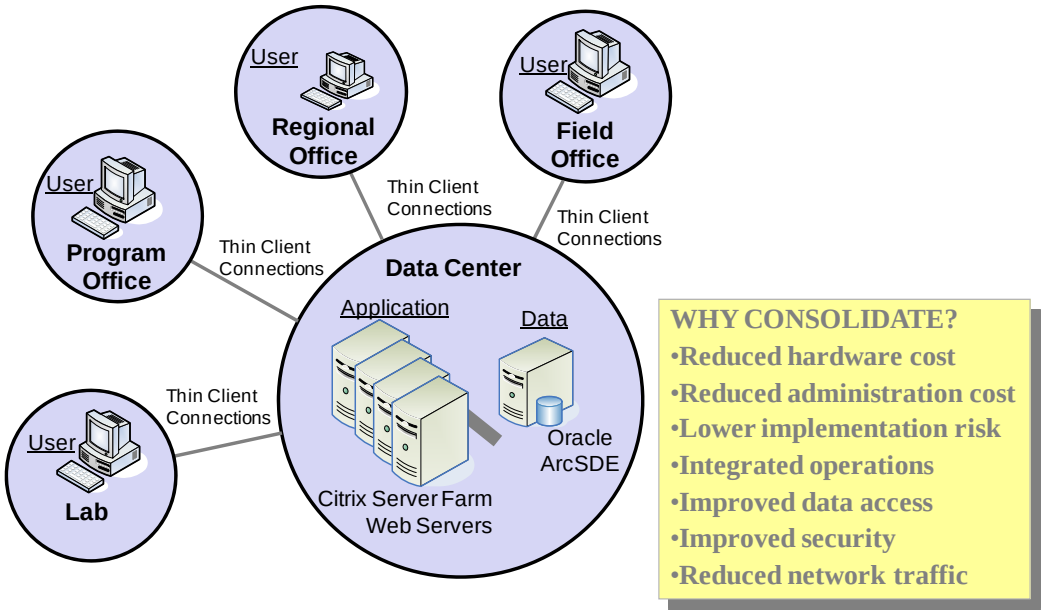
Data can be shared between users in a variety of ways. Most organizations today have user workstations connected to local area network (LAN) environments and locate shared spatial data on dedicated server platforms. User applications connect to shared data sources to support GIS operations.

Centralized Data Configuration Alternative

The most simple system architecture is supported by a central GIS database. A central database architecture supports one copy of the production database environment, minimizing administrative management requirements and ensuring data integrity.

GIS desktop applications can be supported on user workstations located on the central LAN, each with access to central GIS data sources. Data sources can include GIS file servers, geodatabase servers, and related attribute data sources as shown in figure 2-20.

Figure 2-20
Centralized Computing Environment



Remote user access to central data sources can be supported by central Windows Terminal Server (WTS) farms, providing low-bandwidth display and control of central application environments.

Centralized application farms minimize administration requirements and simplify application deployment and support throughout the organization. Source data is retained within the central computer facility, improving security and simplifying backup requirements.

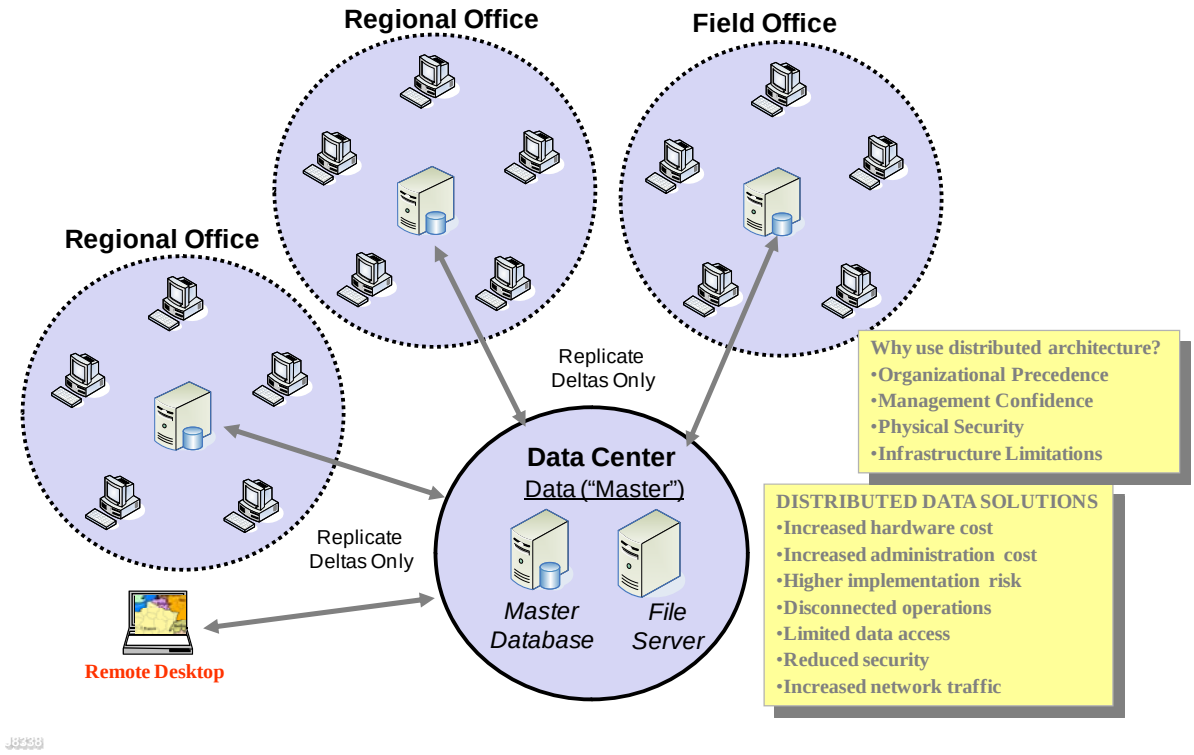
A variety of ArcIMS map services can support standard browser clients throughout the organization. Web mapping services support low-bandwidth access to published GIS information products and services.

Today distributed computing technology can support consolidated architectures at a much lower risk and cost than similar distributed environments. For this reason, many organizations are in the process of consolidating their data and server resources. GIS can benefit from consolidation for many of the same reasons experienced by other enterprise business solutions. Centralized GIS architectures are generally easier to deploy, manage, and support than distributed architectures and provide the same user performance and functionality.

Distributed Data Configuration Alternative

Distributed solutions are supported by replicated copies of the data at remote locations, establishing local processing nodes that must be maintained consistent with the central database environment as shown in figure 2-21. Data integrity is critical in this type of environment, requiring controlled procedures with appropriate commit logic to ensure changes are replicated to the associated data servers.

Figure 2-21
Distributed Computing Environment



Distributed database environments generally increase initial system cost (more hardware and database software requirements) and demand additional ongoing system administration and system maintenance requirements. Distributed solutions are provided to support specific user needs. These generally increase system complexity and cost and lengthen system deployment timelines.

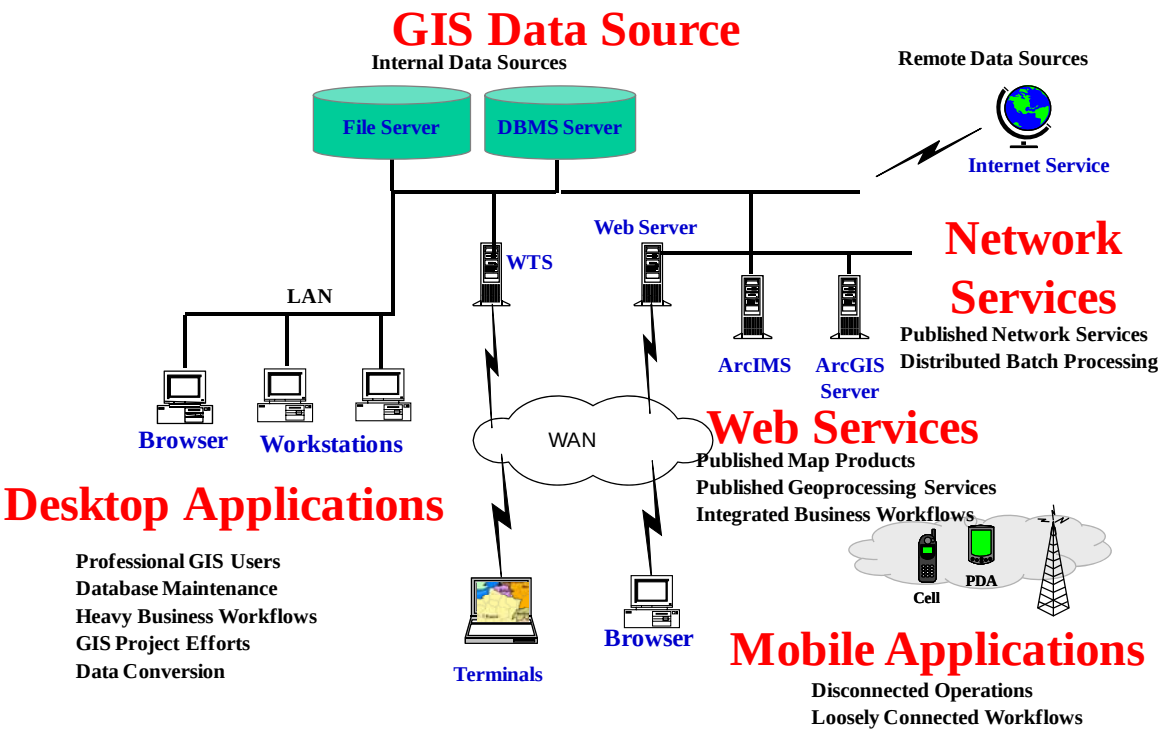
In many cases, standard database solutions do not support replication of spatial data. GIS users with distributed database requirements must modify their data models and establish procedures to administratively support data replication. The complexity of current geodatabase environments has complicated the implementation of an efficient commercial spatial replication solution. Many GIS users are interested in replicating regional or selected versions of a geodatabase, which is not understood by commercial replication technologies. ArcGIS software functions are available to support custom geodatabase replication solutions. ArcGIS 9.2 provides support for distributed geodatabase replication, providing alternative options for supporting distributed operational requirements.

2.7 GIS Software Selection

Selecting the right software and the most effective deployment architecture is very important. ArcGIS technology provides many alternative architecture solutions and a wide variety of software, all designed to support specific user workflow needs.

Figure 2-22 provides an overview of the GIS software technology alternatives. What is the best data source? What user workflows should be supported by GIS desktop applications? What can be supported by cost-effective Web services? What business functions would be best supported by network services? Where will mobile applications improve business operations? Understanding the available technology alternatives and how each will perform and scale within the available user environment can provide the information needed to make the right technology decisions.

Figure 2-22
GIS Software Technology Alternatives



GIS Data Source: Operations can be supported on local disk or CD-ROM, shared file servers, geodatabase servers, or Web data sources. Local data sources support high-performance productivity requirements with minimum network latency. Remote Web services allow connection to a variety of published data sources, with the drawback of potential bandwidth congestion and slow performance. There are other more loosely connected architecture solutions that reduce potential network performance latency and support distributed data integration.

Desktop Applications: The highest level of functionality and productivity is supported with the ArcGIS Desktop applications. Most professional GIS users and GIS power users will be more productive with the ArcGIS Desktop software. These applications can be supported on the user workstation or through terminal access to software executed on central Windows Terminal Server farms. Some of the more powerful ArcGIS Desktop software extensions perform best on the user workstation with a local data source, while most ArcGIS Desktop use workflows that can be supported more efficiently on a terminal server farm. Selecting the appropriate application deployment strategy can have a significant impact on user performance, administrative support, and infrastructure implementations.

Web Services: The ArcIMS and ArcGIS Server technologies provide efficient support for a wide variety of more focused GIS user workflows. Web services also provide a very efficient way to share data to support remote client workflows. ArcIMS provides the most efficient way to publish standard map information products. ArcGIS Server provides enhanced functionality to support more advanced user workflows and services. Web services are a cost-effective way to leverage GIS resources to support users throughout the organization and associated user communities.

Network Services: Intranet applications can access services provided by ArcGIS Server connecting directly through the server object manager. Network services can be used to support a variety of Web and network applications.

Mobile Applications: A growing number of GIS operations are supported by more loosely connected mobile GIS solutions. ArcGIS technology supports continuous workflow operations that include disconnected editing and remote wireless operations. A disconnected architecture solution can significantly reduce infrastructure costs and improve user productivity for some operational workflows. Leveraging mobile services can provide alterative solutions to support a variety of user workflow environments.

Selecting the proper software and architecture deployment strategy can have a significant impact on user workflow performance, system administration, user support, and infrastructure requirements.

3 Network Communications

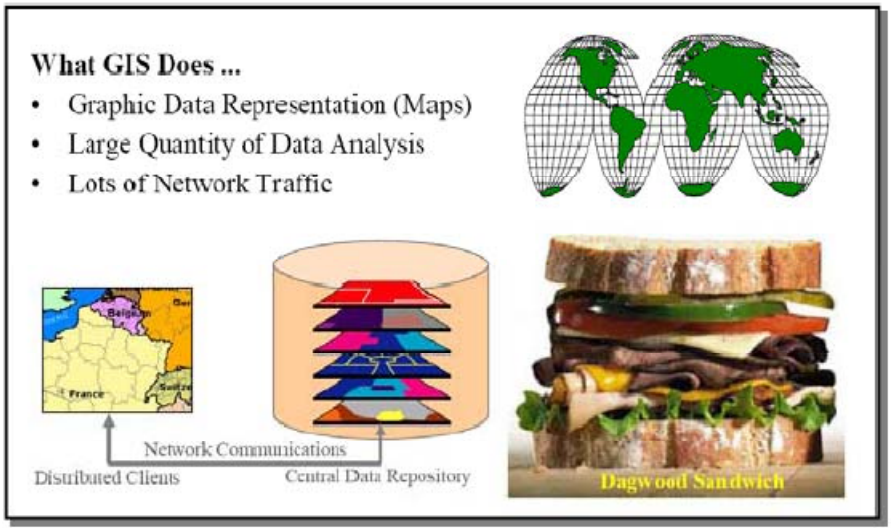
Network communications provide the required connectivity to support distributed computer processing. Network products support a stable and dependable communication protocol for distributed data transport. A variety of communication protocols supports distributed applications and data resources located at sites throughout an organization.

For several years, network technology was a relatively static environment while computer performance was increasing at an accelerating rate. Recent advances in communication technology support a dramatic shift in network solutions, introducing worldwide communications over the Internet and bringing information from millions of sources directly to the desktop in real time.

3.1 Desktop Workstation Environment

GIS applications rate among the heavy users of network traffic along with document management and video conferencing as shown in figure 3-1. GIS technology provides a visual display environment to the user supporting very quick analysis of large amounts of graphic data. Access to distributed data sources for real-time display and analysis puts large demands on network communications. Data must be transported across the network to where the program is executed to display the information.

Figure 3-1
GIS Applications Network Impact



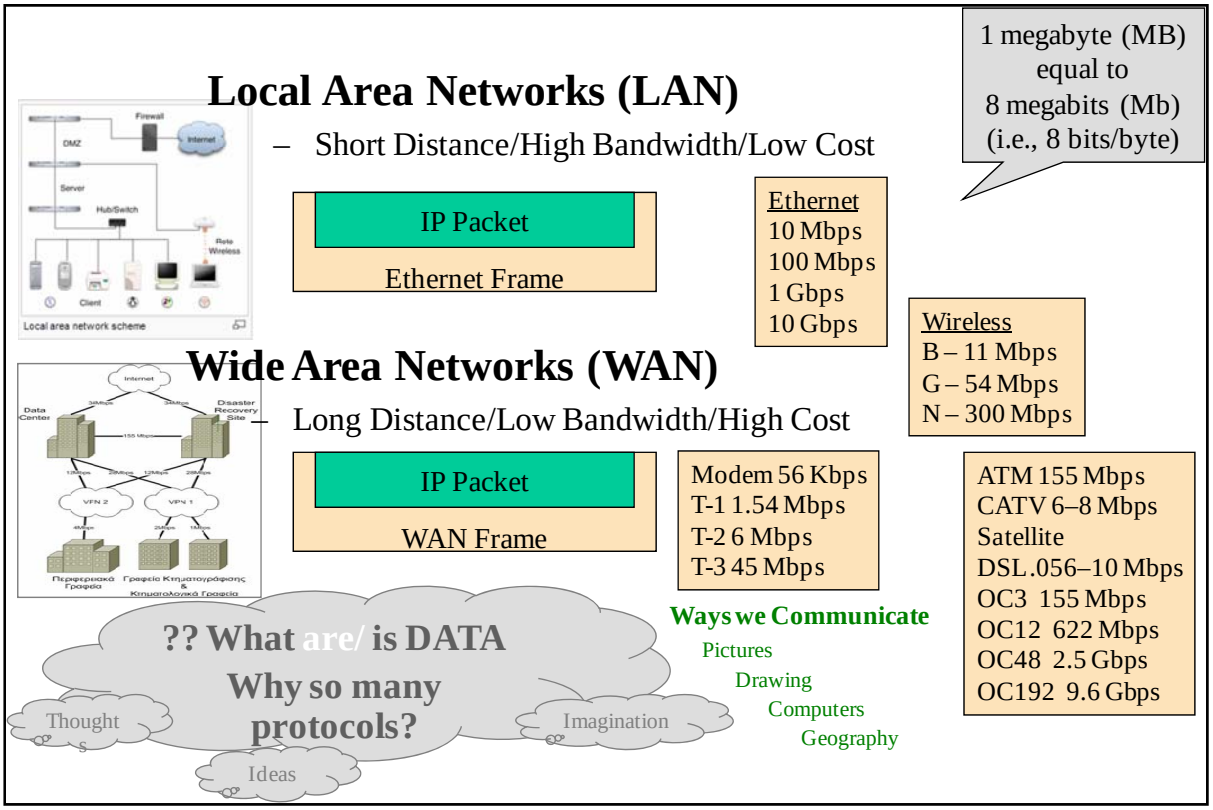
Data is a collection of digital computer information stored in media that have the capability to record and retain the data structure. This data is represented by little pieces of information called bits. Each bit takes up the same space on storage or transmission media. For convenience, these little bits can be grouped into bytes of information with each byte containing eight bits. Data can be transported from one location to another within packets that protect the integrity of this data.

Data is typically transported from one storage location to another over copper wire or glass fiber physical networks. Other types of transport media include microwave, radio wave, and satellite digital transmissions. Each network protocol has limits to its supported rate of data transport based on the technology used to support the transmission.

Network transport solutions can be grouped into two general technology classes. These classes include LANs and WANs. The volume of data (measured in bits) that can be transported per second represents the transport rate (capacity) of a specific network segment. This capacity is called network bandwidth and is typically measured in millions of bits (megabits [Mb]) or billions of bits (gigabits [Gb]) per second.

Figure 3-2 illustrates the different types of networks and following are their descriptions.

Figure 3-2
Types of Networks



Local Area Networks: Local area networks support high-bandwidth communication for short distances. This environment supports local operating environments (usually within a building or local campus environment). Data transport over a single technology is single threaded, which means only one data transmission can be supported on a single LAN segment at any time. The cost for LAN environments is inexpensive relative to other hardware costs supporting the system environment.

Wide Area Networks. Wide area networks support communication between remote locations. Technology supports much lower bandwidth than LAN environments, but transmission is possible over long distances. This is a data transport environment, which means data is packaged in a series of additional packets and transported as a stream of data along the transmission medium. The cost for WAN connections is relatively high compared to LAN environments.

Wireless Communications. Wireless communications use radio frequency bands as a data transport media. Radio frequencies used for wireless transmissions connect user devices to area communication transceivers, and the transceivers connect the communications to the local or wide area communication networks. Wireless communications broadcast over shared public frequencies and tend to experience higher latency than hard wire connections.

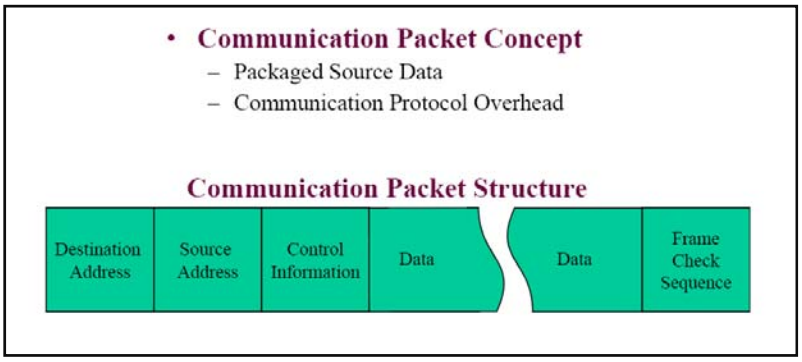
Data Units. Data capacity is measured in terms of megabytes or gigabytes when stored on computer disk. Megabyte is abbreviated using a large "B," while megabit is abbreviated using a small "b." One must remember 1 MB = 8 Mb when converting data volume from disk storage to network traffic. Network traffic also includes some protocol overhead, so a simple rule-of-thumb is to translate 1MB of data to about 10 Mb of traffic.

3.2 Client/Server Communication Concept

Applications move data over the network through proprietary client/server communication protocols. Communication processes located on the client and server platforms define the communication format and address information. Data is packaged in communication packets, which contain communication control information required to transport the data from its source client process to the destination server process.

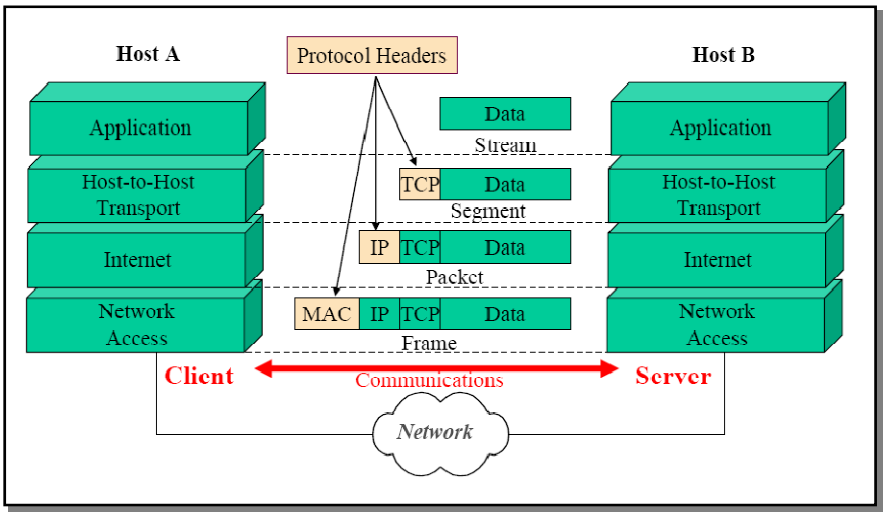
Communication Packet Structure. The basic Internet packet structure as shown in figure 3-3 includes destination and source addresses and a series of control information in addition to the data structure. This information supports delivery of the packet across the network medium. The IP packet size will vary depending on the amount of data. The largest IP packet is around 65 kilobytes (KB). Ethernet frames are limited to 1.5 KB. Data can be distributed across several packets to support a single data transfer.

Figure 3-3
Communication Packet Structure



Network Transport Protocol. The communication packet is constructed at different layers during the transmission process. In figure 3-4, a data stream from the host A application is sent through the protocol layers to establish a data frame for network transmission. The Transmission Control Protocol (TCP) header packages the data at the transport layer, the Internet Protocol (IP) header is added at the Internet layer, and the Media Access Control (MAC) address information is included at the physical network layer. The data frame is then transmitted across the network to host B where the reverse process moves the data to the host application. A single data transfer can include several communications back and forth between the host applications.

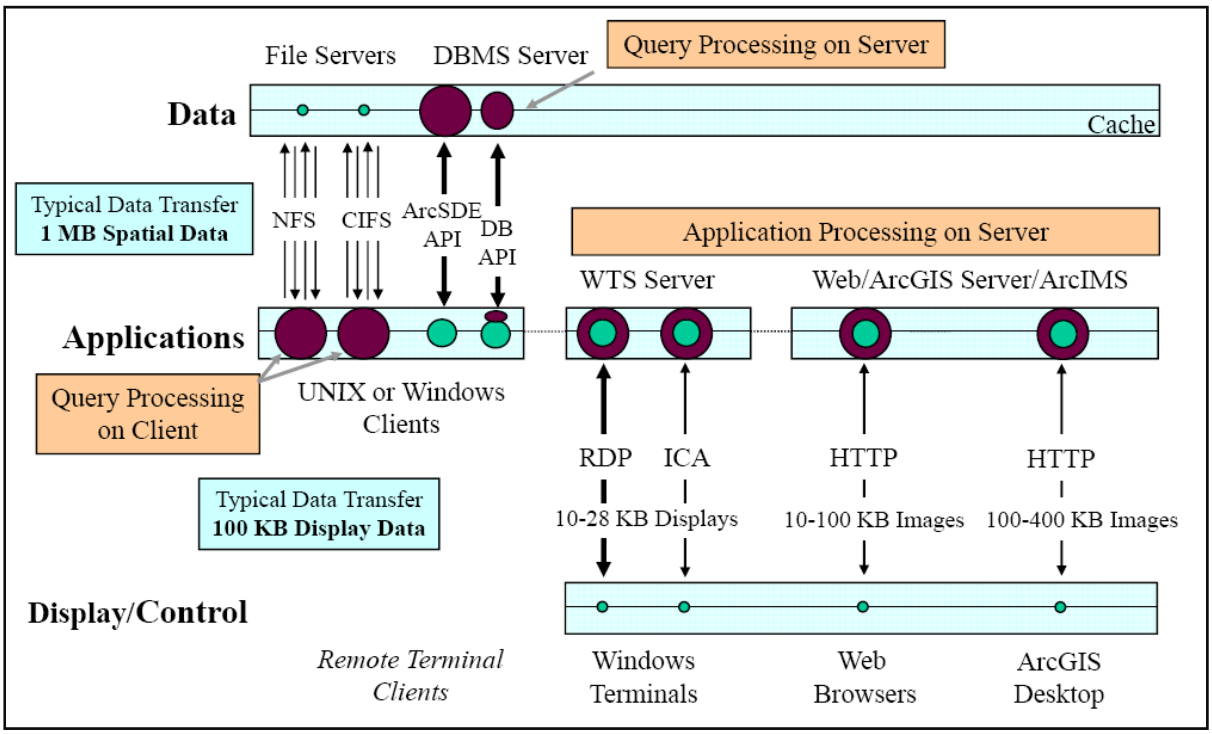
Figure 3-4
Network Transport Protocol



3.3 Client/Server Communications

Figure 3-5 shows several client/server communication protocols available to support network data transfer. Each protocol implementation includes client and server components to support data delivery. The client process prepares the data for transmission, and the server process delivers the data to the application environment. Following are primary protocols used by GIS solutions.

Figure 3-5
Client/Server Communication Protocols



Network file services (NFS) (UNIX) and Common Internet File System (CIFS) (Windows) Protocols: Support remote disk mounting enabling a client application to access data from a distributed server platform. All query intelligence resides in the client application, directing access to data located on the server platform. Data must be transferred to the client application to support analysis and display.

Database Access Protocols: ArcSDE includes client and server communication components. The server component includes intelligence to support query processing. Compressed data is provided for transfer and uncompressed by the client application. Data must be transferred to the client application to support analysis and display.

An alternative ArcSDE client direct connect option is available that connects with a DBMS client application program interface (API) executed on the client desktop. The ArcSDE middleware functions are supported on the client platform, and the DBMS network client supports data transmission to the server. Query processing remains on the DBMS server.

ICA and RDP Protocols: These support remote terminal display and control of applications supported by a host Windows Terminal Server. Both protocols transmit display and control information to the terminal client. Both the Independent Computing Architecture (ICA) protocol and Remote Desktop Protocol (RDP) compress data for transmission.

HTTP Protocol: The Hypertext Transfer Protocol (HTTP) is a standard Web transmission protocol. In this transaction-based environment, product selection and display are controlled by the browser client. Data is compressed for transmission. ArcGIS Desktop applications can also access ArcIMS as a data source. Traffic for the ArcGIS Desktop is higher because of the larger image transfers. Image size is

directly proportional to the physical screen display size; thus, larger image displays can result in higher traffic.

3.4 Client/Server Network Performance

The data transfer volume and the network bandwidth can be used to estimate minimum network transport times for a single map display transaction. A typical GIS application requires up to 1 MB of data to generate a new map display. A typical terminal environment requires 100 KB of data to support the display environment.

Figure 3-6 presents typical data transfer requirements in megabytes, shows the conversion to megabits for transmission, includes any compression of this data performed by the communication protocol, and identifies the total volume of data in megabits that must be transmitted (protocol overhead may be greater than what was used in this sample conversion). The minimum data transport times are calculated for five standard bandwidth solutions (56 kilobits per second [Kbps] for standard dial-up communications; 1.54 megabits per second [Mbps] for typical WAN communications; and 10 Mbps, 100 Mbps, and 1 gigabit per second [Gbps]) for LAN communications. Any existing data traffic on shared network segments would increase these network transfer delays.

Figure 3-6
Client/Server Performance

Client/Server Communications	Network Traffic Transport Time				
	56 Kbps	1.54 Mbps	10 Mbps	100 Mbps	1 Gbps
File Server to Workstation Client (NFS)					
•1 MB => 10 Mb + 40 Mb = 50 Mb	893 Sec	32 Sec	5 Sec	0.5 Sec	0.05 Sec
Geodatabase Server to Workstation Client					
•1 MB => 10 Mb >> 5 Mb	89 Sec	3.2 Sec	0.5 Sec	0.05 Sec	0.005 Sec
Windows Terminal Server to Terminal Client (ICA)					
•Vector 100 KB => 1 Mb >> 280 Kb	5 Sec	0.18 Sec	0.03 Sec	0.003 Sec	0.0003 Sec
•Image 100 KB => 1 Mb	18 Sec	0.6 Sec	0.1 Sec	0.01 Sec	0.001 Sec
Web Server to Browser Client (HTTP)					
•Light 100 KB => 1 Mb	18 Sec	0.6 Sec	0.1 Sec	0.01 Sec	0.001 Sec
•Standard 200 KB => 2 Mb	36 Sec	1.2 Sec	0.2 Sec	0.02 Sec	0.002 Sec
Web Server to ArcGIS Desktop Client (HTTP)					
•Light 200 KB => 2 Mb	36 Sec	1.2 Sec	0.20 Sec	0.02 Sec	0.002 Sec
•Standard 400 KB => 4 Mb	72 Sec	2.4 Sec	0.40 Sec	0.02 Sec	0.004 Sec

Best Practices

User Workflows Traffic Flow Times

During peak work periods, operational workflow performance can slow to a crawl similar to what is experienced in larger cities when driving on major highway arteries during rush hour. This simple illustration identifies the primary cause for many remote client performance problems. Sufficient bandwidth is critical to support productive user workflow performance requirements.

File server configurations support query from the client applications. When selecting data from a file (coverage or shape file), the total file must be delivered to the client for processing. Data not required by the application is rejected at the client location. This accounts for the considerable amount of network overhead experienced by these communications.

ArcSDE client/server configurations support query processing on the server platform. The query processing includes locating the requested data and filtering that data so only the specific data extent requested by the client is returned over the network. If the client limits requests to a small volume of data (e.g., point data or a single parcel in a parcel layer), the resulting data transfer can be very small, and network transport performance would

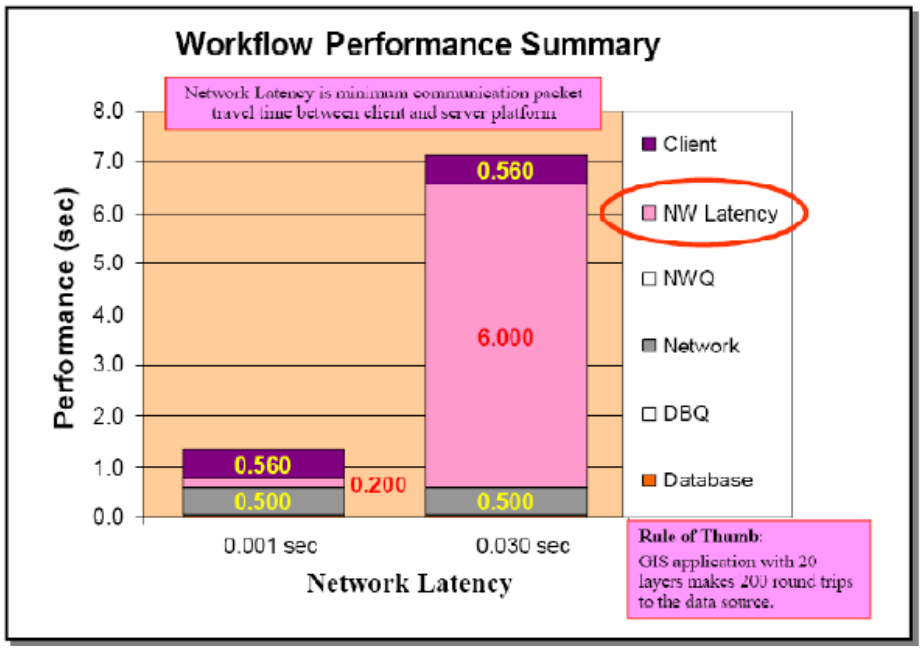
improve accordingly.

Best practices are established for network configuration alternatives. Distributed client/server application environments generally perform better in a local area network environment. Transaction-based services, or persistent Windows Terminal Server connections, provide stable processing environments for processing over less stable wide area network connections.

3.5 Performance Latency Considerations

The maximum bandwidth capacity used by a single user is limited by the total system transaction time. With standard client/server database display transactions, hundreds of data requests are sent to the server spread throughout the display transaction time (ArcGIS Desktop provides sequential requests for each layer in the display, completing each layer transaction before sending requests for the next display layer). A typical map display may have 10–20 data layers supporting the display analysis, which can translate to hundreds of sequential database transactions. Figure 3-7 shows a typical map display profile, showing the client desktop processing time, network transport time, and database processing time required to support a typical map display.

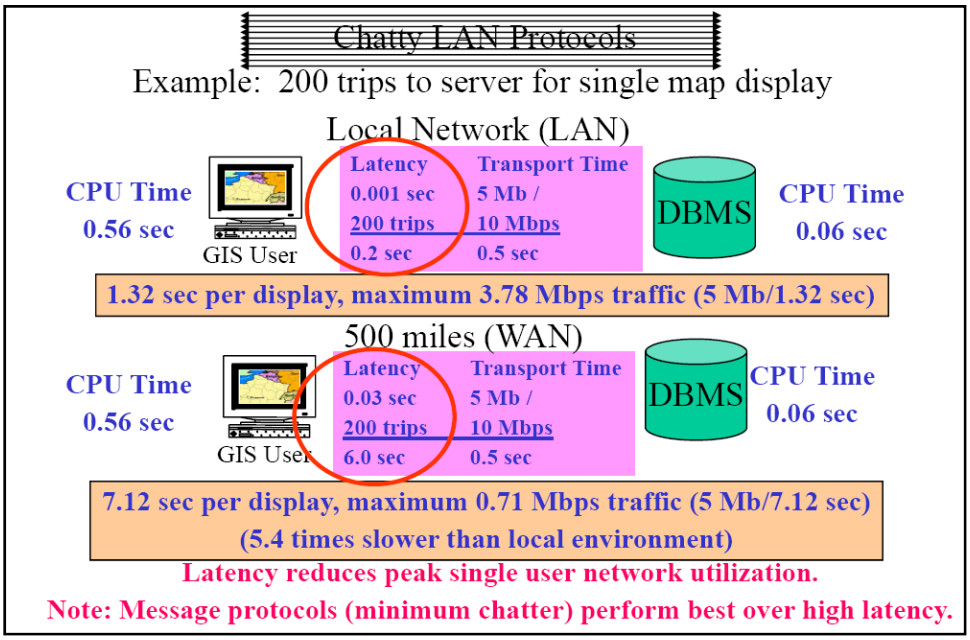
Figure 3-7
System Performance Latency



Bandwidth capacity is typically measured in megabits per second. During a typical local map display transaction, the total transaction time is approximately 1.32 seconds. A total of 5 Mb of data must be transferred from the server to the client application to support a typical map display (see figure 3-6). The average bandwidth utilized by a single user on the LAN is 3.78 Mbps. Increasing the desktop network interface to 1 GB bandwidth would have limited user performance impact when accessing an ArcSDE data source.

Normal database access protocols are "chatty," which means a typical database query requires a large number of trips to and from the server to complete the client display transaction. There are many trips to and from the server (query transactions) for each layer in the map display. Figure 3-8 shows 200 query transactions are required to support a single map display.

Figure 3-8
Performance Latency Considerations



Network latency can impact bandwidth utilization over long WAN distances. Latency is easy to measure, using a simple Windows command prompt (tracert "server host name"). Results provide the number of network hops and the associated network latency time for a single trip.

For LAN environments, network latency is very low (typically < 0.001 milliseconds per trip to the server). Many trips between the server and client have limited performance impact. The primary system latency contribution is client and server processing service times.

For longer WAN distances that involve several router hops, there can be a measurable network latency delay, and for chatty database protocols, network latency can have a measurable performance impact. In the example above, total transaction time over the WAN (including cumulative network latency) is 7.12 seconds. The maximum bandwidth utilized by a single user on this WAN connection is 0.71 Mbps. A single user would not experience performance improvements with increased WAN bandwidth. These days, many global WAN connections include satellite communication links. The fastest communication transfer is limited by the speed of light, which for very long distances will require a minimum bandwidth latency that technology will not overcome. Good performance over WAN environments results from protocols that minimize trips (communication chatter) between the client and server platforms.

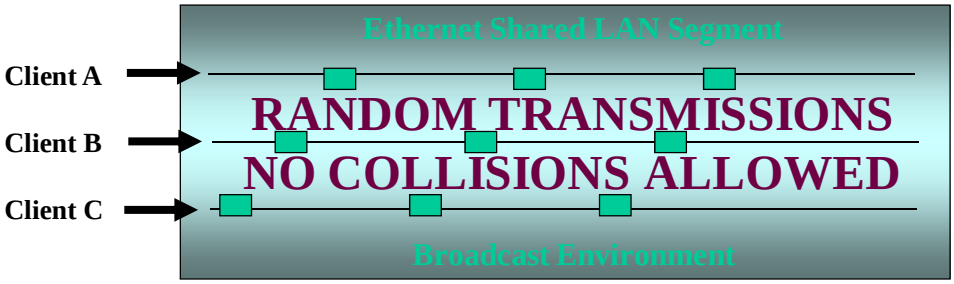
3.6 Shared Network Capacity

The total number of clients on a single network segment is a function of network traffic transport time (size of data transfer and network bandwidth) and the total number of concurrent clients. Only one client data frame can be transmitted over a shared network segment at any time.

With older switch technology, multiple transmissions on the same Ethernet network segment would result in collisions, which require another transmission to complete data frame delivery. Ethernet segments fail rapidly during saturation because of the rapidly increasing number of transmissions. Ethernet switches today include provide options for configuring shared segments in a full duplex mode, which when configured properly take advantage of switch buffer cache and improve transmission efficiency.

Figure 3-9 shows multiple client sessions sharing the same network segment where each data exchange is represented by the small boxes. Only one data exchange can be supported at one time on the same network segment.

Figure 3-9
Shared Network Capacity



Acceptable collision/congestion avoidance
Older broadcast environments (25–35% bandwidth capacity)
Newer switched environments (40—60% bandwidth capacity)

Random arrival times introduce two types of traffic conditions

1. **Network capacity available to support throughput and random transmission delays**
   - **Workflow productivity can be maintained within available network bandwidth.**
   - **Average data transport time = traffic transfer time + average queue time**
2. **Required workflow traffic exceeds network capacity.**
   - **Workflow productivity must be reduced to keep traffic within network capacity.**
   - **Traffic delays increase rapidly as traffic throughput approaches network capacity.**

Shared networks must adjust traffic flow to accommodate random transmission arrival times. Concurrent transactions must wait for network connection, resulting in transmission delays. Delays will increase during heavy traffic loads as the network reaches saturation. Wait time is a function of network transport time (figure 3-6) and bandwidth utilization. Higher traffic per display and busier shared network segments contribute to longer wait times.

A GIS application may require 1 MB of spatial data, or up to 10 Mb of network traffic, to support a single map display as illustrated in figure 3-6. Applications can be tuned to prevent display of specific layers when the map extent exceeds defined thresholds. Only the appropriate data should be displayed for each map extent (e.g., it may not be appropriate to display individual parcel layers in a map showing the full extent of San Diego, California). Proper tuning of the application can reduce network traffic and improve display performance.

Figure 3-10
Typical 1 MB Map Display

(1:3,000 scale [feet], average features = 250)

3.7 Network Configuration Guidelines

Standard published guidelines are used for configuring network communication environments as shown in figure 3-11. These standards are application specific and based on typical user environment needs. Communication environments are statistical in nature, since only a percentage of user processing time requires transmission over the network. Network data transfer time is a small fraction of the GIS users' total response time (on properly configured networks). Network data transfer time can dominate response time when bandwidth is too small or when too many clients are on the same shared network segment.

Figure 3-11
Network Configuration Guidelines

- **Network Design Standards**
  - Represent Typical GIS User Workloads.
  - Address Statistical Application Use.
  - Provide Basis for Initial System Design.
- **Network Management**
  - Ongoing Traffic Management Task
  - Hardware, Application, and User Dependent
  - Strongly Affected by Work Environment and Changes in Computer Technology

The network must be designed to support peak traffic demands. The amount of traffic varies based on the different types of applications and user work patterns. Standard guidelines provide a place to start configuring a network environment. Once the network is operational, network traffic demands should be monitored and necessary adjustments made to support peak user requirements.

3.8 Shared Network Configuration Standards

Figure 3-12 provides recommended design guidelines for network environments. These guidelines establish a baseline for configuring distributed LAN environments. Four separate client/server environments are included for each network bandwidth. The number of recommended clients is based on experience with actual system implementations and does not represent worst-case environments. Networks should be configured with the flexibility to provide special support to power users whose data transfer needs exceed typical GIS user bandwidth requirements.

Figure 3-12
Network Design Guidelines

Local Area Networks	Concurrent Client Loads			
Bandwidth	**File Servers**	**SDE Servers**	**Windows Terminals**	**Web Services**
10 Mbps	1-2	10-20	50-200	50-100
100 Mbps	10-20	100-200	500-2,000	500-1,000
1 Gbps	60-120	600-1200	5,000-20,000	5,000-10,000
10 Gbps	600-1200	6,000-12,000	50,000-200,000	50,000-100,000
Wide Area Networks	**Concurrent Client Loads**			
Bandwidth	**File Servers**	**SDE Servers**	**Windows Terminals**	**Web Services**
56 Kbps Modem	NR	NR	1-2	1-2
256 Kbps DSL	NR	NR	2-5	2-4
768 Kbps DSL	NR	NR	5-15	5-10
1.54 Mbps T-1	NR	1-2	10-30	10-15
2 Mbps E-1	NR	1-2	12-40	12-20
6.16 Mbps T-2	NR	4-6	30-120	30-60
45 Mbps T-3	5-10	25-50	200-900	200-500

Network Bandwidth Peak Clients

3.9 Web Services Configuration Guidelines

Implementation of Web mapping services places additional demands on the network infrastructure. The amount of system impact is related to the complexity of the published mapping services. Map services with small (less than 10 KB) or a limited number of complex images will have little impact on network traffic. Large images (greater than 100 KB) can have significant impact on network performance.

Figure 3-13 provides an overview of network performance characteristics that should be considered when deploying a Web mapping solution. The top portion of the chart shows the maximum number of requests that can be supported over various WAN bandwidths based on average map image size. The bottom portion of the chart shows the optimum transmission time for various map images. Web information products should be designed to support user performance needs, which may be dominated by available bandwidth. Simple high performance map services should produce images from 50 KB to 100 KB in size to minimize network transport time (100 KB image requires more than 36 seconds of network transport time for 28 Kbps modem clients, limiting site capacity to a maximum of 5,544 requests per hour over a single T-1 Internet service provider connection). Higher complexity ArcGIS Server map services can generate images from 100KB to 200KB in size. ArcGIS Desktop users may request image services from 200 KB to 400 KB in size (image size varies with user display size and resolution). Users generally demand reasonable performance, or they will not be satisfied. Proper infrastructure bandwidth and careful map information product design are required to support high-performance Web solutions.

Figure 3-13
Web Services Network Performance

Wide Area Network Bandwidth	Peak Web Map Requests/Hour Based on Average Image Size						
	10 KB	30 KB	50 KB	75 KB	100 KB	200 KB	400 KB
56 Kbps Modem	2,016	672	403	269	202	101	50
1.54 Mbps T-1	55,440	18,480	11,088	7,392	5,544	2,772	1,386
6.16 Mbps T-2	221,760	73,920	44,352	29,568	22,176	11,088	5,544
45 Mbps T-3	1,620,000	540,000	324,000	216,000	162,000	81,000	40,500
155 Mbps ATM	5,580,000	1,860,000	1,116,000	744,000	558,000	279,000	139,500
Note: 1 KB = 10 Kb HTTP traffic							
Wide Area Network Bandwidth	**Image Transfer Time (sec) Based on Average Image Size**						
	10 KB	30 KB	50 KB	75 KB	100 KB	200 KB	400 KB
19 Kbps Modem	5	16	26	39	53	105	211
28 Kbps Modem	4	11	18	27	36	71	143
56 Kbps Modem	2	5	9	13	18	36	71
256 Kbps	0.4	1	2	3	4	8	16
512 Kbps	0.2	1	1	1	2	4	8
1.54 Mbps T-1	0.1	0.2	0.3	0.5	1	1	3
6.16 Mbps T-2	0.02	0.05	0.08	0.1	0.2	0.3	1
45 Mbps T-3	0.002	0.01	0.01	0.02	0.02	0.04	0.1
155 Mbps ATM	0.001	0.002	0.00	0.00	0.01	0.01	0.03

ArcGIS Server can support high performance complex services using a pre-processed data cache. The more intelligent clients (ArcGIS Desktop, ArcGIS explorer, Web applications with adobe flash clients, etc) are able to overlay Web based vector and image services on top of a high performance local cache layer. The local cache data can be sent from the server once and used by the client many time, since the images are stored on the local client machine. Data cache can reduce network transport requirements and increase display performance when configured and used properly.

ArcIMS services can include extract services that will support data download to Web clients over the Internet. The extract service extracts data layers from ArcSDE based on identified extent, zips the data into a compressed file, and downloads the data to the client. Similar Web services can be supported with ArcGIS Server and standard Web server file transfer applications. Figure 3-14 identifies minimum download times based on available bandwidth and the size of compressed data packages. Data downloads should be restricted to protect Web service bandwidth. Data downloads can very easily dominate available bandwidth and impact performance to other Web mapping clients.

Figure 3-14
Data Download Performance

Wide Area Network Bandwidth	Peak FTP Downloads/Hour Based on Average File Size				
	1 MB	5 MB	10 MB	20 MB	50 MB
56 Kbps Modem	17	3	2	1	0
1.54 Mbps T-1	462	92	46	23	9
6.16 Mbps T-2	1,848	370	185	92	37
45 Mbps T-3	13,500	2,700	1,350	675	270
155 Mbps ATM	46,500	9,300	4,650	2,325	930
Note: 1 KB = 10 Kb FTP traffic					
Wide Area Network Bandwidth	**File Transfer Time (sec) Based on Average File Size**				
	1 MB	5 MB	10 MB	20 MB	50 MB
19 Kbps Modem	526	2,632	5,263	10,526	26,316
28 Kbps Modem	357	1,786	3,571	7,143	17,857
56 Kbps Modem	179	893	1,786	3,571	8,929
128 Kbps	78	391	781	1,563	3,906
256 Kbps	39	195	391	781	1,953
1.54 Mbps T-1	6	32	65	130	325
6.16 Mbps T-2	2	8	16	32	81
45 Mbps T-3	0.2	1	2	4	11
155 Mbps ATM	0.1	0.3	1	1	3

Many network administrators establish and maintain metrics on network utilization, which help them estimate increased network demands when planning for future user deployments. Figure 3-15 identifies standard network design planning factors for typical GIS clients based on their target data source. These numbers are used in the following sections to project network bandwidth requirements to support planned GIS user deployments.

Figure 3-15
Network Design Planning Factors

Client Platform	Data per display		Traffic per display		Kbps Traffic per user	
	KBpd	Adj KBpd	Kbpd	Mbpd	6 dpm	10 dpm
File Server Client	1,000	5,000	50,000	50.000	5,000	8,333
Geodatabase Client	1,000	500	5,000	5.000	500	833
Terminal Client (vector)	100	28	280	0.280	28	47
Terminal Client (raster)	100	100	1,000	1.000	100	167
Web Browser Client	200	200	2,000	2.000	200	333
Web Browser Client	100	100	1,000	1.000	100	167
Web GIS Desktop Client	200	200	2,000	2.000	200	333
Web GIS Desktop Client	400	400	4,000	4.000	400	667

KBpd = Kilobytes per display Mbpd = Megabits per display
Kbpd = Kilobits per display dpm = Displays per minute productivity

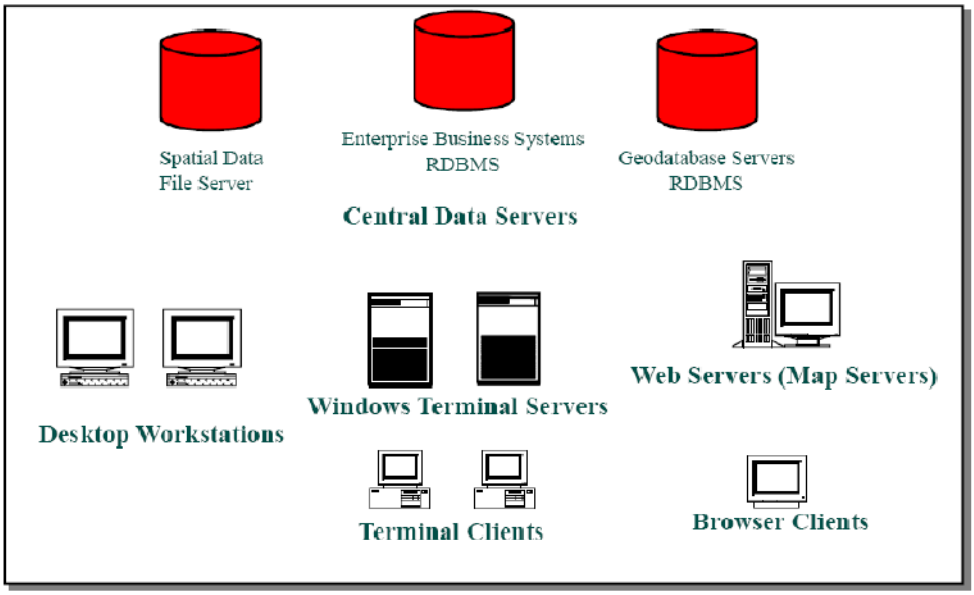
Capacity Planning Tool
Metrics

4 GIS Product Architecture

This section provides a foundation for understanding the software components and platform configuration options available to support distributed GIS operations. Understanding application architecture alternatives and associated configuration strategies provides a foundation for selecting an appropriate distributed GIS design.

Enterprise-level GIS applications support a variety of users throughout an organization, all requiring access to shared spatial and attribute data sources. System hardware and software environments for distributed GIS applications are supported by a multitier client/server or Web services architecture. A simple overview of the various architecture components is provided in figure 4-1.

Figure 4-1
GIS Multitier Architecture



Central Data Servers: Shared spatial and tabular database management systems provide central data repositories for shared geographic data. These database management systems can be located on separate data servers or on the same central server platform.

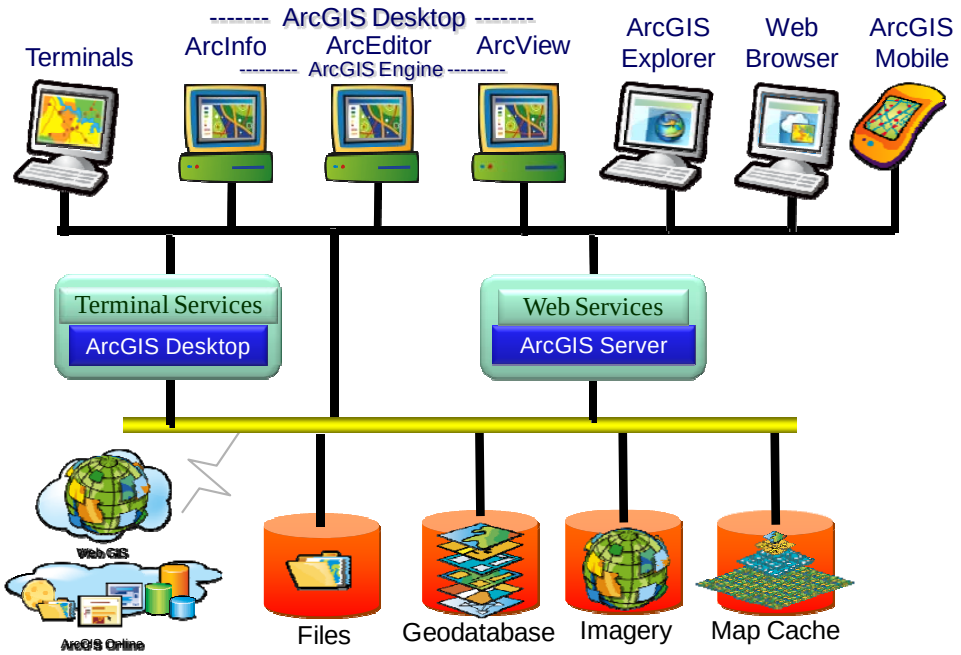
Application Servers: GIS applications are supported within the distributed configuration by hardware platforms that execute GIS functions. In a centralized solution, Windows Terminal Server and Web application server platforms can provide host compute services to a large number of concurrent GIS clients. Windows Terminal Servers host GIS desktop applications on centrally managed server farms allowing remote terminal clients to display and control applications executed on the terminal server platforms. Web application servers support a variety of Web applications and services accessed by standard browser clients or other desktop applications.

Desktop Workstations: Display and execution of application processes are supported by desktop workstations which, in many cases, are PCs that also can function as Windows terminal clients or Web browser clients. In many GIS solutions, the client application server and desktop user workstation are the same platform.

4.1 ArcGIS System Software Architecture

ArcGIS is an integrated collection of software products for building a complete geographic information system. The ArcGIS family of software products is used to deploy GIS functionality and business logic where needed—in desktops, servers, custom applications, Web services, and mobile devices. The ArcGIS applications are supported by a common set of ArcObjects components developed using Microsoft Component Object Model (COM) programming technology. Figure 4-2 provides an overview of the ArcGIS system environment.

Figure 4-2
ESRI ArcGIS System Environment

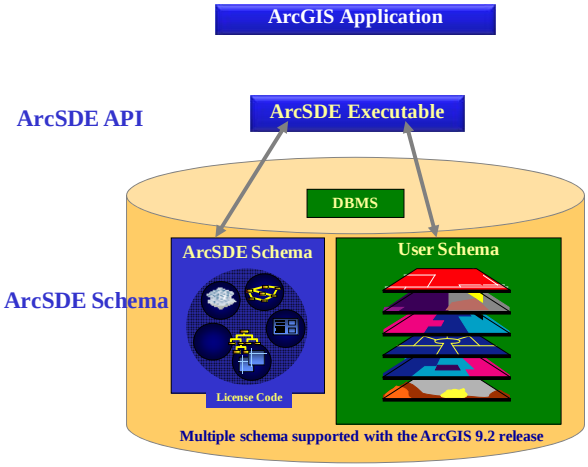


GIS is technology used for the creation, management, integration, analysis, display, and dissemination of spatial data. Spatial data includes any information that can be associated with a location on the earth's surface or data that can be associated with a person or place that has a location. Vector data types are used to represent geographic points, lines, and areas (polygons). Other spatial data types include scanned drawings, global positioning system (GPS) coordinates, satellite imagery, survey measurements, photogrammetry, and aerial photography, all of which can be georeferenced to establish proper placement within a geographic map display. GIS technology is currently being used in business, government, public safety, defense and intelligence, health and human services, utilities and transportation, education, and natural resources to manage and understand spatial relationships.

How the spatial data is maintained and published within the organization contributes to the performance and scalability of the system design. The volume of spatial data used to support a GIS has grown exponentially over the last 10 years. Many operational GIS environments maintain and support several terabytes of active GIS data resources with megabytes of data reviewed and processed within a typical user display session. This data must be organized and managed to support effective and efficient GIS operations.

The ArcGIS technology includes a spatial database engine (ArcSDE) for managing and publishing GIS data. Figure 4-3 provides an overview of the supported ArcSDE components.

Figure 4-3
ArcSDE Components



Note: ArcSDE executables are included in direct connect API.

Every ESRI software product includes an ArcSDE communication client. ESRI publishes standard data models that incorporate schema for a variety of GIS user environments. Geodatabase is a term used to describe spatial data resources stored in an ArcSDE user schema. ESRI provides a published API supporting an open development interface to the ArcSDE schema.

Figure 4-4 provides an outline of the system architecture design configuration strategies supporting current GIS enterprise operations.

Figure 4-4
ESRI Software Environments

ArcGIS Desktop Client/Server Configurations
- Distributed Workstation Architecture
- Centralized Windows Terminal Server Architecture

ArcGIS Server Web Services Architecture
- ArcGIS Server Component Architecture
- Platform Configuration Strategies

ArcGIS Desktop can be deployed on client workstations or hosted by a Windows Terminal Server. Custom ArcGIS Engine applications include the same ArcObjects components supporting the ArcGIS Desktop commercial software and share common configuration strategies. Different configuration alternatives are available to support communications between the client application and the GIS data source.

ArcGIS Server is deployed in a scalable Web application server architecture. The Web server solutions include a software component architecture supporting application development and system performance and scalability. Recommended platform configuration strategies will be provided for both standard and high availability Web solutions.

GIS applications support an open systems architecture. The GIS enterprise architecture combines a variety of closely integrated commercial products to establish a fully supported system solution. All commercial software products must be maintained to support evolving communication interface standards. The importance of selecting well established (popular) software architecture solutions based on standard design practices cannot be overemphasized, since all parts of the distributed configuration are critical and must work together to ensure communication interfaces are properly maintained and supported.

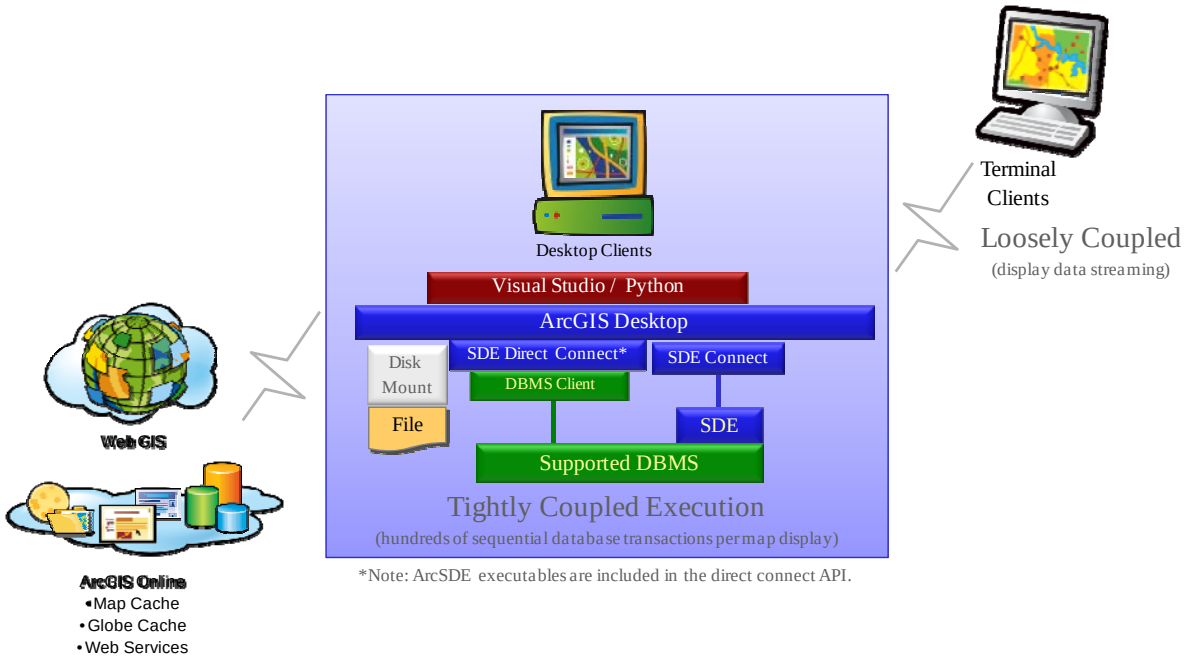
4.2 ArcGIS Desktop Client/Server Configurations

ArcGIS Desktop software is supported on Microsoft Windows desktop and terminal server platform environments. ArcGIS Engine is a software development kit providing ArcObjects components for custom desktop application development. Visual Studio is used as the primary programming language for ArcGIS Desktop applications and Python is the primary scripting language.

ArcGIS Desktop is deployed in a client/server architecture. The client applications are tightly coupled with the GIS data source, exchanging hundreds of sequential data requests to complete each user map display transaction. A typical map display is refreshed in less than a second, requiring a very chatty protocol exchange with the connected data source. Communications between the ArcGIS Desktop application and the GIS data source should be supported over stable high-bandwidth local network environments with minimum communication latency. Remote clients should be supported using terminal access to a central Windows Terminal Server located with the GIS data source. ArcGIS Desktop maintenance includes user licensed Web access to worldwide high quality image cache basemaps (Microsoft Virtual Earth - Bing Imagery).

Figure 4-5 provides an overview of the primary software components supporting the ArcGIS Desktop application workflow.

Figure 4-5
Client/Server Software Architecture

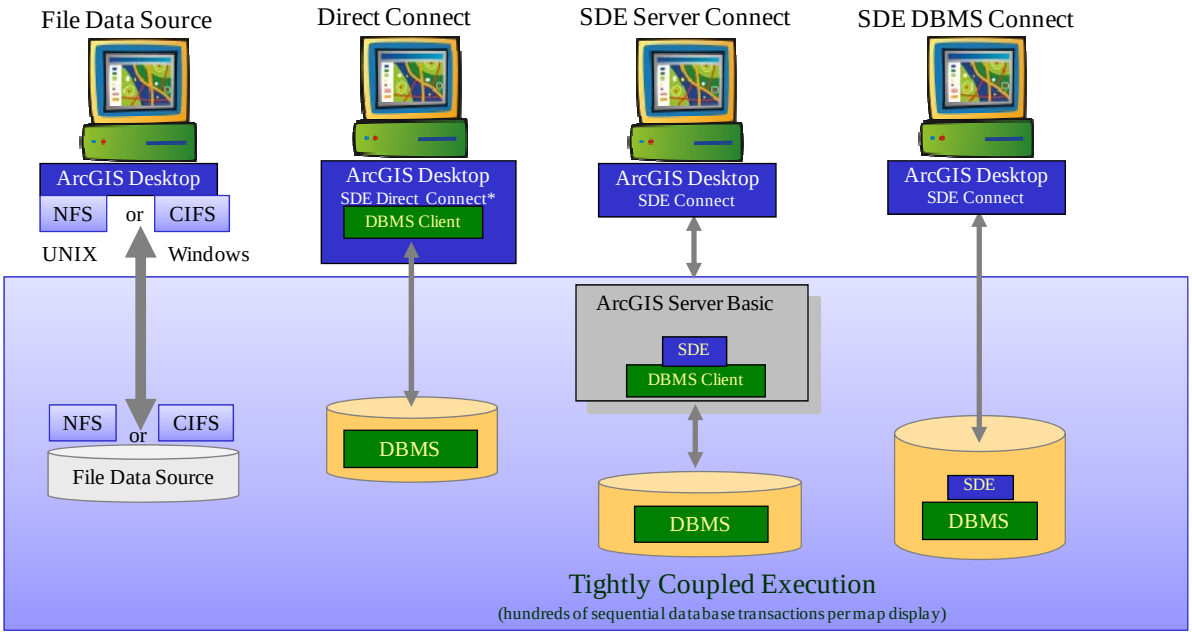


ArcGIS Desktop software supports connection to local file data sources, DBMS data sources through an ArcSDE interface, and published Web data services. Web data services can be integrated with local data in a standard GIS map display.

4.2.1 Distributed Workstation Architecture

Four distributed ArcGIS Desktop client configuration alternatives are identified in figure 4-6. These configurations include access to a network file data source, direct connect access to an ArcSDE data source, and two ArcSDE connect options to an ArcSDE data source (one connecting through a middle tier ArcGIS Server and the other connecting to ArcSDE installed on the database server).

Figure 4-6
Distributed ArcGIS Desktop Client



*Note: ArcSDE executables are included in the direct connect API.

The ArcGIS Desktop software will provide native file access to GIS data located on local disk. GIS applications can access a remote file data source by using Microsoft's CIFS or similar UNIX NFS. When mounting the remote disk, the remote file would appear as a local file share to the desktop application. Query processing for a file data source is supported by the ArcGIS Desktop application.

ArcGIS Desktop software provides two options for accessing geodatabase. The direct connect option includes the ArcSDE executables as part of the direct connect API and will communicate with a database client installed on the same machine. The database client will support network communications to the database server. The ArcSDE connect option supports network communication with a remote geodatabase server (ArcSDE can be supported as an ArcGIS Server geodatabase service or installed on the DBMS server). Query requests are sent to the data server and processed by the supported DBMS software. All data is stored and maintained in the DBMS repository.

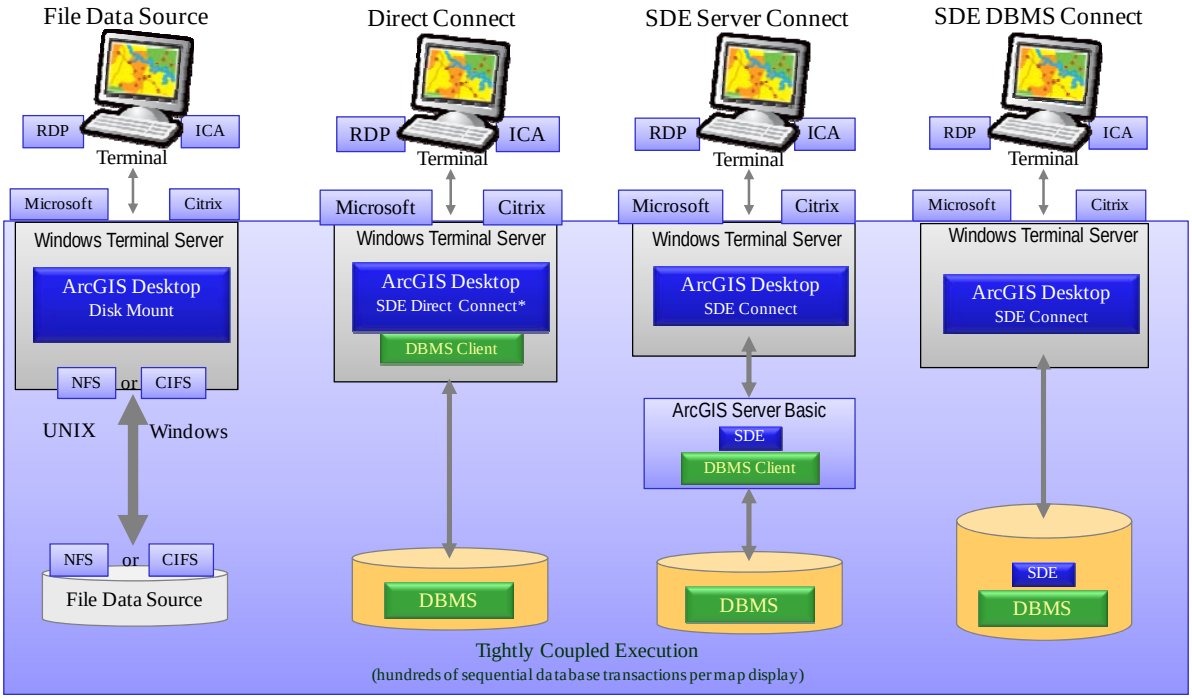
4.2.2 Centralized Windows Terminal Server Architecture

The Microsoft Windows Terminal Server product establishes a multiuser environment on a Windows server host. A Windows terminal client provides display and control of applications executed on the Windows Terminal Server. Microsoft uses a standard remote desktop protocol (RDP) to support communication between the terminal server and the Windows client.

The Citrix Presentation Server extension product enables a more efficient independent computing architecture (ICA) communication protocol to support communication between the terminal server and client Windows platform. The ICA protocol requires less than 28 Kbps bandwidth (rendering vector data information products) to support full Windows display and control of GIS applications supported on a Windows Terminal Server. Traffic can increase to 100 Kbps bandwidth when accessing a raster data source. Presentation Server includes client software for Windows, UNIX, Macintosh, and embedded Web client applications.

Four distributed ArcGIS Desktop configuration alternatives are identified in figure 4-7. All configurations are supported by remote terminal client access to ArcGIS Desktop applications supported on a central Windows Terminal Server. ArcGIS Desktop configuration options include access to a network file data source, direct connect access to a geodatabase, and two ArcSDE connect options to a geodatabase (one connecting through a middle tier ArcGIS Server and the other connecting to ArcSDE installed on the database server).

Figure 4-7
Centralized ArcGIS Desktop Client



RDP: Microsoft Remote Desktop Protocol
ICA : Citrix Independent Computing Architecture
*Note: ArcSDE executables are included in the direct connect API.

The Windows terminal client communicates with the Windows Terminal Server through a compressed message-oriented communication protocol. Terminal clients have a persistent connection with the Windows Terminal Server ArcGIS Desktop session; lost connections are reinstated without losing the session. The application display is provided over the network to the terminal client, requiring much less data transfer than the spatial data query chatter between the application and the data source. The terminal client display traffic requirements are very small; supporting good application performance over 28 Kbps modem dial-up connections (displays with an image backdrop may require more bandwidth).

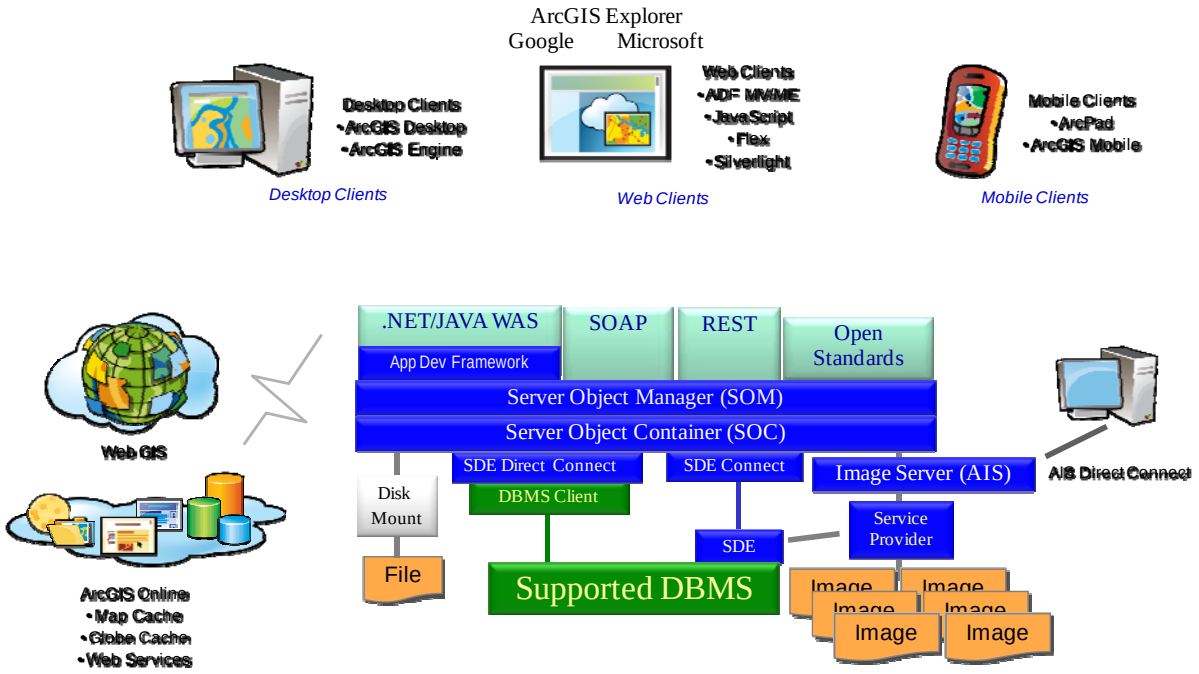
Each ArcGIS Desktop user session hosted on the Windows Terminal Server connects to each data source the

same way as a distributed client workstation session. Most current ESRI customers using Windows Terminal Server also use the Citrix Presentation Server software. The Windows Terminal Server farm is supported by commodity Windows server platforms (Intel or AMD). Client session load balancing across the terminal server farm is managed by the Citrix software. Client profiles and security options are provided to support an optimum GIS user display experience.

4.3 ArcGIS Server Web Services Architecture

Web mapping services provide an efficient and effective approach to serving map products and services over the Internet. The ArcGIS Desktop architecture presented earlier in this section requires tightly coupled client/server processes that demand stable high-bandwidth communications supported over relatively short distances. Web client communications are supported using a transaction-based HTTP, which supports optimum communications over long distances and less stable communication environments. Critical components associated with the Web services communication architecture are identified in figure 4-8.

Figure 4-8
Web Services Software Architecture



Web services are published on a Web server. Web server clients are presented with a catalog of published services when accessing the Web site. Web applications consume map services and render a client presentation layer to support the published application workflow. Client and Web servers are loosely connected in which each client communication represents a complete transaction. Transactions are processed by the appropriate Web-based GIS server and returned to the client.

ESRI Web GIS services are hosted by ArcGIS Server software. ArcGIS Server provides an ArcObjects software-based server development environment for deploying GIS server-based ArcGIS applications and services. ArcGIS Server can be deployed in a Web architecture or as a LAN/WAN network service for desktop clients. ArcGIS Server is also used to deploy "smart client" mobile GIS technology. Smart clients are loosely connected, lightweight, handheld or desktop computers that support persistent data cache and disconnected GIS client operations. Client application deployment and data synchronization are managed by ArcGIS Server parent services.

The software architecture components for ArcGIS Server include Web applications (WAS), service object manager (SOM), server object container (SOC), and data server (DS) component functions that can be deployed

on different platform combinations to support scalable capacity and system availability requirements. Location of the various software components and the selected software configuration can directly impact system capacity, service reliability, and overall output performance.

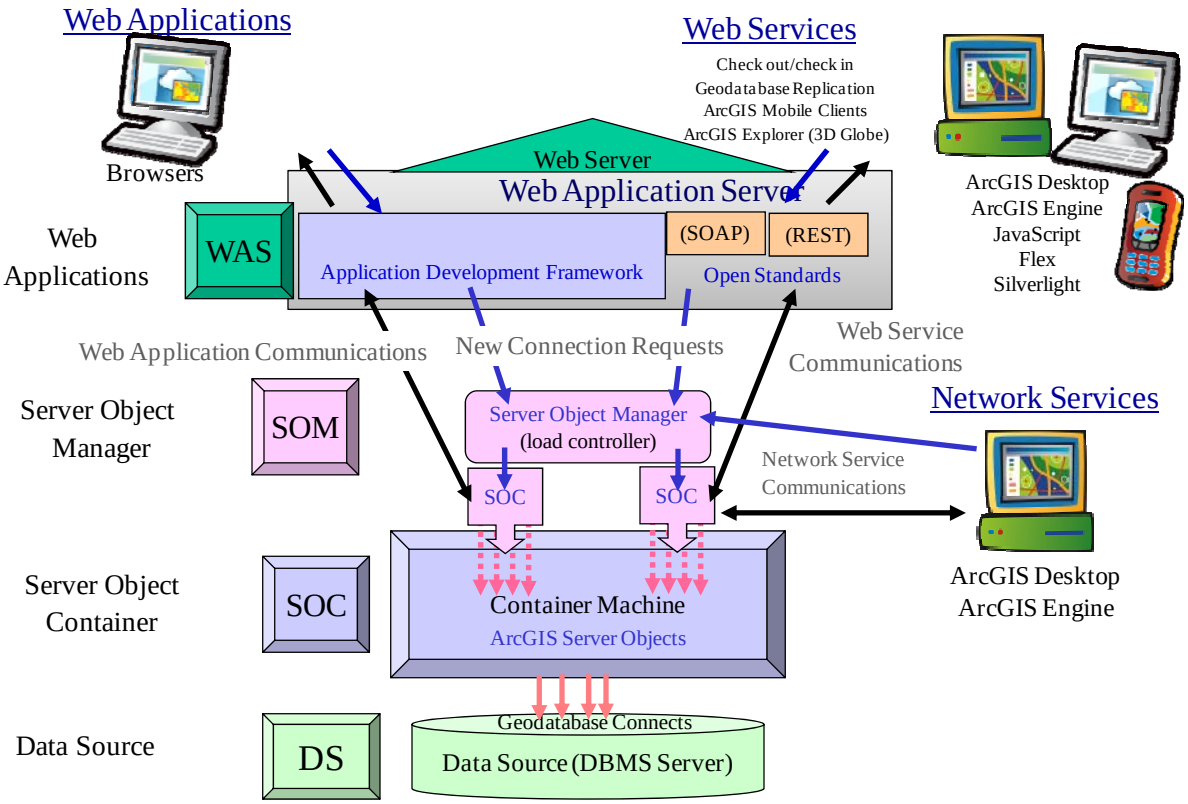
4.3.1 ArcGIS Server Component Architecture

ArcGIS Server is a set of objects, applications, and services that make it possible to run ArcObjects components on a server platform environment. Server objects are managed and run within the GIS server. Server applications make use of server objects and may also use other ArcObjects components that are installed on the GIS server.

The Web server hosts server applications and Web services developed using the ArcGIS Server application programming interface. These Web services and Web applications can be developed using the ArcGIS Server ADF, which is available for both .NET and Java developers and supported within the associated Web application server development environments.

Figure 4-9 provides an overview of the ArcGIS Server component architecture and the associated software functional locations. The ArcGIS Server architecture includes four configuration groups of software identified as Web applications, service manager, spatial services, and data source.

Figure 4-9
ArcGIS Server Component Architecture



The ArcGIS Server configuration groups include the following:

Web Applications: The ArcGIS Server WAS components include a commercial Web HTTP server supporting communications with the Web clients and a Web application server development environment for .NET or Java Web applications or Web service catalogs. ArcGIS Server includes a Web tier that includes a .NET and Java Application Development Framework. ArcGIS Server also includes software components for simple object access protocol (SOAP) and representational state

transfer (REST) Web service application program interfaces (APIs). ArcGIS Server includes translation of mapping services to a variety of Open Geospatial Consortium APIs, including WMS (Web Mapping Service for simple map images), WFS (Web Feature Service for streaming points, lines, and polygons), WCS (Web Coverage Service for raster and image steaming), CWS (Catalog Web Services for metadata searches), and KML (Keyhole Markup Language) for integration with Google Earth clients.

Server Object Manager: A server object manager (SOM) controls service object deployment and initial application assignment to deployed server object containers. SOM performs as a parent process, controlling service load balancing and managing deployment of published service configuration instances based on active inbound service request loads.

Server Object Container: The container machines (one or more depending on peak transaction requirements) host the server object containers (SOCs) that are managed by the SOM. Each service configuration is supported by dedicated SOCs. The server objects hosted within each SOC are supported by ArcObjects components installed on the container machine.

Data Source: The data server (GIS data source) is where the GIS data is stored. An ArcSDE data source supports the query processing functions. Standard GIS image or file data sources are also represented at this level.

Services available on ArcGIS Server can be published for use by intranet LAN or WAN applications. Application service assignment can be provided with direct access to the SOM without using the Web server interface.

4.3.2 Web Platform Configuration Strategies

A Web site can be supported with as few as one platform or as many as six or more platforms, depending on site capacity and availability requirements. ArcGIS Server platform configuration strategy can adapt to the required system reliability, availability, and security needs. This section will address the Web software component configuration strategies. Web security requirements will be addressed in more detail in section 5. ArcGIS Server performance metrics will be included later in sections 7 and 8 to address system performance and capacity planning.

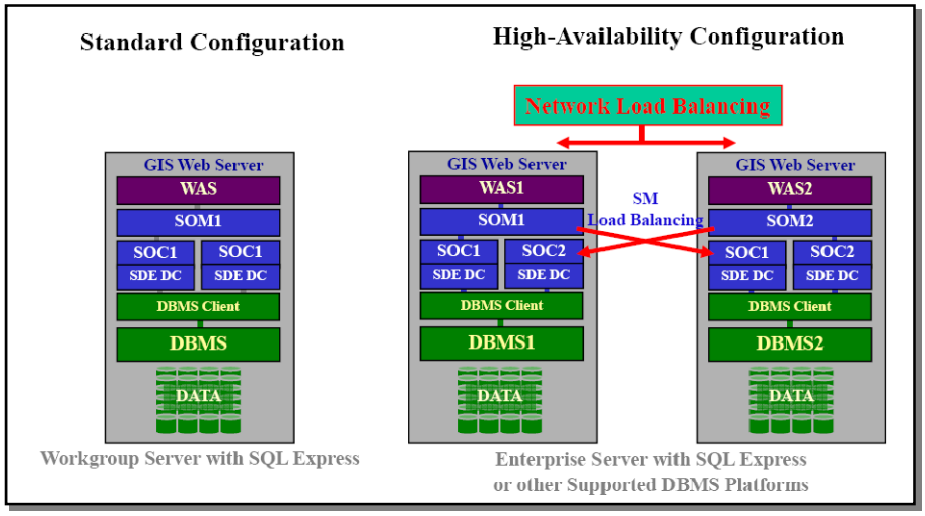
The Web system architecture design alternatives are grouped as single-tier, two-tier, and three-tier configurations. Simple configurations are easier to maintain and support. More complex configurations satisfy higher-capacity and system availability requirements. Production operations are normally supported with high availability configurations (configurations that will continue providing services following any single platform failure).

ArcGIS Server is designed to support a scalable Web architecture. Optimum platform environments are configured using standard commodity server platform technology. ArcGIS Server is licensed based on the number of platform processor core supporting the primary ArcGIS Server software components. Following are recommended platform configuration strategies for supporting standard GIS Web services.

4.3.2.1 Single-Tier Platform Configuration

Figure 4-10 provides an overview of single-tier platform configurations. Single-tier configurations provide one or two platforms capable of supporting all Web service components. Most initial customer deployments with a small database can be supported by a single-tier architecture.

Figure 4-10
Single-Tier Platform Configurations



Standard Configuration: A complete Web site can be supported on a single hardware platform. This configuration is appropriate for Web service development and testing, sites with a limited number of service requests, and initial prototype deployments. A special single chip workgroup server license bundled with a Microsoft SQL Server database is available for customer sites that can be supported by a single platform configuration.

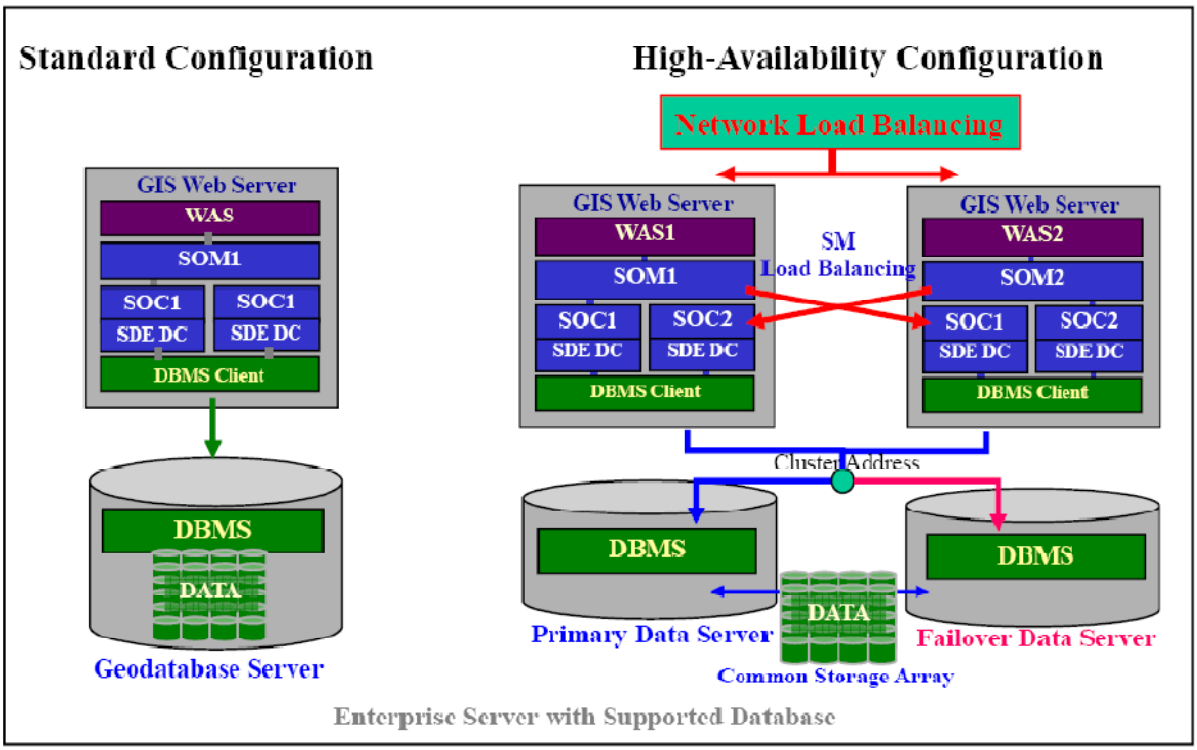
High-Availability Configuration: Most GIS server production operations require redundant server solutions, configured so the site remains operational in the event of a single platform failure. This configuration will continue to support production operations during single platform maintenance and upgrade and while configuring and publishing new services. This configuration includes (1) network load balancing to route the traffic to each of the servers during normal operations and only to the active server if one of the servers fails, (2) service manager load balancing to distribute spatial services processing load between the two platforms to avoid having requests back up on one server when extra processing resources are available on the other server (separate SOC containers are required on each platform to support this configuration), and (3) duplicated data servers that require a complete copy of the data.

4.3.2.2 Two-Tier Platform Configuration

A two-tier architecture provides an optimum solution for sites supported with a separate database server. The two-tier high-availability option may become the most popular and practical configuration supporting most ArcGIS Server deployments.

The two-tier architecture in figure 4-11 includes GIS server and data server platforms. The Web server and GIS server components are located on the GIS server platform, and the data server is located on a separate data server platform. This is a popular configuration for sites with large volumes of data resources or existing data servers. A single copy of the data can support multiple server components in conjunction with other enterprise GIS data clients.

Figure 4-11
Two-Tier Platform Configurations
(Separate Data Servers)



Standard Configuration: The standard configuration includes one GIS Web server platform with a separate single data server platform. The Web server is installed on the GIS Web server platform.

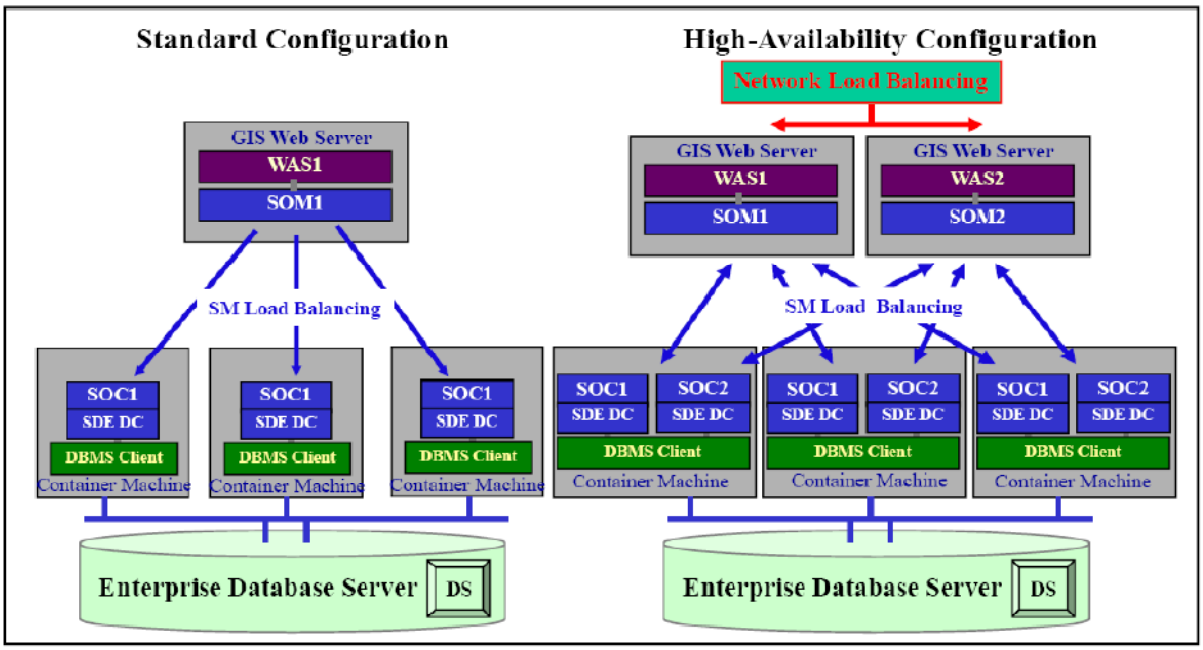
High-Availability Configuration: High-availability operations require redundant server solutions, configured so the site remains operational in the event of any single platform failure. This configuration includes (1) network load balancing to route the traffic to each of the GIS Web servers during normal operations and only to the active GIS Web server if one of the servers fails, (2) SOM load balancing to distribute SOC processing load between the two GIS Web server platforms to avoid having requests back up on one server when extra processing resources are available on the other server (two SOC container groups are required on each GIS Web server platform to support this configuration), and (3) two data servers that are clustered and connected to a common storage array data source. The primary data server supports query services during normal operations, and the secondary data server takes over query services when the primary server fails. Data server clustering is not required if availability requirements are satisfied with a single data server.

4.3.2.3 Three-Tier Platform Configuration

Three-tier configurations include Web server, map server/container machine, and data server tiers. Two configuration options are provided, based on the location of the ArcGIS Server SOM.

Figure 4-12 shows a three-tier configuration with the service manager located on the Web server tier. This configuration provides the simplest three-tier architecture (network load balancing handles Web tier failover), and would likely be the most popular solution. The three-tier configuration provides a scalable architecture, where the middle tier can support two or more platforms as required to support capacity requirements.

Figure 4-12
Three-Tier Platform Configuration—SOM on Web Tier

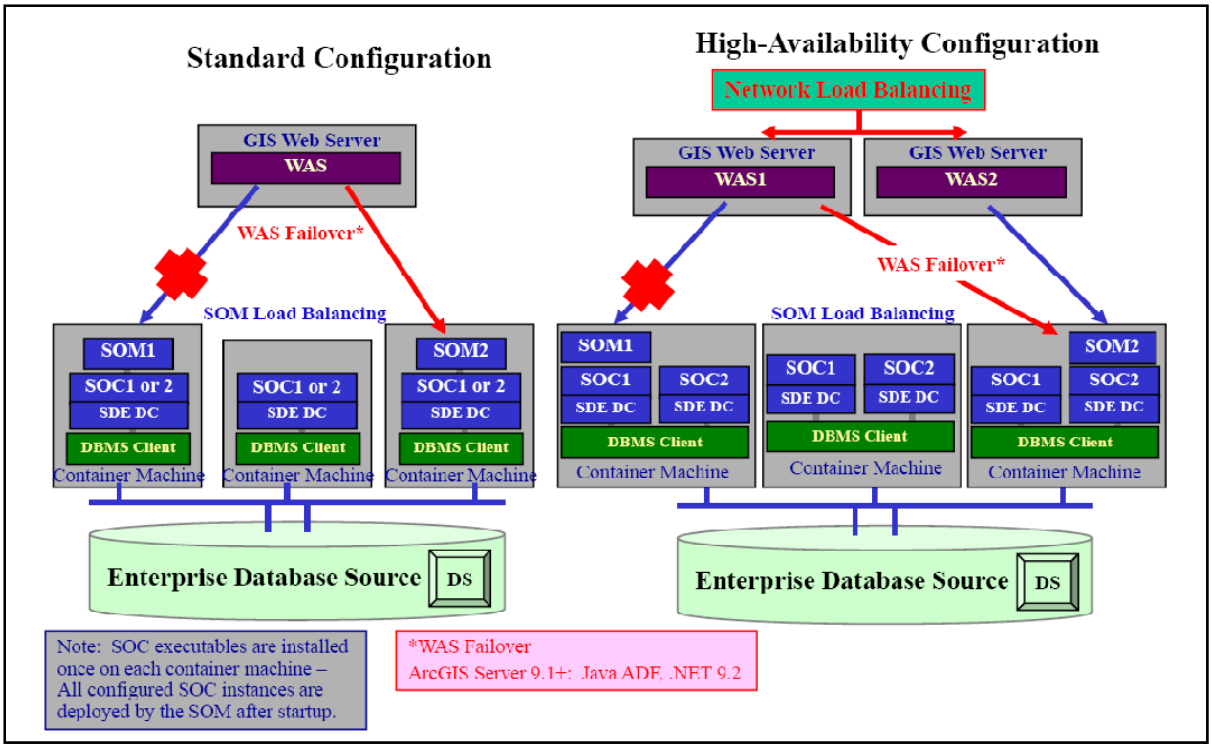


Standard Configuration: The standard configuration includes a single Web server with a separate container machine layer and a separate data server. The container machine layer can be a single platform or can be expanded to support several platforms, depending on the required site capacity. SOM load balancing is provided by the GIS Web server service manager.

High-Availability Configuration: High-availability operations require redundant server solutions, configured so the site remains operational in the event of any single platform failure. This configuration includes (1) network load balancing to route the traffic to each of the GIS Web servers during normal operations and only to the active GIS Web server if one of the servers fails, (2) SOM load balancing to distribute SOC processing load between the two map server/container machine platforms to avoid having requests back up on one server when extra processing resources are available on the other server (separate SOC containers are required on each map server/container machine platform to support this configuration), and (3) data server configuration supporting enterprise requirements.

Figure 4-13 shows a three-tier configuration with the SOM located on the container machine tier. The Web server and SOM connectors are located on the Web server platform, and the SOM and SOC components are located on the container machine platforms. This could be a preferred configuration when supporting Java applications on Linux-based Web servers. In this configuration, all the COM-based software is located on the container machine tier. The failover scenarios are more complicated. If the SOM1 software fails, WAS1 will send maps to SOM2. SOM2 will return results to the parent WAS2 output file. Client will return to WAS1 to get results and will need a virtual disk mount to receive the result from the SOM2 output directory. It will still be necessary to configure SOM load balancing for optimum capacity during peak loads.

Figure 4-13
Three-Tier Platform Configuration—SOM on SOC Tier



Standard Configuration: The standard configuration includes a single Web server with a separate container machine layer. The container machine layer can be a single platform or can be expanded to support several platforms, depending on the required site capacity. Web application traffic balancing is supported by the GIS Web server SOM connectors. The ArcGIS Server implementation can be configured in a failover mode (SOM2 would be activated only if SOM1 fails). SOM load balancing is provided by the server manager components (preferably no more than two SOM components on the container machine tier). All of the container machines can host SOM1 and SOM1 instances, so both SOM1 and SOM2 will deploy dedicated SOC instances on each host platform. Separate data server is provided as a common data source. Administration of this architecture can become increasingly complex as additional container machines are deployed - most of this complexity is managed by the SOM dispatch software.

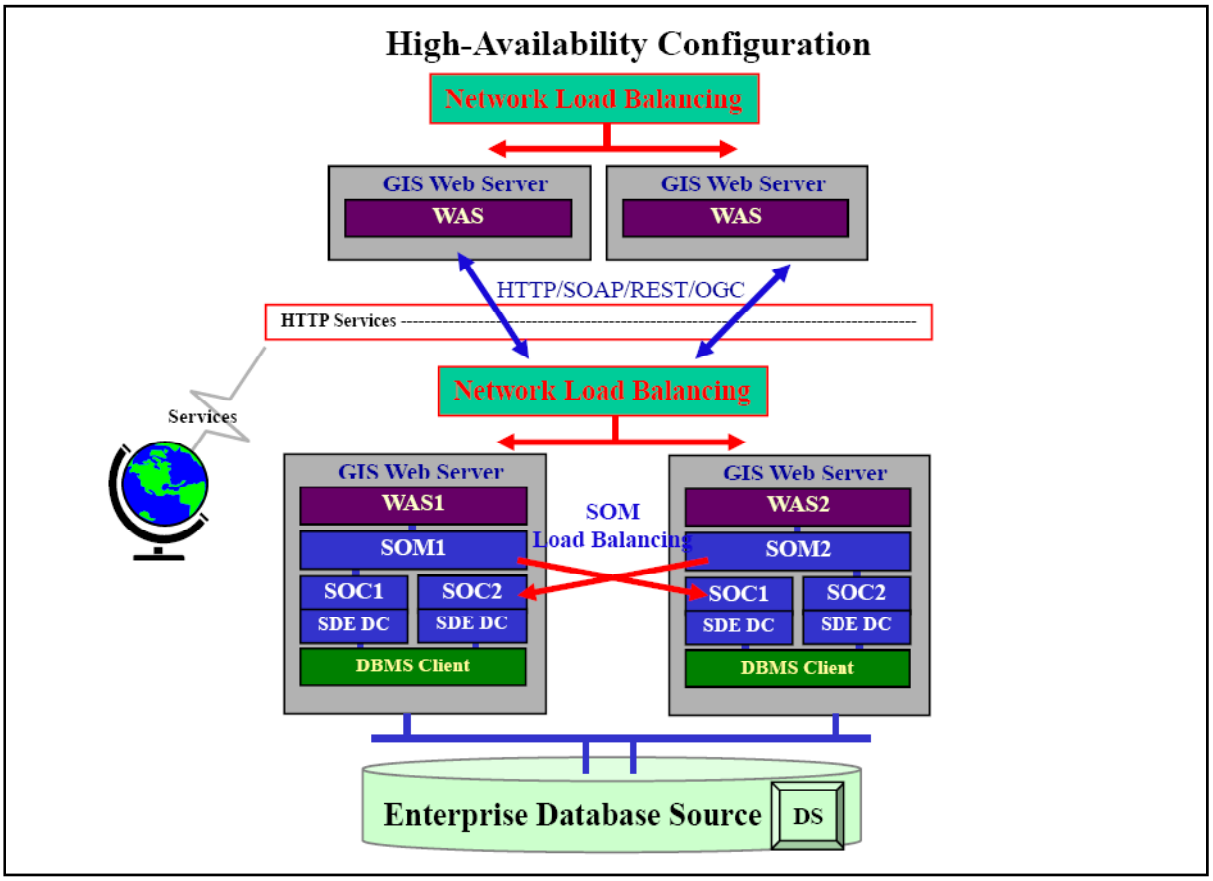
High-Availability Operations: High-availability operations require redundant server solutions, configured so the site remains operational in the event of any single platform failure. This configuration includes (1) network load balancing to route the traffic to each of the GIS Web servers during normal operations and only to the active GIS Web server if one of the servers fails, (2) Web application traffic load balancing to distributed inbound load between the two SOM located on the container machine tier. (3) SOM load balancing to distribute SOC processing load between the container machine platforms to avoid having requests back up on one server when extra processing

resources are available on the other server (dedicated SOC containers are required on each container machine platform to support this configuration, with each SOC assigned to a parent SOM), and (4) data server configuration supporting enterprise requirements. Administration of this architecture becomes increasingly complex as additional map servers/container machines are deployed - most of this complexity is managed by the SOM dispatch software.

4.3.2.4 Three-Tier Service-Oriented Platform Configuration

ArcGIS Server Web Applications can be developed and deployed supported entirely by remote Web data services. It is also possible to provide these HTTP SOAP based services from a separate local ArcGIS Server Web site. Figure 4-14 provides an example of a ArcGIS Server Web configuration that supports an enterprise services architecture configured across a firewall.

Figure 4-14
Three-Tier Platform Configuration—Web Services Architecture



The internal GIS Web Servers are configured as identified in the earlier configurations. Web data services can be published by the internal GIS Web Servers supporting enterprise applications deployed on a separate Web application tier. Web services can be passed through the firewall using standard HTTP service protocols.

Many of the more powerful ArcGIS Server applications benefit from a more tightly coupled DCOM communications. Each application is directly coupled to an assigned SOC to support each transaction. Results from these applications can be provided as services to more loosely coupled enterprise applications supported in a separate security zone by using standard HTTP protocols.

ArcGIS Server provides a broad range of functionality that can be used to support standard Web mapping or for more complex geospatial workflows that previously were not available through open standard Web protocols.

Use of the ArcGIS 9.3.1+ optimized map document (MSD) or preprocessed cached data layers available with ArcGIS Server can improve user performance beyond what we saw with ArcIMS technology and reduce network traffic requirements. Careful attention to user requirements and proper deployment of the technology can make a difference.

4.3.3 ArcGIS Server Image Extension Platform Configuration

ArcGIS Server Image Extension provides an efficient and effective approach to store and serve imagery. ArcGIS Server Image Extension is part of the ArcGIS Server 9.3 release. Critical components associated with the Web services communication architecture are identified in figure 4-15.

Figure 4-15
ArcGIS Server Image Extension Software

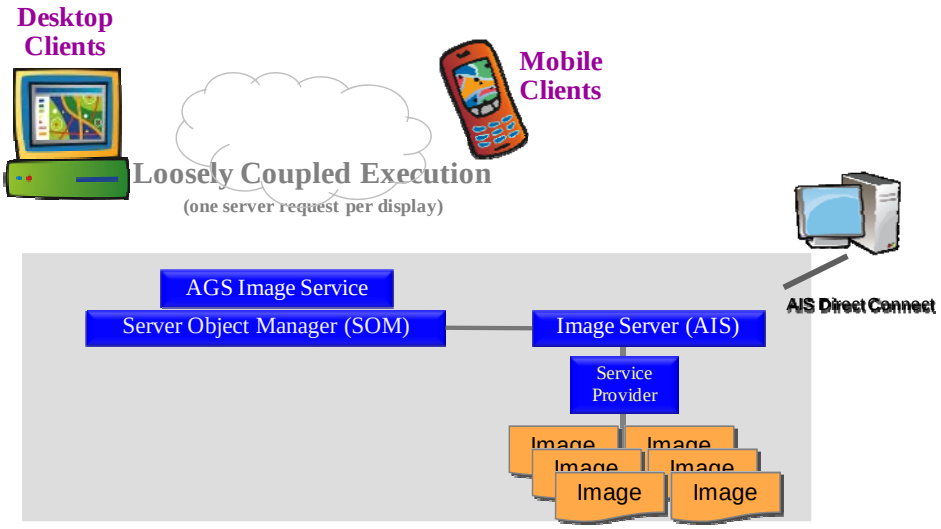
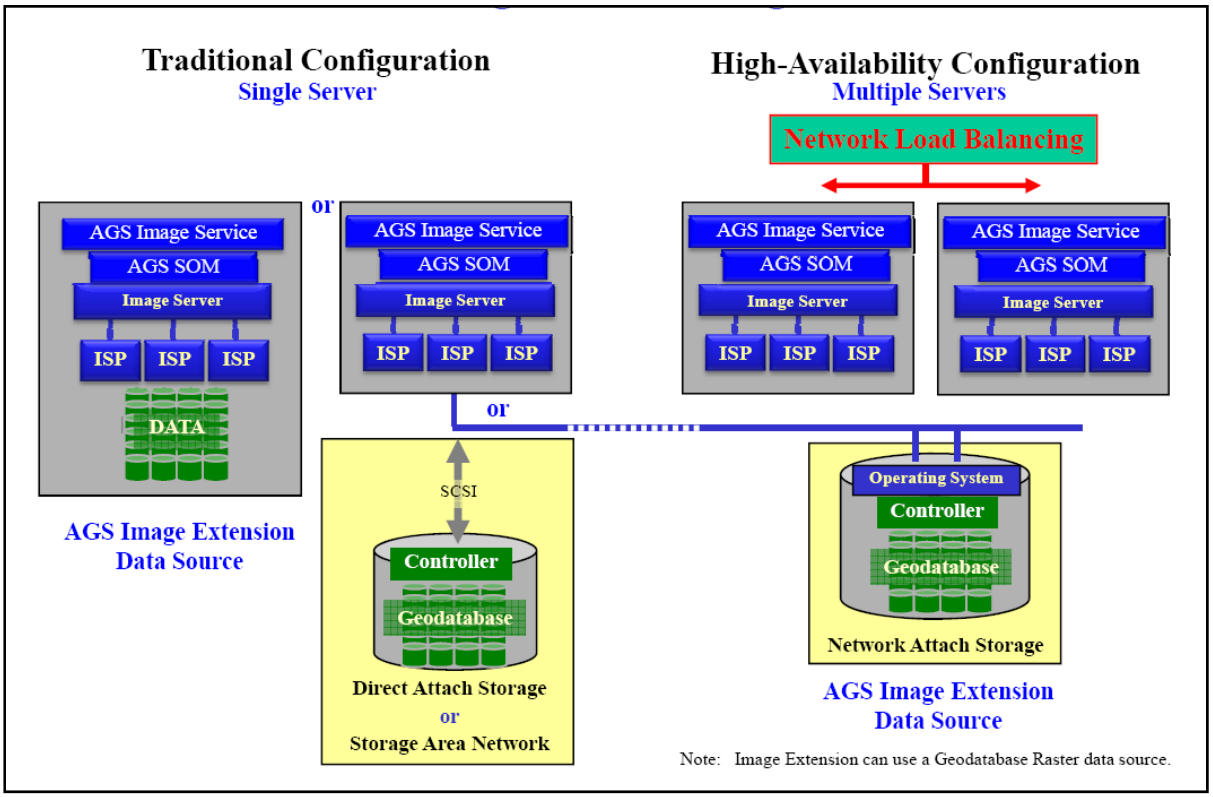


Figure 4-16 provides an overview of the ArcGIS Server Image Extension platform configurations. The Image Extension software can be configured as a standalone server with image files supported on local drives, direct attached storage, or a storage area network. Network attached storage can provide a common data source form multiple image service provider platforms required for high availability or high capacity configurations.

Figure 4-16
ArcGIS Server Image Extension Platform Configurations



ArcGIS Server Image Extension will allow legacy ArcGIS Image Server clients to connect direct to the Image Server interface for the ArcGIS 9.3.1 release. Image Server and Service Provider software components will be fully integrated with ArcGIS Server SOM and SOC instances with future releases.

5 Enterprise Security

This section provides an introduction to the types of security measures that should be considered in supporting enterprise operations. Implementation needs vary based on the type of operations and the associated threat environment.

Enterprise security can be a challenge for IT architects and security specialists. Until the last few years, entire IT systems were frequently designed around a single mission objective and a single "community of interest" normally supported with physically isolated systems, each with its own data stores and applications. New emerging standards are supported with more mature communication environments, more intelligent operating systems, and a variety of standard integration protocols enabling IT architects to design and maintain comprehensive organization-wide interactive enterprise solutions.

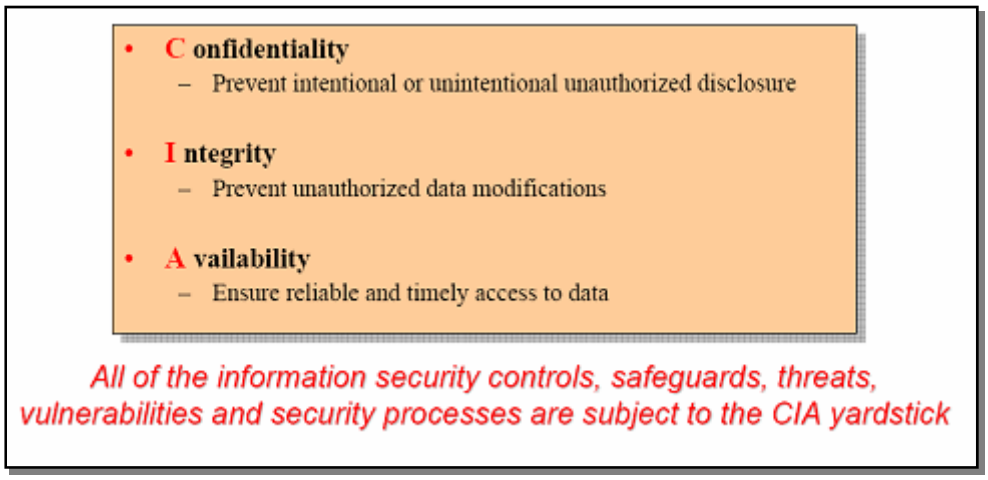
Recent industry advancements, especially in the areas of Web service standards and service-oriented architectures, are enabling architects to more effectively satisfy enterprise security objectives. ESRI's careful attention to these standards, coupled with an overall philosophy of providing highly interoperable software, provides security architects with a high level of flexibility, thus establishing trust for all ESRI components contained in an enterprise solution.

A full discussion on enterprise security is beyond the scope of this document. An ESRI white paper, ArcGIS Enterprise Security: Delivering Secure Solutions (July 2005), addresses ArcGIS configuration strategies to support enterprise operations. This white paper is available at http://www.esri.com/library/whitepapers/pdfs/arcgis-security.pdf

5.1 Security and Control

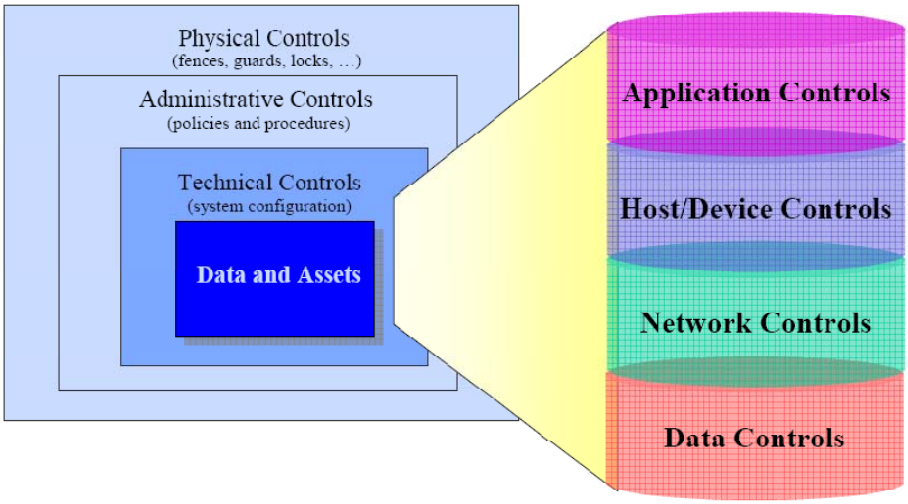
Information security controls, safeguards, threats, vulnerabilities and security processes can all be measured in terms of their impact on Confidentiality, Integrity, and Availability (CIA). Figure 5-1 provides an overview of the three CIA tenets. The primary focus of information security is to avoid unauthorized information disclosure, prevent unauthorized data modifications, and ensure reliable and timely access to data.

Figure 5-1
"THE" InfoSec Tenets–CIA



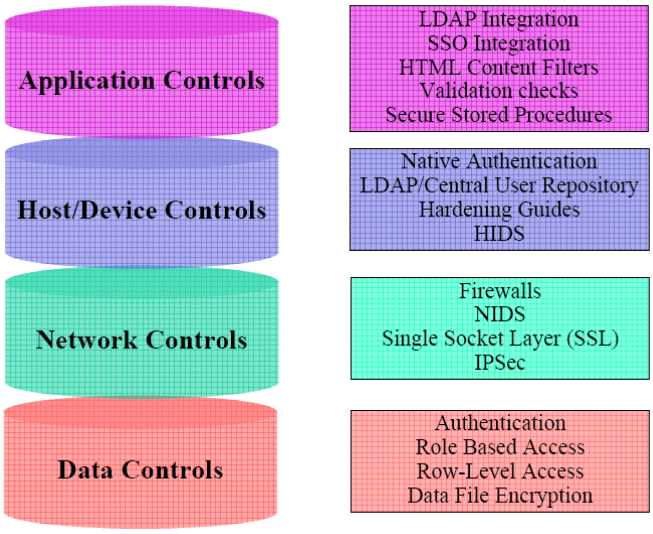
Enterprise protection is provided through multiple levels of security controls. No security solution is infallible, and protection can only be achieved through a layered defense. Levels of defense include physical controls, administrative controls, and technical controls that work together to provide a secure environment. Figure 5-2 presents an overview of the types of security controls.

Figure 5-2
Security Control Types



Information security includes multiple layers of technical controls, implemented through several layers of authentication and validation defense measures. The configuration layers can be grouped as application controls, host/device controls, network controls, and data controls. Figure 5-3 provides examples of the types of controls available for each system technical layer.

Figure 5-3
Technical Control Examples



Application Security encompasses measures taken to prevent exceptions in the security policy of applications or the underlying system (vulnerabilities) through flaws in the design, development, or deployment of the application. The development of security and control procedures for the custom applications are based on COTS functionalities provided by Windows OS, ArcGIS, RDBMS, and HTTP protocols:

Windows Access Control List (ACL), which provides for mandatory system wide access control through role based access control where permissions are assigned to roles and roles are assigned to users

Thorough ACL's for file systems Access Control Entries (ACE) can be defined for group rights to specific system objects, such as applications, processes or files. These privileges or permissions determine specific access rights, such as whether a user can read from, write to, execute or delete an object.

ArcGIS controls for client or web applications are mechanisms implemented either through ArcGIS out-of-the-box configuration, custom application enhancement (using ArcObjects) or ArcGIS Web client. The following application controls are available in the ArcGIS enterprise environment:

Custom Control extensions can be utilized to implement technologies such as Identity Management (IM) and access control. ArcGIS custom control extensions are developed using the ArcObjects development interface. ArcGIS gives the user the ability to restrict ArcGIS client operations (edit, copy, save, print) or controls users access to various data assets based on their role.

GML is an XML schema used for modeling, transporting, and storing geographic information. ArcObjects, utilizing GML and RDBMS storage functionality, offers a framework and method for auditing controls in ArcGIS multi-user geodatabase environments. A detailed history of GIS workflow activities can be recorded in a GML structure and stored in the RDBMS. In addition to recording who performed the edit, activities can be supplemented with comments and notes to provide a traceable, documented, activity log containing before-edit, after-edit, and edit justification history.

Integrated operating system authentication and single sign-on (SSO) are two security infrastructures that can be leveraged by ArcObjects applications to authenticate against and connect to ArcGIS products using user names and passwords managed in a centralized location. This location can be an encrypted file, an RDBMS table, a Lightweight Directory Access Protocol (LDAP) server, or a combination of RDBMS tables and an LDAP server. The primary intent is to insulate users from having to continually authenticate themselves. This technique relies on users' authentication into their desktop workstation (integrated operating system authentication) or the organization's SSO infrastructure.

Native authentication by ArcSDE and RDBMS; Strong authentication controls can be established between ArcGIS and system components through the use of native authentication allowing the user to be authorized by downstream systems. ArcSDE utilizing the direct connect architecture supports native Windows authentication from the ArcGIS client connecting to the RDBMS. The direct connect configuration allows ArcGIS clients to leverage RDBMS connectivity functionality. Deployed utilizing two-tier ArcSDE architecture configured with a RDBMS SSL transport layer, native authentication provides an encrypted communication channel between the trusted operating system and the RDBMS.

SSL is a protocol that communicates over the network through the use of public key encryption. SSL establishes a secure communication channel between the client and server. Encryption functionality of the RDBMS converts clear text into cipher text that is transmitted across the network. Each new session initiated between the RDBMS and the client creates a new public key, affording increased protection. Utilizing ArcSDE in a direct connect configuration eliminates the use of the ArcSDE application tier by moving the ArcSDE functionality from the server to the ArcGIS client. By moving the ArcSDE functionality from the server to the client (dynamic link library), the client application is enabled to communicate directly to the RDBMS through the RDBMS client software. ArcSDE interpretations are performed on the client before communication to the RDBMS. This provides the client application the ability to leverage network encryption controls supplied by the RDBMS client.

IPSec is a set of protocols that secures the exchange of packets between the ArcGIS client and the RDBMS server at the IP level. IPSec uses two protocols to provide IP communication security controls: authentication header (AH) and encapsulation security payload (ESP). The AH offers integrity and data origin authentication. The ESP protocol offers confidentiality.

Intrusion detection is available for ArcGIS users: Network based intrusion detection analyzing network packages flowing through the network or host based intrusion monitoring operation on a specific host.

Feature level security implemented in parallel with ArcSDE allows the Lands Department to assign privileges at the feature level, restricting data access within the geodatabase object. RDBMS Feature-level security is based on the concept of adding a column to a table that assigns a sensitivity level for that particular row. Based on the value in that column, the RDBMS determines, through an established policy, whether the requesting user has access to that information. If the sensitivity level is met, the RDBMS allows access to the data; otherwise, access is denied

Data file encryption can be used by the ArcSDE direct connect architecture by using a data encryption "add-in" in the RDBMS which works with ArcGIS products accessing an RDBMS as a data store, custom ArcObjects applications, and custom non-ESRI technology-based applications using the ArcSDE C and Java APIs to access non-versioned data.

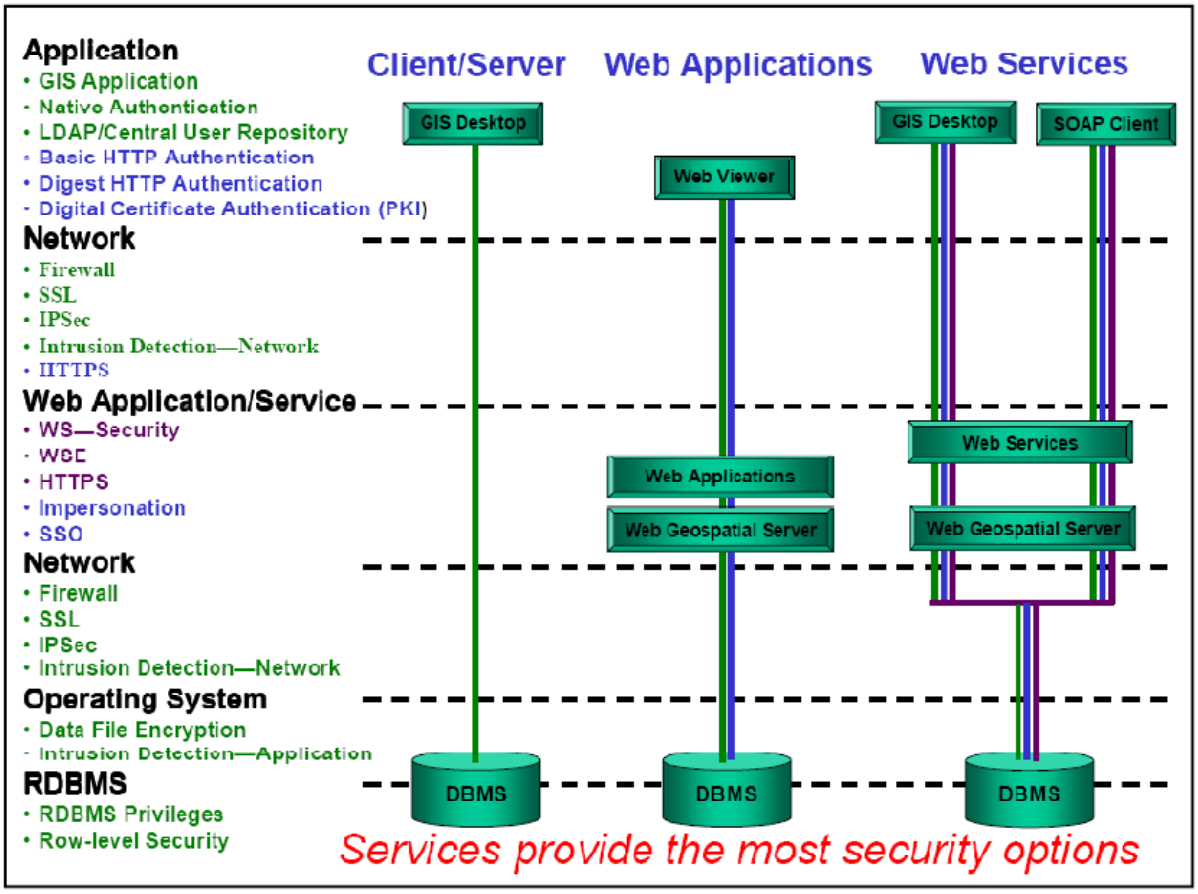
RDBMS privileges; RDBMS assigns SELECT, UPDATE, INSERT, and DELETE privileges to either a user or role. The ArcSDE command line and ArcCatalog leverage the RDBMS privilege assignment functionality and provide an interface that allows the administrator to assign privileges.

HTTP authentication is a mechanism by which an HTTP authentication method is used to verify that someone is who they claim to be. The standard methods of HTTP authentication integrated with ArcGIS Web applications are the basic, digest, form, and client certificate methods. Basic authentication involves protecting an HTTP resource and requiring a client to provide a user name and password to view that resource. Digest authentication also involves protecting an HTTP resource by requesting that a client provide user name and password credentials; however, the digest mechanism encrypts the password provided by the client to the server. Form-based authentication is identical to basic except that the application programmer provides the authentication interface using a standard HTML form. Client certificate is the most secure authentication method in that it uses the organizational PKI environment to provide and authenticate digital certificates for both client and server.

5.2 Enterprise Security Strategies

Business operations today are exposed to a variety of information security threats. These threats can be generated by friendly and unfriendly sources and may include both internal and external users. Threats can be intentional or inadvertent, but in either case, they can result in loss of resources, compromise of critical information, or denial of service. Figure 5-4 provides an overview of security options available for client/server, Web application, and Web services architecture.

Figure 5-4
Security in Depth (ArcGIS Architecture)



5.2.1 Client/Server Architecture

Desktop and network operating systems should require user identification and password based on defined system access privileges. Networks can include firewalls that restrict and monitor content of communications, establishing different levels of access criteria for message traffic. Communication packets can be encrypted (Secure Sockets Layer [SSL]) to deny unauthorized information access, even if the data is captured or lost

during transmission. Specific content exchange criteria can be established between servers (IPSec) to restrict communication flow and to validate traffic sources. Traffic activity can be monitored (intrusion detection) to identify attempts to overcome security protection. Data can be protected on disk to avoid corruption or prevent access as appropriate (encryption). Database environments provide access control (privileges) and row-level security. A combination of these security techniques throughout the information flow can provide the highest level of protection.

5.2.2 Web Application Architecture

Standard firewall, SSL, IPSec, intrusion detection, data file encryption, and RDBMS security solutions continue to support Web operations. Additional security can be implemented to protect and control HTTP communications; basic and digest authentication and implementation of digital certificate authentication (PKI) procedures promote secure communications and support restricted user access to published Web applications. Secure HTTP protocols (HTTPS) encrypt data transfers supporting a higher level of communication protection. Web applications can assume user rights for data access (impersonation), and options for passing user authentication (single sign-on [SSO]) for database access enhance security and control access to the data source.

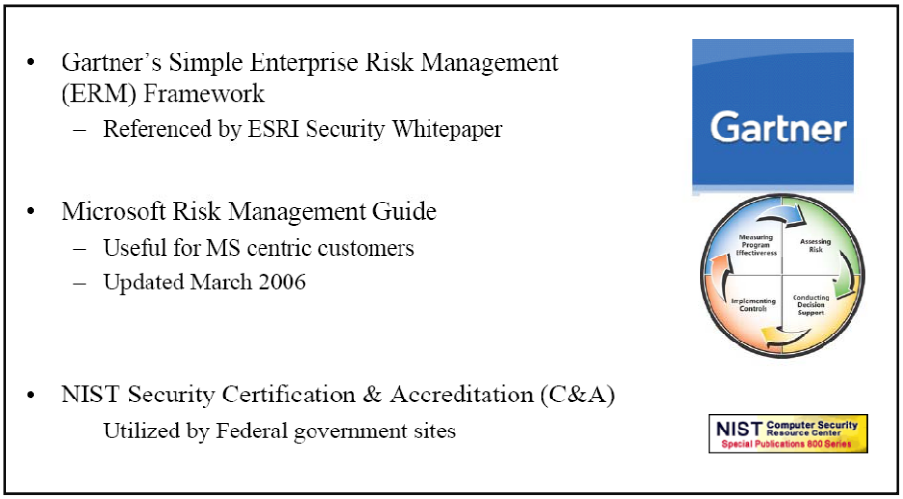
5.2.3 Web Services Architecture

The most security controls are available when deploying an enterprise service-oriented architecture. Protection provided by the Web application architecture supports an SOA, and additional options are available to enhance access controls. Client applications can include additional security features to ensure proper use and control. Additional Web services security (WS-Security) solutions can be implemented to support user authentication and restrict access to Web services. Web services extensions (WSE) are specific Web services security implementations supported through Web server technology. Secure HTTP communications encrypt data transmissions and improve communication security.

5.3 Selecting the Right Security Solution

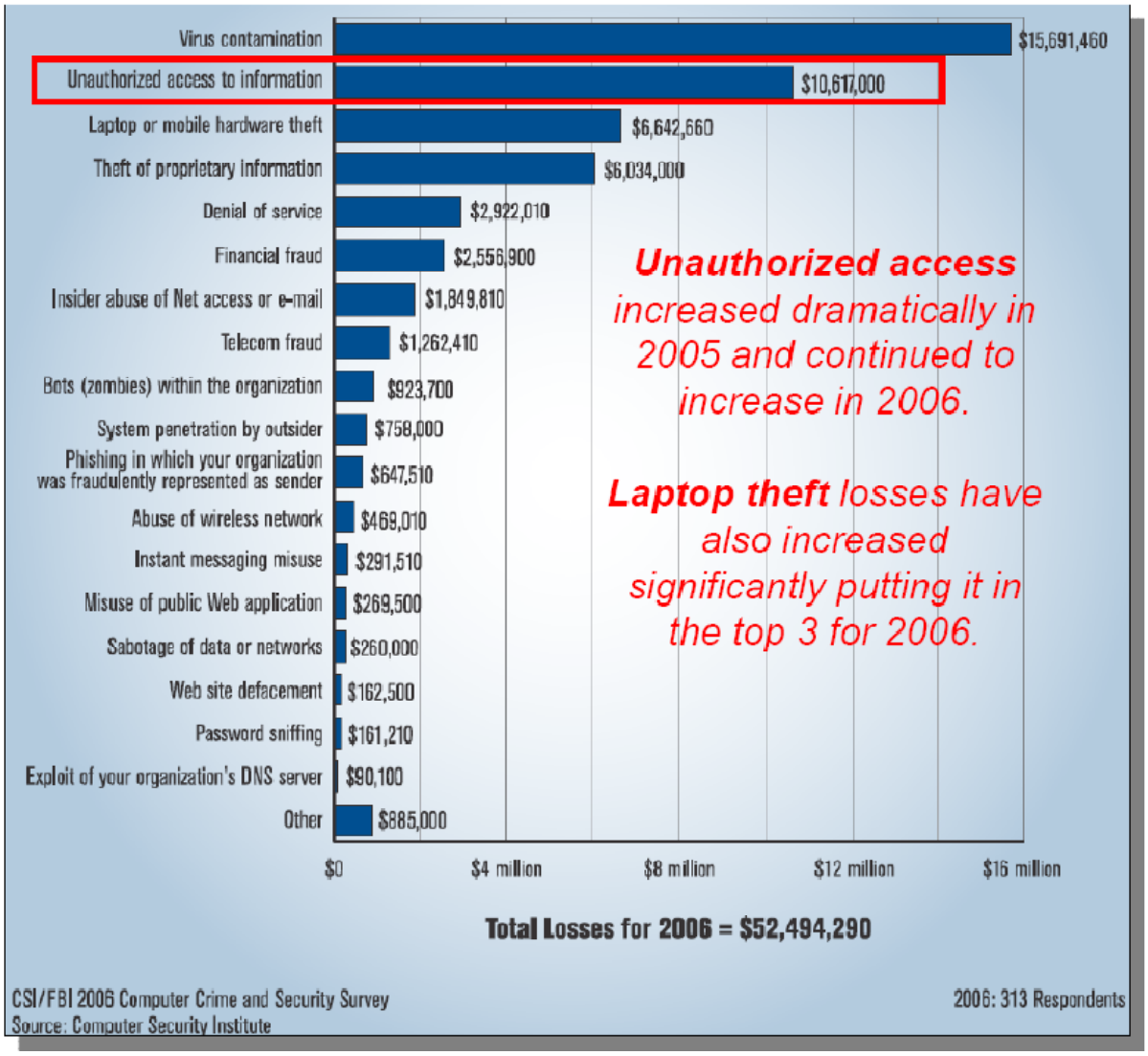
Security solutions are unique to each client situation. The right security solution depends on your enterprise risks and your selection of enterprise controls. The challenge is to implement reasonable and appropriate security controls. It is important to maintain and support a current security risk assessment, establish security guidelines and controls, and perform on-going security audits to ensure objectives are being maintained.
Figure 5-5 provides a list of standard risk management frameworks that can be used to develop and support a security risk management program. These are provided by trusted industry experts, and include Gartner's Simple Enterprise Risk Management Framework, Microsoft Risk Assessment Guide for Microsoft centric customers and the National Institute for Standards and Technology providing a baseline for security certification and accreditation of federal government sites.

Figure 5-5
Risk Management Frameworks



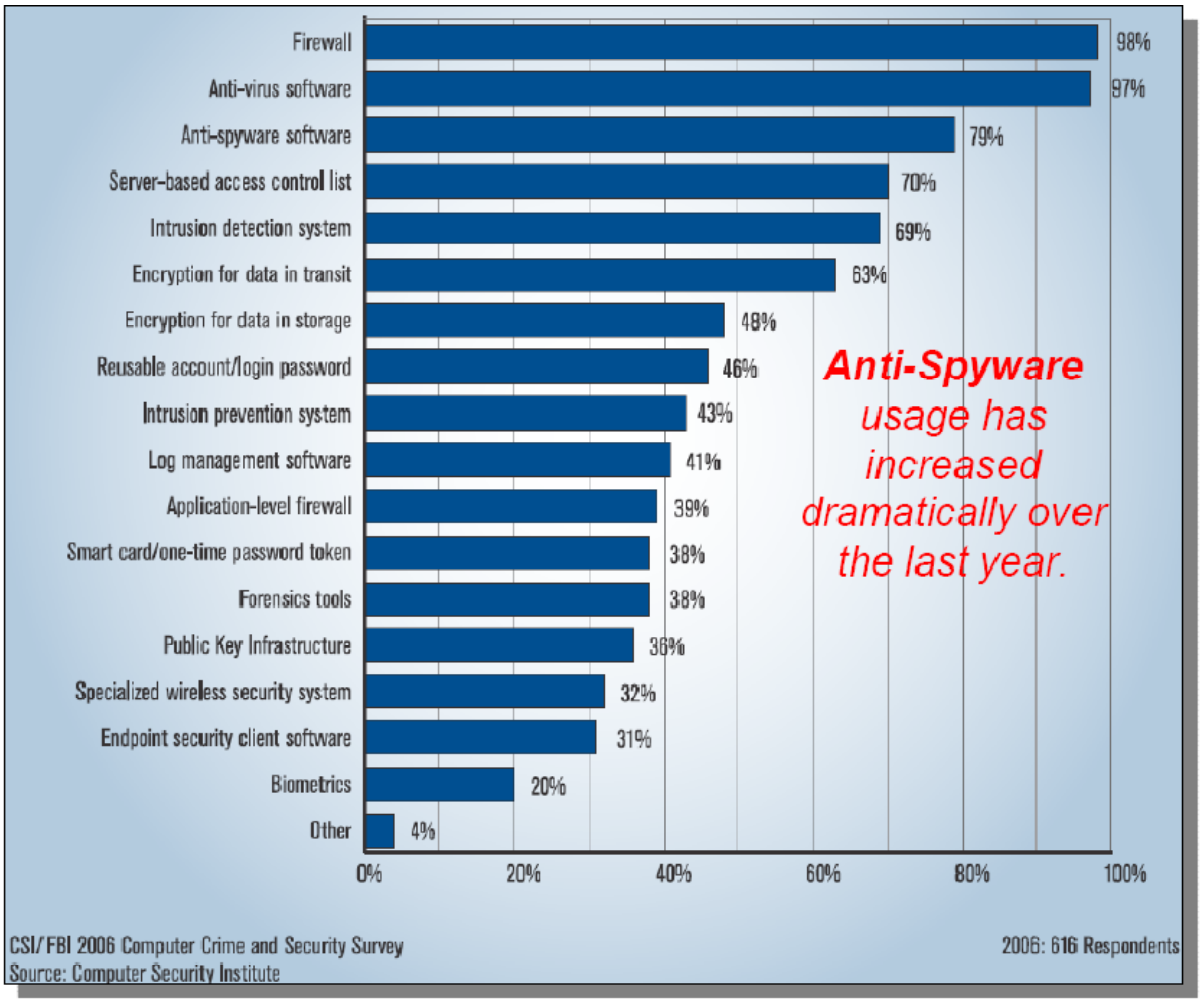
The most common security risks are identified in Figure 5-6. Over 30 percent of the 2006 risk was virus contamination, followed with over 20 percent from unauthorized access to information. Unauthorized access increased dramatically in 2005 and continued to increase in 2006. Laptop theft has also increased making it rise to the number 3 risk in 2006.

Figure 5-6
Dollar Amount Losses by Threat



The most common security controls are highlighted in Figure 5-7. Firewall technology holds a slight lead over anti-virus software. Anti-spyware technology has increased dramatically over the past year. A diverse range of controls are available to address security concerns.

Figure 5-7
Security Technologies Utilized



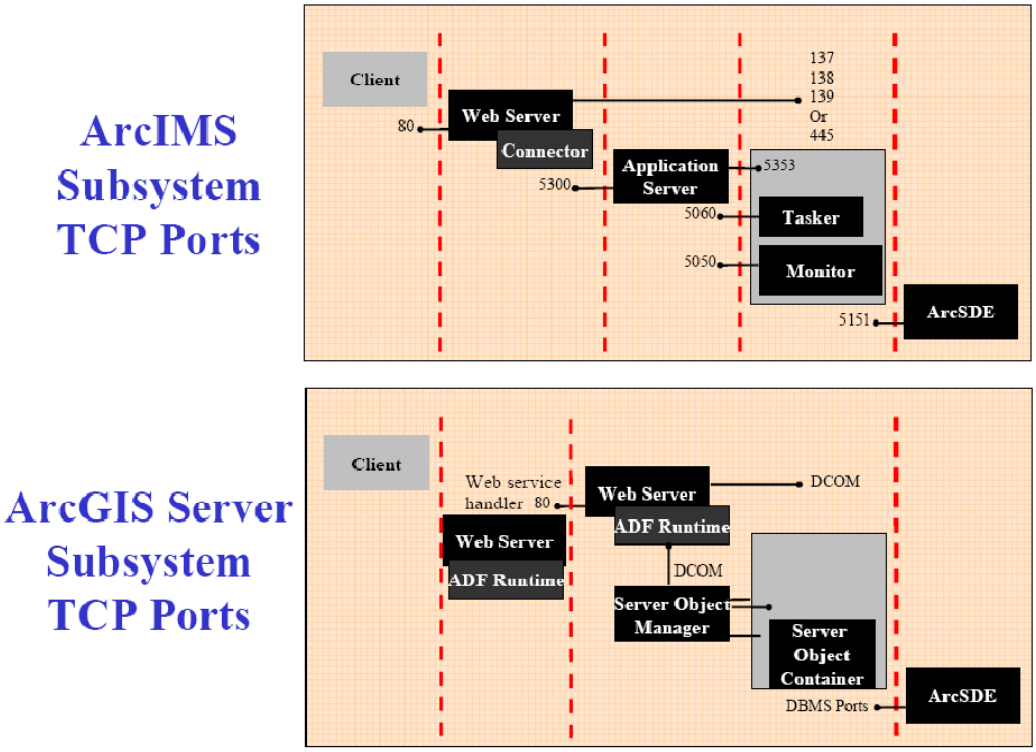
Security comes with a price. Understanding the specific security risk and applying the appropriate security controls can reduce overall cost and provide the best operational solution.

5.4 Web Firewall Configuration Alternatives

Firewall configurations are provided to support communication between various levels of security. A number of firewall configuration options are identified here, based on the location of the ArcIMS or ArcGIS Server software components. An ESRI white paper, Security and ArcIMS, addresses configuration options for secure ArcIMS environments. This paper is available at http://www.esri.com/library/whitepapers/pdfs/securityarcims.pdf.

Figure 5-8 provides an overview of default TCP ports used with ArcIMS and ArcGIS Server firewall configurations. ArcIMS firewall configuration ports are provided between each of the software configuration layers. ArcGIS Server communications between the Web application server and server object container use the Distributed Component Object Model (DCOM) protocols. The use of DCOM involves dynamically acquiring TCP/IP ports for communication between components, and separation of these components over a firewall configuration is not recommended.

Figure 5-8
Firewall Communications

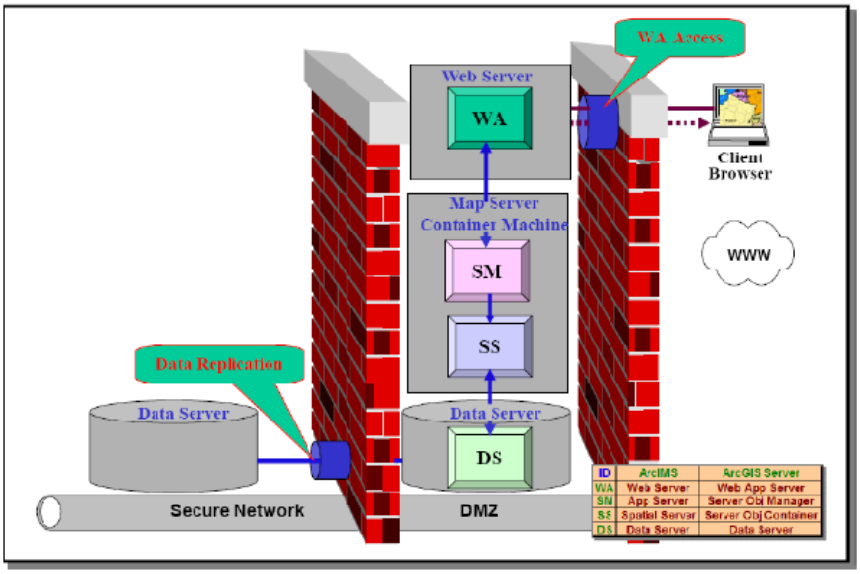


The remaining discussion addresses available Web services firewall configuration strategies. Advantages and disadvantages of each configuration are discussed. Understanding the available configuration options and associated implications can help the security architect select the best solution for supporting enterprise security needs.

5.4.1 All Web Services Components in DMZ

The most secure solution provides physical separation of the secure network from all ArcIMS software components. Figure 5-9 shows the Web application, service manager, spatial services, and data source are all located outside the secure network firewall and within the demilitarized zone (DMZ). This configuration requires maintenance of duplicate copies of the GIS data. Data must be replicated from the internal GIS data server to the external data server supporting the ArcIMS services.

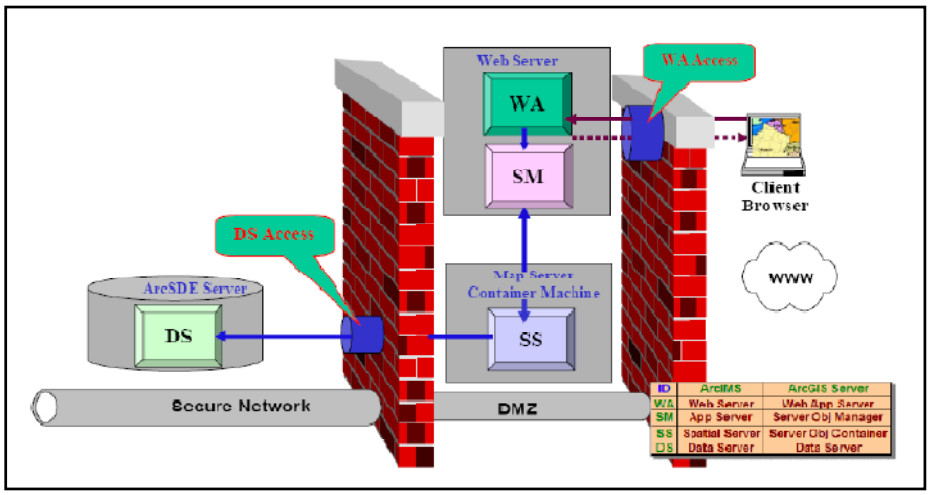
Figure 5-9
All Web Services Components in DMZ



5.4.2 All Web Services Components in DMZ except Data Server

Figure 5-10 shows the Web application, service manager, and spatial services located in the DMZ, accessing the internal ArcSDE data server located on the secure network. Port 5151 access through the secure firewall allows limited access to the ArcSDE DBMS data server. A high volume of traffic must be supported between the spatial services and the data source. Any network disconnects with the data server would generate delays while all publish service connections are reestablished.

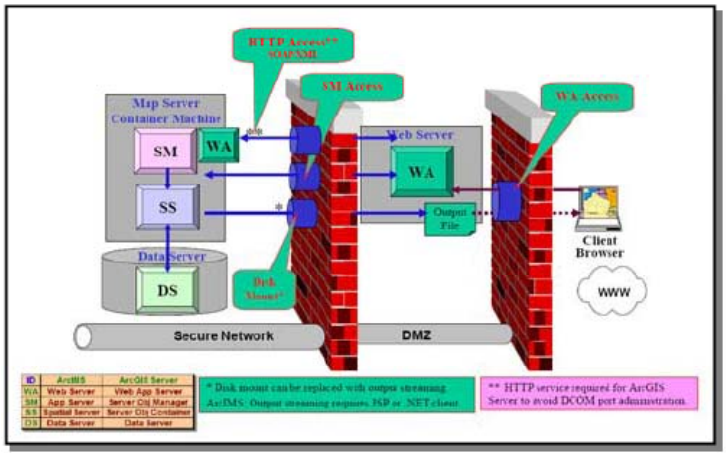
Figure 5-10
All Web Services Components in DMZ except Data Source



5.4.3 Web Application in the DMZ, Remainder of the Web Services Components on the Secure Network

Figure 5-11 shows the Web application server located in the DMZ, with the map server/container machine and data server located on the secure network. The service manager and spatial services must be located on the internal network for this configuration to be acceptable. The output file, located on the Web server, must be shared with the map server. This disk mount will support one-way access from the map server through the firewall to the Web server. This configuration is not recommended for ArcGIS Server.

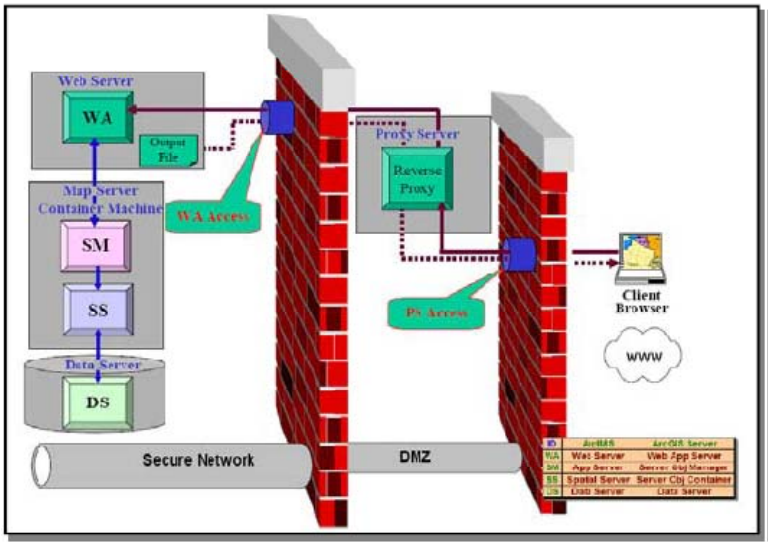
Figure 5-11
Web Application in DMZ, Remainder of Web Services Components on Secure Network



5.4.4 Web Services with Proxy Server

Figure 5-12 shows interface with intranet Web application configuration supported by a proxy server. This solution provides private network security through a reverse proxy server and supports the complete Web services configuration on the private network. This configuration enables full management of the Web site on the private network. This is the preferred configuration for ArcGIS Server deployment.

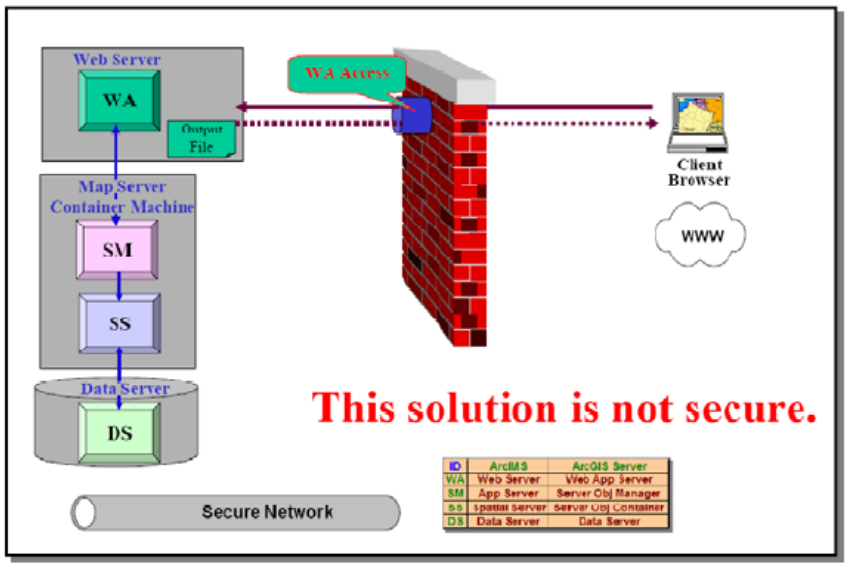
Figure 5-12
Web Services with Proxy Server



5.4.5 All Web Services Components on the Secure Network

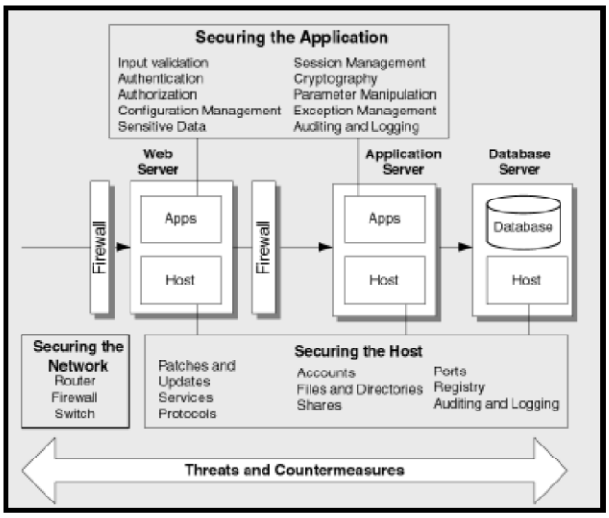
Figure 5-13 shows the Web application, map server/container machine, and data server components all inside the firewall on the secure network. Port 80 must be open to allow HTTP traffic to pass through the firewall. Many organizations are not comfortable with this level of security.

Figure 5-13
All Web Services Components on Secure Network



Security is everybody's job, there is no exception. The world is not a secure environment, and we need to keep our eyes and minds open to the threats around us. There is no single solution for security. There are costs and tradeoffs that must be made to support an optimum solution. Too much security controls can reduce productivity and increase cost. To little attention and control can result in loss of property and the ability to perform. Finding the right balance is important, and the right solution can be a moving target.

Figure 5-14
Security in Depth



We all must Contribute

6 Data Administration

Data management is a primary consideration when developing enterprise GIS architectures. Enterprise GIS normally benefits from efforts to consolidate agency GIS data resources. There are several reasons for supporting data consolidation. These reasons include improving user access to data resources, providing better data protection, and enhancing the quality of the data. Consolidation of IT support resources also reduces hardware cost and the overall cost of system administration.

The simplest and most cost-effective way to manage data resources is to keep one copy of the data in a central data repository and provide required user access to this data to support data maintenance and operational GIS query and analysis needs. This is not always practical, and many system solutions require that organizations maintain distributed copies of the data. Significant compromises may have to be made to support distributed data architectures.

This section provides an overview of data management technology. Several basic data management tasks will be identified along with the current state of technology to support these tasks. These data management tasks include the following:

Ways to store spatial data

Ways to protect spatial data

Ways to back up spatial data

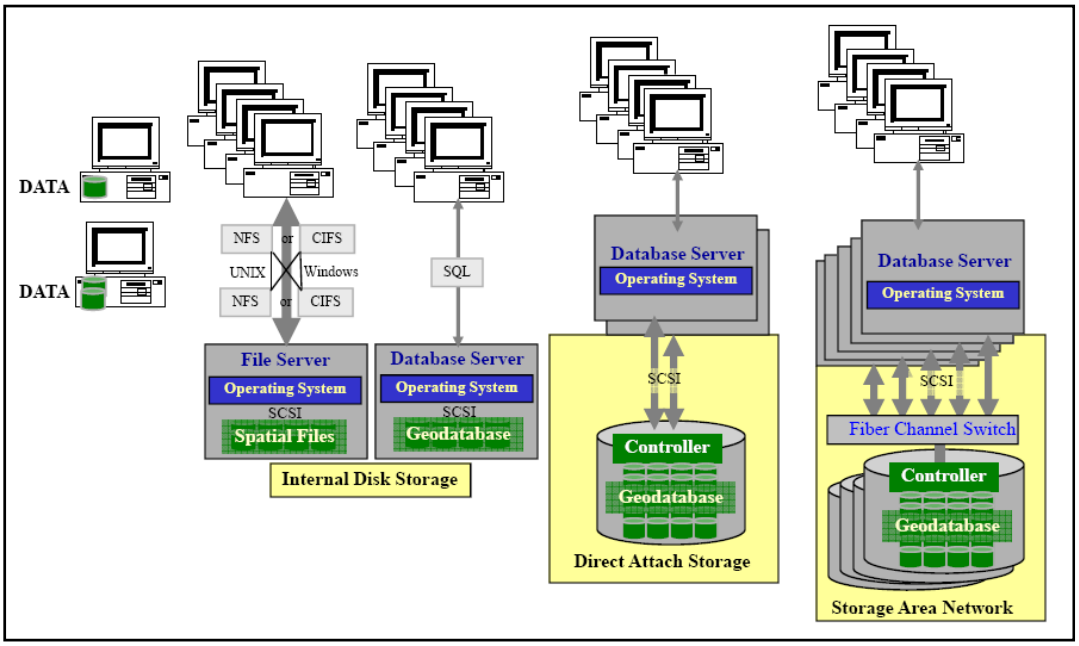
Ways to move spatial data

New ways to manage and access spatial data

6.1 Ways to Store Spatial Data

Storage technology has evolved over the past 20 years to improve data access and provide better management of available storage resources. Understanding the advantages of each technical solution will help you select the storage architecture that best supports your needs. Figure 6.1 provides an overview of the technology evolution from internal workstation disk to the storage area network architecture.

Figure 6-1
Advent of the Storage Area Network



Internal Disk Storage. The most elementary storage architecture puts the storage disk on the local machine.

Most computer hardware today includes internal disk for use as the storage medium. Workstations and servers can both be configured with internal disk storage. The fact that access to it is through the local workstation or server can be a significant limitation in a shared server environment: if the server operating system goes down, there is no way for other systems to access the internal data resources.

File server storage provides a network share that can be accessed by many client applications within the local network. Disk mounting protocols (NFS and CIFS) provide local application access over the network to the data on the file server platform. Query processing is provided by the application client, which can involve a high amount of chatty communications between the client and server network connection.

Database server storage provides query processing on the server platform, and significantly reduces the required network communication traffic. Database software improves data management and provides better administration control of the integrity of the data.

Internal storage can include RAID mirror disk volumes that will preserve the data store in the event of a single disk failure. Many servers include bays that support multiple disk drives for configuring RAID 5 configurations and support high capacity storage needs. The internal storage access is limited to the host server, so a many data center environments grew larger in the 1990s customers would have many servers in their data center with too much disk (disk not being used), and other servers with too little disk making disk volume management a challenge (data volumes could not be shared between server internal storage volumes). External storage architecture (Direct Attached, Storage Area Networks, and Network Attached Storage) gives a way for organizations to “break out” from these “silo based” storage solutions and build a more manageable and adaptive storage architecture.

Direct Attached Storage. A direct attached storage (DAS) architecture provides the storage disk on an external storage array platform. Host bus adaptors (HBA) connect the server operating system to the external storage controller using the same block level protocols that were used for Internal Disk Storage, so from an application and server perspective the direct attached storage appears and functions the same as internal storage. The external storage arrays can be designed with fully redundant components (system would continue operations with any single component failure), so a single storage array can support high available storage requirements.

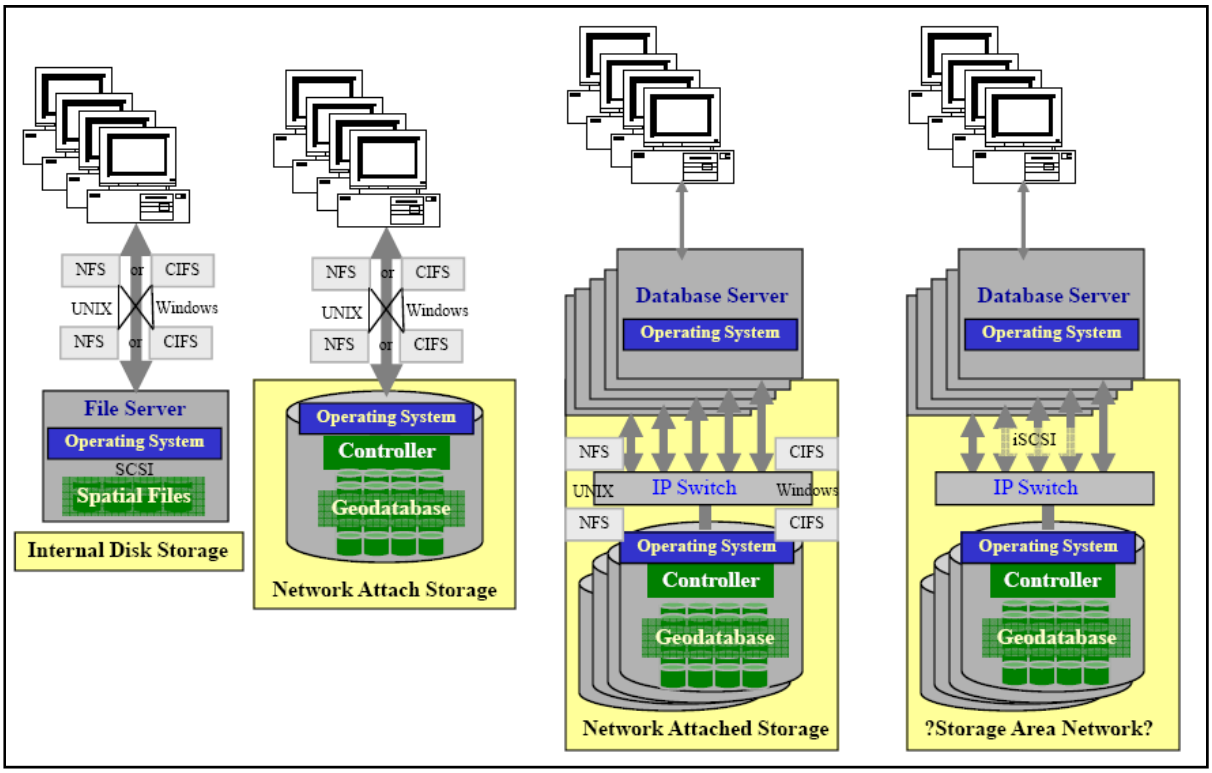
Direct attached storage technology can provide several fiber channel connections between the storage controller and the server HBAs. For high availability purposes, it is standard practice to configure two HBA fiber channel connections for each server environment. Direct Attached Storage solutions provide from 4 to 8 fiber channel connections, so you can easily proved up to 4 servers with two redundant fiber channel connections from a single direct connect storage array controller. The disk storage volumes are allocated to specific host servers, and the host servers control access to the assigned storage volumes. In a server failover scenario, the primary server disk volumes can be reassigned to the failover server.

Storage Area Networks. The difference between direct attached storage and a storage area network is the introduction of a Fiber Channel Switch to provide network connectivity between the Servers and the external Storage Arrays. The storage area network (SAN) improves administrative flexibility for assigning and managing storage resources when you have a growing number of server environments. The Server HBAs and the External Storage Array controllers are connected to the Fiber Channel Switch, so any Server can be assigned storage from any Storage Array located in the storage farm (connected through the same storage network). Storage protocols are still the same as with Direct Attached or Internal Storage – so from a software perspective, these storage architecture solutions appear the same and are transparent to the application and data interface.

Network Attached Storage. By the late 1990s, many data centers were using servers to provide client application access to shared file data sources. High available environments require clustered file server storage, so if one of the servers fail users would still have access to the file share. Hardware vendors now provide a highbred appliance configuration to support network file shares (called Network Attached Storage or NAS) – the network attached storage provides a file server and storage in a single high available storage platform. The file server can be configured with a modified operating system that provides both NFS and CIFS disk mount protocols, and a storage array with this modified file server network interface is deployed as a simple network attached hardware appliance. The storage array includes a standard Network Interface Card (NIC) interface to

the local area network, and client applications can connect to the storage over standard disk mount protocols. The network attached storage provided a very simple way to deploy a shared storage for access by a large number of UNIX and Windows network clients. Figure 6-2 shows the evolution of the Network Attached Storage architecture.

Figure 6-2
Advent of the Network Attached Storage



Network attached storage provides a very effective architecture alternative for supporting network file shares, and has become very popular among many GIS customers. As GIS data moves from early file based data stores (coverages, LIBRARIAN, ArcStorm, Shapefiles) to a more database centric data management environment (Geodatabase servers), the network attached storage vendors suggest you can use a network file share to support a database server storage. There are some limitations: It is important to assign dedicated data storage volumes controlled by the host database server to avoid data corruption. Other limitations include slower database query performance over the chatty IP disk mount protocols than with the traditional fiber channel SCSI protocols, and the bandwidth over the IP network is lower than the Fiber Channel switch environments (1 Gbps IP networks vs 2 Gbps Fiber Channel networks) – implementation of Network Attached Storage as an alternative to Storage Area Networks is not an optimum storage architecture for geodatabase server environments. At the same time, it is an optimum architecture for file based data sources and use of the NAS technology alternative continues to grow.

Because of the simple nature of network attached storage solutions, you can use a standard local area network (LAN) Switch to provide a network to connect your servers and storage solutions; this is a big selling point for the NAS proponents. There is quite a bit of competition between Storage Area Networks and Network Attached Storage technology, particularly when supporting the more common database environments. The SAN community will claim their architecture is supported by higher bandwidth connections and the use of standard storage block protocols. The NAS community will claim they can support your storage network using standard LAN communication protocols and provide support for both database server and network file access clients from the same storage solution.

The network attached storage community is providing a more efficient iSCSI communication protocol for their

storage networks (basically SCSI storage protocols over IP networks). GIS architectures today include a growing number of file data sources (examples include ArcGIS Image Server imagery, ArcGIS Server pre-processed 2-D and 3-D file caches, and the file geodatabase). For many GIS operations, a mix of these storage technologies provides the optimum storage solution.

6.2 Ways to Protect Spatial Data

Enterprise GIS environments depend heavily on GIS data to support a variety of critical business processes. Data is one of the most valuable resources of a GIS, and protecting data is fundamental to supporting critical business operations.

The primary data protection line of defense is provided by the storage solutions. Most storage vendors have standardized on redundant array of independent disks (RAID) storage solutions for data protection. A brief overview of basic storage protection alternatives includes the following:

Just a Bunch of Disks (JBOD): A disk volume with no RAID protection is referred to as just a bunch of disks configuration, or (JBOD). This represents a configuration of disks with no protection and no performance optimization.

RAID 0: A disk volume in a RAID 0 configuration provides striping of data across several disks in the storage array. Striping supports parallel disk controller access to data across several disks reducing the time required to locate and transfer the requested data. Data is transferred to array cache once it is found on each disk. RAID 0 striping provides optimum data access performance with no data protection. One hundred percent of the disk volume is available for data storage.

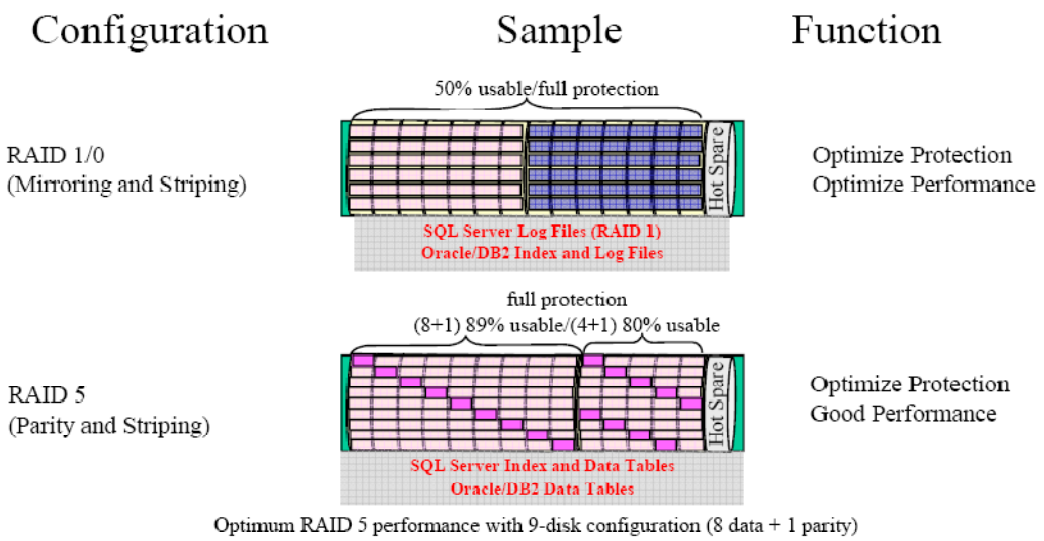
RAID 1: A disk volume in a RAID 1 configuration provides mirror copies of the data on disk pairs within the array. If one disk in a pair fails, data can be accessed from the remaining disk copy. The failed disk can be replaced and data restored automatically from the mirror copy without bringing the storage array down for maintenance. RAID 1 provides optimum data protection with minimum performance gain. Available data storage is limited to 50 percent of the total disk volume, since a mirror disk copy is maintained for every data disk in the array.

RAID 3 and 4: A disk volume in a RAID 3 or RAID 4 configuration supports striping of data across all disks in the array except for one parity disk. A parity bit is calculated for each data stripe and stored on the parity disk. If one of the disks fails, the parity bit can be used to recalculate and restore the missing data. RAID 3 provides good protection of the data and allows optimum use of the storage volume. All but one parity disk can be used for data storage, optimizing use of the available disk volume for data storage capacity.

There are some technical differences between RAID 3 and RAID 4, which, for our purposes, are beyond the scope of this discussion. Both of these storage configurations have potential performance disadvantages. The common parity disk must be accessed for each write, which can result in disk contention under heavy peak user loads. Performance may also suffer because of requirements to calculate and store the parity bit for each write. Write performance issues are normally resolved through array cache algorithms on most high-performance disk storage solutions.

The following RAID configurations are the most commonly used to support ArcSDE storage solutions. These solutions represent RAID combinations that best support data protection and performance goals. Figure 6-3 provides an overview of the most popular composite RAID configuration.

Figure 6-3
Ways to Protect Spatial Data
(Standard RAID Configurations)



Additional hybrid RAID solutions are available.

RAID 1/0: RAID 1/0 is a composite solution including RAID 0 striping and RAID 1 mirroring. This is the optimum solution for high performance and data protection. This is also the costliest solution. Available data storage is limited to 50 percent of the total disk volume, since a mirror disk copy is maintained for every data disk in the array.

RAID 5: RAID 5 includes the striping and parity of the RAID 3 solution and the distribution of the parity volumes for each stripe across the array to avoid parity disk contention performance bottlenecks. This improved parity solution provides optimum disk utilization and near optimum performance, supporting disk storage on all but one parity disk volume.

Hybrid Solutions: Some vendors provide alternative proprietary RAID strategies to enhance their storage solution. New ways to store data on disk can improve performance and protection and may simplify other data management needs. Each hybrid solution should be evaluated to determine if and how it may support specific data storage needs.

ArcSDE data storage strategies depend on the selected database environment.

SQL Server: Log files are located on RAID 1 mirror, and index and data tables are located on RAID 5 disk volume.

Oracle, Informix, and DB2: Index tables and log files are located on RAID 1/0 mirror, and striped data volumes and data tables are located on RAID 5.

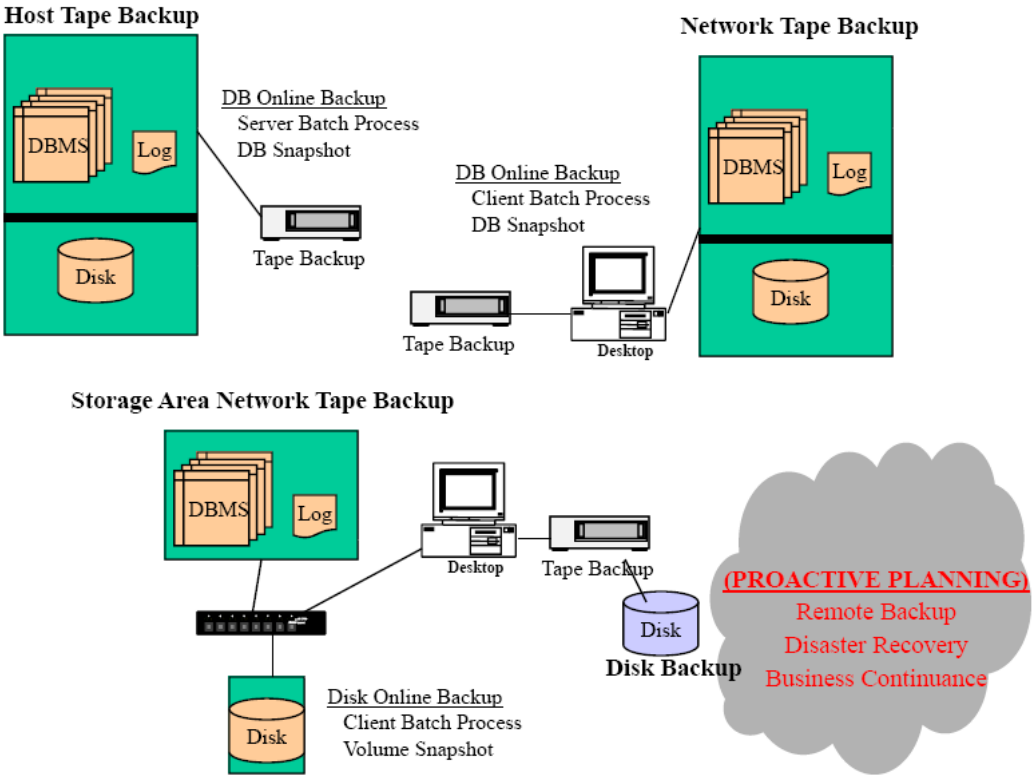
6.3 Ways to Back Up Spatial Data

Data protection at the disk level minimizes the need for system recovery in the event of a single disk failure but will not protect against a variety of other data failure scenarios. It is always important to keep a current backup copy of critical data resources at a safe location away from the primary site.

Data backups typically provide the last line of defense for protecting data investments. Careful planning and attention to storage backup procedures are important factors to a successful backup strategy. Data loss can

result from many types of situations, with some of the most probable situations being administrative or user error. Figure 6-4 provides an overview of the different ways to back up spatial data.

Figure 6-4
Ways to Back Up Spatial Data



Host Tape Backup: Traditional server backup solutions use lower-cost tape storage for backup. Data must be converted to a tape storage format and stored in a linear tape medium. Backups can be a long drawn out process taking considerable server processing resource (typically consume a CPU during the backup process) and requiring special data management for operational environments.

For database environments, these backups must occur based on a single point in time to maintain database continuity. Database vendors support online backup requirements by establishing a procedural snapshot of the database. A copy of the protected snapshot data is retained in a snapshot table when changes are made to the database, supporting point-in-time backup of the database and potential database recovery back to the time of the snapshot.

Host processors can be used to support backup operations during off-peak hours. If backups are required during peak-use periods, backups can impact server performance.

Network Client Tape Backup: The traditional online backup can often be supported over the LAN with the primary batch backup process running on a separate client platform. DBMS snapshots may still be used to support point-in-time backups for online database environments. Client backup processes can contribute to potential network performance bottlenecks between the server and the client machine because of the high data transfer rates during the backup process.

Storage Area Network Client Tape Backup: Some backup solutions support direct disk storage access without impacting the host DBMS server environment. Storage backup is performed over the SAN or through a separate fiber channel access to the disk array with batch process running on a separate client platform. A disk-level storage array snapshot is used to support point-in-time backups for online database environments. Host platform processing loads and LAN performance bottlenecks can be

avoided with disk-level backup solutions.

Disk Copy Backup: The size of databases has increased dramatically in recent years, growing from tens of gigabytes to hundreds of gigabytes and, in many cases, terabytes of data. Recovery of large databases from tape backups is very slow, taking days to recover large spatial database environments. At the same time, the cost of disk storage has decreased dramatically providing disk copy solutions for large database environments competitive in price to tape storage solutions. A copy of the database on local disk, or a copy of these disks to a remote recovery site, can support immediate restart of the DBMS following a storage failure by simply restarting the DBMS with the backup disk copy.

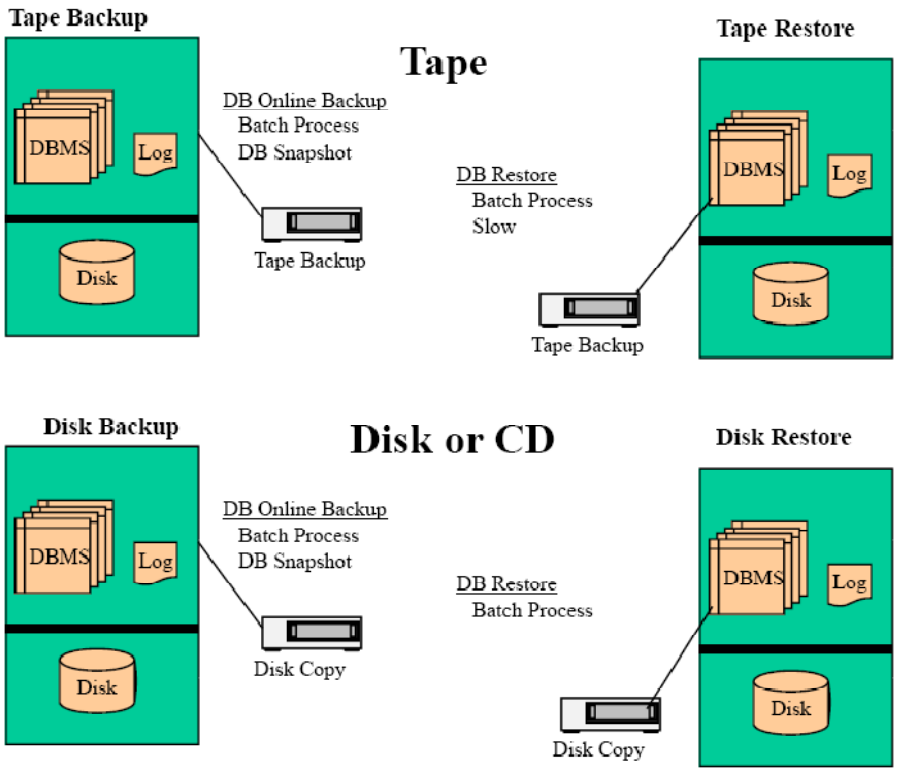
6.4 Ways to Move Spatial Data

Many enterprise GIS solutions require continued maintenance of distributed copies of the GIS data resources, typically replicated from a central GIS data repository or enterprise database environment. Organizations with a single enterprise database solution still have a need to protect data resources in the event of an emergency such as fire, flood, accidents, or other natural disasters. Many organizations have recently reviewed their business practices and updated their plans for business continuance in the event of a major loss of data resources. The tragic events of September 11, 2001, demonstrated the value of such plans and increased interest and awareness of the need for this type of protection.

This section reviews the various ways organizations move spatial data. Traditional methods copy data on tape or disk and physically deliver this data to the remote site through standard transportation modes. Once at the remote site, data is reinstalled on the remote server environment. Technology has evolved to provide more efficient alternatives for maintaining distributed data sources. Understanding the available options and risks involved in moving data is important in defining optimum enterprise GIS architecture.

Traditional Data Transfer Methods: Figure 6-5 identifies traditional methods for moving a copy of data to a remote location.

Figure 6-5
Ways to Move Spatial Data
(Traditional Tape Backup/Disk Copy)



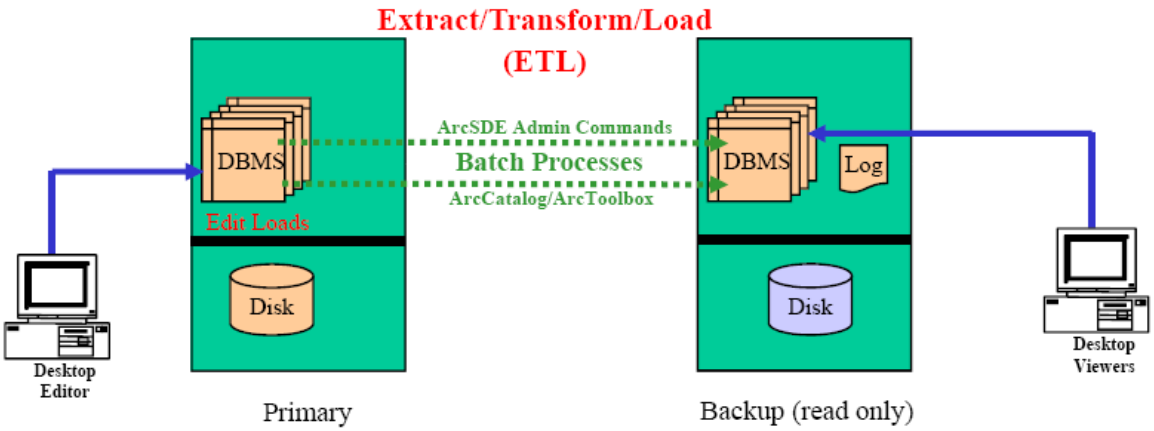
Traditional methods include backup and recovery of data using standard tape or disk transfer media. Moving data using these methods is commonly called "sneaker net." These methods provide a way to transfer data without the support of a physical network.

Tape Backup: Tape backup solutions can be used to move data to a separate server environment. Tape transfers are normally very slow. The reduced cost of disk storage has made disk copy a much more feasible option.

Disk Copy: A replicated copy of the database on disk storage can support rapid restore at a separate site. The database can be restarted with the new data copy and online within a short recovery period.

ArcGIS Geodatabase Transition: Moving subsets of a single database cannot normally be supported with standard backup strategies. Data must be extracted from the primary database and imported into the remote database to support the data transfer. Database transition can be supported using standard ArcGIS export/import functions. These tools can be used as a method of establishing and maintaining a copy of the database at a separate location. Figure 6-6 identifies ways to move spatial data using ArcGIS data transition functions.

Figure 6-6
Ways to Move Spatial Data
(Geodatabase Transition)



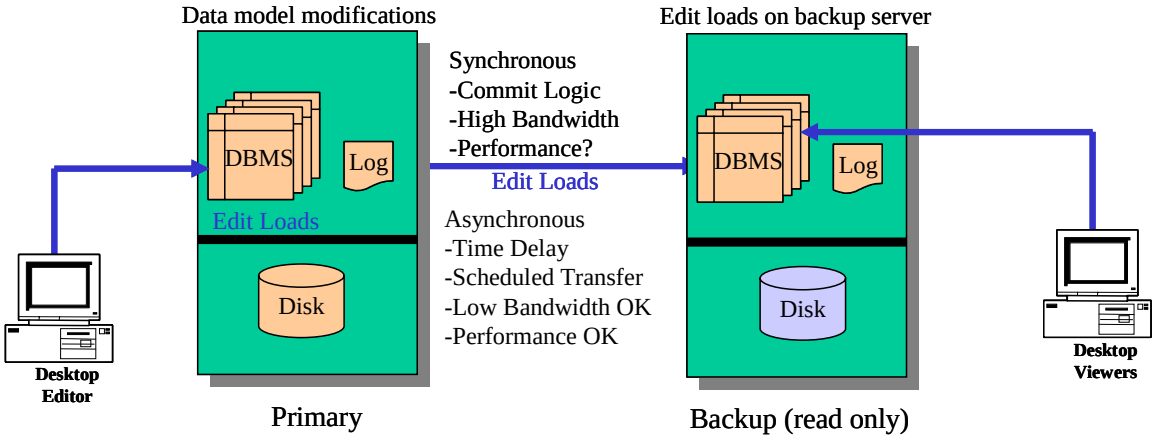
ArcSDE Admin Commands: Batch process can be used with ArcSDE admin commands to support export and import of an ArcSDE database. Moving data using these commands is most practical when completely replacing the data layers. These commands are not optimum solutions when transferring data to a complex ArcSDE geodatabase environment.

ArcCatalog/ArcTools Commands: ArcCatalog supports migration of data between ArcSDE geodatabase environments, extracts from a personal geodatabase, and imports from a personal geodatabase to an ArcSDE environment.

Database Replication: Customers have experienced a variety of technical challenges when configuring DBMS spatial data replication solutions. ArcSDE data model modifications may be required to support DBMS replication solutions. Edit loads will be applied to both server environments, contributing to potential performance or server sizing impacts. Data changes must be transmitted over network connections between the two servers, causing potential communication bottlenecks. These challenges must be overcome to support a successful DBMS replication solution.

Customers have indicated that DBMS replication solutions can work but require a considerable amount of patience and implementation risk. Acceptable solutions are available through some DBMS vendors to support replication to a read-only backup database server. Dual master server configurations significantly increase the complexity of an already complex replication solution. Figure 6-7 presents the different ways to move spatial data using database replication.

Figure 6-7
Ways to Move Spatial Data
(Database Replication)



Not appropriate when moving portion of versioned geodatabase

Synchronous Replication. Real-time replication requires commitment of data transfer to the replicated server before releasing the client application on the primary server. Edit operations with this configuration would normally result in performance delays because of the typical heavy volume of spatial data transfers and the required client interaction times. High-bandwidth fiber connectivity (1000 Mbps bandwidth) is recommended between the primary server and the replicated backup server to minimize performance delays.

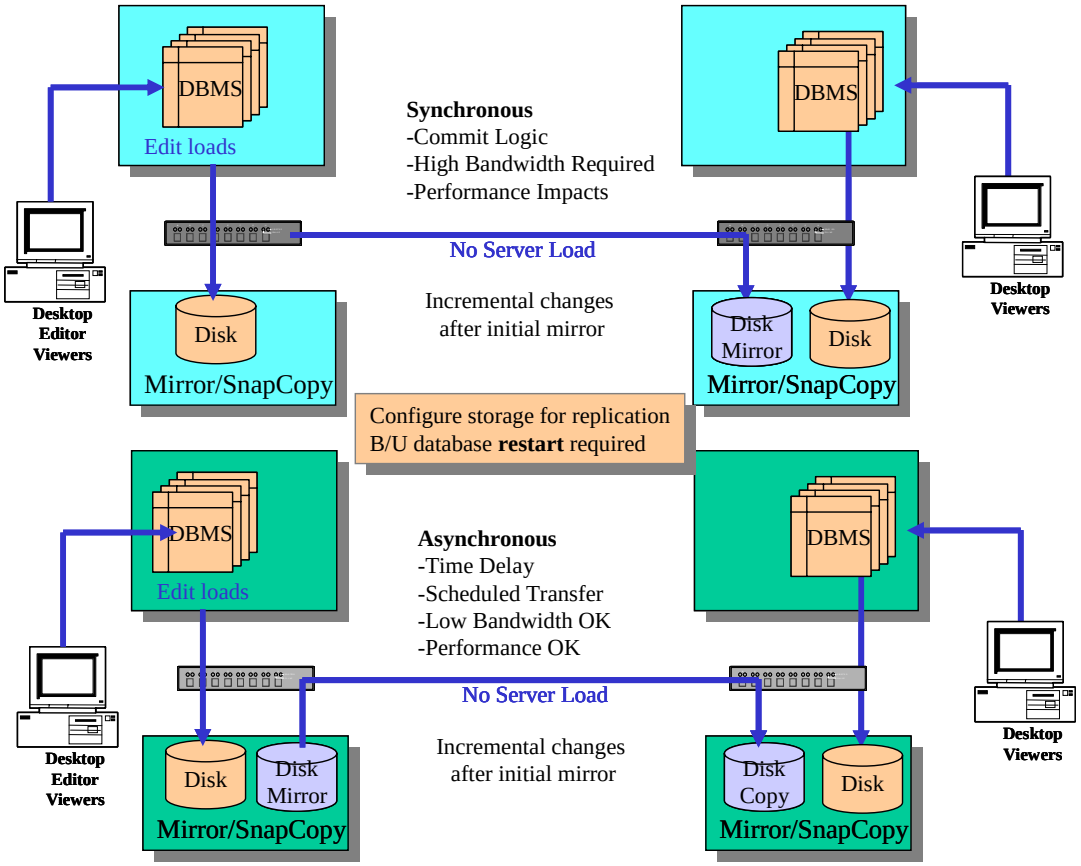
Asynchronous Replication. Near real-time database replication strategies decouple the primary server from the data transfer transaction to the secondary server environment. Asynchronous replication can be supported over WAN connections, since the slow transmission times are isolated from primary server performance. Data transfers (updates) can be delayed to off-peak periods if WAN bandwidth limitations dictate, supporting periodic updates of the secondary server environment at a frequency supporting operational requirements.

Disk-Level Replication: Disk-level replication is a well-established technology, supporting global replication of critical data for many types of industry solutions. Spatial data is stored on disk sectors very similar to any other data storage and, as such, does not require special attention beyond what might be required for other data types. Disk volume configurations (data location on disk and what volumes are transferred to the remote site) may be critical to ensure database integrity. Mirror copies are refreshed based on point-in-time snapshot functions supported by the storage vendor solution.

Disk-level replication provides transfer of block-level data changes on disk to a mirror disk volume located at a remote location. Transfer can be supported with active online transactions with minimum impact on DBMS server performance capacity. Secondary DBMS applications must be restarted to refresh the DBMS cache and processing environment to the point in time of the replicated disk volume.

Figure 6-8 presents different ways to move spatial data using disk-level replication.

Figure 6-8
Ways to Move Spatial Data
(Disk-Level Replication)



Not appropriate when moving portion of versioned geodatabase

Synchronous Replication—Real-time replication requires commitment of data transfer to the replicated storage array before releasing the DBMS application on the primary server. High-bandwidth fiber connectivity (1000 Mbps bandwidth) is recommended between the primary server and the replicated backup server to avoid performance delays.

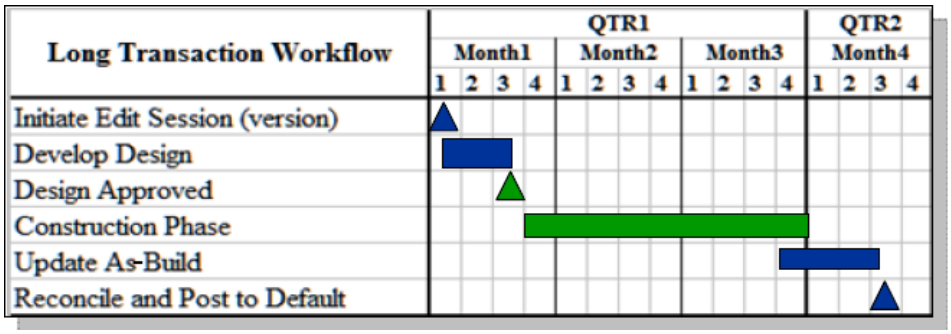
Asynchronous Replication: Near real-time disk-level replication strategies decouple the primary disk array from the commit transaction of changes to the secondary storage array environment. Asynchronous replication can be supported over WAN connections, since the slow transmission times are isolated from primary DBMS server performance. Disk block changes can be stored and data transfers delayed to off-peak periods if WAN bandwidth limitations dictate, supporting periodic updates of the secondary disk storage volumes to meet operational requirements.

6.5 Ways to Manage and Access Spatial Data

Release of the ArcGIS technology introduced the ArcSDE geodatabase, which provides a way to manage long transaction edit sessions within a single database instance. ArcSDE supports long transactions using versions (different views) of the database. A geodatabase can support thousands of concurrent versions of the data within a single database instance. The default version represents the real world, and other named versions are proposed changes and database updates in work.

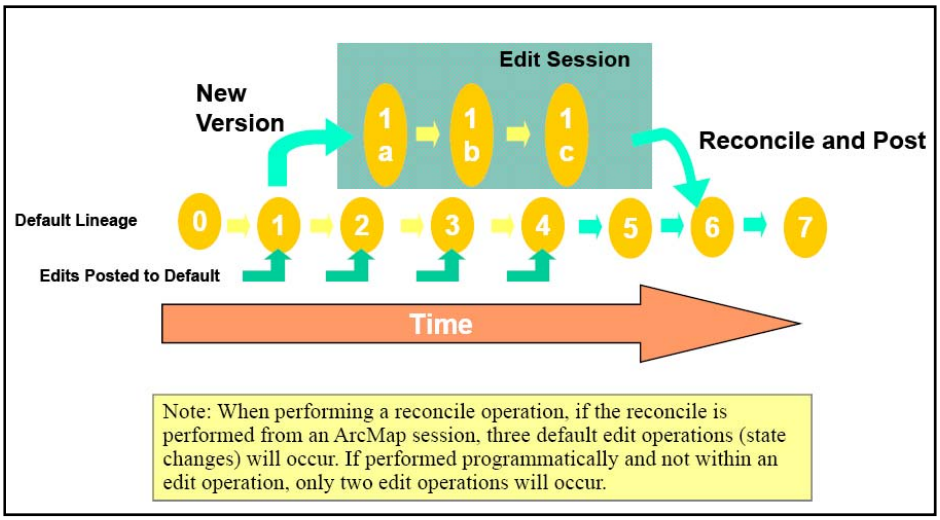
Figure 6-9 shows a typical long transaction workflow life cycle. The workflow represents design and construction of a typical housing subdivision. Several design alternatives might initially be represented as separate named versions in the database to support planning for a new subdivision. One of these designs (versions) is approved to support the construction phase. After the construction phase is complete, the selected design (version) is modified to represent the as-built environment. Once the development is completed, the final design version will be reconciled with the geodatabase and posted to the default version to reflect the new subdivision changes.

Figure 6-9
Long Transaction Workflow Life Cycle



The simplest way to introduce the versioning concept in the geodatabase is by using some logical flow diagrams. Figure 6-10 demonstrates the explicit state model represented in the geodatabase. The default version lineage is represented in the center of the diagram, and a new default version state is added each time edits are posted to the default view. Each edit post represents a state change in the default view (accepted changes to the real-world view). There can be thousands of database changes (versions) at a time. As changes are completed, these versions are posted to the default lineage.

Figure 6-10
Explicit State Model



The "new version" on the top of the diagram shows the life cycle of a long transaction. The transaction begins as changes from "state 1" of the default lineage. Maintenance updates reflected in that version are represented by new states in the edit session (1a, 1b, and 1c). During the edit session, the default version accepts new changes from other completed versions. The new version active edit session is not aware of the posted changes to the default lineage (2, 3, 4, and 5) since it is referenced from default state 1. Once the new version is complete, it must be reconciled with the default lineage. The reconcile process compares the changes in the new version (1a, 1b, and 1c) with changes in the default lineage (2, 3, 4, and 5) to make sure there are no edit conflicts. If the reconcile process identifies conflicts, these conflicts must be resolved before the new version can be posted to the default lineage. Once all conflicts are resolved, the new version is posted to the default lineage forming state 6.

Figure 6-11 shows a typical workflow history of the default lineage. Named versions (t1, t4, and t7) represent edit transactions in work that have not been posted back to the default lineage. The parent states of these versions (1, 4, and 7) are locked in the default lineage to support the long edit sessions that have not been posted. The default lineage includes several states (2, 3, 5, and 6) that were created by posting completed changes.

Figure 6-11
Default History

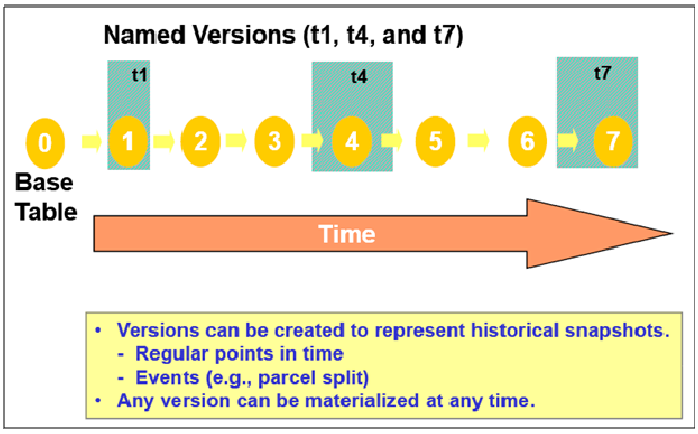
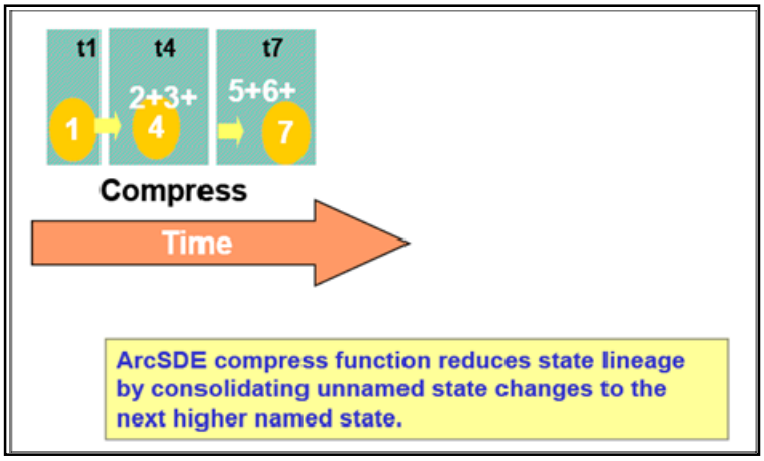


Figure 6-12 demonstrates a geodatabase compress. Very long default lineages (thousands of states) can impact database performance. The geodatabase compress function consolidates all default changes into the named version parent states, thus decreasing the length of the default lineage and improving database performance.

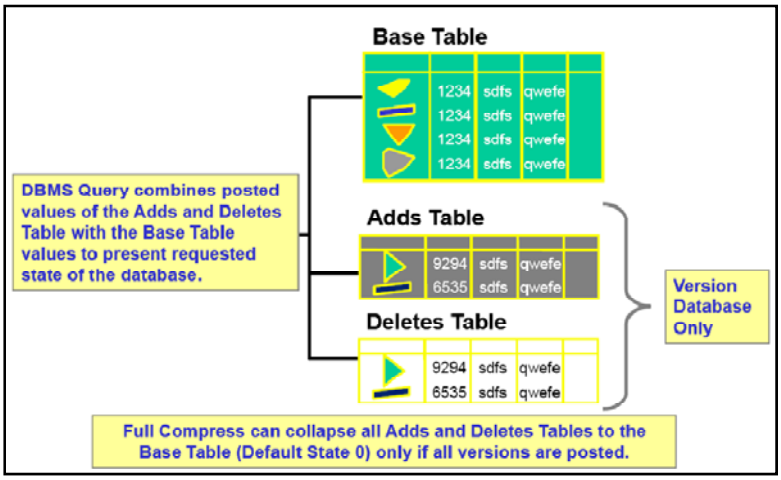
Figure 6-12
Geodatabase Compress



Now that the geodatabase versioning concept is understood, it is helpful to recognize how this is physically implemented within the database table structure. When a feature table within the geodatabase is versioned, two new tables are created to track changes to the base feature table. An Adds Table is created to track additional rows added to the base feature table, and a Deletes Table is created to record deleted rows from the Base Table . Each row in the Adds and Deletes tables represents change states within the geodatabase. As changes are posted to the default version, these changes are represented by pointers in the Adds and Deletes tables. Once there is a versioned geodatabase, the real-world view (default version) is represented by the Base Table plus the Adds and Deletes tables included in the default lineage (the Base Table does not represent default). All outstanding versions must be reconciled and posted to compress all default changes back to the Base Table (zero state). This is not likely to occur for a working maintenance database in a real-world environment.

Figure 6-13 provides a representation of the Base Table, Adds Table, and Deletes Table in a versioned geodatabase.

Figure 6-13
Geodatabase Tables



ArcSDE manages the versioning schema of the geodatabase and supports client application access to the

appropriate views of the geodatabase. ArcSDE also supports export and import of data from and to the appropriate database tables and maintains the geodatabase scheme defining relationships and dependencies between the various tables.

Geodatabase Single-Generation Replication: The ArcGIS 8.3 release supports a disconnected editing solution. This solution provides a registered geodatabase version extract to a personal geodatabase or separate database instance for disconnected editing purposes. The version adds/deletes values are collected by the disconnected editor and, on reconnecting to the parent server, can be uploaded to the central ArcSDE database as a version update.

Figure 6-14 presents an overview of the ArcGIS 8.3 disconnected editing with checkout to a personal geodatabase (PGD). The ArcGIS 8.3 release is restricted to a single checkout/check-in transaction for each client edit session.

Figure 6-14
ArcGIS 8.3 Disconnected Editing—Personal Geodatabase

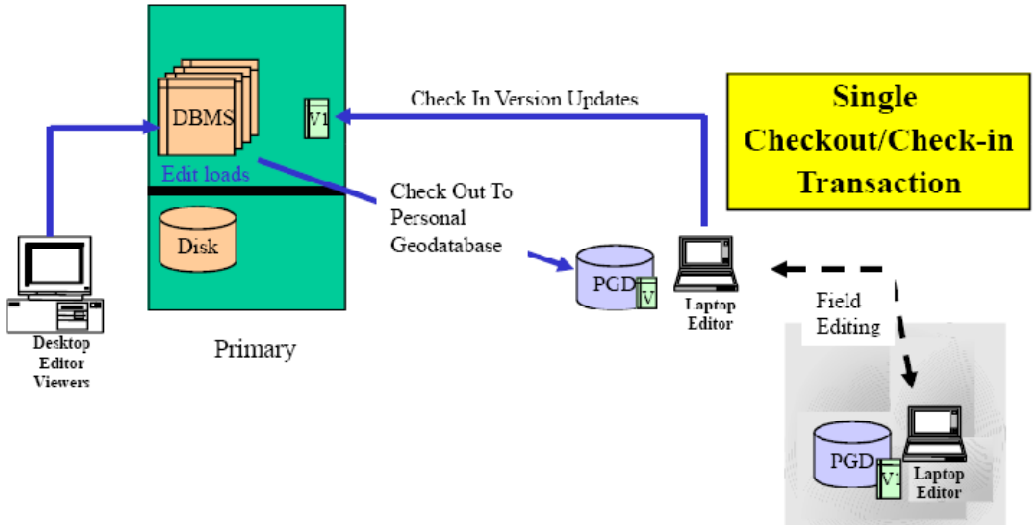
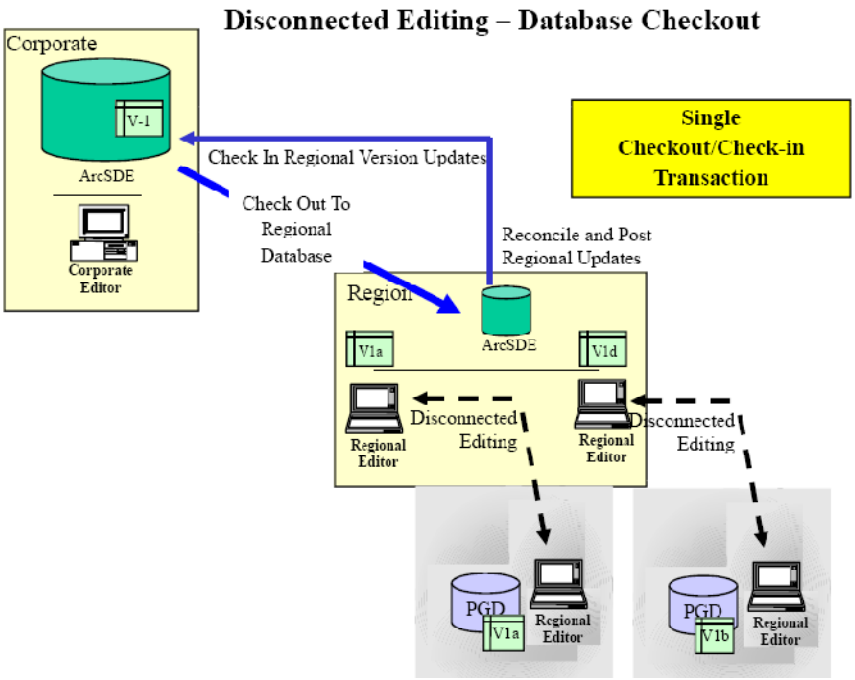


Figure 6-15 presents an overview of the ArcGIS 8.3 disconnected editing with checkout to a separate ArcSDE geodatabase. The ArcGIS 8.3 release is restricted to a single checkout/ check-in transaction for each child ArcSDE database. The child ArcSDE database can support multiple disconnected or local version edit sessions during the checkout period. All child versions must be reconciled before check-in with the parent ArcSDE database (any outstanding child versions will be lost during the child ArcSDE database check-in process).

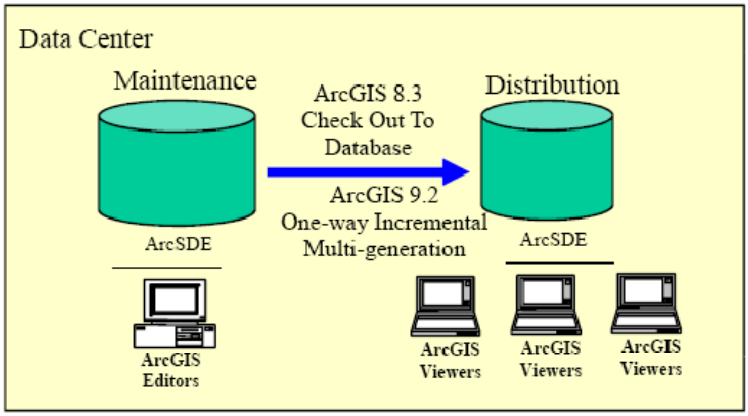
Figure 6-15
ArcGIS 8.3 Disconnected Editing—Database Checkout



The ArcGIS 8.3 database checkout functions provided with disconnected editing can be used to support peer-to-peer database refresh. Figure 6-16 shows a peer-to-peer database checkout, where ArcSDE disconnected editing functionality can be used to periodically refresh specific feature tables of the geodatabase to support a separate instance of the geodatabase environment. This functionality can be used to support a separate distribution view-only geodatabase that can be configured to support a nonversioned copy of the default version.

Figure 6-16
ArcGIS 8.3 Peer-to-Peer—Database Checkout

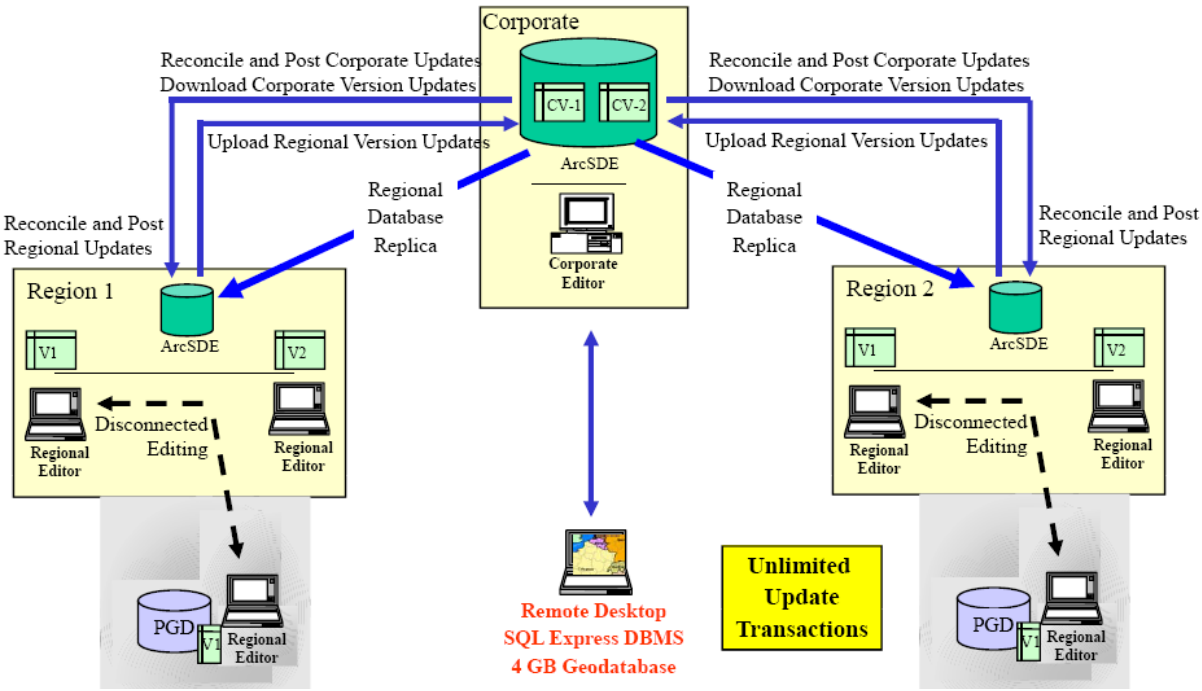
Database Check Out (one-way read only) with ArcGIS 8.3
Multi-generation one-way incremental (read only) with ArcGIS 9.2
Incremental updates to File Geodatabase with ArcGIS 9.3



The ArcGIS 9.2 software incorporates support for incremental updates between ArcSDE geodatabase environments.

Geodatabase Multigeneration Replication: The ArcGIS disconnected editing functionality will be expanded in future ArcGIS 9 releases to support loosely coupled ArcSDE distributed database environments. Figure 6-17 presents an overview of the future loosely coupled ArcSDE distributed database concept.

Figure 6-17
Distributed Geodatabase Architecture



Multigeneration replication supports a single ArcSDE geodatabase distributed over multiple platform environments. The child checkout versions of the parent database supports an unlimited number of update transactions without losing local version edits or requiring a new checkout. Updates are passed between parent and child database environments through simple datagrams that can be transmitted over standard WAN communications. This new geodatabase architecture supports distributed database environments over multiple sites connected by limited bandwidth communications (only the reconciled changes are transmitted between sites to support database synchronization).

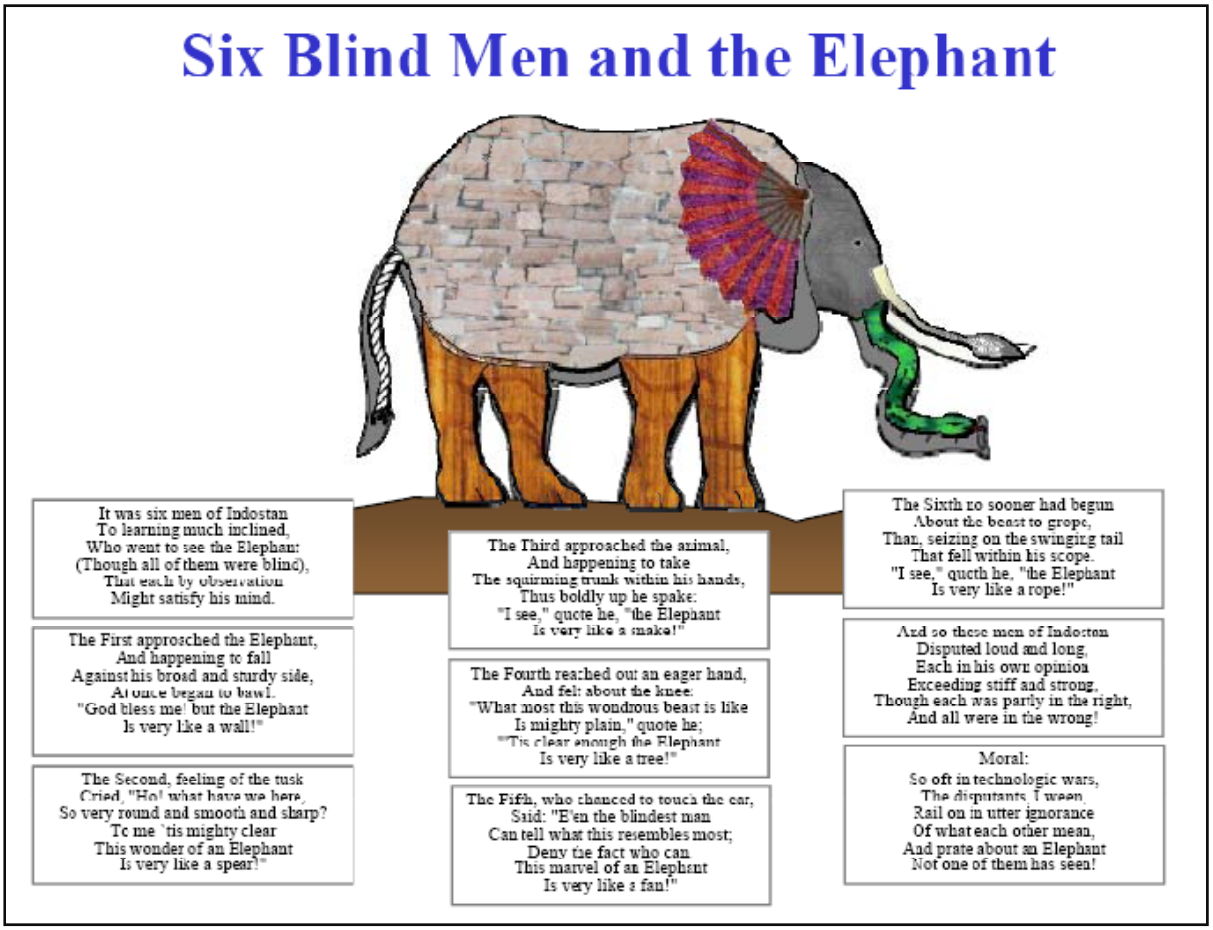
6.6 Data Management Overview

Support for distributed database solutions has traditionally introduced high-risk operations, with potential for data corruption and use of stale data sources in GIS operations. There are organizations that support successful distributed solutions. Their success is based on careful planning and detailed attention to their administrative processes that support the distributed data sites. More successful GIS implementations support central consolidated database environments with effective remote user performance and support. Future distributed database management solutions may significantly reduce the risk of supporting distributed environments. Whether centralized or distributed, the success of enterprise GIS solutions will depend heavily on the administrative team that keeps the system operational and provides an architecture solution that supports user access needs.

7 Performance Fundamentals

Computer platforms must be configured properly to support system performance requirements. There are many factors that contribute to user performance and productivity. Enterprise GIS solutions include distributed processing environments where user performance can be the product of contributions from several hardware platform environments. Many of these platform resources are shared with other users. Understanding distributed processing technology provides a fundamental framework for deploying a successful enterprise GIS. The importance of working together to understand the technology is illustrated in figure 7-1.

Figure 7-1
Understanding the Technology



Technology is changing very rapidly, and we all see and try to understand what these changes mean with the help of our own experience. It is important to listen to the experience of others and incorporate their experience with our own as we move technology forward. I have found that we all have information to contribute and the questions we have can help others understand the technology in a better and more complete way. Modeling our experience, and joining these models with the experience of others, can facilitate our learning process.

7.1 Understanding the Technology

ESRI has implemented distributed GIS solutions since the late 1980s. For many years, distributed processing environments were not well understood, and customers relied on the experience of technical experts to identify hardware requirements to support their implementation needs. Each technical expert had a different perspective on what hardware infrastructure might be required to support a successful implementation, and recommendations were not consistent. Many hardware decisions were made based on the size of the project budget rather than a clear understanding of user requirements and the associated hardware technology.

System performance models were developed in the early 1990s to document what was understood about distributed processing systems. These system performance models have been used by ESRI consultants to support distributed computing hardware solutions since 1992. These same performance models have also been used to identify potential performance problems with existing computing environments.

The initial performance models were developed to support desktop GIS applications with file and GIS database data sources. UNIX and Windows application computer servers were used to provide remote terminal access to GIS applications supported in centralized data centers. A simple concurrent user model was used to support capacity planning.

Web mapping services were introduced in the late 1990s, and transaction-based sizing models were developed to support capacity planning and proper hardware selection. Transaction rates were identified in terms of map displays per hour. The transaction-based capacity planning models proved to be much more accurate and measurable than the previous concurrent user models, although in many cases customers were more comfortable identifying sizing requirements in terms of peak concurrent user load than using peak map requests per hour.

The release of ArcGIS Server 9.2 in 2006 introduced some new challenges for the traditional sizing models, and an effort to review lessons learned and take a close look at the road ahead was in order. The result is a new approach to capacity planning that incorporates the best of the traditional client/server and Web services sizing models and provides an adaptive sizing methodology to support future enterprise GIS operations. The new capacity planning methodology is much easier to use and provides metrics to manage performance compliance during development, initial system implementation, and delivery.

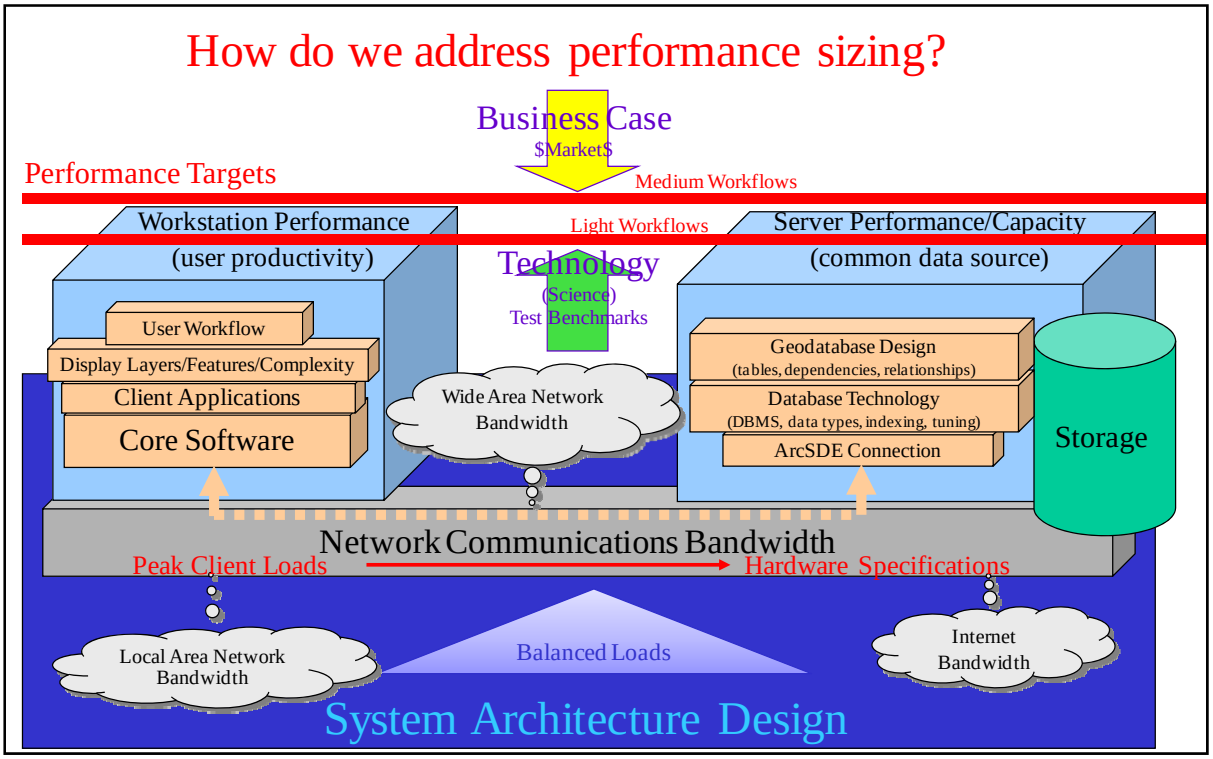
This new capacity planning model was developed and shared with the objective of helping software developers, business partners, technical marketing specialists, and ESRI distributors better understand the performance and scalability of ESRI technology to provide customers with the best possible GIS solutions to support their enterprise GIS operations.

This section presents a basic overview of the system performance fundamentals. The terms and relationships introduced in this section describe fundamental relationships we can all use to better understand performance. The following sections on Software and Platform performance provide additional insight on the processing demands GIS software can place on our system environment, and identify the processing capabilities of current vendor hardware technology. An understanding of these performance fundamentals can provide a framework for building and maintaining more effective real-world GIS operations.

7.1.1 What Is Capacity Planning?

Figure 7-2 identifies some key factors that contribute to overall system performance. Proper hardware and architecture selection is one primary component of the overall system performance equation. There are many other performance factors that contribute to overall user productivity.

Figure 7-2
System Performance Factors



Capacity planning is selecting the right software and hardware to meet user workflow performance needs. Enhancements in any of the system performance factors can improve user productivity and impact total system capacity. Performance cannot be guaranteed by proper hardware selection alone. The performance fundamentals described in this section can help identify appropriate hardware selection based on customer business requirements. Our understanding of GIS processing requirements and how this workload is supported by vendor platform technology is based on more than 20 years of experience helping customers deploy ESRI GIS technology. A balanced hardware investment, based on projected peak user workflow loads, supports system performance requirements and saves money and time through properly targeted hardware purchases.

7.1.2 What Is System Performance?

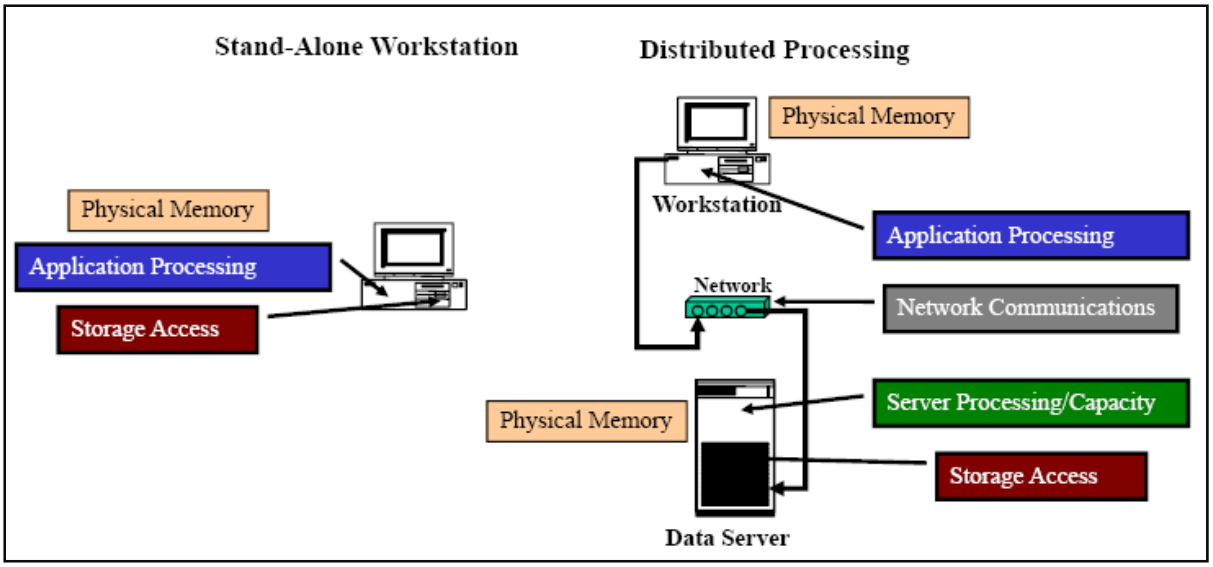
Computer platforms are supported by several component technologies. Each component technology is important - the weakest component will limit overall platform performance. Some software applications require lots of data transfer and dynamic graphics display, while others require heavy computing. Hardware vendors build computers with a balance of component resources that optimize performance for a broad range of customers. Compute intensive software, like GIS, find the hardware platform processor core the computer technology which limits server performance.

In much the same way, distributed computing solutions (enterprise computing environments) are supported by several hardware platforms that contribute to overall system performance. Each hardware component contributes - the weakest component will limit overall system performance. Hardware platforms supporting a computing environment must be carefully selected to satisfy peak processing needs.

The primary objective of the system architecture design process is to provide the highest level of user performance for the available system hardware investment. Each hardware component must be selected with sufficient performance to support processing needs. Current technology can limit system design alternatives. Understanding the distributed processing loads at each hardware component level provides a foundation for establishing an optimum system solution.

Figure 7-3 provides a simplified overview of the components in a stand-alone workstation and in a distributed processing configuration. Each component participates sequentially in the overall program execution. Processing is supported in platform memory—sufficient memory must be available to support the software execution.

Figure 7-3
Platform Performance Components



The total response time of a particular application display will be a collection of the processing and wait times from each of these components. A computer vendor optimizes the component configuration within the hardware to support the fastest computer response to an application query. The customer IT/systems department has the responsibility of optimizing the organization's hardware and network component investments to provide the highest system-level response at the user desktop. System performance can directly contribute to user productivity.

GIS users have experience significant performance and productivity improvements over the past eight years. Time for computers to process a dynamic map display is over 10 times faster with 2009 hardware than what was possible with 2001 technology - this change has a significant impact on user productivity and the opportunities for use of GIS technology.

Software technology selection has also made a difference in display performance. The earlier scripted program technologies of the 1990s (ArcIMS) performs faster than the standard component object based software technology (ArcGIS Server) used today. Performance is a function of the amount of work (processing) required by the software and the performance of the hardware technology (how fast the processing workload can be performed). Significant performance gains improved user experience with the ArcGIS Server 9.3.1 optimized map service, using a new graphics rendering engine to improve quality and performance of standard Web mapping services - generating dynamic maps faster than the ArcIMS Image service. ArcGIS Server client user of pre-processed map cache offloads real time server processing loads and improves user productivity.

Selecting the right software technology pattern and the right hardware architecture is more important today than it ever was before. This chapter is about understanding the fundamental concepts about performance and

scalability. The following two chapters will look more closely at software and hardware contributions to overall system performance.

7.2 System Performance Fundamentals

The study of work performance is not new, and there are a considerable amount of theories and ideas published on this topic. Understanding the fundamental terms and relationships that define work performance and applying these fundamentals to computer processing helps us better understand the technology and make more appropriate design choices.

Figure 7-5 provides a summary of the fundamental performance terms and relationships used in the ESRI system design models.

Figure 7-5
What is Performance?

Terms

- **Work Transaction (W_t)**
  - **Concurrent Users**
  - **Transaction Rates (TPH)**
- **Throughput (T)**
- **Capacity (T_{peak})**
- **Utilization (U)**
- **Processor core (C_p)**
- **Service time (S_t)**
- **Queue time (Q_t)**
- **Response time (R_t)**

Relationships

- Capacity = T / U
- Service time (sec) = 60 x #C_p / T_{peak}
- Queue time $Q_t = \underbrace{\left\{\frac{1}{[1+k^1 U (C_p-1)]}\right\}}_{\text{Multi Core: Service factor}} \times \underbrace{\left\{\frac{S_t U}{(1-U)}\right\}}_{\text{Single Core: Queue}}$

  [1] k factor will depend on the arrival time distribution. Factors are set based on consulting experience. A k factor of 1 shows the best match in comparing capacity planning models to benchmark test results.

- Response time = $S_t + Q_t$

The most important terms include definition of the average work transaction (display), work throughput, system capacity, and system utilization. The processor core is the hardware that executes the computer program instructions, so knowing the number of processor core available in a selected hardware platform configuration is important. Service time is a measure of the average work transaction processing time, and queue time is a measure of the time waiting to be processed (waiting in line for service). Response time is the overall measure of system performance, and includes all of the component service times plus any additional wait or travel times required to refresh the client display (complete a work transaction).

The relationships between these performance terms are quite simple and many times misunderstood. The relationship between throughput, capacity, and utilization are true based on how these terms are defined. Throughput is the number of work transactions being processed per unit time, capacity is the maximum throughput that can be supported by a specific hardware configuration, and utilization is the ratio of the current throughput to the system capacity (expressed as percentage of capacity). If you know the current throughput (users working on the system) and you measure the system utilization (average computer CPU utilization), then you can know the capacity of the server.

Work transaction service time is a key term used to measure software performance. The software program provides a set of instructions that must be executed by the computer to complete a work transaction. The processor core executes the instructions defined in the computer program to complete the work transaction. Transactions with more instructions represent more work for the computer, while transactions with fewer instructions represent less work for the computer. The complexity of the computer program workflow can be defined by the amount of work (or processing time) required to complete an average work transaction. Service

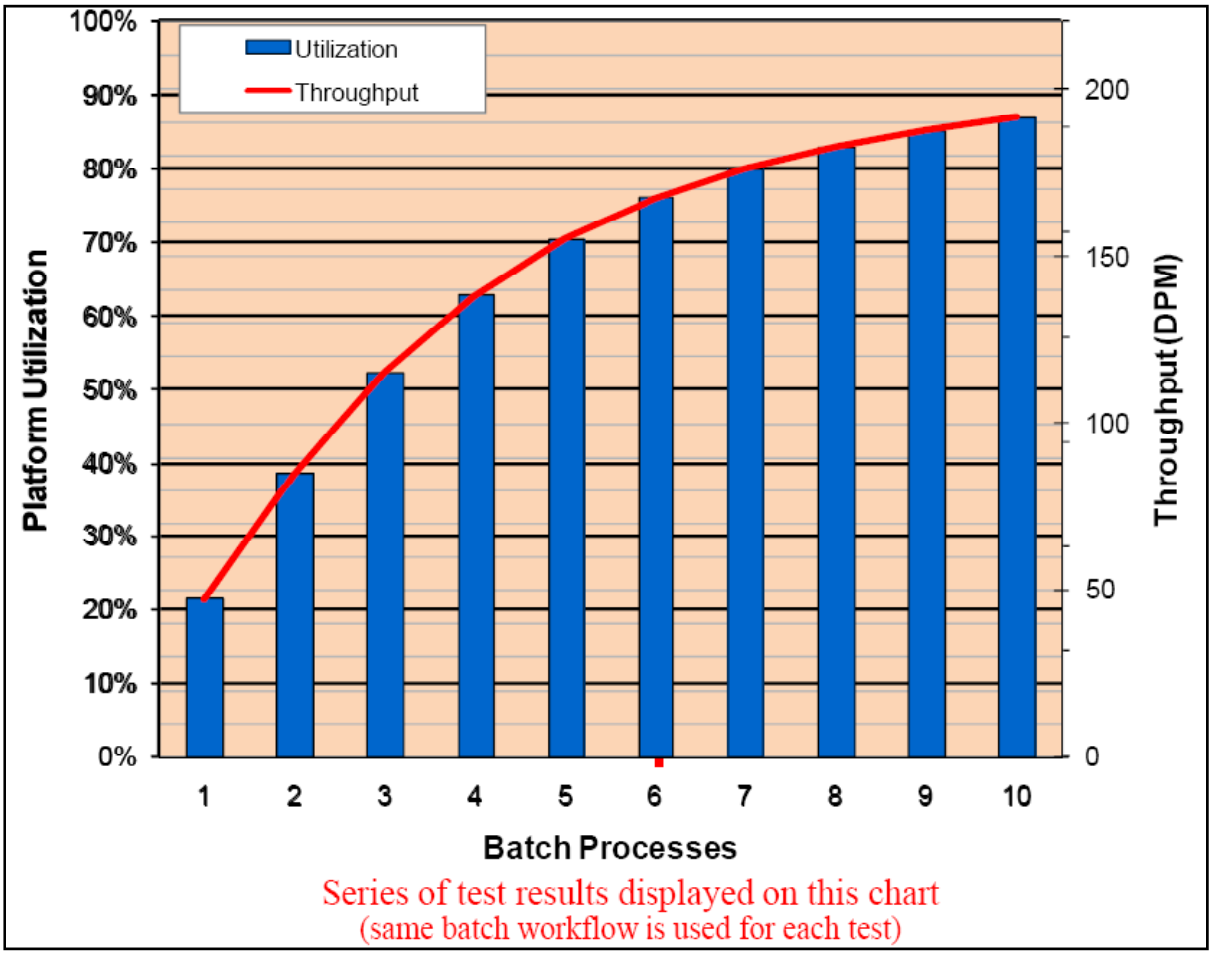
time is measured relative to a platform performance baseline. Faster platform processor core execute program instructions in less time than slower processor core.

Queue time is any time the software program instructions must wait in line to be processed. A single processor core is able to execute only one instruction at a time, so if two requests arrive for processing by a single core at the same time one must wait while the other is processed. This wait time is called Queue time.

7.2.1 Platform Throughput

Figure 7-6 shows the relationship between platform utilization and throughput. The chart shows performance of a four (4) core test platform running a series of batch tests. Platform utilization is shown by the vertical bars and the left vertical axis, and throughput is shown by the RED line and the right vertical axis.

Figure 7-6
What is Platform Throughput?
(4 core 2 chip test platform)

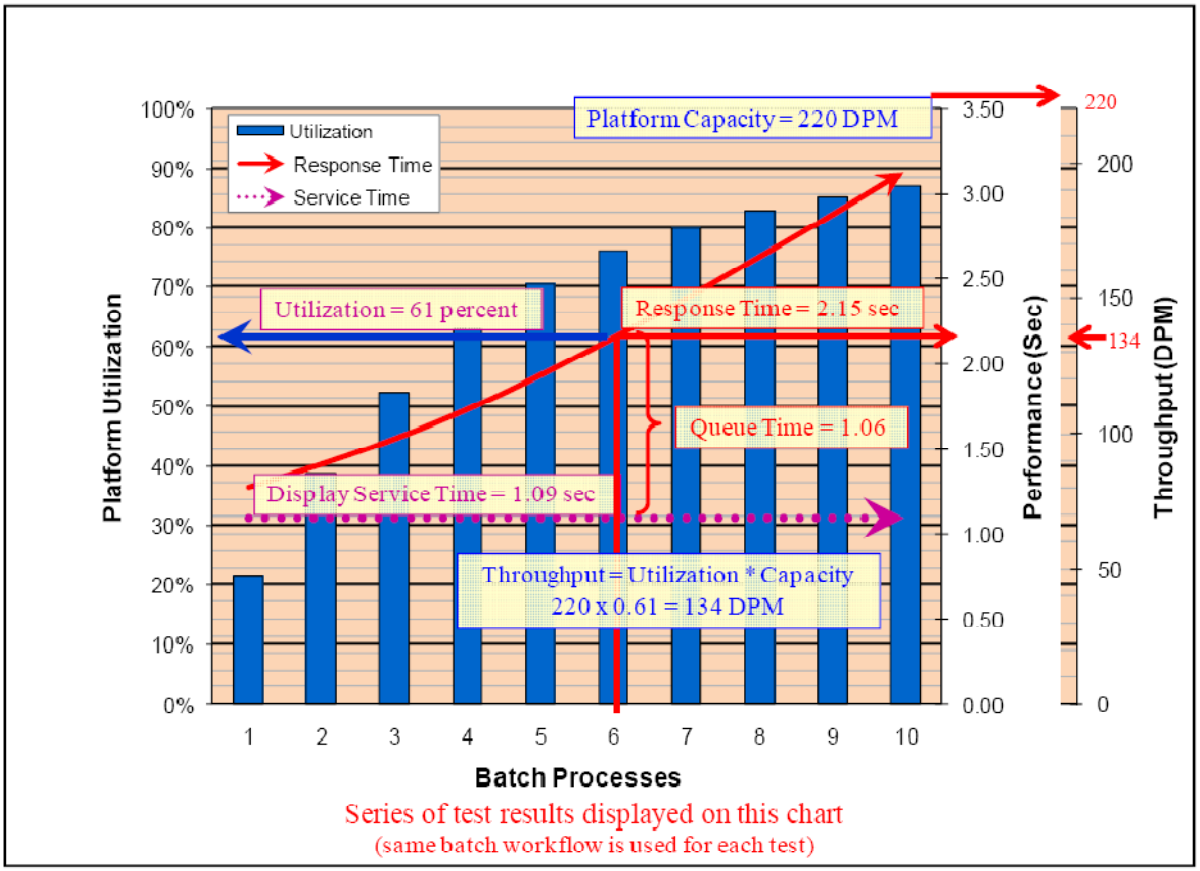


There is a direct relationship between Utilization and Throughput since utilization is defined relative to peak throughput or capacity.

7.2.2 Platform Utilization

Figure 7-7 shows the relationship between platform utilization and performance. During initial processing loads, work transaction response time and service time are the same. As processing loads increase, inbound work transactions start to arrive at the same time. Each processor core can only process one program instruction at a time, so if two different work transaction requests arrive for processing at the same time, one must wait. This wait time (Queue time) increases as utilization increases, and as throughput approaches platform capacity queue time will increase to unacceptable values (throughput will never reach full capacity). Display response time will continue to slow down as throughput increases to full capacity levels.

Figure 7-7
What is Platform Utilization?
(4 core 2 chip test platform)

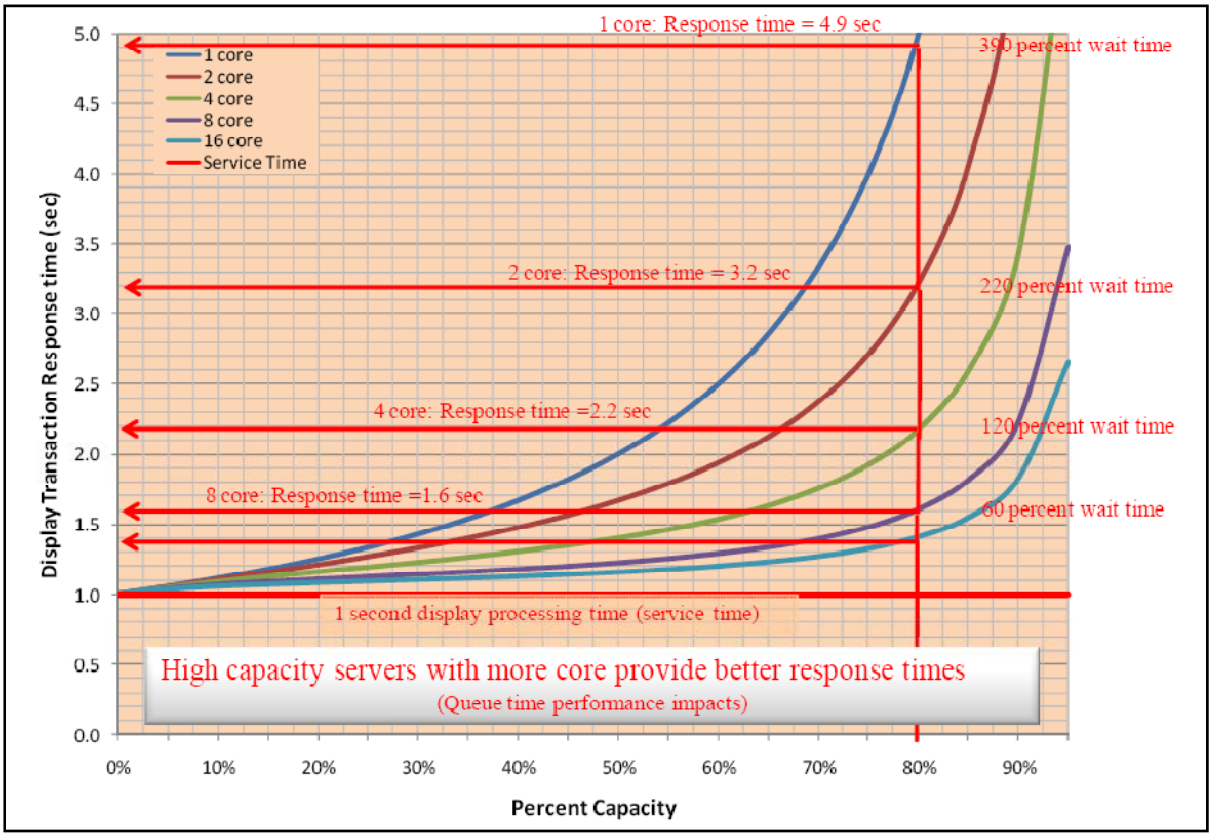


The relationship between service time and response time is demonstrated on the graphic. Response time is service time plus queue time. Initially response time is equal to service time, and during higher system loads response time will increase as queue time increases - display service time will stay the same for all loads.

7.2.3 Transaction Queue Time Sensitivity

Transaction queue time is more significant for heavier processing loads (larger transaction service times increase the probability of concurrent service transactions). Queue time is also sensitive to the number of available service providers, or processor core, available to service the transaction request (eight core server can process high volumes of service requests much more effectively than a single core server). Figure 7-8 shows the relationship between queue time and the number of server processor core.

Figure 7-8
Transaction Queue Time Platform Core Sensitivity

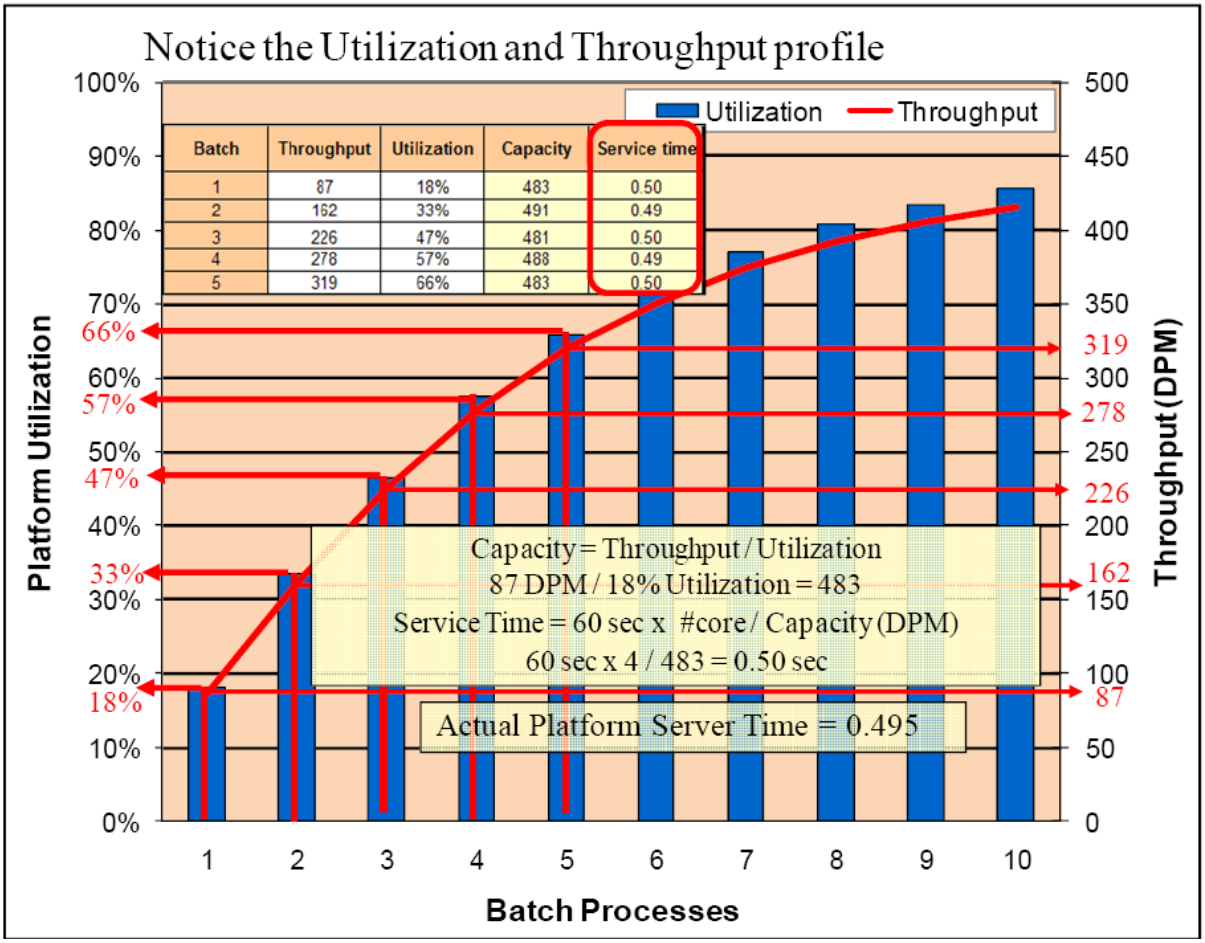


This engineering chart can be used to identify the display response time for a transaction with 1 second service time. Response time as a function of platform utilization is plotted for five different server platform configurations (1 core, 2 core, 4 core, 8 core, and 16 core). A response time comparison is made at 80 percent platform utilization (demonstrates how to use this chart) for the same 1 second display transaction. The single core server response time is 4.9 seconds; while the eight core server response time is 1.6 seconds (over three times faster display response time). This chart can be used to translate display service time to expected user response time for any transaction (multiply the results from this chart by the display transaction service time). Transactions with higher service times will have much higher response times.

7.2.4 Platform Service Time

Once you identify a specific throughput and utilization, you can identify service time. Figure 7-9 shows the direct relationship between utilization, throughput, and service time.

Figure 7-9
Computing Platform Service Times
(4 core 2 chip test platform)



Batch	Throughput	Utilization	Capacity	Service time
1	87	18%	483	0.50
2	162	33%	491	0.49
3	226	47%	481	0.50
4	278	57%	488	0.49
5	319	66%	483	0.50

Platform capacity is a constant value, and can be calculated from any throughput level (you do not need to conduct a peak load test to identify 100 percent platform capacity). If you identify the platform utilization (CPU utilization) during a known throughput level, you can then calculate the average work transaction service time for software components running on that platform. The work transaction service time (transaction processing time per core) is the number of core divided by the platform capacity (capacity measured in work transactions per second).

The spreadsheet in the upper left of the chart shows service time calculations for 1, 2, 3, 4, and 5 batch process loads respectively. Capacity and service time calculated for each throughput were all very close (this assumes you can accurately measure utilization at the lighter test loads). Knowing how to calculate service time any defined platform throughput is important for establishing proper workflow performance targets and verifying capacity planning targets are satisfied when monitoring real system workflow loads.

7.2.5 Performance Processing Delays (calculating display response time)

Most of the performance factors used for system design capacity planning involve simple terms and relationships. A work transaction (display) is an average unit of work, throughput is a measure of the average work transactions completed over a period of time (displays per minute or transactions per hour), capacity is the maximum rate at which a platform can do work, and utilization is the percentage of capacity represented by a give throughput rate. You can calculate display service time if you know the platform throughput and corresponding utilization, calculated at any throughput level.

Calculating user display response time for shared system loads is a little bit more difficult. Only one user transaction can be services at a time on each processor core. If lots of user transaction requests arrive at the same time, some of the transactions must wait in line while the others are processed first. Waiting in line for processing contributes to system processing delays. User display response time must account for all the system delays, since the display is not complete until the final processing is done.

Fortunately, computing transaction service response time is a common problem for many business applications. The theory of queues or waiting in line has its origin in the work of A. K. Erlang, starting in 1909.[1] There are a variety of different queuing models available for estimating queue time, and I went back to one of the textbooks used during my graduate school days to incorporate these models for use in system design capacity planning. The simplest models were for large populations of random arrival transactions, which should certainly be the case in a high capacity computer computation (we are dealing with thousands of random computer program instructions being executed within a relatively small period of time - i.e. minutes).

Figure 7-10 provides an overview of the model used in the Capacity Planning Tool for estimating component queue times. The second half of the model (single core section) was quite strait forward, and there is general agreement that this simple model would identify wait times in the case of a single service provider (single core platform or single network connection). The multi-core case was a little more complicated, and unfortunately was the case when you have a multi-core server platform configuration.

Figure 7-10
Queue Time Performance Factors

$$\text{Queue time } Q_t = \underbrace{\left\{\frac{1}{[1+k^1 U (C_p-1)]}\right\}}_{\text{Multi Core: Service factor}} \times \underbrace{\left\{\frac{S_t U}{(1-U)}\right\}}_{\text{Single Core: Queue}}$$

[1] k factor will depend on the arrival time distribution. Factors are set based on consulting experience. A k factor of 1 shows the best match in comparing capacity planning models to benchmark test results.

In the multi-core service provider case, it was important to include the probability of a service provider (processor core) being available to service the request (not busy) and then multiply this value by the single core factor. The more processor core in the server, the more likely one of these core will not be busy when the next service transaction arrives - thus this is a fraction that grows larger for platforms with more server core. There were some other constraints to consider. The total equation must also be zero when there is no load on the system (queue time = 0 when utilization = 0) and reduce to the simple single core formula when the number of processor core = 1. These considerations were all made in developing the queue time formula presented above.

The k factor was introduced as a variable that could be adjusted to match benchmark test results (the actual

[1] Robert J. Thierauf and Richard A. Groosse, "Queuing Models", Decision Making through Operations Research, *John Wiley and Sons*, Inc, 1970, p. 430 - 454.

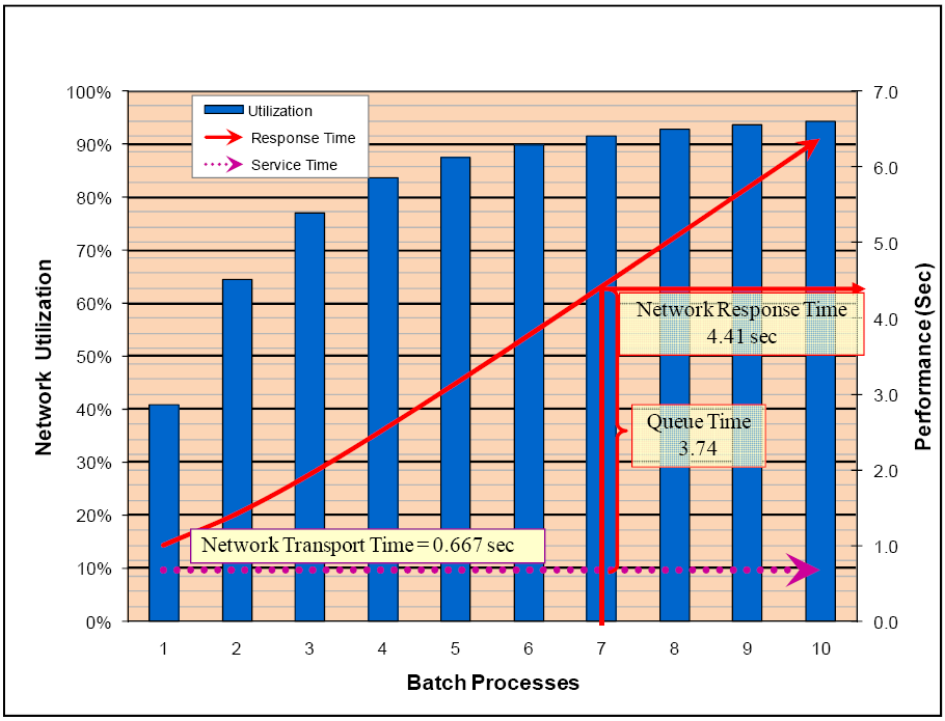
multiple service provider formulas included in the reference textbook included a number of factorial and summation calculations which were far too complex for use in a simple capacity planning tool). This simple formula provided above has been compared against several benchmark test results, and the computed response time was reasonably close to the measure test results (k factor of 1 showed conservative response times - slightly higher than measured values). Increasing the k factor above 1 reduced the computed response time to more optimistic levels. I recommend using a k factor of 1 for capacity planning purposes.

It is important to recognize that the accuracy of the queue time calculation impacts only the expected user response time, and does not reduce the accuracy of the platform capacity calculations provided by the earlier simple relationships. For many years, capacity planning did not include estimates for user response time. If display response times were too slow, the peak throughput estimates would not be achieved and the capacity estimates would be conservative. Including user response time in the capacity planning models can provide more accurate and less conservative platform specifications, and provide customers with a better understanding of user performance and productivity.

7.2.6 Network Communication Performance Factors

Performance models used to support network communications follow the same type of terms and relationships identified above for server platforms, with the only difference being the names of these same terms. Figure 7-11 shows the terms and relationships used for network capacity planning.

Figure 7-11
How do we Size the Network?
(Network bandwidth = 1.5 Mbps)

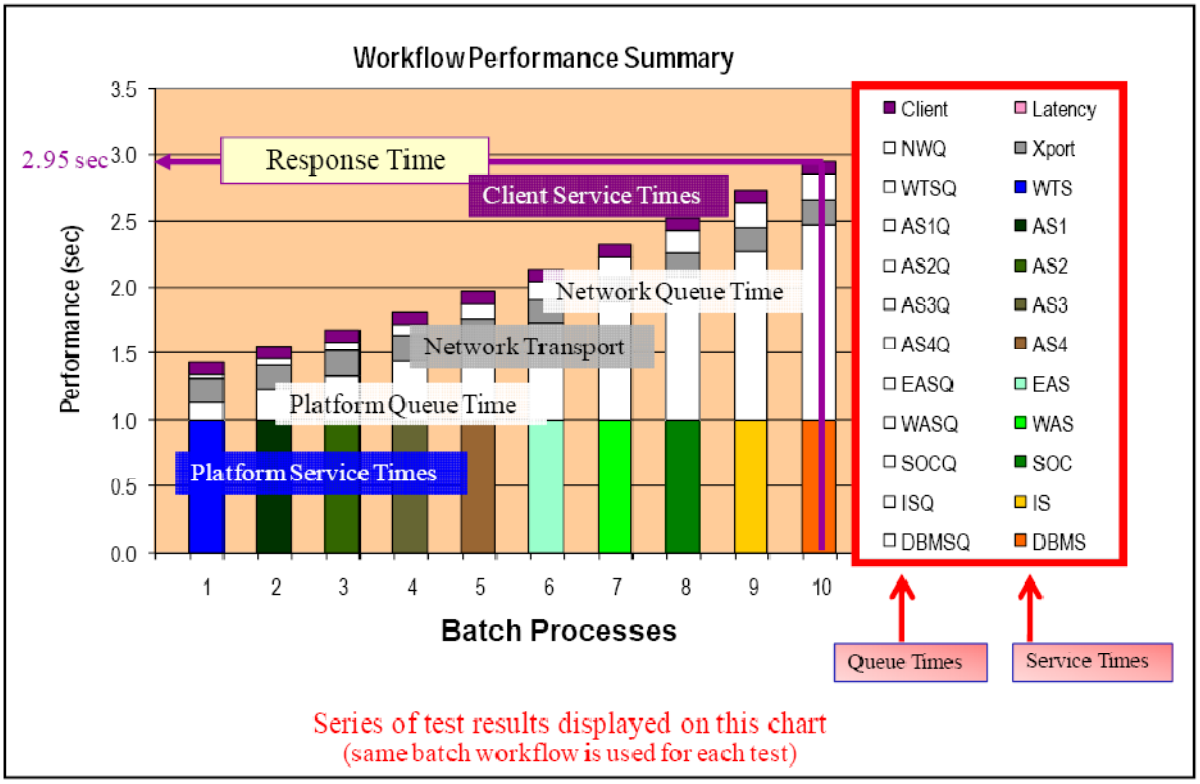


The most important terms include definition of the average work transaction (display), work throughput (traffic), system capacity (bandwidth), and system utilization (network utilization). The network connection (switch port, router port, network interface card, hardware bus adapter, etc) is the hardware that processes the network traffic, so knowing the number of NIC cards available in a selected hardware platform configuration or the number of WAN or Internet connections and aggregate bandwidth is important. Service time (network transport time) is a measure of the average work transaction processing time, and queue time is a measure of the time waiting to be processed (waiting in line for service).

7.3 Work Transaction (display) Response time

Response time is the total time required to complete the work transaction. The software program instructions (work procedure) is executed sequential, thus each instruction in the program must be executed before the next step in the program can be completed (results of the first step in the procedure often must be known before completing the second step, etc). Response time includes all of the processing times and queue times experienced in completing an average work transaction. Figure 7-12 provides a system performance summary that shows service times, network transport time, and associated component service times displayed as a stacked bar chart for each workflow. Display response time is represented at the top of each workflow stack, representing the sum of all service and queue times for each workflow display.

Figure 7-12
What is System Performance?
(4 core 2 chip platform)



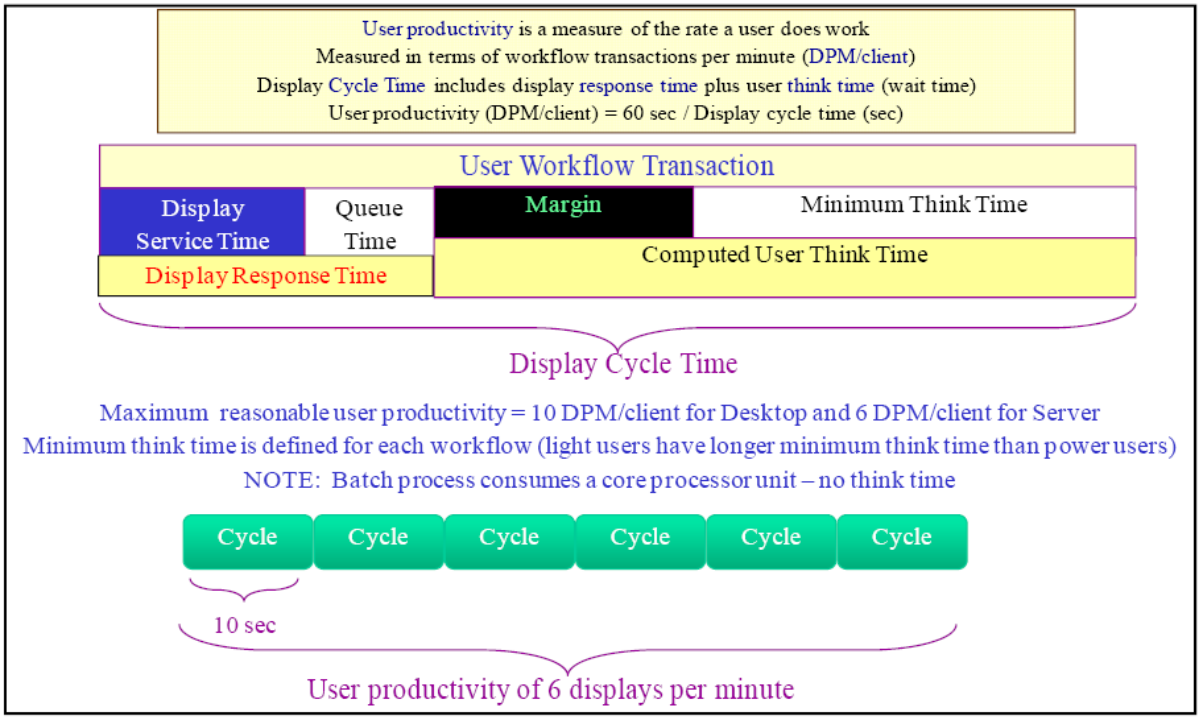
This Workflow Performance Summary chart is included in the Capacity Planning Tool (CPT) showing performance of each of the workflows identified in the CPT requirements module. The chart above shows results of a series of batch load tests performed on a 4 core server platform. Service times for each system component (10 server tier and network with their associated queue time, plus network latency and client display processing time)

Response time is close to the total processing time for the first batch process, and the queue time grows on the server and network components as utilization increases. Queue time is more than processing time with the system is running more than two batch processes per core (total of 8 batch processes on the 4 core server).

7.4 User Productivity

For many years we use concurrent users as our standard unit for quantifying system processing throughput load profiles. In time we found it much more effective to represent user workflows as a combination of peak concurrent users and user productivity, defining the processing loads for an average display as our standard unit of work. The relationship between a concurrent user and a user display is shown in Figure 7-13.

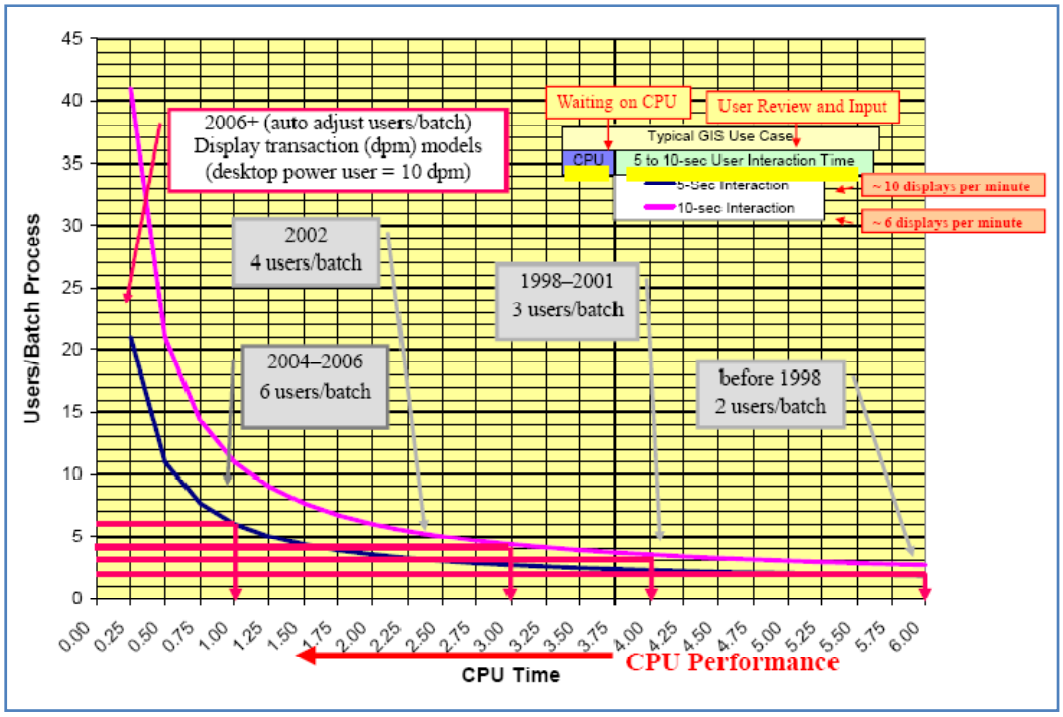
Figure 7-13
User Productivity
(Client Displays per Minute)



A GIS user workflow transaction includes a period of computer work time (display response time) and a period of user work time (user think time). The purpose of a user workflow is to collect input from real people to develop a required information product, update system data repositories, or modify and maintain system data resources. Software programs should be developed to optimize user input time (user think time) and minimize user wait time (display response time). Valid user workflows are based on optimum user productivity (10 displays per minute for desktop applications and 6 displays per minute for Web services) and a reasonable minimum think time (user needs at least 3 seconds to input changes between each display transaction). Zero think time is a batch process - this is not a user workflow. As the display response time increases, the available user think time for a given productivity is reduced. If the computed user think time is less than the minimum think time for a given workflow, the user productivity should be reduced to maintain the minimum think time.

GIS user work profiles have changed considerably over the years, and this change would directly impact the accuracy of our early client/server sizing models. Figure 7-14 shares our experience in maintaining the concurrent user models, and how they had to be modified over the years to keep up with technology change.

Figure 7-14
What is a Real User?



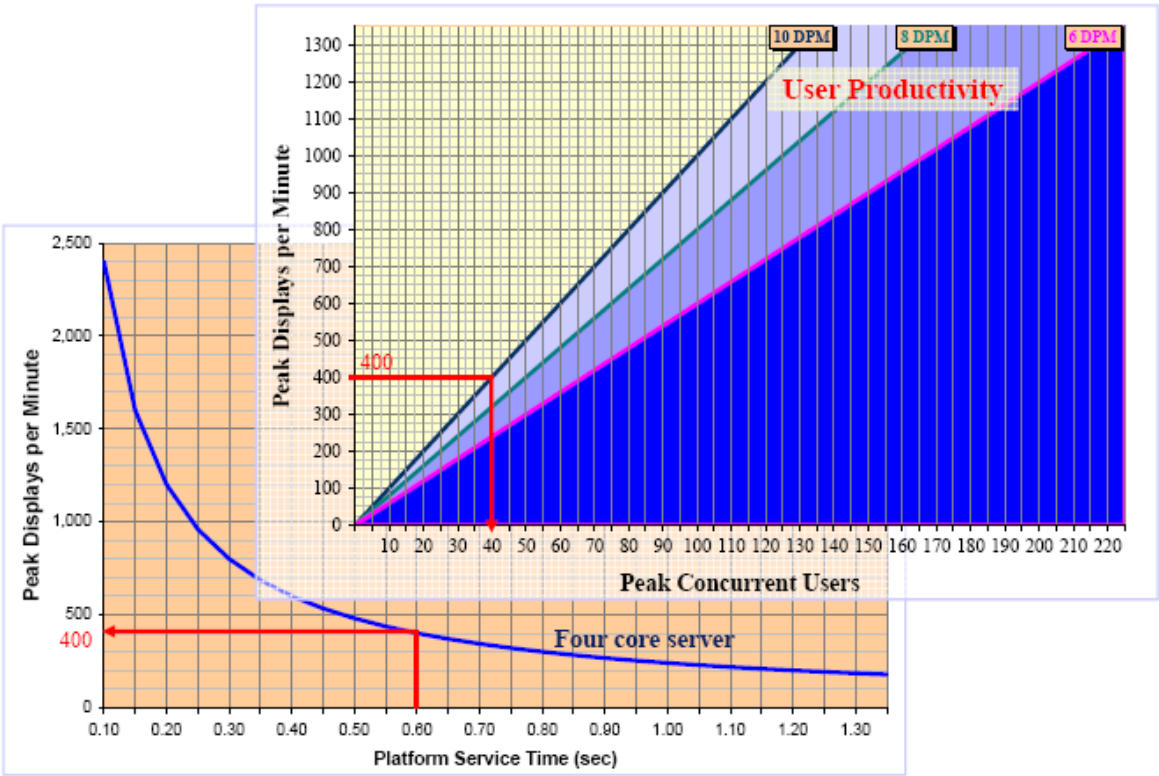
When I joined ESRI in 1990, a typical ARC/INFO workstation map display service time could take as long as 60 seconds. Once the computer completed the map display, the user would think carefully before entering the next map request - a think time that also would take about 60 seconds. In those days, a single CPU could support two (2) concurrent users, each user taking turns using the CPU while the other user was thinking. Since a batch process would consume a CPU, we could represent 2 concurrent users per batch process in our sizing models. Users were quite impressed that the computer could make a map in 60 seconds that used to take hours to build manually from physical mylar overlay sheets - an impressive productivity gain.

As computer performance improved, user productivity also improved. In 1998 our models supported 3 users per batch process, increased to 4 users per batch process in 2002 and 6 users per batch process in 2006. Our new display transaction models were introduced in 2006 - display transaction rates (dpm) were calculated by multiplying peak concurrent users by 10 displays per minute (user productivity). Workflow displays per minute can be modified to ensure users have adequate user think time to support the specific workflow requirements.

7.5 Capacity Planning

The models supporting ESRI capacity planning today are based on the performance fundamentals introduced in this section. Figure 7-15 provides a simple overview of the concepts driving these models, using some simple graphic engineering charts to illustrate some basic performance relationships.

Figure 7-15
Platform Capacity Planning



Platform capacity is determined by the software processing time (platform service time) and the number of platform core, and is expressed in terms of peak displays per minute. Platform capacity (DPM) can be translated to supported concurrent users by dividing by the user productivity (DPM/client).

The performance fundamentals discussed in this chapter do not change with changing technology, and an understanding of these fundamentals will provide a solid foundation for understanding system performance and scalability. Software and hardware technology will continue to change, and the terms and relationships identified in this section can help us normalize these changes and help us understand what is required to support our system performance needs. The next section will discuss Software Performance, providing some insight on how to build GIS applications that will support our system performance needs. The following section will discuss Platform Performance, sharing ways to quantify relative performance of available platform technology and identify how to select the right technology to support your processing needs.

8 Software Performance

This section is provided to share some of our lessons learned about building effective GIS solutions to support operational performance and scalability needs. Software technology allows us to script the procedures of our work, and provide these procedures to computers to optimize our workflow performance. The complexity of these procedures, how applications are orchestrated to support use workflows, and the functions selected to support our GIS display needs have a significant impact on system performance and scalability.

For many years we focused our system design consulting efforts toward identifying and establishing a hardware infrastructure that would support a standard implementation of ESRI software technology. We developed platform sizing models based on hardware performance feedback from successful customer implementations. We would modify our sizing models based on relative performance tests which focused on quantifying any additional processing loads introduced with each new software release.

Many customers deployed GIS solutions focusing on functional requirements with limited attention to performance and scalability. Systems were deployed with existing performance issues, and scalability was not well understood. In some cases performance issues were not identified until peak system loads were needed, and the system failed to support capacity requirements. Consultants were hired to test and tune the final system design, modifying the production applications and database to until performance goals could be achieved.

I noticed over the years that our test and tuning consultants were finding and fixing the same performance issues over and over again, and with some experience we were able to document some best practices for building high performance scalable system. The ArcGIS Server 9 technology today includes access to a broad range of functionality, from simple cached map services that support very high performance and scalability to heavy geoprocessing services that may take hours consuming all available server resources with a single request. Developers need to understand what functional are appropriate, and how to publish these functions in a manner that supports user performance expectations.

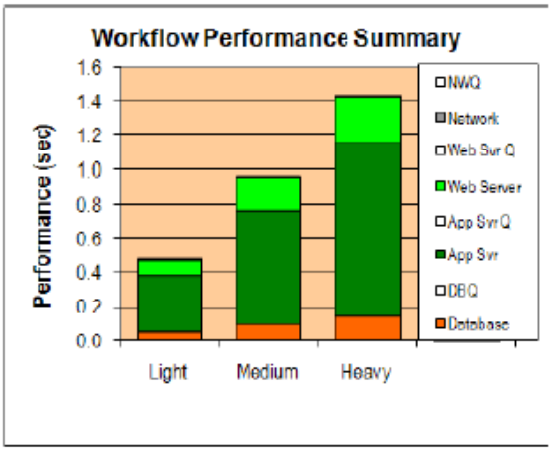
This section shares our lessons learned, and identifies functional areas within the software that can make a difference in meeting customer performance needs. Our new capacity planning tools allow consultants and developers to set appropriate workflow performance targets based on their selected software technology, and include performance measures that can be used during development to ensure performance goals are met during initial system implementation. These same tools can be used to maintain and support deployed operational system performance needs, identifying every day platform performance thresholds that can be used to validate a proper system is deployed to meet peak performance needs.

8.1 Map Display Performance

GIS provides users with a geographic view of their business environment, and for many GIS workflows the map display is used as a common user information resource. The software procedures and functions required to generate the user map display in many cases represents the heavies processing loads required within the user workflow. The map data resources are often shared across a network from a common geodatabase data source, generating a relatively high network traffic and server processing load with each display. The map displays are simple to understand, so often the average user productivity may include up to 10 displays per minute for an active GIS power user. The types of functions, data sources, and design of the user display can make a big difference on the level of processing and network loads required to support a GIS user workflow.

Figure 8-1 shows the processing performance for three different ArcGIS Server dynamic map displays, all deployed on the same platform environment. The performance difference can be traced back to the complexity of the Web application map display document.

Figure 8-1
Optimum Display will make a Difference



The Web light display is generated in less than 0.5 seconds. The Web medium display requires twice the amount of display processing (less than 0.1 second), and the Web heavy display requires over three times the display processing (approximately 1.4 seconds). The design and complexity of the map display can make a very big difference in system performance and scalability. Faster server platforms would result in faster map displays.

The following best practices share lessons learned in designing and tuning for high performance Web services. An optimum display can make a big difference on performance and scalability.

- **Only show relevant data**
  - Start simple
  - Use field visibility
- **Use Scale Dependencies**
  - Appropriate data for given scale
  - Same number of features at all scales
- **Select the right Point representation**
  - Use single layer simple or character markers
  - Use EMF instead of bitmaps
  - Use integer fields for symbol values
  - Avoid halos, complex shapes, masking
- **Select the right Lines and Polygons**
  - Use ESRI Optimized style
  - Avoid cartographic lines and polygon outline
- **Use appropriate Text and labeling**
  - Use annotation instead of labels
  - Use indexed fields
  - Use label and feature conflict weights sparingly
  - Avoid special effects (fill patterns, halos, callouts, backgrounds)
  - Avoid very large text size (60+ pts)
  - Avoid Maplex for dynamic labeling (avoid overuse)

8.1.1 Quality vs. Speed

Figure 8-2 shows the classic tradeoff between quality and performance.

Figure 8-2
Quality versus Speed Tradeoff

- Shaded Relief
- Transparent Layers
- Maplex Labeling

Expensive Functions

- Low-res relief
- Solid colors
- Annotation

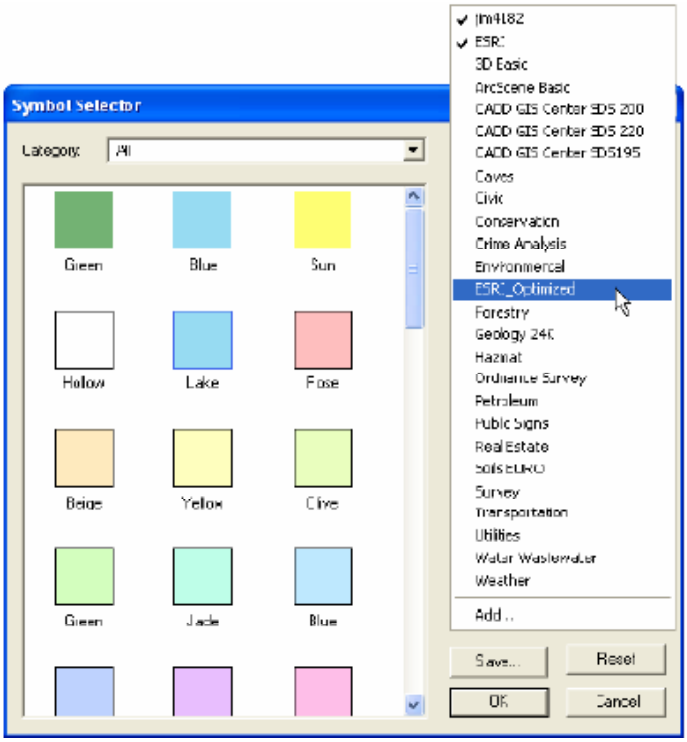
Good Performance

The high quality map above is a shaded relief with transparent layers and dynamic Maplex labeling. These are expensive functions that require extra computer processing for each user display. In contrast, the same display on the right uses low resolution relief, solid colors, and simple annotation providing similar information with good display performance.

8.1.2 Optimizing lines and polygons

We discussed earlier the importance of keeping the display functions simple. The ArcGIS software includes a symbol selection called ESRI Optimized to guide users to the more simple display symbols. Outlines for all fills are simple instead of cartographic lines. Picture fills are EMF-based instead of BMP-based. Figure 8-3 shows the location of the ESRI Optimized symbol selection.

Figure 8-3
ArcGIS Desktop Display Performance



Using ESRI Optimized symbols can improve display drawing performance by up to 50 percent.

8.1.3 GIS Dynamic Map Display Process

GIS displays are normally created one layer at a time, similar to the procedure geographers would follow to layout a map display on a table using Mylar sheets. The technology has changed, but the procedure for building a map is much the same. Maps with a few layers require less processing than maps with many layers (computer programs are more sensitive about the number of layers (feature tables) in a display than about the number of features in a single layer (rows in a feature table).

Figure 8-4 shows the software procedure for building a map display, one layer at a time, joining the features (points, polygons, lines) in each layer sequentially one on top of the other until the final display is complete.

Figure 8-4
Sequential Processing

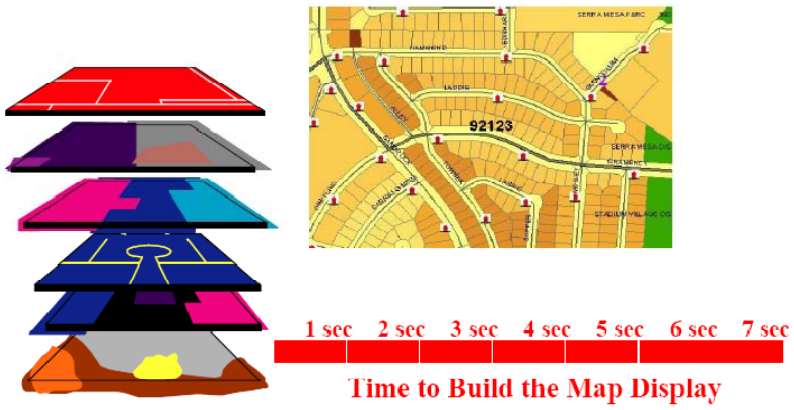


Figure 8-5 shows a software process for building the layers of a map display using a parallel processing procedure. The procedure initiates three separate display requests, each building a third of the display layers. An additional server process then brings the three primary layers together to build the final display.

Figure 8-5
Parallel Processing

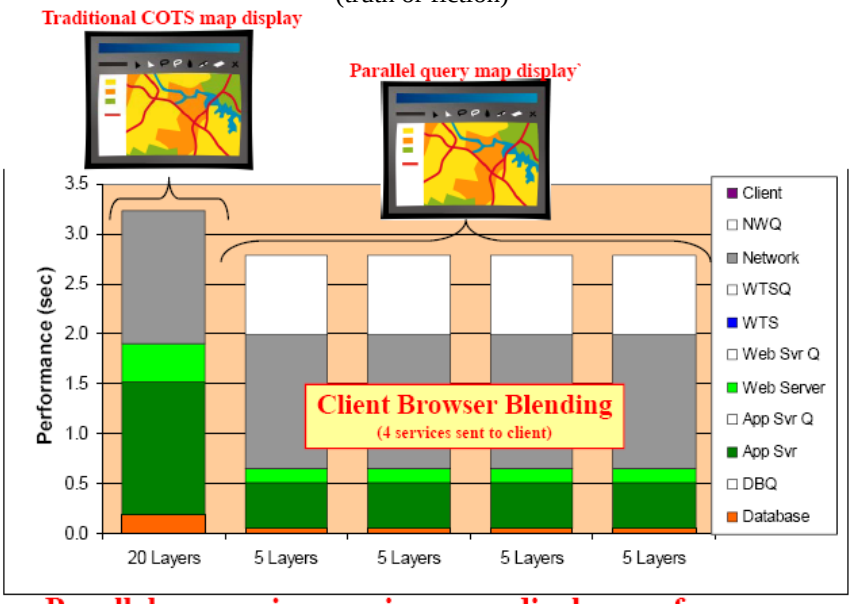
In theory, the second approach generates the display faster. The same amount of processing is required for both methods - in fact the parallel approach requires additional procedures for establishing the parallel display request and then bringing those results back together to produce the final map display.

Hardware vendors are providing computers with an increasing number of processor core per chip, expanding the capacity of the server platforms with reducing expectations for increased performance gains per processor core. Vendors have encourages software developers to take advantage of this increased server capacity by increasing the number of concurrent processes used in generating each user display. Most heavy processing workloads today require sequential processing and a single display generation will not take advantage of multiple processor core. The actual user display performance gains for reprogramming software to take advantage of parallel processing may not be worth the extra programming effort and additional processing loads.

Figure 8-6 compares a traditional COTS map display performance with the performance of a parallel query map display, where the display layers are blended together on the client browser. These types of displays can be published with the current ArcGIS Server 9.3 technology - but is the performance gain worth the use of extra shared infrastructure resources.

Figure 8-6
Parallel Processing Performance Gains

(truth or fiction)



Parallel processing can improve display performance

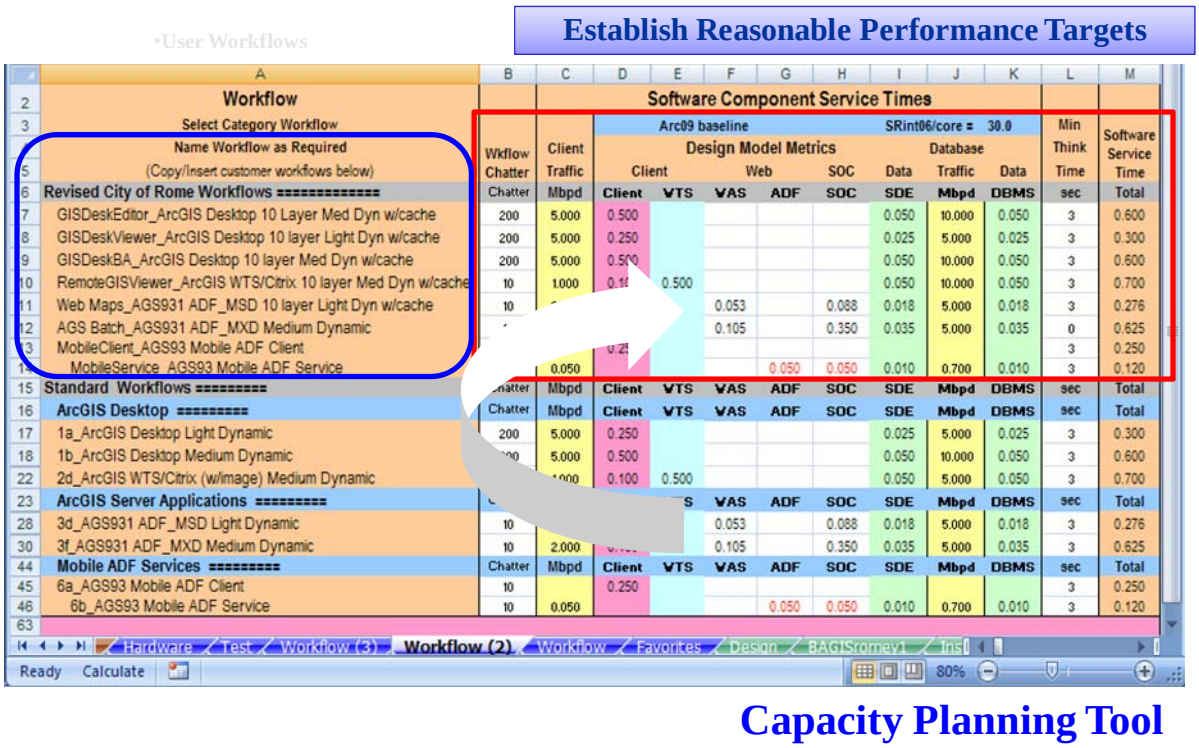
The parallel implementation is supported by three ArcGIS Server REST API map services mashed together in a JavaScript API client browser application. Client access was over a 1.5 Mbps DSL Internet connection, requiring over 1 sec to deliver each 200 KB map image over the network connection to the client browser display. The extra network transport time and queue time to support the parallel display build consumed most of the parallel processing display performance gain. Parallel processing may not always improve system performance, and in some cases could reduce overall system performance and scalability.

8.1.4 Establishing appropriate map display workflow service times

Standard ESRI Workflow software service times are published to provide a common reference for initial capacity planning and to help ESRI customers establish appropriate performance targets when deploying ESRI technology. These software service times represent what many customers are able to do with the ESRI core software technology.

Figure 8-7 provides an overview of the Standard ESRI Workflow software services times provided with the Capacity Planning Tool. Standard workflows are provided for the most common deployment scenarios. This is an example how workflow performance service times are selected from Standard ESRI Workflows to establish reasonable performance targets for workflows used later in the City of Rome use case.

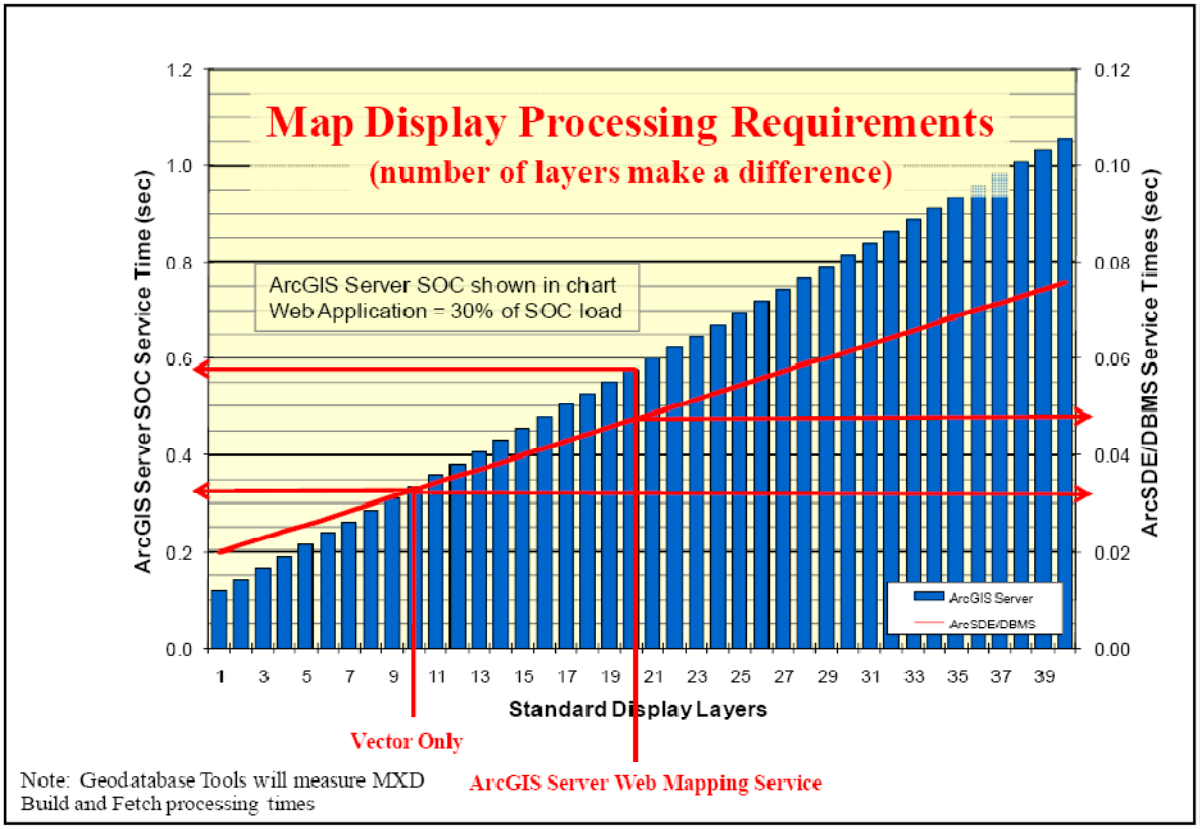
Figure 8-7
ArcGIS Standard ESRI Workflows



	A	B	C	D	E	F	G	H	I	J	K	L	M
2	Workflow		Software Component Service Times										
3	Select Category Workflow			Arc09 baseline					SRint06/core = 30.0			Min	Software
4	Name Workflow as Required	Wkflow	Client	Design Model Metrics					Database			Think	Service
5	(Copy/Insert customer workflows below)	Chatter	Traffic	Client		Web		SOC	Data	Traffic	Data	Time	Time
6	Revised City of Rome Workflows ==========	Chatter	Mbpd	Client	VTS	VAS	ADF	SOC	SDE	Mbpd	DBMS	sec	Total
7	GISDeskEditor_ArcGIS Desktop 10 Layer Med Dyn w/cache	200	5.000	0.500					0.050	10.000	0.050	3	0.600
8	GISDeskViewer_ArcGIS Desktop 10 layer Light Dyn w/cache	200	5.000	0.250					0.025	5.000	0.025	3	0.300
9	GISDeskBA_ArcGIS Desktop 10 layer Med Dyn w/cache	200	5.000	0.500					0.050	10.000	0.050	3	0.600
10	RemoteGISViewer_ArcGIS WTS/Citrix 10 layer Med Dyn w/cache	10	1.000	0.1[illegible]	0.500				0.050	10.000	0.050	3	0.700
11	Web Maps_AGS931 ADF_MSD 10 layer Light Dyn w/cache	10	[illegible]			0.053		0.088	0.018	5.000	0.018	3	0.276
12	AGS Batch_AGS931 ADF_MXD Medium Dynamic	[illegible]	[illegible]			0.105		0.350	0.035	5.000	0.035	0	0.625
13	MobileClient_AGS93 Mobile ADF Client			0.25[illegible]								3	0.250
14	MobileService_AGS93 Mobile ADF Service	[illegible]	0.050				0.050	0.050	0.010	0.700	0.010	3	0.120
15	Standard Workflows =========	[illegible]atter	Mbpd	Client	VTS	VAS	ADF	SOC	SDE	Mbpd	DBMS	sec	Total
16	ArcGIS Desktop =========	Chatter	Mbpd	Client	VTS	VAS	ADF	SOC	SDE	Mbpd	DBMS	sec	Total
17	1a_ArcGIS Desktop Light Dynamic	200	5.000	0.250					0.025	5.000	0.025	3	0.300
18	1b_ArcGIS Desktop Medium Dynamic	[illegible]00	5.000	0.500					0.050	10.000	0.050	3	0.600
22	2d_ArcGIS WTS/Citrix (w/image) Medium Dynamic	[illegible]	[illegible]000	0.100	0.500				0.050	5.000	0.050	3	0.700
23	ArcGIS Server Applications =========	[illegible]	[illegible]	[illegible]	[illegible]S	VAS	ADF	SOC	SDE	Mbpd	DBMS	sec	Total
28	3d_AGS931 ADF_MSD Light Dynamic	10				0.053		0.088	0.018	5.000	0.018	3	0.276
30	3f_AGS931 ADF_MXD Medium Dynamic	10	2.000	[illegible]		0.105		0.350	0.035	5.000	0.035	3	0.625
44	Mobile ADF Services =========	Chatter	Mbpd	Client	VTS	VAS	ADF	SOC	SDE	Mbpd	DBMS	sec	Total
45	6a_AGS93 Mobile ADF Client	10		0.250								3	0.250
46	6b_AGS93 Mobile ADF Service	10	0.050				0.050	0.050	0.010	0.700	0.010	3	0.120
63													

Simplifying the configuration and design of the user map display is one way to reduce workflow processing loads. Processing base map layers in a map cache can reduce the number of dynamic layers required when requesting each map display. Cached layers are pre-processed, and sent directly to the client browser with negligible server processing. Figure 8-8 provides a conceptual example of how ArcGIS Desktop workflow loads can be adjusted by reducing the number of display layers.

Figure 8-8
ArcGIS Server Display Performance



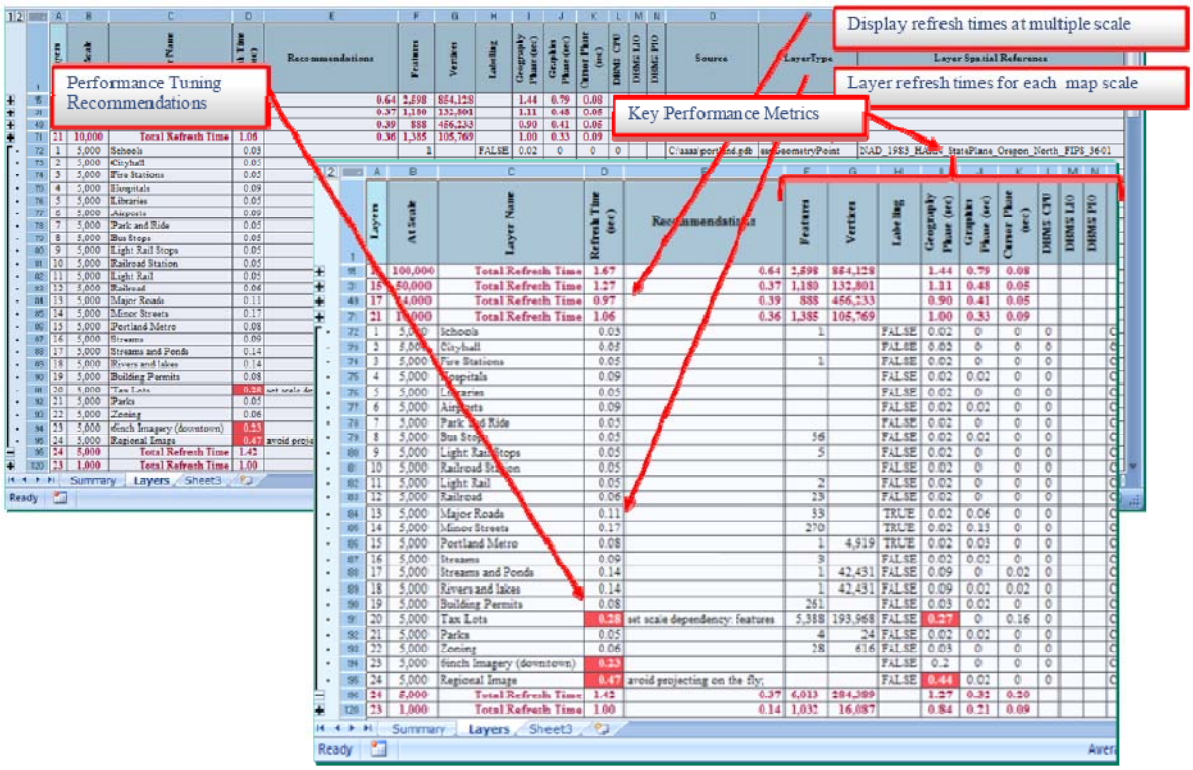
The example ArcGIS Server Web Mapping Service display software service times (SOC = 0.58 sec, ADF = 0.17 (0.58 x 30%), SDE = 0.048 sec, DBMS = 0.048 sec) can be reduced by simplifying the display. Ten of the display layers are relatively static, and can be pre-processed as a basemap and delivered to the browser from a map cache. Creating the map cache reduces the number of dynamic map layers from 20 to 10, which reduces the processing requirements represented by the adjusted service times (SOC = 0.33 sec, ADF = 0.01, SDE = 0.032 sec, DBMS = 0.032 sec). This shows an example in concept only, in real environments all display layers are not created equal - in fact the number of features in a layer and the complexity of the layer functions will make a difference on processing time for each layer. Performance tools for measuring individual MXD layer processing performance can be very useful for identifying performance issues and making the most effective display modifications.

It is important for me to state that the average workflow service times used for capacity planning may not be the same value as a specific workflow map display. Some workflow displays will be heavier, while other displays are lighter. User productivity (displays per minute) is also a factor in defining workflow processing loads. The workflow service times and user productivity together should represent the average user loads applied to the system over time. It is best to establish actual workflow service time loads from platform throughput and CPU utilization measurements collected during real user operations - it is very difficult to accurately simulate or estimate these loads by measuring single map display processing times. This does not preclude setting

appropriate workflow performance targets for capacity planning purposes. These performance targets will be directly influenced by the number of map display layers, the features per layer, and the complexity of the functions used in building the display.

Figure 8-9 provides a sample output generated by the MXDPerfmon performance measurement tool. This tool is available for download as an ArcScript published on http://arcscripts.esri.com. Results below were copied into Microsoft Excel for display purposes.

Figure 8-9
MXDperfstat ArcGIS Map Display Performance Results



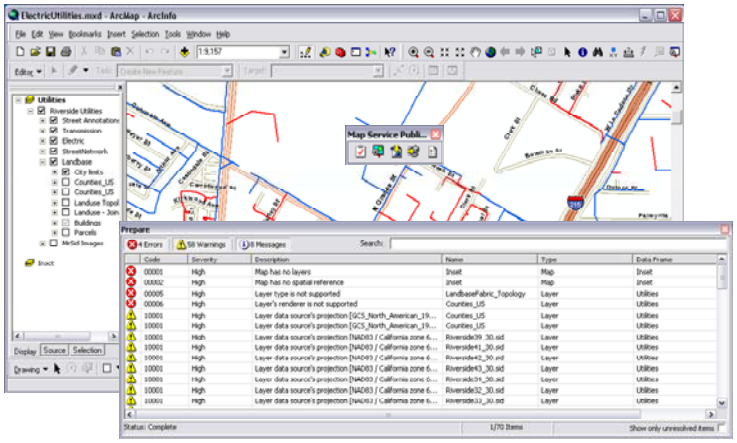
The MXDperfstat tool identifies display refresh times at multiple scales, shows layer refresh times for each map scale, provides layer performance statistics such as number of features, vectors, labeling, and breaks out display time for several key rendering phases (geography, graphics, cursor, and database). The tool also provides some high level recommendations for performance tuning (actually, once you see the layer processing time and the performance metrics the display problems are quite obvious).

ArcGIS 9.3.1 introduced some new map optimization tools incorporated in the Map Service Publishing tool bar. These tools are similar to MXDperfstat in that they render the selected ArcGIS Desktop MXD and analyze each display layer. With the ArcGIS 9.3.1 optimization tools, the results will be provided for the current map scale only - so you would need to move the display to each area you would like to evaluate. Figure 8-10 provides an overview of the optimizing tool output.

Figure 8-10
ArcGIS 9.3.1+ Map Display Optimization Tools

Analyze map directly from ArcMap

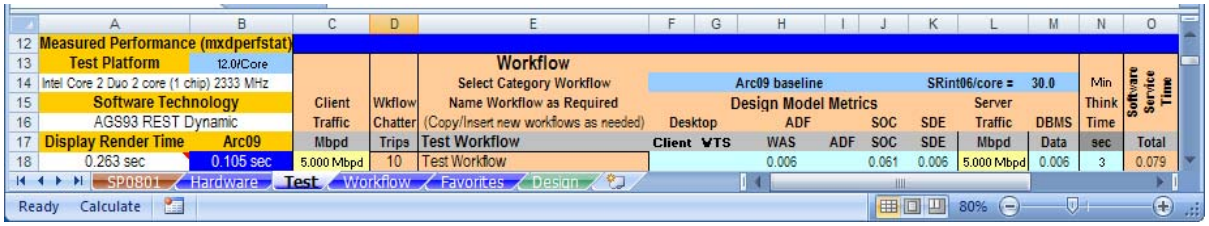
- Analyze function included in Map Service Publishing tools
- Preview allows for visual analysis of performance and graphic quality



Some of the key benefits of the ArcMap optimization tools are that they are directly linked to ESRI Help and functions that will direct you to fixes.

The mxdperfstat and ArcMap optimization tools both generate MXD map document rendering times that represent complexity of the current ArcMap display. The Capacity Planning Tool Test Tab includes a workflow service time calculator function that can generate custom workflow service times based on the measured display rendering times provided by the mxdperfstat or ArcMap optimization performance measurements. Figure 8-11 provides an overview of the Capacity Planning Tool MXDperfstat workflow service time generator tool.

Figure 8-11
Capacity Planning Tool MXDperfstat Workflow Service Time Calculator



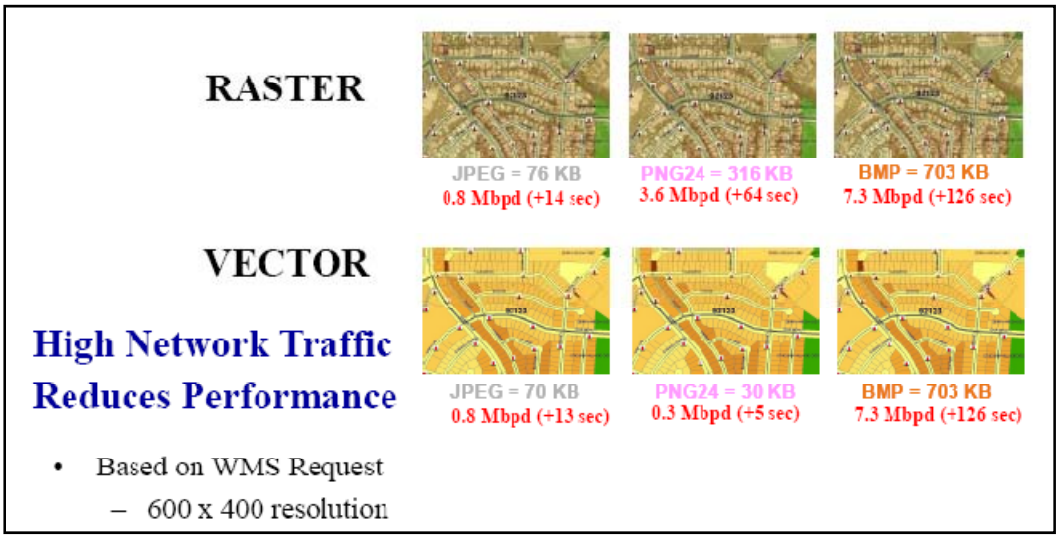
We are making a big effort to establish a transfer function between ArcGIS Desktop MXD display rendering times and appropriate service times for the various ArcGIS mapping software deployment patterns. Map display complexity includes a high number of performance variables that impact system performance, and these tools provide a simple way to measure map complexity. The capacity planning tool can translate measured results from the test platform to baseline service times that can be used for capacity planning.

8.2 Selecting the right image format

Web mapping services produce map images that are sent to the client browser for display. Each user request will generate a new map image that must be delivered to the client browser. The selected image type can have a direct impact on the volume of network traffic. Lighter images require less display traffic and heavier images require more display traffic. The required amount of traffic per display can have a significant impact on user performance over lower bandwidth.

Figure 8-12 identifies the amount of data required to support three common image types (JPEG, PNG24, BMP). The volume of data (KB) required to support the same map display varies with each image type. Transport times are provided to represent display performance impacts over low bandwidth (56 Mbps client connection). A common resolution of 600 x 400 pixels was used for image type comparison. Vector only images would compress better than images than include a Digital Ortho raster layer.

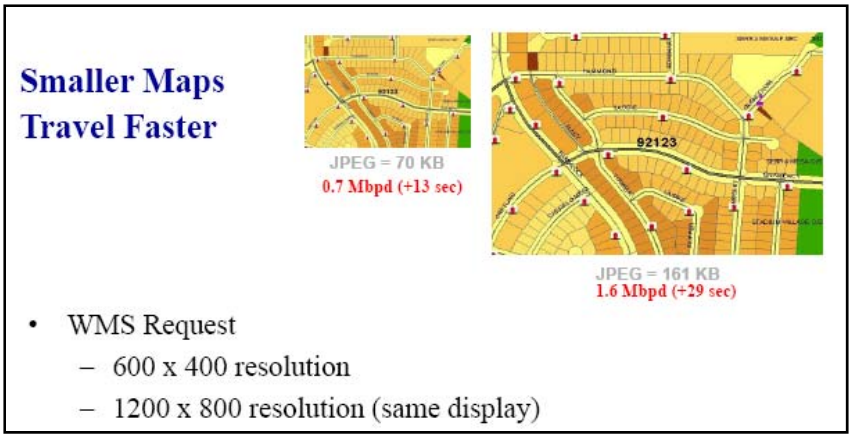
Figure 8-12
Selecting the right Web Map Image Type



JPEG image types provide the most consistent compression, with a slight variation between raster and vector images. PNG24 images do much better with vector data than with raster - PNG supports transparencies and is the default ArcGIS Server format. BMP is a very heavy image format and should not be used with Web services.

The resolution of the image is also a very important consideration. Figure 8-13 compares the JPEG data volume (KB) when increasing the display extent from 600 x 400 to 1200 x 800 resolution. Doubling the display resolution more than doubles the data volume.

Figure 8-13
Configuring the right Web Map Image Resolution



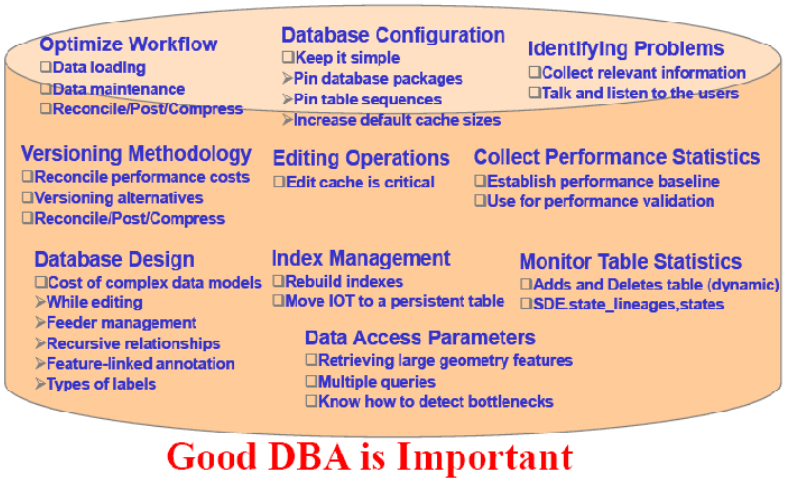
The resolution of the map display image can be an important consideration for many GIS user workflows. A high resolution display is important for users that work with map displays throughout the day. A display resolution of 1280 x 1024 would be common for a standard GIS power user or maintenance workflow. Using a simple Web browser display interface for these heavier GIS workflows may not be practical.

8.3 Providing the right data source

There are several types of data sources available for GIS users. The most common enterprise data source is the ArcSDE Geodatabase. The geodatabase provides the most efficient and effective way to integrate and manage your enterprise GIS operational data resources.

Building and maintaining a high performance geodatabase is very important. The enterprise geodatabase is a central repository shared by all enterprise workflows. A properly tuned database can support many users with minimum processing load (less than 10 percent of most workflow processing happens on the database server), yet the database has been the primary system component for testing and tuning when addressing system performance problems. Figure 8-14 provides a high level presentation that highlights the importance of a trained database administrator responsible for tuning and managing the ArcSDE Geodatabase.

Figure 8-14
Geodatabase Performance



There are some simple things to keep in mind when building and maintaining a geodatabase. Simple database models perform better than complex data models. Optimize database workflows, collect and monitor performance statistics, use reconcile/post/compress functions periodically, use edit cache for edit workflows, and rebuild indexes to maintain optimum queries. An enterprise geodatabase needs an administrator that is trained on how to support and maintain a high performance database. If a complex data model is needed for an maintenance database, it might be wise to use geodatabase replication to include a separate distribution database that can perform more efficiently than the primary maintenance environment. Selecting and maintaining the right data source can make a difference on system performance and scalability.

8.4 Building high performance Web applications

There are variables within the Web services architecture that can be measured and modified to improve site performance. Figure 8-15 identifies the associated performance measurements and configuration variables available in tuning an ArcIMS configuration.

Figure 8-15
Web Services Performance

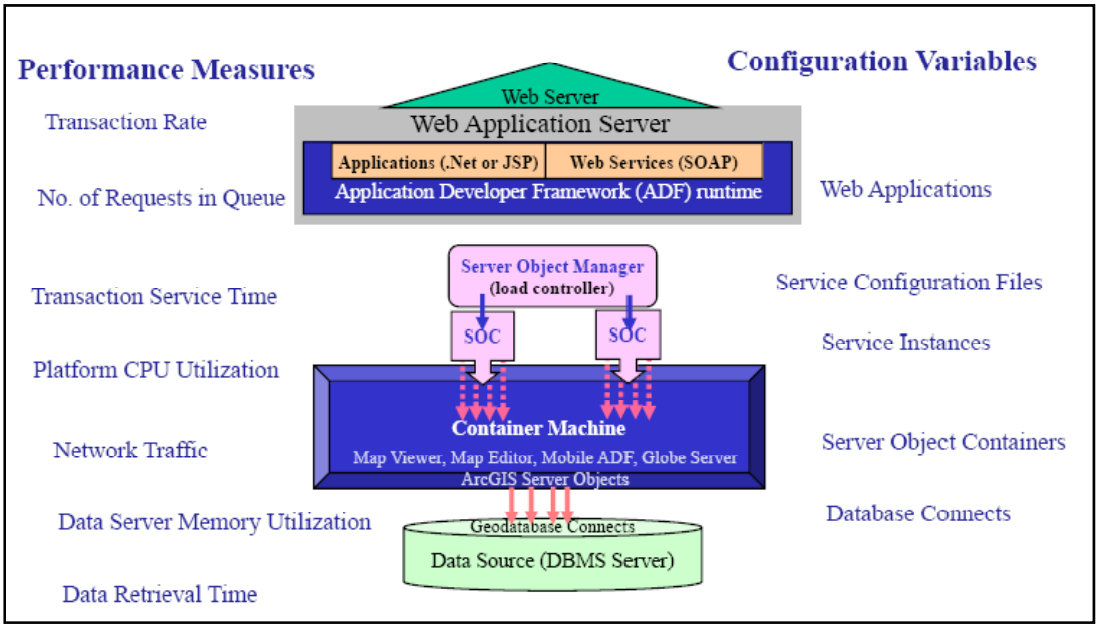


Figure 8-16 identifies performance variables you can measure and how you can modify the map services or component variables to optimize site performance.

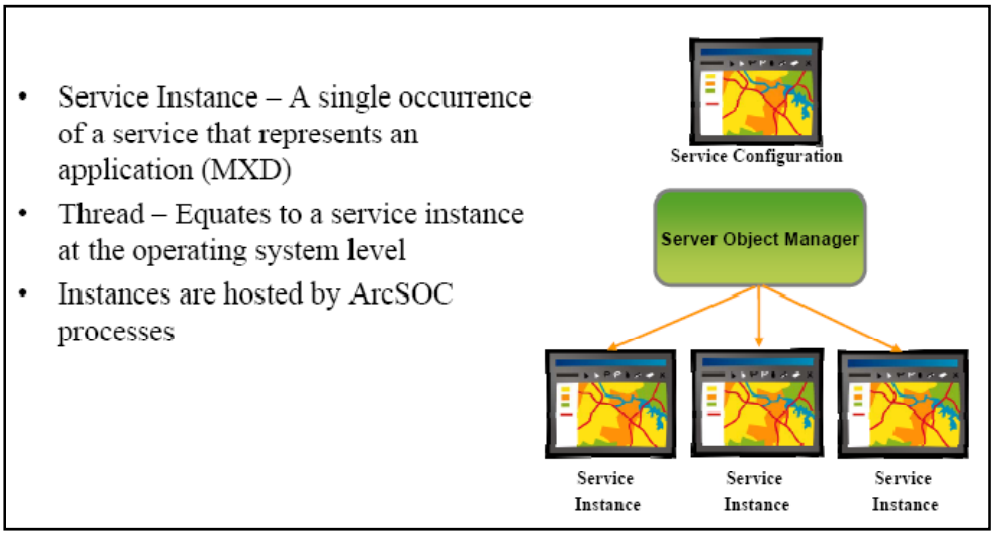
Figure 8-16
Web Mapping Services Performance Tuning Guidelines

Performance Measure	Tuning Options
Transaction Rate	Transaction rate identifies the number of requests supported by the site configuration. Peak transaction rate is the maximum capacity of the site to support incoming requests. Peak capacity can be increased by reducing the time required to generate each map service (simpler information products) or by increasing the number of CPUs (more processing power). Sufficient service agent threads should be included to take full advantage of the available CPUs (usually 2–3 threads per CPU is sufficient). Once CPUs are fully utilized, additional threads will not improve site capacity.
No. of Requests in Queue	The server object manage acts as a processing queue for inbound requests. Requests are held in the queue until there is a service thread available to process the request. If a request arrives and the service queue is full, the browser will receive a server busy error. Queue depth can be increased to avoid rejecting browser requests.
Transaction Service Time	This is the CPU time required to process the published service once it has been assigned to a service instance for processing. Long service times can significantly reduce site capacity and should be avoided if possible. Simple map services (minimum number of layers and simple display functions) can significantly improve site capacity.
CPU Utilization	Sufficient service instances should be configured to support maximum CPU utilization. Maximum site capacity is reached when CPU utilization reaches a peak level (close to 100% utilization). Increasing service instances beyond this point without increasing the number of processor core will increase the average client reponse time during peak system loads.
Network Traffic	Sufficient network bandwidth must be available to support display transport to the client browser. Network bottlenecks can introduce serious client response delays. Bandwidth utilization can be improved by publishing simple map services, keeping image size from 100 KB to 200 KB, and ensuring sufficient bandwidth to support peak transaction loads.
Data Server Memory	Sufficient physical memory must be available to support all processing and adequate caching for optimum performance. Memory utilization should be checked once the system is configured to ensure more physical memory exists than what is being used to support the maximum production configuration.
Data Retrieval Time	This is the CPU processing time on the ArcSDE server. Query time can be optimized by proper indexing and tuning of the ArcSDE geodatabase.

8.4.1 Configuring the service instances

Web services must be configured with the appropriate number of service instances to take full advantage of the licensed hardware. Each map service is identified by a service configuration, which is represented by a map document (MXD). The Server Object Manager (SOM) deploys the service instances for each service configuration. Figure 8-17 shows the relationship between service instances, threads and ArcSOC processes.

Figure 8-17
Service Instances, Processes, and Threads

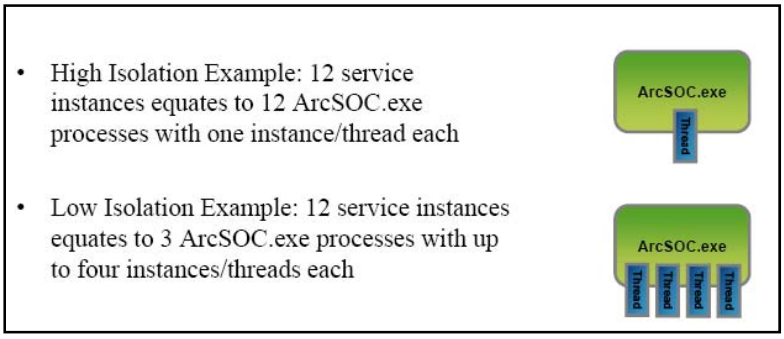


From a user perspective, service instances and threads mean the same thing. The service instances managed by the SOM are hosted by the ArcSOC processes (SOC) on the container machine tier.

8.4.2 Selecting high isolation or low isolation

Figure 8-18 shows two different service configurations for an ArcSOC process (high isolation or low isolation). A high isolation SOC supports only one service thread (instance). A low isolation SOC supports more than one thread (ArcGIS Server 9.2 supports up to 4 threads, and ArcGIS Server 9.3 supports up to 8 threads).

Figure 8-18
Server Object Container (SOC) Isolation



The low isolations SOC are supported by a multi-threaded ArcSOC process. A multi-threaded process is one set of executables shared by more than one service instance, each service instance supported by separate pointers within the ArcSOC process. The ArcSOC process is designed to support execution of the available service instances in parallel (no processing conflicts). Available SOC Threads can support execution of assigned service requests in parallel on separate processor core located on the same platform. The primary

advantage of using a low isolation SOC is to reduce the server platform memory footprint (one set of ArcSOC executables supporting multiple service instances). Single SOC thread failure will kill the complete SOC process, so the high isolation SOC configuration is preferred when sufficient physical memory is available to support the isolation configuration. Standard memory recommendations (2 GB per platform core) should be adequate to support high isolation SOC instances for most ArcGIS Server customer configurations.

8.4.3 Selecting a pooled or non-pooled service model

ArcGIS Server provides options for two different service models (Pooled and Non-pooled). The pooled service model published service instances that can be shared to provide optimum service for large populations of users. The non-pooled service model assigns each concurrent user session to separate dedicated ArcSOC processes.

Figure 8-19 provides an example of the pooled service model. Many Web application users are accessing the deployed pool of shared service instances. This is the most efficient way to support a large population of users accessing standard Web mapping services.

Figure 8-19
Pooled Service Model

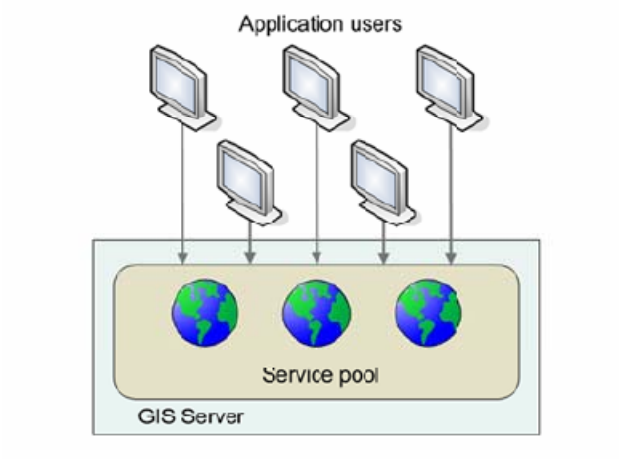
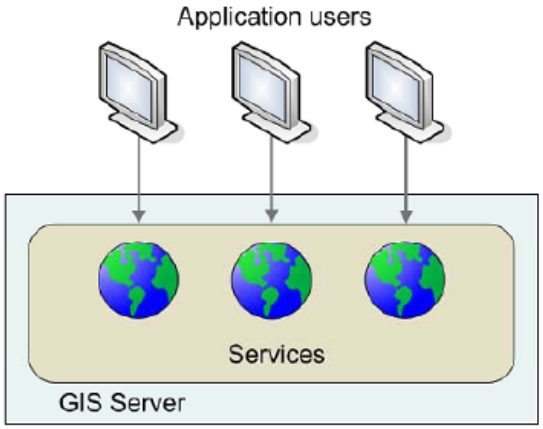


Figure 8-20 provides an example of the non-pooled service model. Each Web application user is assigned a dedicated ArcSOC process to support their workflow. A non-pooled service model would normally be used only when the workflow application requires state changes be maintained at the ArcSOC process level.

Figure 8-20
Non-pooled Service Model

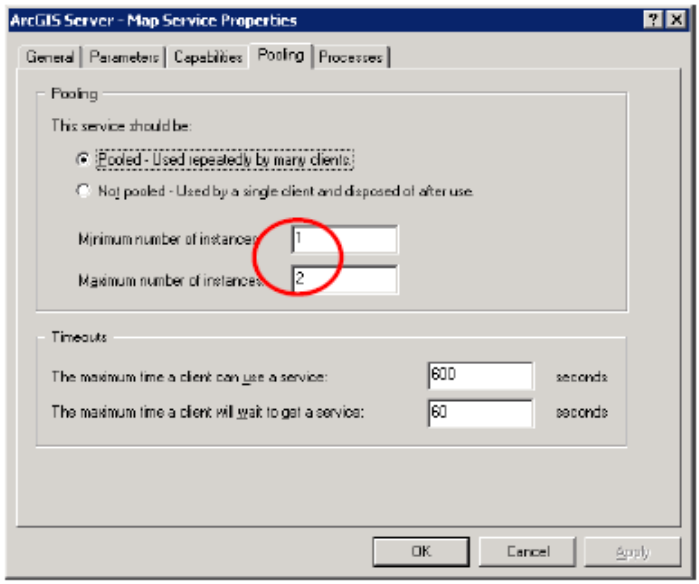


The pooled service model should be used for all service configurations except those that require a non-pooled session. Non-pooled sessions are map editor workflows where MXD context changes are maintained within the ArcSOC process.

8.4.4 Configuring SOM service instances

The SOM service configuration instance settings are very important system capacity parameters. Understanding how the SOM uses these parameters to manage Web site service capacity will help optimize utilization of the deployed hardware platform configuration. Figure 8-21 provides a view of the ArcGIS Server Map Service Properties template Pooling tab where the service instance configuration is defined.

Figure 8-21
Configuring a Pooled Service



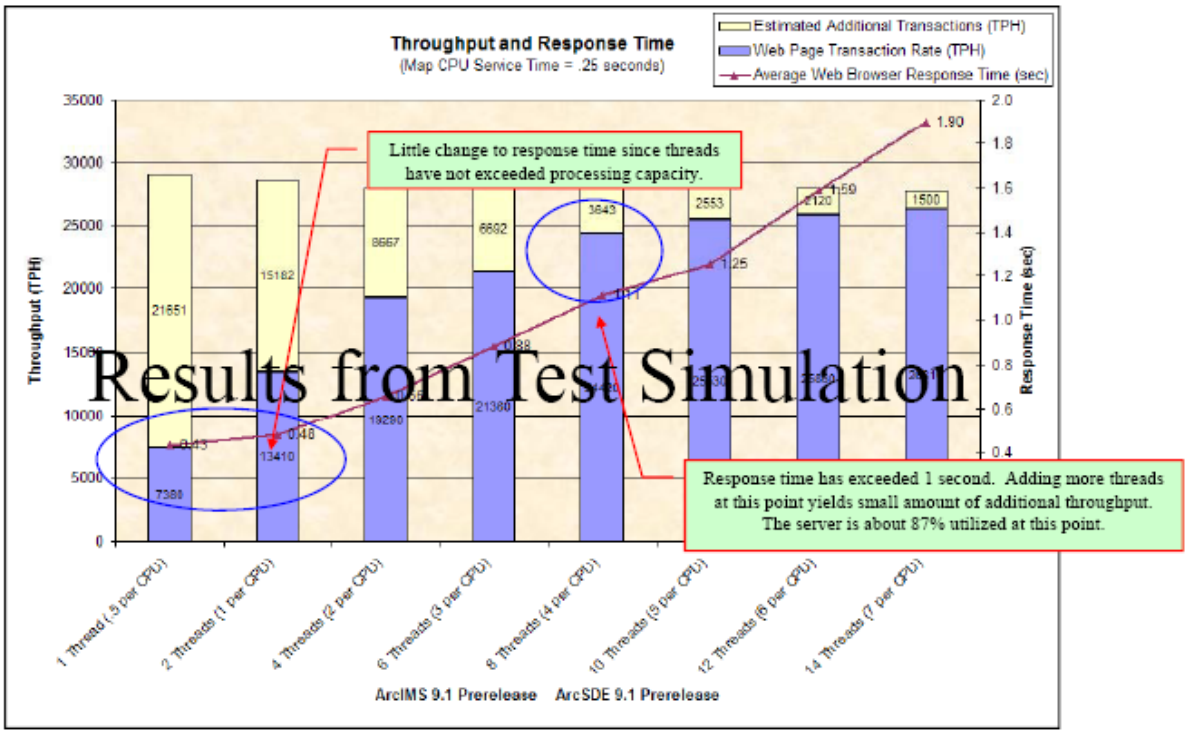
Map Service Properties must be defined for each published service configuration (map service). The minimum number of instances property identifies the number of service instances deployed by the SOM during startup and maintained on the host platforms during light operations. The maximum number of instances property identifies the maximum instances the SOM will deploy during peak load conditions. The SOM will distribute the assigned instances across the available SOC platforms, and will increase and decrease the number of deployed instances within the identified minimum and maximum instance properties based on demand for this service.

8.4.5 Configuring host instance capacity

Each host container machine will have a limited number of processor core (4 core, 8 core, etc) that can be used for processing service requests. Each processor core can handle one service request at a time. Sufficient number of service instances must be deployed to take full advantage of the available processor core. A configuration that allows too many concurrent instances on a limited number of processor core can experience performance degradation during heavy processing loads. Having the right number of service instances to support peak service loads can optimize peak site capacity and user display performance.

Figure 8-22 shows results from a series of tests that demonstrate the optimum map service configuration to optimize system capacity and user display performance.

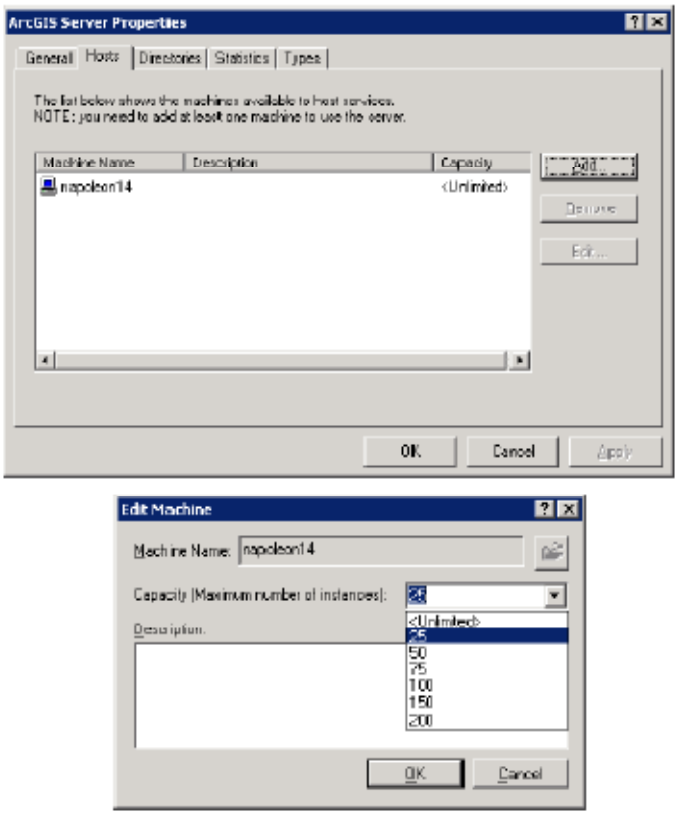
Figure 8-22
Performance Test Results
(Number of Threads makes a difference)



The results show a tradeoff between optimum capacity and user display performance. As the number of available service instances (threads) increased the peak system throughput and utilization would also increase. As utilization increased above 50 percent of available platform capacity, user display response time would increase; slowing more and more as additional service instances were included in the configuration. The optimum configuration was achieved with 4 service instances (threads) per server core (CPU).

Figure 8-23 shows a view of the Map Services Properties Hosts tab and how the host capacity properties are defined. The default host capacity value is unlimited, which would allow the SOM to deploy the maximum number of instanced defined in all the service configurations at the same time. The host capacity value should be set to optimize system performance during peak system loads. Our recommended setting would be four (4) instances per host platform core (16 instances for 4 core server, 32 instances for 8 core server, etc).

Figure 8-23
Configuring Host Capacity



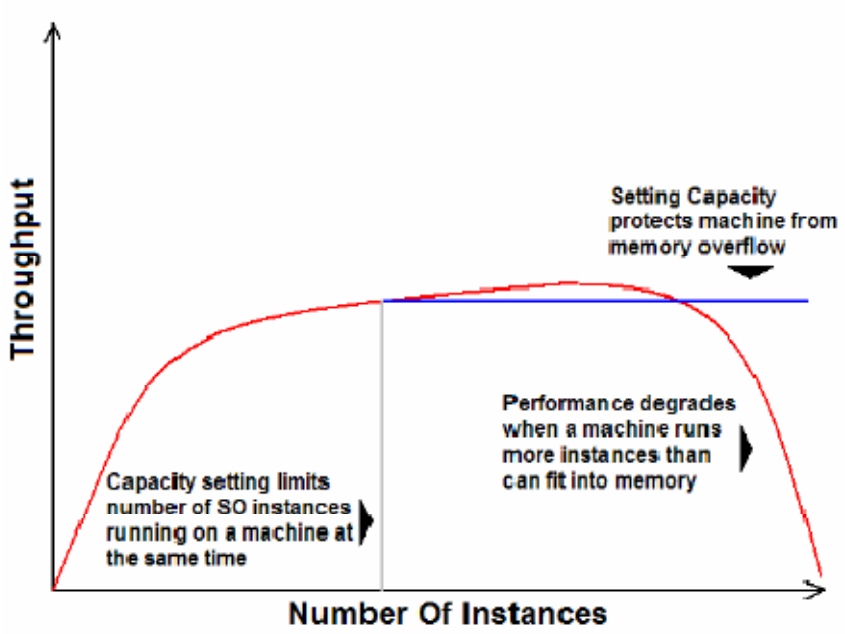
When configuring high available Web configurations with multiple SOM, separate host capacity configurations must be defined for each SOM environment (each SOM will deploy separate service instances on the assigned host machines - the SOM are not aware of each other and will function as separate environments). System load balancing is managed by each SOM to optimize system performance and scalability - system should be configured as discussed earlier in Section 4 (Product Architecture).

8.5 Selecting the right physical memory

The platform configuration must include sufficient physical memory to accommodate all active process executions. Systems that do not have sufficient memory will experience performance degradation during peak load conditions. If memory is not sufficient to support concurrent server processing requirements, system will start to experience random process failures.

Figure 8-24 shows memory performance considerations for an ArcGIS Server host machine deployment.

Figure 8-24
Memory Recommendations

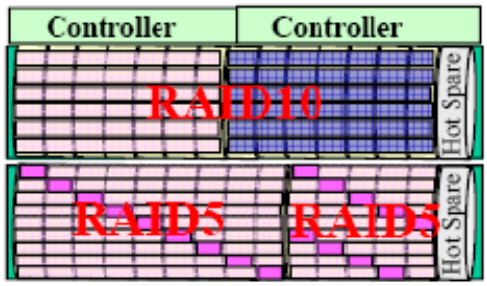


Sufficient number of instances must be deployed to take advantage of processor core capacity. If the ArcSOC process memory requirements exceed available physical memory, performance will start to degrade and at some point start to fail. Setting host capacity can limit the number of SOC instances deployed to make sure memory is sufficient to support peak deployed instance memory requirements.

8.6 Avoiding disk performance bottlenecks

Storage technology changes have resulting in larger disk capacity at lower cost, while disk access performance (seek time) has not improved to keep up with the growing disk capacity. The disk drive is a mechanical device that must search the entire disk volume to complete a data output transaction. Each disk can process one transaction at a time, so multiple requests for data located on the same disk must wait for processing. Disk wait time is called disk contention, and disk contention causes data access delays. The probability of disk contention decreases when data is located on several different disk across a storage volume - disk contention can be resolved by stripping data across multiple disk storage volumes. Figure 8-25 shows the recommended RAID storage configurations discussed earlier in Section 6 (Data Administration).

Figure 8-25
Avoid Disk Contention



GIS technology is also changing, providing higher capacity systems serving a growing number of service requests. Web output files, Image cache, geoprocessing services, and growing number of file data sources (ShapeFile, Personal Geodatabase, File geodatabase, Imagery) increase the volume of data stored on disk, the amount of users sharing access to the same data sources, and increase the probability of disk contention.

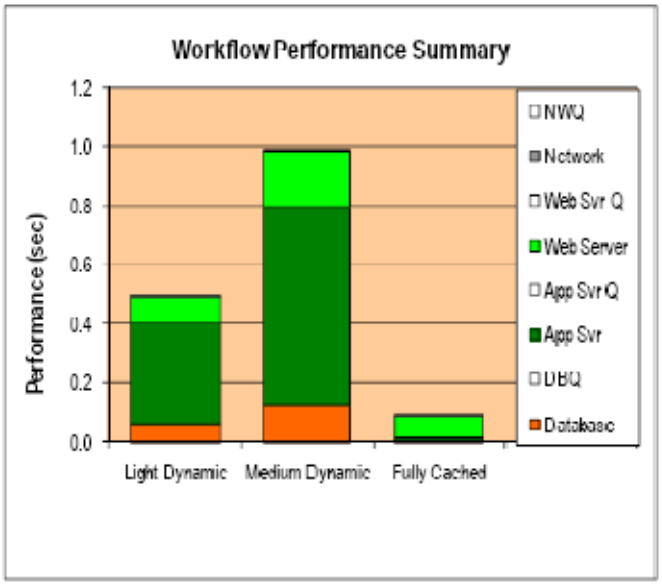
The risk of disk contention can be reduced by using RAID technology to stripe data across multiple disks within each storage volume, increase use of RAID10 mirroring to provide multiple data sources for critical data files, improve utilization of memory cache through, and expand use of distributed cache solutions (disk access is not required if data files are found on storage controller or computer memory cache).

8.7 ArcGIS Server map and globe cache: the Performance Edge

ArcGIS Server technology provides a variety of enhanced solutions to improve production system performance and scalability. The concept of pre-processing map outputs for customer delivery is not new - we used to do this with paper map books before we had computers. The ArcGIS Server 9.3 release introduces a variety of ways to maintain and support pre-processed maps, and to organize the map files in a map cache structure optimized for map publishing.

Figure 8-26 shows the display time for both light and medium complexity dynamic map products in comparison to the display time for a fully cached map. The quality of the fully cached map can be much higher than the medium dynamic display, the difference is that the fully cached map processing was completed before posting on the Web site, and the final processing time is minimal. Pre-cached maps performance and scale much faster than previous Web map services that were processed on demand.

Figure 8-26
Take Advantage of Caching



8.7.1 Selecting the right technology: A case study

Selecting the right software technology can make a big difference in performance and scalability, and cost of the production system. The following case study shares an experience with a real customer implementation which clearly represents the value of selecting the right software technology.

Our customer had a requirement to design a Web application solution that would be used to collect national property location and census information during a three month national citizen declaration period. Citizens would report to regional government centers, use a local desktop computer to locate their home residence on a map display generated from a national imagery and geospatial features repository. Citizen would place a point

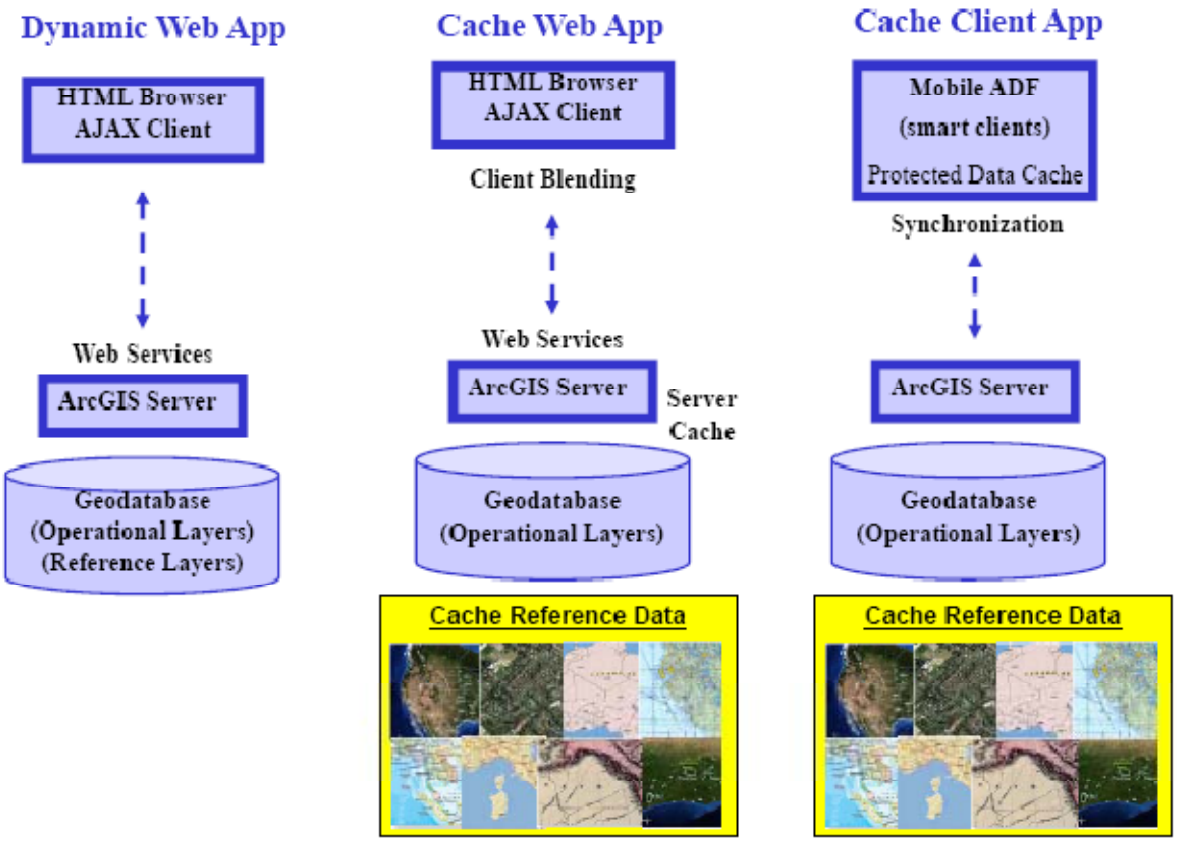
on the map identifying their residence, and then fill out a reference table identifying their census information. The citizen input would be consolidated at a centralized national data center and shared with all regional government centers throughout the declaration process.

The initial system design was developed using the ArcGIS Server 9.1 technology, using a centralized ArcGIS Server dynamic Web application technology to support browser clients located at 50 regional national sites. Following contract award, the customer reviewed available technology options to finalize the system design.

Figure 8-27 shows three possible software technology options that were considered during the review. Technology had progressed since the initial proposal, and there were three possible solutions that would support the customer operational requirements

Figure 8-27
Cache Performance Advantage

(System Design Use Case)



The ArcGIS Server dynamic Web application was the solution provided in the initial design proposal two years earlier, and the ArcGIS Server 9.2 technology included improvements in Web application performance and user experience incorporating a Map Viewer application development framework for deploying images to an AJAX browser client.

ArcGIS Server 9.2 also included cache option where the reference map layers could be pre-processed and stored map cache pyramid data source. Pre-processing the reference layers would significantly reduce server processing loads during the production deployment.

ArcGIS Server 9.2 also supports a Mobile ADF client that can be deployed from a centralized Web server. The mobile client can support the required editing functionality and can pre-cache the reference layers local to

distributed client workstations. The point declaration layer point features can be exchanged with the centralized data center database using background network processing with current point displays maintained in a local client cache.

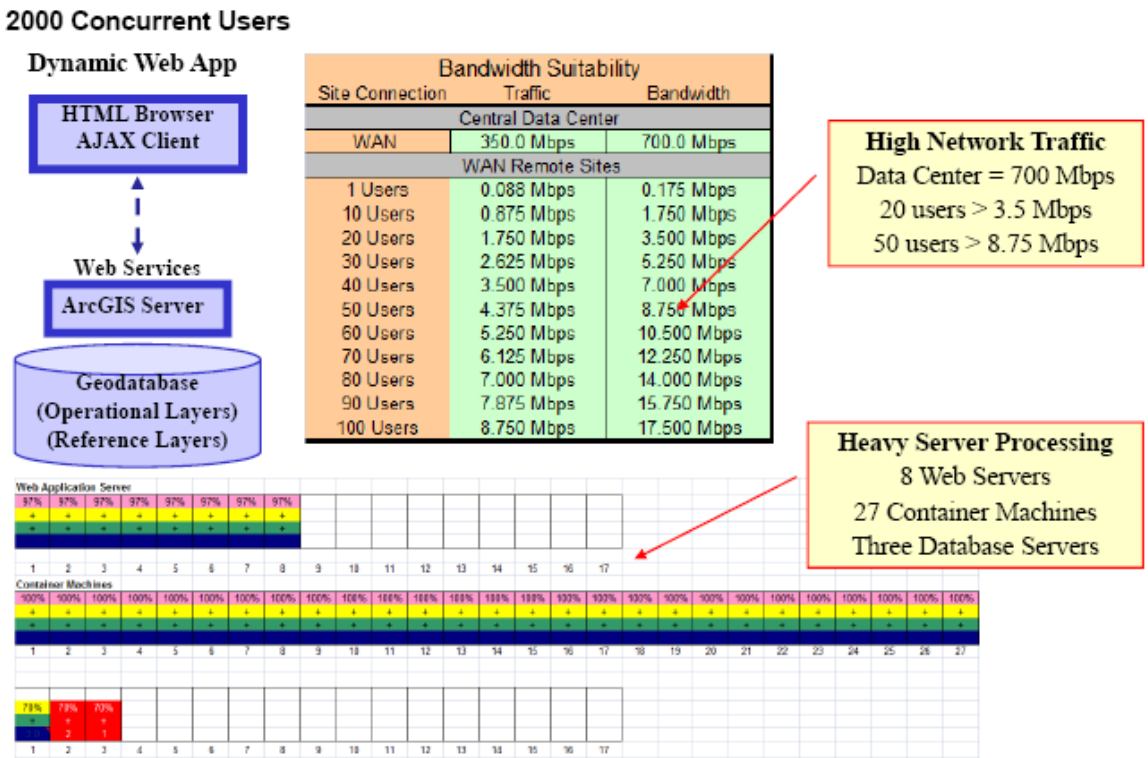
The ESRI Capacity Planning Tool was used to evaluate the architecture for the three different workflow technologies identified above. Peak system loads were estimated at 2000 concurrent users with standard Web productivity of 6 displays per minute. System design results are provided in the following paragraphs.

8.7.1.1 Dynamic Web Application

The ArcGIS Server ADF light Dynamic standard ESRI workflow was used to generate hardware requirements and traffic loads required to support the dynamic web application solution. Figure 8-28 provides the results of the capacity planning analysis.

Figure 8-28
ArcGIS Server ADF light Dynamic

(System Design Analysis Summary)



Peak central data center traffic loads were estimated to reach 350 Mbps, well beyond the bandwidth available with the current data center Wide Area Network (WAN) service connection. Smaller regional office sites (10 concurrent users) WAN connections would need over 1.5 Mbps bandwidth. Larger regional office sites (50 concurrent users) would require WAN connections would need 9 Mbps bandwidth to support the peak citizen declaration traffic. Major infrastructure bandwidth increases would be needed to handle projected traffic flow requirements.

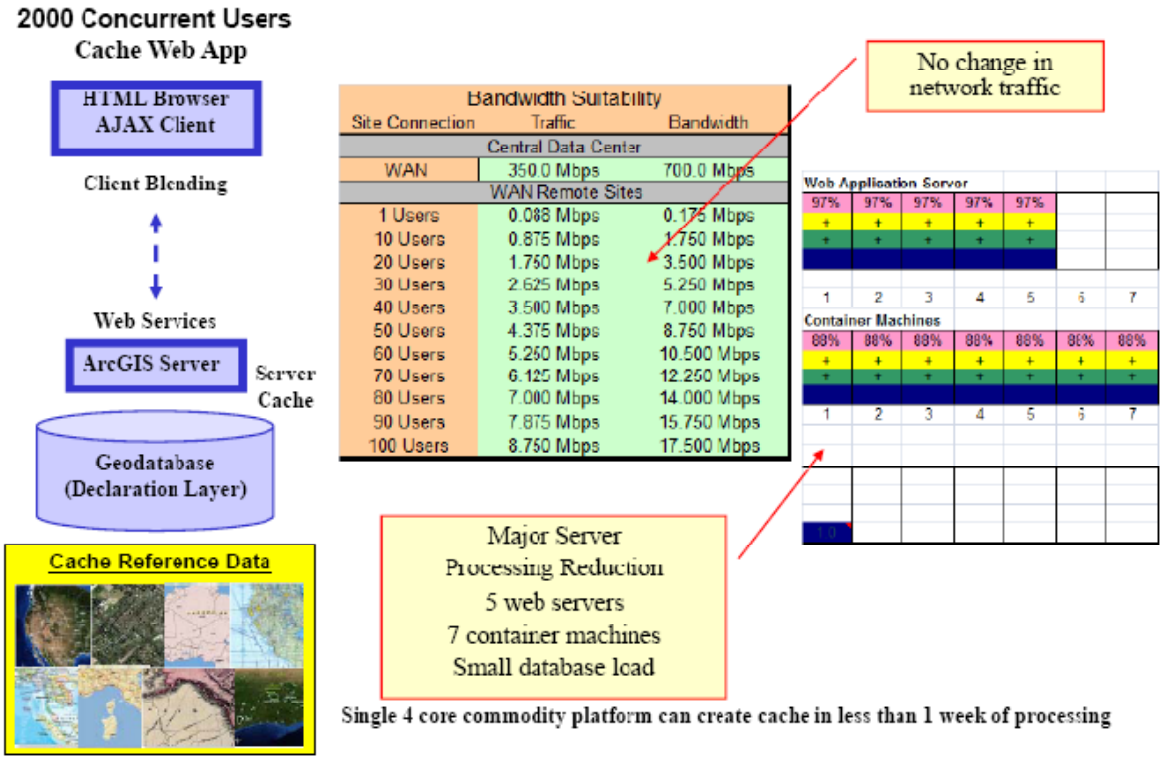
The selected central hardware solution was supported by Intel Xeon 5160 4 core (2 chip) 3000 MHz Windows

64-bit Servers each with 16 GB memory. Total of 8 servers were required for the Web Application Server tier, 27 servers for the Container Machine tier, and three geodatabase servers (higher capacity data server could be used to support the geodatabase on a single machine).

8.7.1.2 Cached Web Application

The custom ArcGIS Server ADF light 1 layer Dynamic with Cached Reference layers workflow was used to support the cached web application workflow analysis. Software service time and network traffic performance targets were set for the custom workflow. Figure 8-29 shows the results of the system design analysis.

Figure 8-29
ArcGIS Server AJAXlight 1 layer Dynamic with Cached Reference Layers
(System Design Analysis Summary)



Peak central data center traffic load estimates remained at 350 Mbps, the AJAX application traffic display would still be generated on the server leaving traffic requirements unchanged .

The selected central hardware solution was reduced to 5 servers required for the Web Application Server tier, 7 servers for the Container Machine tier, and minimum load on a single geodatabase server. This was a significant cost reduction from the initial proposal.

A sample data set was used to evaluate map caching timelines, and the complete country reference map cache could be generated within one week of processing time. This would be well worth the effort, since there would be no need to update or change this data cache during the peak citizen declaration period (data would be static).

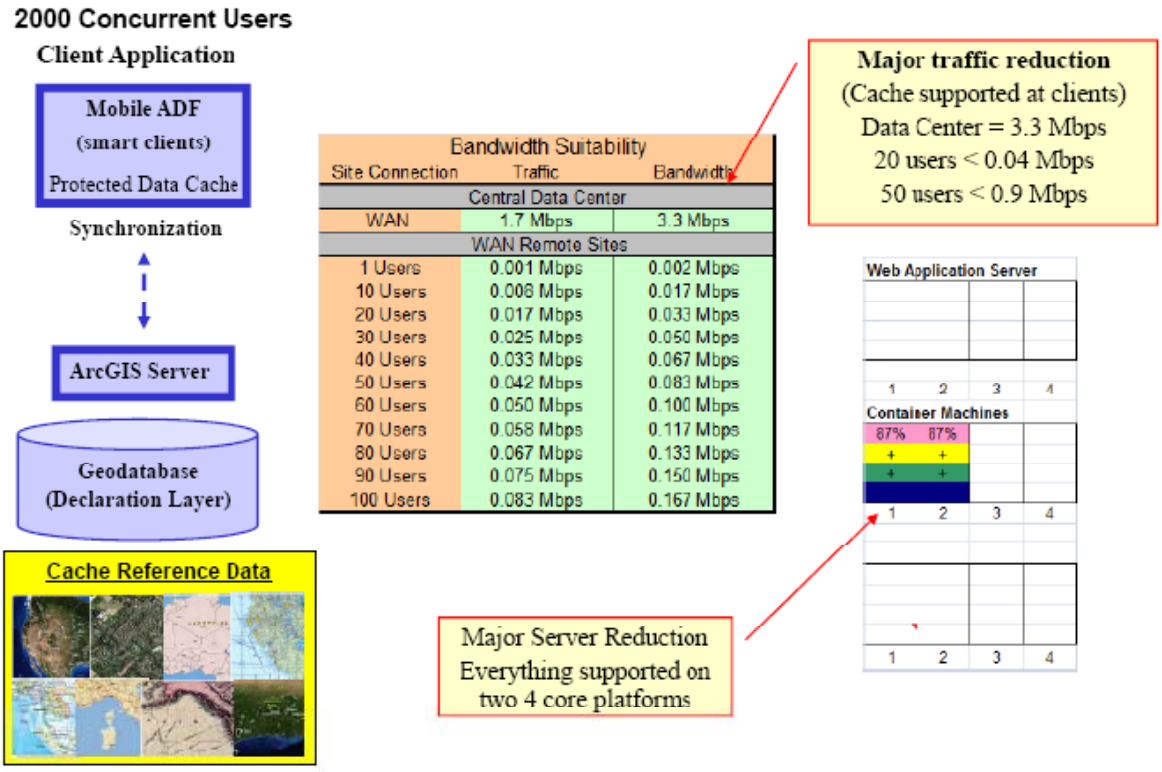
8.7.1.3 Cached Client Application

The third design options was to use the ArcGIS Server Mobile ADF application with a local cache data source. A demo of the moble ADF client was provided on a Windows desktop platform to demonstrate feasibility of supporting the required editing functions with this client technology. The mobile ADF client technology

performed very well on a standard Windows display environment and performed all the functions needed to support the citizen declaration requirements.

The ArcGIS Server Mobile ADF standard ESRI workflow was used to support the design analysis. Cached reference layers would be provided to each regional site in advance, and access would be provided over the a file share to the Mobile ADF client running on the local workstations. The ArcGIS Mobile client would exchange changes to the dynamic citizen declaration layer over the government LAN. User display performance would be very fast, supported by the local reference map cache and the point layer in the Mobile ADF cache. The point layer cache was updated from the central data center geodatabase with each display refresh, and point layer edits were sent to the central server as a background data exchange. Figure 8-30 shows the results of our capacity planning analysis.

Figure 8-30
ArcGIS Server Mobile ADF 1 dynamic layer with Cached Reference Layers
(System Design Analysis Summary)



Bandwidth Suitability		
Site Connection	Traffic	Bandwidth
Central Data Center		
WAN	1.7 Mbps	3.3 Mbps
WAN Remote Sites		
1 Users	0.001 Mbps	0.002 Mbps
10 Users	0.008 Mbps	0.017 Mbps
20 Users	0.017 Mbps	0.033 Mbps
30 Users	0.025 Mbps	0.050 Mbps
40 Users	0.033 Mbps	0.067 Mbps
50 Users	0.042 Mbps	0.083 Mbps
60 Users	0.050 Mbps	0.100 Mbps
70 Users	0.058 Mbps	0.117 Mbps
80 Users	0.067 Mbps	0.133 Mbps
90 Users	0.075 Mbps	0.150 Mbps
100 Users	0.083 Mbps	0.167 Mbps

Peak central data center traffic loads were reduced to 1.7 Mbps, well within the bandwidth available with the current data center Wide Area Network (WAN) service connection. Regional office sites would function well within the available 1.5 Mbps WAN connections, actual traffic less than 0.1 Mbps for the larger site loads. The existing infrastructure would be able to support peak WAN traffic loads with guaranteed service to each of the remote desktop locations (Mobile ADF client would continue to function as a standalone system if WAN communication were lost, and edits would be sent to the central server when communication was restored).

The central hardware requirements were reduced to2 composite Web/container machines servers and the data server load was minimal.

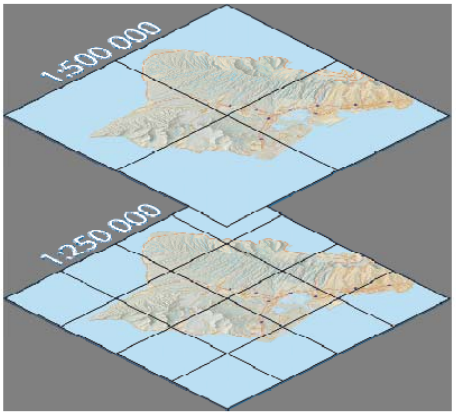
It was very clear that the cached client application provided significant cost and performance benefits over the centralized Web application dynamic solution included in the initial proposal. Pre-processing of map reference layers as an optimized map cache pyramid can significantly improve display performance. Use of an intelligent desktop client that can access reference layers from a local map cache can further reduce network traffic and

improve display performance even more. Selecting the right technology can make a big difference in total system cost and user productivity.

8.7.2 Building the data cache

ArcGIS Server 9.2 introduced automated processing functions to build and maintain an optimized map cache pyramid for service pre-processed images as a map service. The ArcGIS Server Globe service would stream pre-processed 3D Globe cache imagery to an ArcGIS Explorer or ArcGIS Desktop with 3D Analysis client for 3D visualization - cached imagery would be stored at the client for high performance display. ArcGIS Server also provided a 2-D map cache service that could be used as a data source for 2D Web applications and ArcGIS Desktop or custom desktop clients developed with the ArcGIS Engine software. Figure 8-31 provide an overview of the cached map service image structure.

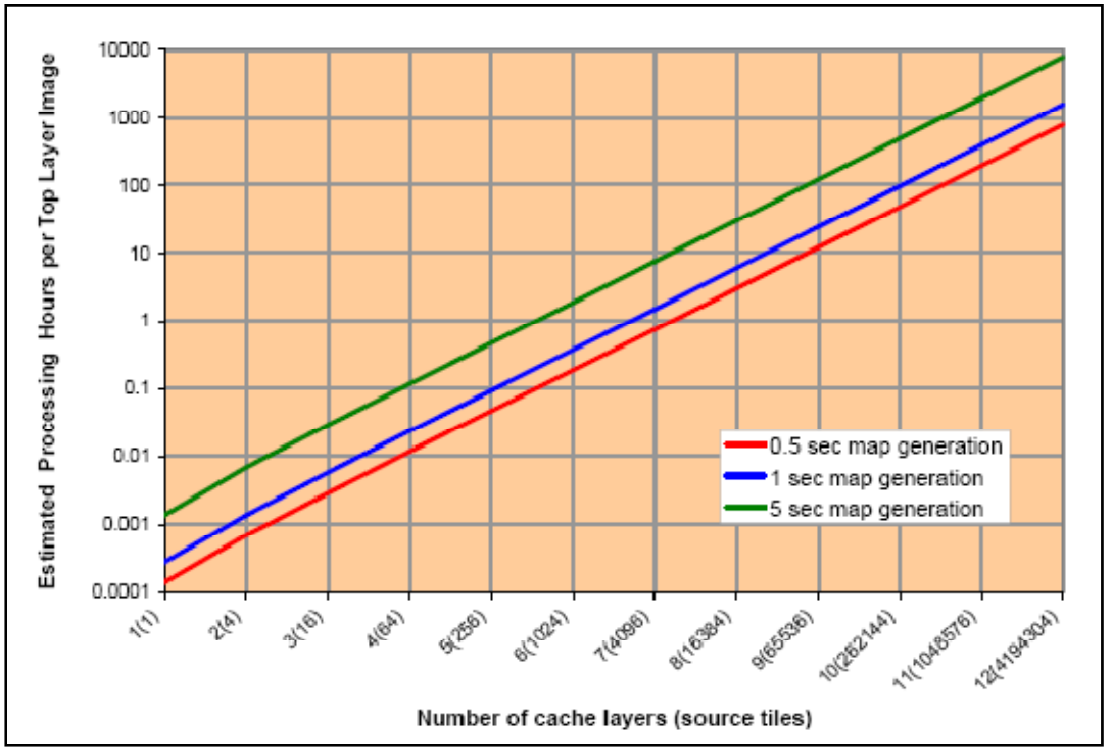
Figure 8-31
Cached Map Service



The cached map service would consist of a pyramid of pre-processed data imagery, starting at a single map resolution at the highest layer and breaking each image into four images at twice the resolution for each additional layer included in the pyramid. Client access to the cached data would deliver only tiles that correspond to the resolution of the map display request. Tiles would be combined together as a single reference layer by the client application. Total pre-processing time would depend on the total number of images in the map cache and the map service time required to generate an average map display.

Figure 8-32 can be used to get a rough estimate of the expected map cache generate time.

Figure 8-32
Generating the Map Cache



The chart above shows the estimated processing hours starting with one tile at the top layer and building the required number of layers to reach a resolution level. Three map generation times are plotted - 0.5 seconds for a simple map display, 1 second for a medium map display, and 5 seconds for a heavy map display. Charts shows about 100 hours to generate 9 layers with average map generation time of 5 seconds. It would take over 2000 hours to generate just two more layers (11 layers at 5 sec per map tile). This chart is generated simply by multiplying the time to generate one map tile by the total number of tiles required to complete the map cache pyramid.

The recommended procedure for estimating map cache generation times is to select a same area dataset that represents a small percentage of the total caching area. Build a map cache of the sample area and test the output symbology, labeling, and performance with your primary map client. Use the cache time for the small sample area to estimate processing time for the remaining map cache.

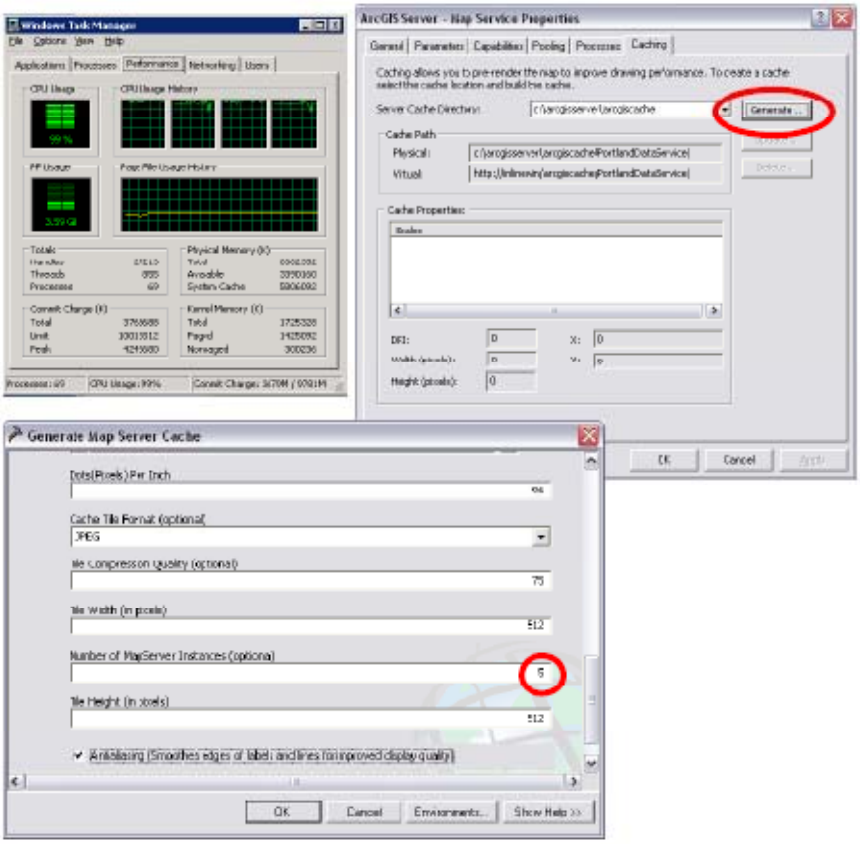
The ArcGIS 9.3 release additional options for building and maintaining map cache that make this technology adaptive to a broad range of operational scenarios.

Partial Data Cache. A partial data cache can be defined for high priority areas within your display environment. The partial data cache can be specified by both area and level of resolution, providing optimum flexibility for pre-processing custom areas and levels of map cache.

Cache on Demand. The initial map query can be generated from a dynamic data source, and the resulting tiles can be added to the map cache. The next request for these tiles would come from the map cache, significantly improving display performance for popular map display environments.

Figure 8-33 shows a view of the ArcGIS Server Map Service Properties Caching tab. The map cache process is compute intensive, automatically generating tile after tile based on the defined map cache properties. Map caching can be performed by several parallel service instances, and ArcGIS Server coordinates activities of these services to build and store the prescribed map cache.

Figure 8-33
Optimize number of SOC Instances for Building a Map Cache

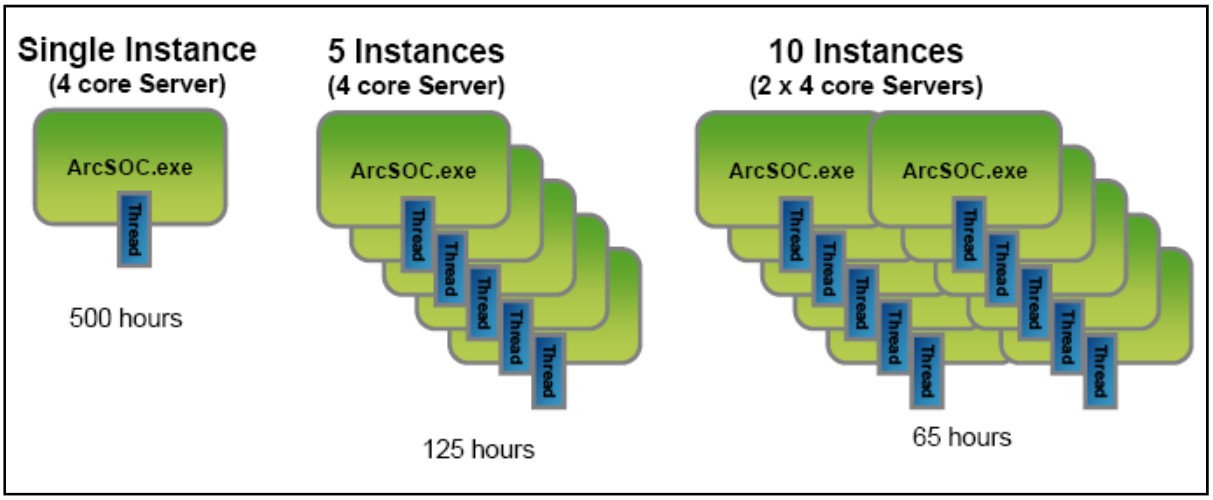


When using a local image file data source, each map cache instance will consume a platform core. The optimum service configuration would specify 5 instances to make sure the server is operating at 100 percent capacity. The windows performance task monitor can be used to identify the machine is operating at 100 percent utilization. If your map cache include a DBMS data source, you may be able to include one or two additional service instances to reach 100 percent utilization. It is important to configure sufficient instances to take full advantage of your hardware platform processing resources.

It is a best practice to execute longer cache jobs in sections. It is good to plan cache areas that can be completed within an 8 or 10 hour period, providing an opportunity to complete the jobs in a stable platform environment. This will also provide an opportunity to periodically reboot system between each job section to maintain a reliable platform environment.

Figure 8-34 provides an example of taking advantage of the hardware, as described above. In this example, the cache job required 500 hours of processing. A single instance would take 500 hours to complete. Five (5) instances running on a 4 core machine will complete the same job in 125 hours. Ten (10) instances, using two 4 core machines each with 5 instances, completed the same job in 65 hours.

Figure 8-34
Sample Map Cache Processing Profile
(500 Hours of Processing)



A variety of caching scenarios are being evaluated to expand the feasibility of using pre-processed data cache to improve performance and scalability of future GIS operations. Experience shows pre-processing map data can make a difference for our customers. ArcGIS Server is leading the way toward more productive and higher quality Enterprise level GIS operations.

8.8 Software performance summary

Experience suggests we can do a better job selecting and building better software solutions. Understanding software performance can reduce implementation risk and save customer time and money. Best practices provided in this section include the following:

Publish and optimum map display
Reduce layers to improve performance
Optimum image displays reduce network traffic
Maintain a high performance data source
Configure and maintain high performance Web services
Avoid disk contention
Take advantage of caching
Select the right technology solution
Pre-processing: Configure to leverage available processing resources

The next section will take a closer look at hardware platform performance - what server platforms do you need to support your software processing requirements. Understanding your processing needs is the first step in selecting the right software. Understanding platform performance is the rest of the story.

9 Platform Performance

Chapter 7 (Sizing Fundamentals) provided an overview of the system configuration performance models assuming all hardware platforms were the same. Chapter 8 (Software Performance) discussed some best practices for building software solutions that performance and scale, and the importance of selecting the right software technology that will support your performance needs. This chapter will focus on hardware platform performance, and share the importance of selecting the right computer technology to support your system performance needs.

9.1 Platform Performance Baseline

The world we live in today is experiencing the benefits of rapid technology change. Technology advancements are directly impacting GIS user productivity—the way we all deal with information and contribute to our environment. Our ability to manage and take advantage of technology benefits can contribute to our success in business and in our personal life.

To develop a system design, it is necessary to identify user performance needs. User productivity requirements are represented by the workstation platforms selected by users to support computing needs. GIS users have never been satisfied with platform performance, and each year power users are looking for the best computer technology available to support their processing needs. Application and data servers must be upgraded to continue support for increasing user desktop processing requirements.

GIS user performance expectations have changed dramatically over the past 10 years. This change in user productivity is enabled primarily by faster platform performance and lower hardware costs.

Figure 9.1 identifies the hardware desktop platforms selected by GIS users as their performance baseline since the ARC/INFO 7.1.1 release in February 1997.

Figure 9-1
User Performance Expectations

- **ARC/INFO 7.1.1 (Feb. 1997)**
  - Pentium Pro 200 MHz, 64 MB Memory
- **ARC/INFO 7.2.1 (April 1998)**
  - Pentium II 300 MHz, 128 MB Memory
- **ArcInfo 8 (July 1999)**
  - Pentium III 500 MHz, 128 MB Memory
- **ArcInfo 8.0.2 (May 2000)**
  - Pentium III 733 MHz, 256 MB Memory
- **ArcInfo 8.1 (July 2001)**
  - Pentium III 900 MHz, 256 MB Memory
- **ArcInfo 8.2 (July 2002)**
  - Intel Xeon MP 1500 MHz, 512 MB Memory
- **ArcInfo 8.3 (July 2003)**
  - Intel Xeon 2400 MHz, 512 MB Memory
- **ArcInfo 9.0 (May 2004)**
  - Intel Xeon 3200 MHz, 512 MB Memory
- **ArcInfo 9.1 (June 2005)**
  - Intel Xeon 3200 MHz, 1 GB Memory
- **ArcInfo 9.2 (August 2006)**
  - Intel Xeon 3800 MHz, 1 GB Memory
- ArcInfo 9.2 (June 2007)
  - Intel Xeon 2 core (1 socket) 3000 (4MB L2) MHz, 2 GB Memory
- ArcInfo 9.3 (July 2008)
  - Xeon X5260 2 core (1 chip) 3333 (6MB L2) MHz, 2 GB Memory
- **ArcInfo 9.3.1 (July 2009)**
  - Xeon X5570 4 core (1 chip) 2933 MHz, 4 GB Memory

In February 1997, ESRI released ARC/INFO 7.1.1. The first Windows ARC/INFO software was supported on the Intel Pentium Pro 200 MHz platform. The new Windows workstation environment supported GIS users with performance almost twice what was available the previous year on UNIX platforms and at a fraction of the cost.

In April 1998, ESRI released ARC/INFO 7.2.1. Intel was selling the Pentium II 300 MHz platform, which was 50 percent faster than the Pentium Pro 200. GIS users quickly moved to the new platform environment to improve productivity.

ARC/INFO application performance improved throughout this same period as the code was optimized with new ARC/INFO 7 incremental releases. A script developed with ARC/INFO 7.1.1 would run faster using the ARC/INFO 7.2.1 release, both running on the same platform. The change in the ARC/INFO performance baseline was not a software-driven requirement but a change in user performance expectations brought about by the faster and less expensive workstation and server technology.

In July 1999, ESRI provided the first release of ArcInfo 8 (ArcGIS Desktop software). The Intel Pentium III 500 MHz platform was the popular ArcInfo Workstation candidate. The Pentium III 500 MHz platform was more than 2.5 times faster than the Pentium Pro 200.

In May 2000, ESRI released ArcInfo 8.0.2. The Pentium III 733 MHz platform was selected as the performance baseline supporting the summer 2000 deployments. The Pentium III 733 MHz platform is more than 4.7 times faster than the Pentium Pro 200.

Modest performance improvements continued into 2001. Rapid performance enhancements were delayed because of vendor development of the next-generation processor technology. The Pentium III 900 MHz platform was selected as the performance baseline supporting the ArcInfo 8.1 release.

In June 2002, the Intel Xeon MP 1500 MHz server provided a significant performance gain. Intel had some performance problems with the 1500 MHz chips, and deployment was delayed until the fall. The Intel Xeon MP 1500 MHz platform was selected as the performance baseline for the ArcInfo 8.2 release.

In June 2003, the Intel Xeon MP 2000 MHz platform was the current technology Windows server. Performance issues for the Intel Xeon MP were resolved, and the 2000 MHz processors supported the current server platforms. This server provided an impressive gain over the previous year, matching performance of the Intel Xeon 2400 MHz workstation (almost 11 times faster than the Pentium Pro 200). Intel was releasing the Intel Xeon 3060 MHz workstation platforms before midyear and was planning to release a 2800 MHz server version in the fall. The Intel Xeon 2400 MHz platform was selected as the 2003 performance baseline.

In June 2004, the Intel Xeon MP 3000 MHz platform was the current technology Windows server. This server provided an impressive gain over the previous year, matching performance of the Intel Xeon 3200 MHz workstation. Intel was releasing the Intel Xeon 3600 MHz workstation platforms before midyear. The Intel Xeon 3200 MHz platform was selected as the 2004 performance baseline (more than 19 times faster than the Pentium Pro 200).

In 2005, hardware performance improved very slightly because of CPU technology thermal limits. The Intel Xeon MP 3000 MHz platform continued to be the server of choice through summer 2005. This was the first time in more than 10 years we did not see a change in the GIS user platform performance baseline.

Hardware vendors made good performance enhancements in 2006. Dual-core processor technology was introduced doubling the capacity of commodity server platforms and reducing hardware cost. The Intel Xeon 4 CPU (2 dual core) 3773 MHz servers matched the performance of earlier Intel Xeon 2 CPU 3800 MHz servers and supported twice the compute capacity. The 3800 MHz servers (dual-core 3773 MHz servers) were selected as the 2006 GIS user performance baseline.

Advanced Micro Devices (AMD) is a chip development company that competes with Intel, and their platforms grew in popularity as customers experienced growing concerns about the high power consumption and heat generated by the high MHz Intel servers. AMD was able to provide a lower MHz platform solution that matched the performance of the higher MHz Intel processor platforms. Hewlett-Packard favored the cooler temperature and lower power consumption of the new AMD platform and promoted the AMD platforms as its server technology choice when it introduced the new HP Blade servers. Dell decided to start selling AMD

server technology as an alternative to Intel. By midyear, Intel released a new Intel Xeon 3000 MHz commodity server with dual-core sockets, 1333 MHz front side bus, and 4 MB L2 cache that performed 40 percent faster than the Intel Xeon 3800s.

The dual core processor technology challenged the continued use of the acronym CPU as the abbreviation for central processing units, and vendors introduced a new set of names for identifying the core processing units used to enable compute intensive platform performance and capacity. A new naming convention was adopted for this document in 2007 replacing the use of CPU by core, and introducing the use of socket to refer to the processor (socket was the term used in ESRI software pricing models). Thus the new Intel Xeon 3000 MHz platform with dual core processors was referenced as the Intel Xeon 4 core (2 socket) 3000 MHz platform.

The Intel Xeon 4 core (2 socket) 3000 MHz platform with the 4 MB L2 cache was selected as the 2007 performance baseline. This platform was introduced by hardware vendors in mid 2006. Since that time vendors have released different configurations of this same chip technology, ranging from 1600 MHz to 3000 MHz dual core socket and 1600 MHz – 2666 MHz quad core socket configurations.

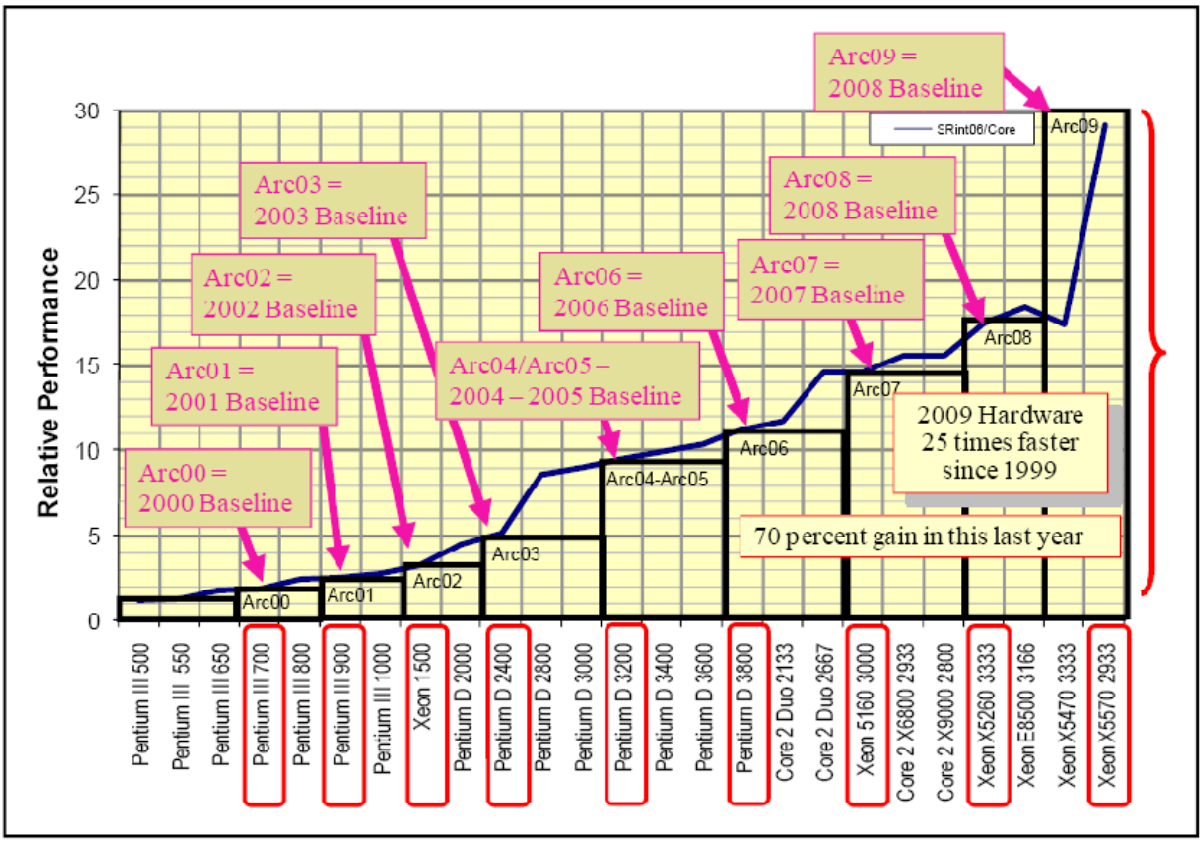
There was a broad mix of hardware configurations provided by vendors in CY2007, some designed for lower cost and throughput and others for performance, making it challenging to identify the specific hardware configuration that would best meet customer performance needs. Vendors started to include the processor model number with each platform to more clearly identify the specific chip configuration. The 2007 baseline was more specifically identified as the Xeon 5160 4 core (2 socket) 3000 MHz platform.

Platform performance continued to improve in CY2008, and quad-core platforms were growing in popularity. ESRI software pricing was based on number of processor core, so the term for processor was changed to chip. The Xeon 5260 4 core (2 chip) 3333 MHz platform was selected as the CY2008 performance baseline.

Intel introduced a new chip design in CY2009 that was 70 percent faster than the previous year. Dual core chip technology was no longer an option, replaced by the new Quad core chip technology for workstations and servers. The Xeon X5570 8 core (2 chip) 2933 MHz platform was selected as the CY2009 performance baseline.

Figure 9-2 provides a graphic overview of Intel workstation performance over the past ten years. The chart shows the radical change in relative platform performance since 2000. Technology change introduced by the hardware platform manufacturers represented a major contribution to performance and capacity enhancements over the past 10 years.

Figure 9-2
Platform Performance Baseline

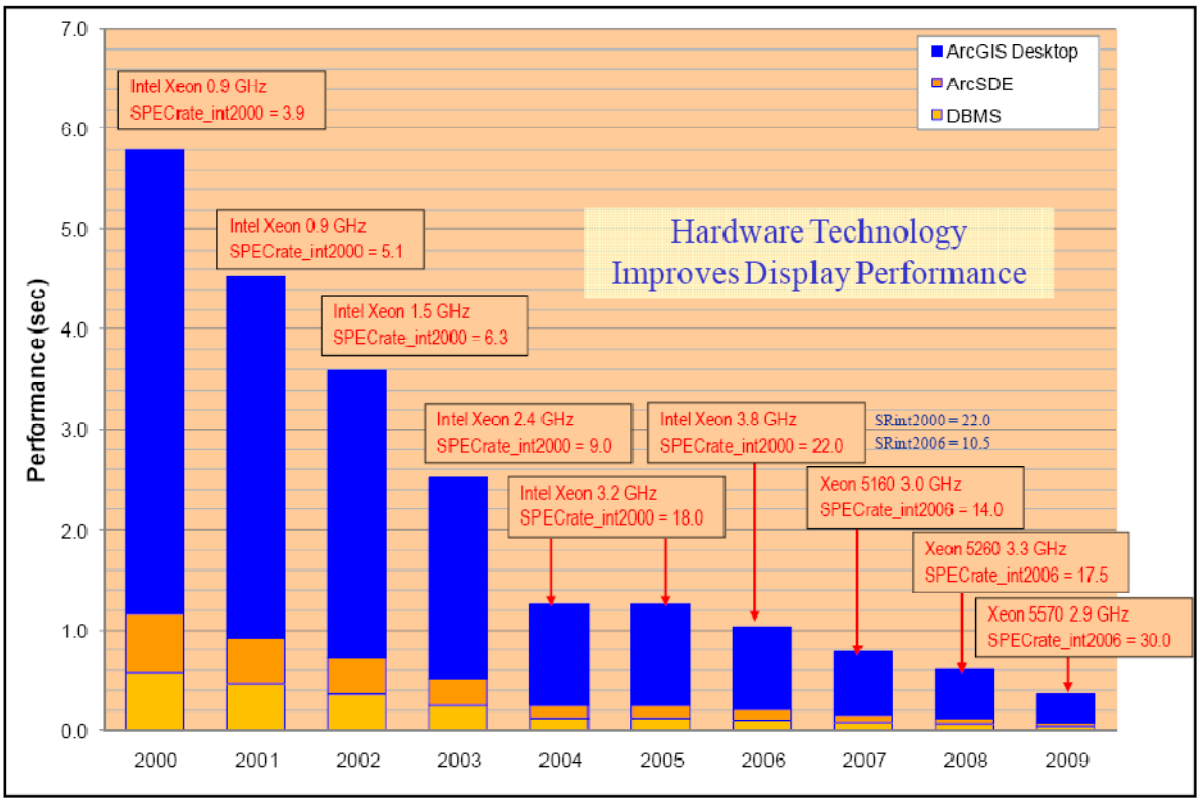


The boxes at the bottom of the chart represent the performance baselines selected to support ESRI capacity planning models. These performance baselines were reviewed and updated each year to keep pace with the rapidly changing hardware technology.

The change in hardware performance over the years has introduced unique challenges for customer capacity planning and for software vendors trying to support customer performance and scalability expectations. Understanding how to handle hardware performance differences is critical when addressing capacity planning, performance, and scalability issues.

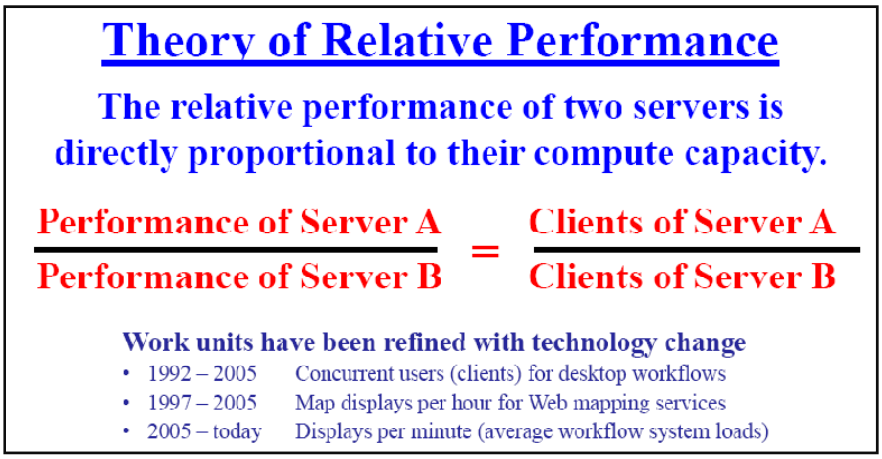
Figure 9-3 shows how user expectations have changed over the past eight years. An ArcGIS Desktop simple dynamic map display processing time in CY2000 would take almost 6 seconds. That same map display today can be rendered in less than 0.6 seconds - 10 times faster than just 8 years earlier. All of this performance gain can be accounted for by the change in platform technology.

Figure 9-3
Time to Product a Map



Understanding how to account for platform technology change is fundamental to understanding capacity planning. Figure 9-4 identifies a simple relationship that we have used since 1992 to relate platform performance with capacity planning.

Figure 9-4
How do we Handle Platform Performance Change?

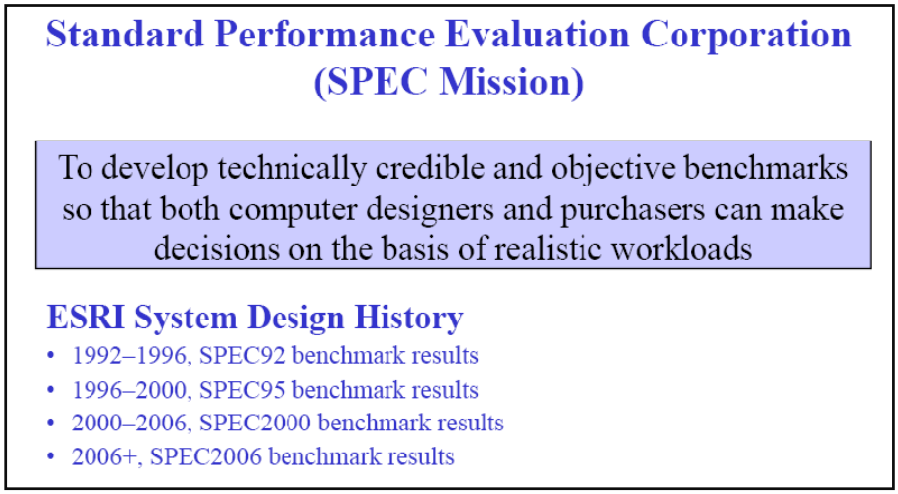


The relationship simply states that if one can determine the amount of work that can be supported by server A (display transactions supported by server A) and identify the relative performance between server A and server B, then one can identify the work that can be supported by server B. This relationship is true for single-core servers (servers with a single computer processing unit) and for multi-core servers with the same number of cores. This relationship is also true when comparing the relative capacity of server A and server B.

Identifying a fair measure of relative platform performance and capacity is very important. Selection of an appropriate performance benchmark and agreement on how the testing will be accomplished and published are all very sensitive hardware marketing issues.

Figure 9-5 shares the mission statement published by the Standard Performance Evaluation Corporation (SPEC), a consortium of hardware vendors established in the late 1980s for the purpose of establishing guidelines for conducting and sharing relative platform performance measures.

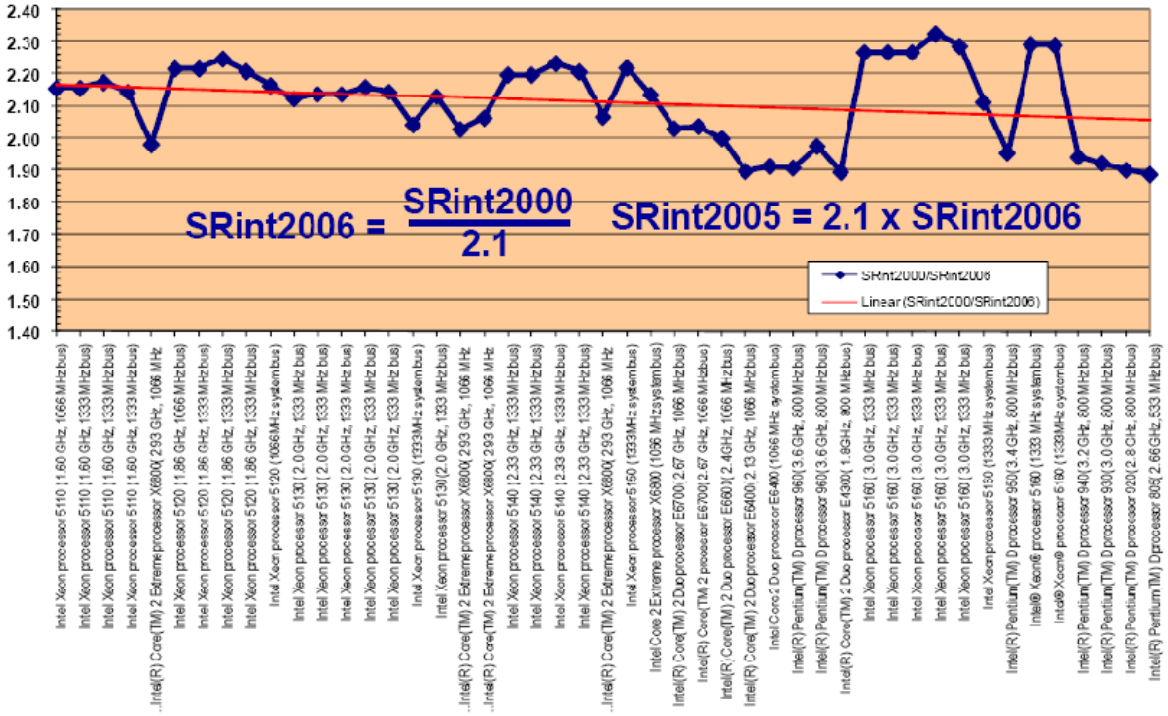
Figure 9-5
How do we Measure Relative Platform Performance



The SPEC compute-intensive benchmarks have been used by ESRI as a reference for relative platform capacity metrics since 1992. The system architecture design platform sizing models used in conjunction with these relative performance metrics have supported ESRI customer capacity planning since that time. The SPEC benchmarks were updated in 1996 and 2000 to accommodate technology changes and improve metrics. A new SPEC2006 release was published in 2006 and provides the platform baseline metrics for ESRI system architecture design sizing models starting with the Arc07 2007 baseline (there is normally a 6–12 month overlap in testing and published results once SPEC introduces the new benchmarks).

Test results for over 45 vendor platforms were published on the SPEC2000 and the SPEC2006 benchmark sites, and these test results were used to calculate a transfer function between the two benchmark sets for capacity planning purposes. The ratios of the published benchmark results (SPECrate_2000/SPECrate_2006) were plotted on a graph to identify a mean value for the transfer function. Figure 9-6 provides the results of this analysis.

Figure 9-6
SPECrate2000 to SPECrate2006 Translation



A published SPECrate_int2000 benchmark divided by 2.1 will provide an estimate for the equivalent SPECrate_int2006 benchmark value. A published SPECrate_int2006 benchmark multiplied by 2.1 will provide a translated estimate for the equivalent SPECrate_int2000 benchmark value. It is interesting to notice that the older platform technology has better benchmark values on the SPEC2006 benchmarks than the newer technology (the conversion ratio for the new Intel Xeon 51xx series platforms is over 2.2 while the conversion factor for the older Intel Pentium D platforms are under 2.0).

SPEC provides a separate set of integer and floating point benchmarks. Computer processor core are optimized to support integer or floating point calculations, and performance can be very different between these environments. Testing with the ESRI software since the ArcGIS technology release has followed the integer benchmark results, suggesting the ESRI ArcObjects software predominantly uses integer calculations. The Integer benchmarks should be used for relative platform performance calculations when using ArcGIS software technology.

SPEC also provides two methods for conducting and publishing benchmark results. The SPECint2006 benchmarks measure execution time for a single benchmark instance and use this measure for calculating relative platform performance. The SPECint_rate2006 benchmarks are supported by several concurrent benchmark instances (maximum platform capacity) and measure executable instance cycles over a 24-hour period. The SPECint_rate2006 benchmark results are used for relative platform capacity planning metrics in the ESRI system architecture design sizing models.

There are two results published on the SPEC site for each benchmark, the conservative (baseline) and the aggressive (result) values. The conservative baseline values are generally published first by the vendors, and the aggressive values are published later following additional tuning efforts. Either published benchmark can be used to estimate relative server performance, although the conservative benchmarks would provide the most conservative relative performance estimate (removes tuning sensitivities).

Figure 9-7 provides an overview of the published SPEC2006 benchmark suites. The conservative SPECint_rate2006 benchmark results are used in the ESRI system architecture design documentation as a vendor-published reference for platform performance and capacity planning.

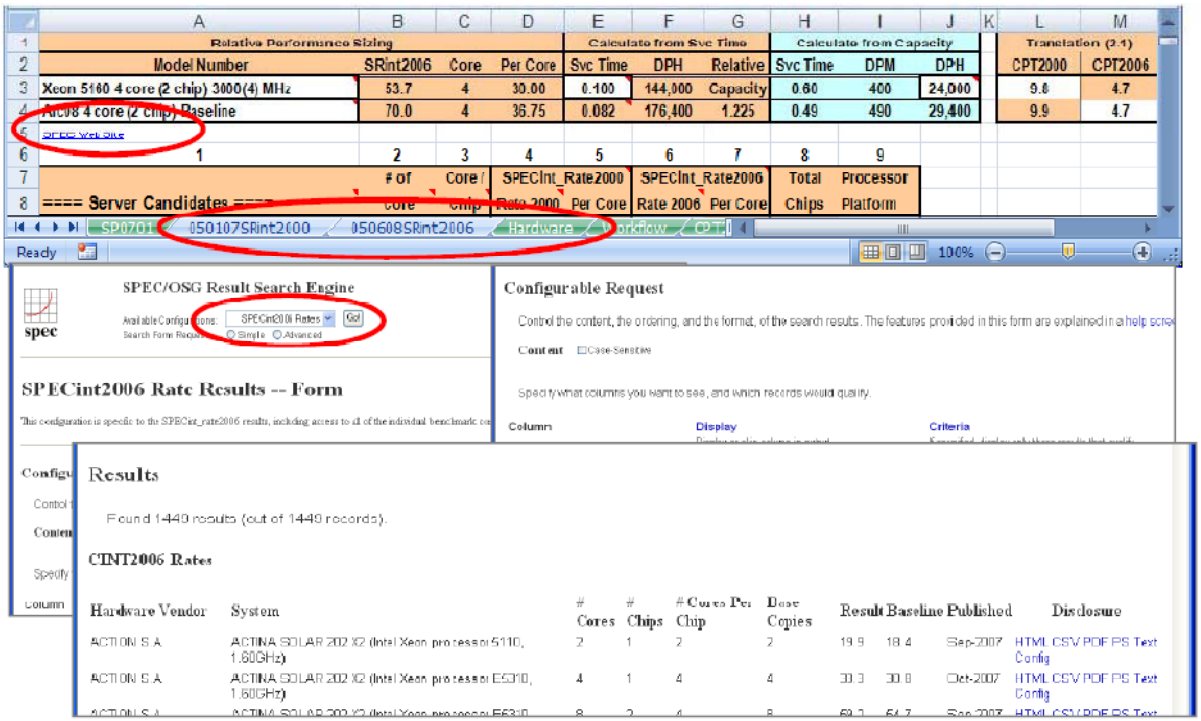
Figure 9-7
Platform Relative Performance
(SPEC2006 Benchmark Suites)

- **SPEC2006 Comprises Two Suites of Benchmarks**
  - **CINT2006: Compute-Intensive Integer**
    - Twelve CPU-intensive integer benchmarks (C and C++ Languages)
    - Base (SPECint_base2006, SPECint_rate_base2006)
    - Peak (SPECint2006, SPECint_rate2006)
  - **CFP2006: Compute-Intensive Floating-Point**
    - Seventeen CPU-intensive Floating-Point Benchmarks (C++, FORTRAN, and a mixture of C and FORTRAN)
    - Base (SPECfp_base2006, SPECfp_rate_base2006)
    - Peak (SPECfp2006, SPECfp_rate2006)
- **Sun UltraSPARC-II 296 MHz Reference Platform**

The SPEC performance benchmarks are published on the Web at www.spec.org. The ESRI Capacity Planning Tool (the Capacity Planning Tool will be introduced in Chapter 10) includes a HardwareSPEC workbook that provide a list of published SPECrate_integer benchmarks. The SRint2000 tab includes all vendor published SPECrate_int2000 benchmarks available on the SPEC site. SPEC stopped publishing the SRint2000 benchmarks in January 2007. All the new platform benchmarks are now published on the SPECrate_integer2006 site (SRint2006 tab). The last date the benchmark tab was updated is shown with the link name. A hot link to the SPEC site is included on the top of the Capacity Planning Tool (CPT) hardware tab.

Figure 9-8 identifies the location of the SPEC link on the CPT hardware tab and provides some views of the SPEC site.

Figure 9-8
SPEC Web Site



Several benchmarks are published on the SPEC Web site. You will need to select and go to the SPECrate2006 Rates and the scroll down to configurable request selection - you can then select specific items that you want included in your display query. I like to include the processor MHz in my display, which was not included in the default selection.

The SRint2006 results tab in the CPT includes an additional column (baseline/core) that I add to the table. This identifies the processing performance of an individual core, a value that is used to estimate relative platform processing performance for a single sequential display. The relative processing performance per core values will be used in comparing user display performance.

9.2 Platform Performance

Hardware vendor technology has been changing rapidly over the past 10 years. Improved hardware performance has enabled deployment of a broad range of powerful software and continues to improve user productivity. Most business productivity increases experienced over the past 10 years have been promoted by faster computer technology. Technology today is getting fast enough for most user workflows, and faster compute processing is becoming less relevant. Most user displays are generated in less than a second. Access to Web services over great distances is almost as fast. Most of a user's workflow is think time—the time a user spends thinking about the display before requesting more information.

Most future user productivity gains will likely come from more loosely coupled operations, higher capacity network communications, disconnected processing, mobile operations, pre-processed cached maps, and more rapid access and assimilation of distributed information sources. System processing capacity becomes very important. System availability and scalability are most important. The quality of information provided by the technology can make a user's think time more productive.

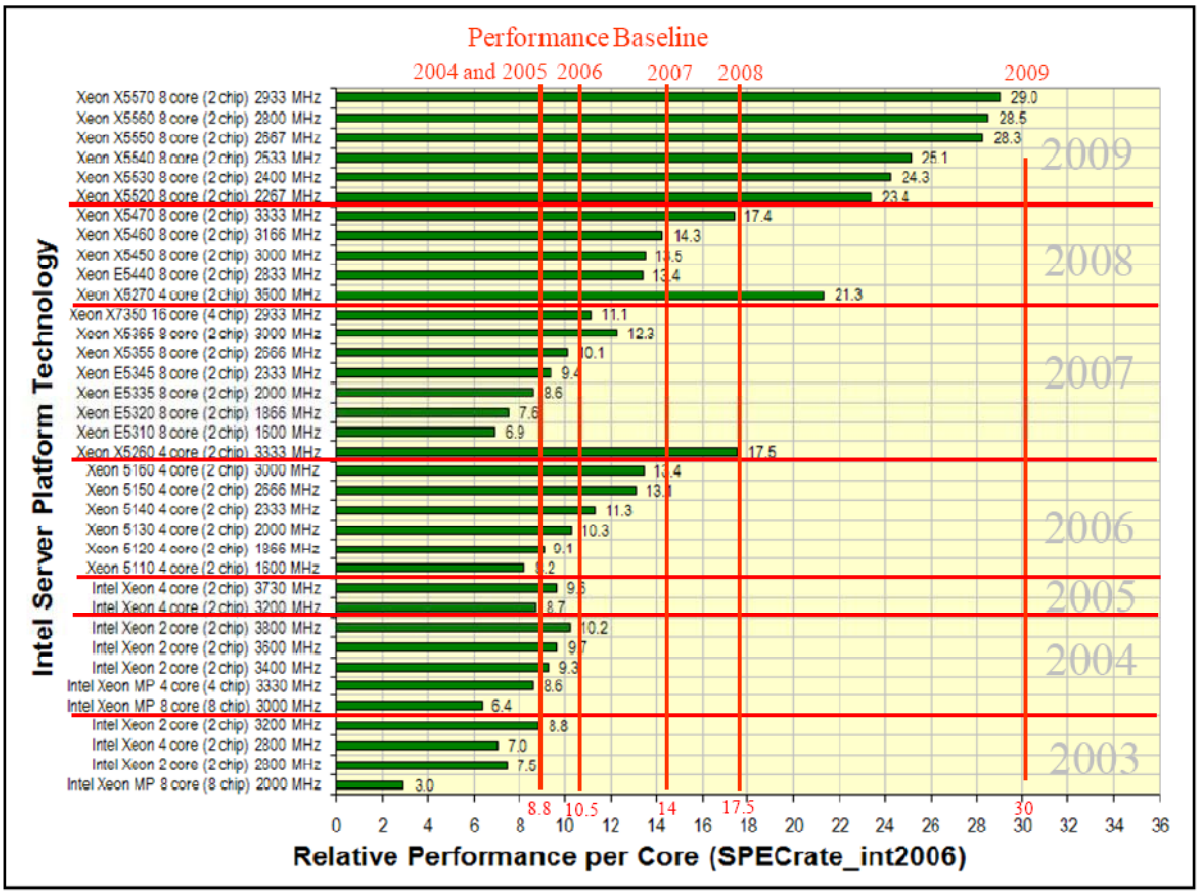
Hardware processing encountered some technical barriers during 2004 and 2005 which slowed the performance gains experienced between platform releases. There was little user productivity gain by upgrading to the next platform release (which was not much faster), so as a result, computer sales were not growing at the pace

experienced in previous years. Hardware vendors searched for ways to change the marketplace and introduced new technology with a focus on more capacity at a lower price. Vendors also focused on promoting mobile technologies, wireless operations, and more seamless access to information. Competition for market share was brutal, and computer manufacturers tightened their belts and their payrolls to stay on top. CY2006 brought some surprises with the growing popularity of the AMD technology and a focus on more capacity for less cost. Intel provided a big surprise with a full suite of new dual-core processors (double the capacity of the single-core chips) while at the same time significant processing performance gains at a reduced platform cost. Hardware vendor packaging (Blade Server technology) and a growing interest in virtual servers (abstracting the processing environment from the hardware) is further reducing the cost of ownership and provide more processing capacity in less space.

Figure 9-9 provides an overview of vendor-published single-core benchmarks[1] for hardware platforms using Intel processor technology.

Figure 9-9
Platform Performance Makes a Difference—Intel

Supported Intel Platforms



The Intel Xeon 3200 MHz platform (single-core SPECrate_int2000 = 18 / SPECrate_int2006 = 8.8) was released in 2003 and remained one of the highest-performing workstation platforms available through CY2005. The SPECint_rate2000 benchmark result of 18 was used as the Arc04 and Arc05 performance baseline.

[1] SPECrate_int2006 throughput benchmarks are published for a range of server configurations. A single GIS display is a sequential execution and display performance is a function of processor speed, which can be represented in terms of platform throughput per core. Platform throughput per core is calculated by dividing the published throughput by the total core included in the platform configuration.

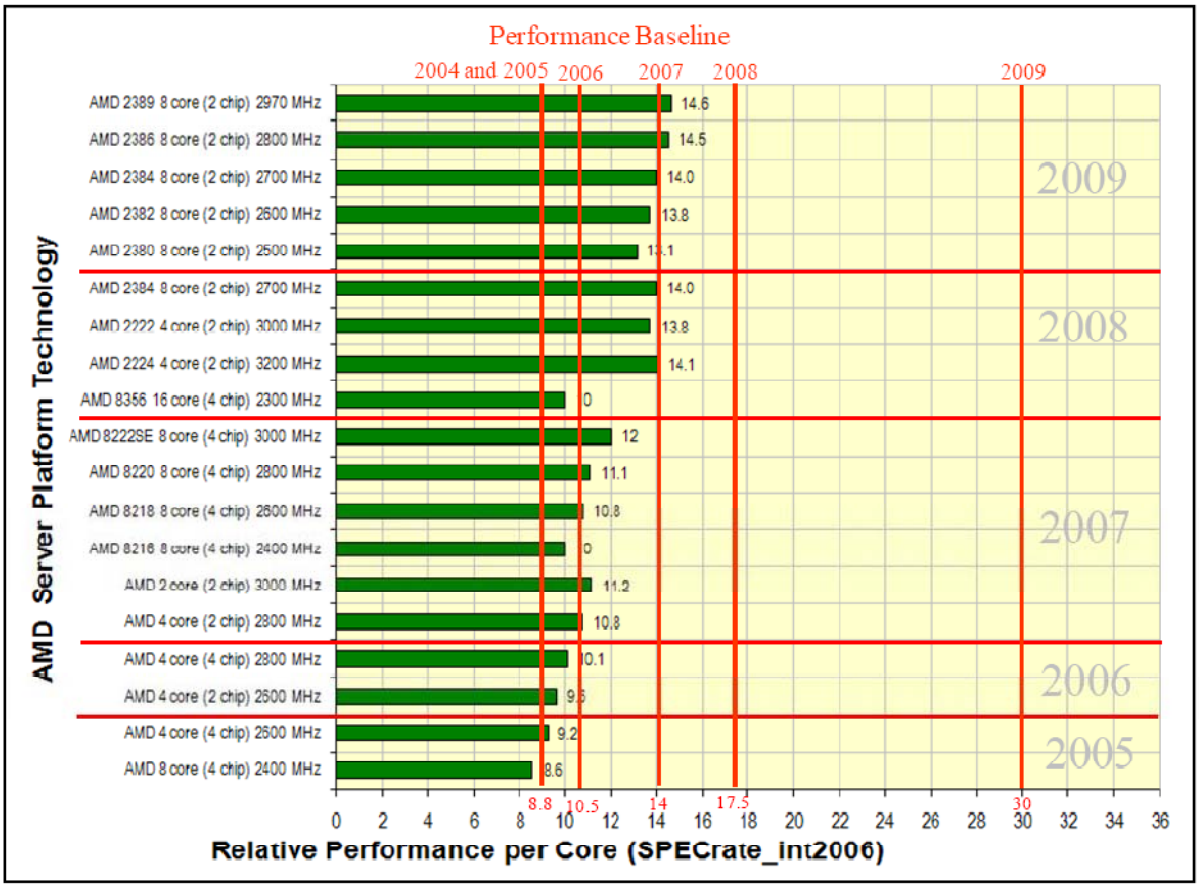
CY2005 was the first year since CY1992 that there was no noticeable platform performance change (most GIS operations were supported by slower platform technology).

There were some noticeable performance gains early in CY2005 with the release of the Intel Xeon 3800 MHz and the AMD 2800 MHz single-core socket processors. An Arc06 performance baseline of 22 (SPECrate_2006 = 10.5) was selected in May 2006. Since May, Intel released the Intel Xeon 5160 4 core (2 chip) 3000 MHz processor, a dual-core chip processor with a single core SPECrate_int2000 benchmark of 30 (SPECrate_int2006 = 13.4) and operating much cooler (less electric consumption) than the earlier 3.8 MHz release. The Arc07 performance baseline of 14 (SPECrate_int2006 = 14) was selected based on the Intel X5160 technology.

Figure 9-10 provides an overview of vendor-published single-core benchmarks for hardware platforms using AMD processor technology.

Figure 9-10
Platform Performance Makes a Difference—AMD

Supported AMD Platforms



AMD platforms were very competitive with Intel in the 2004 - 2005 timeframe. Since that time, Intel processor performance improvements have been much more impressive than available AMD alternatives.

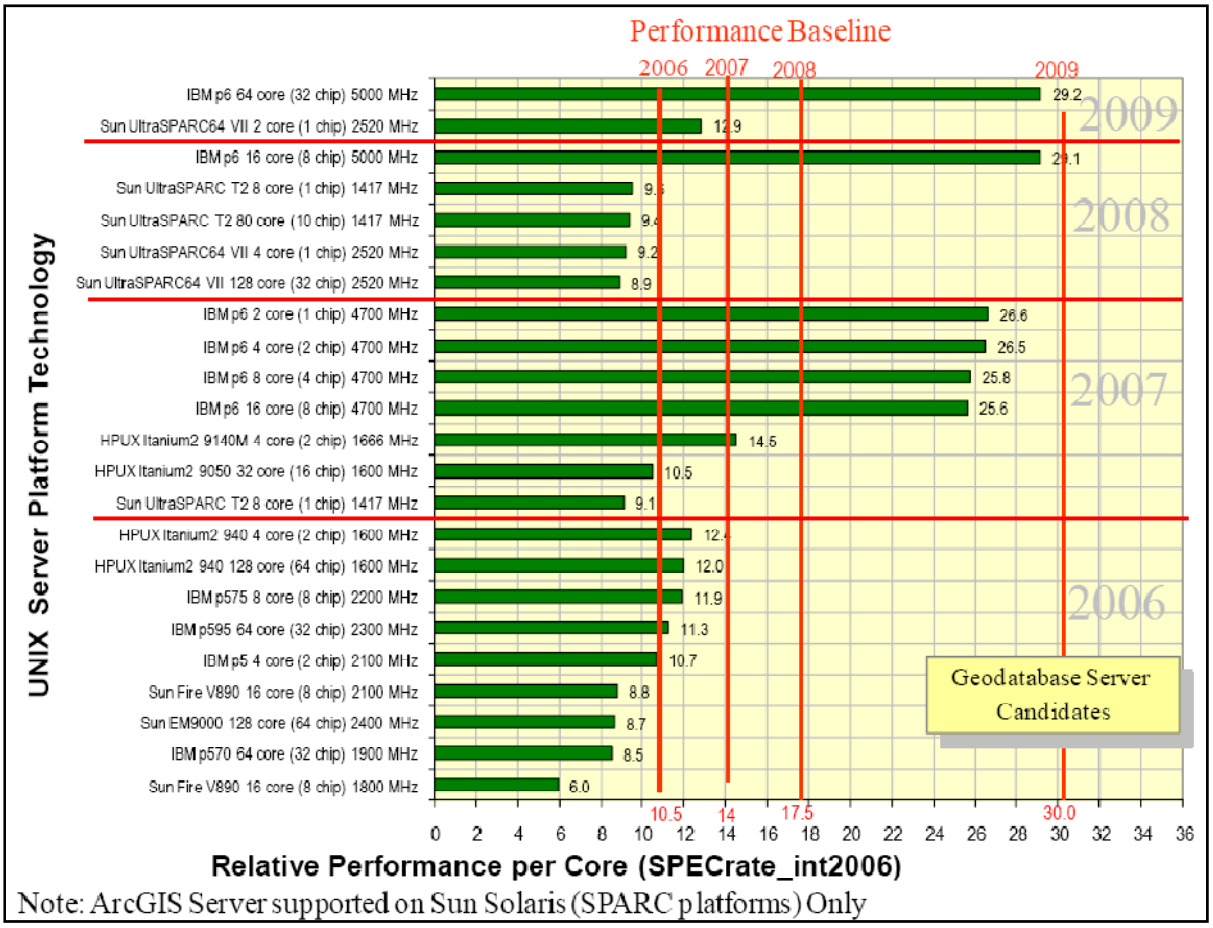
Intel technology continued to improve in CY2008 and server pricing was even more competitive. Hardware vendors were promoting platforms with dual core chips and reducing the price on lower performance low power configurations. The Xeon 5260 4 core (2 chip) platform (SPECrate_int2006 = 17.5) was selected as the 2008 baseline.

2009 was another great year for performance gains. Intel released a new chip technology that was over 70 percent faster per core than their 2008 release. Hardware vendors stop providing Dual core chip options, and all

entry level commodity servers include Quad core or higher capacity chips. The Intel Xeon 5570 8 core (2 chip) 2933 MHz platform was over 3.3 times the capacity of the 2008 baseline at about the same platform cost.

Figure 9-11 provides an overview of vendor-published per core benchmarks for hardware platforms supporting UNIX operating systems.

Figure 9-11
Platform Performance Makes a Difference—UNIX Supported UNIX Platforms



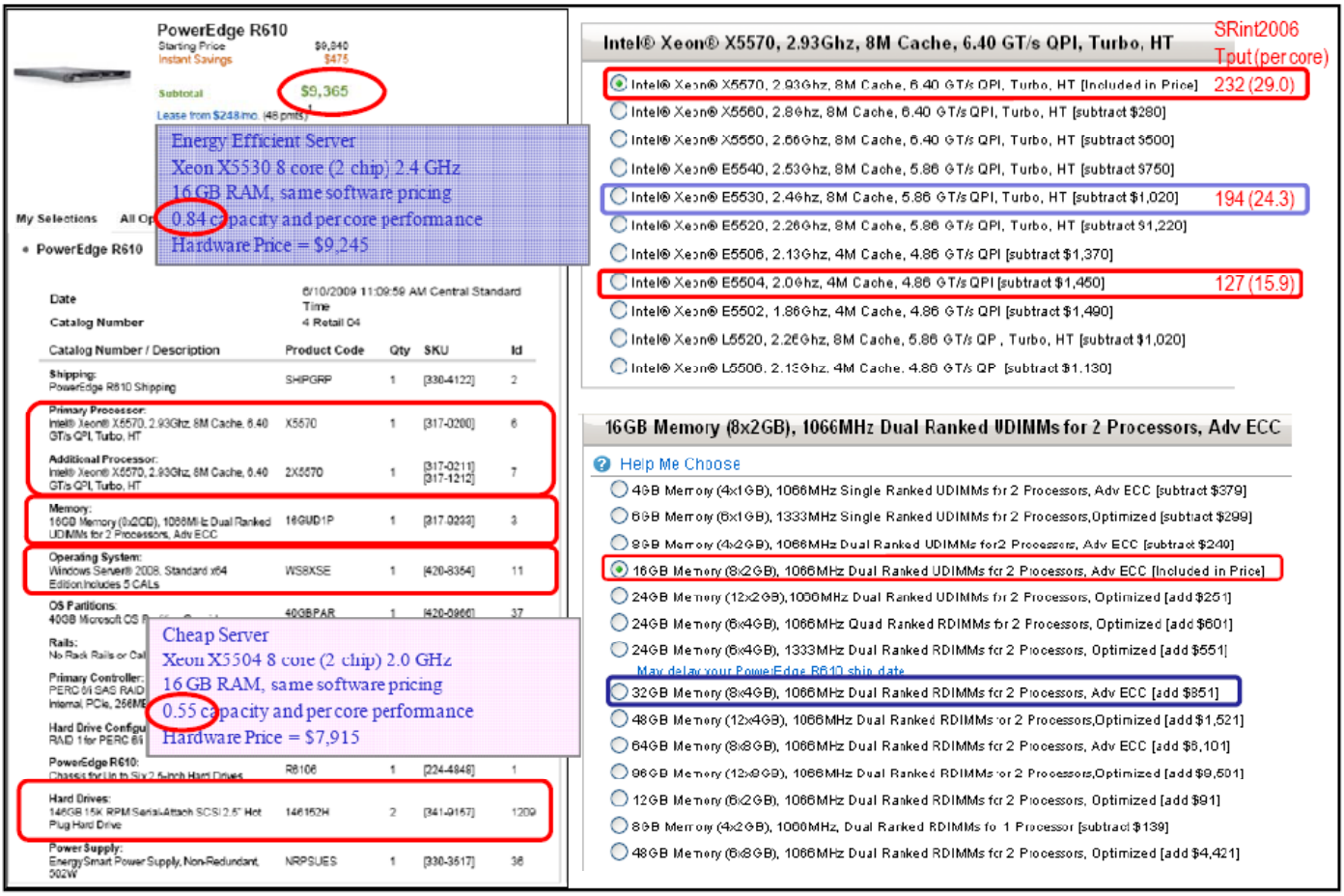
The UNIX market has focused for many years on large "scale up" technology (expensive high-capacity server environments). These server platforms are designed to support large database environments and critical enterprise business operations. UNIX platforms are traditionally more expensive than the Intel and AMD "commodity" servers, and the operating systems typically provide a more secure and stable compute platform.

IBM (PowerPC technology) is an impressive performance leader in the UNIX environment. Sun has also retained a significant hardware market share with many loyal customers, particularly in the GIS marketplace. Many GIS customers continue to support their critical enterprise geodatabase operations on UNIX platforms.

Hardware vendor efforts to reduce cost and provide more purchase options make it important for customers to understand their performance needs and capacity requirements. In the past new hardware included the latest processor technology, and customers would expect new purchases would increase user productivity and improve operations. In today's competitive market place, new platforms do not ensure faster processor core technology. You must understand your performance needs and consider relative hardware performance is selecting the right platform.

Figure 9-12 provides an overview of platform configuration options available on a DELL site. The selected platform is the 2009 recommended optimum ArcGIS Server container machine configuration.

Figure 9-12
Identifying the Right Platform
How do we select the platform we want?



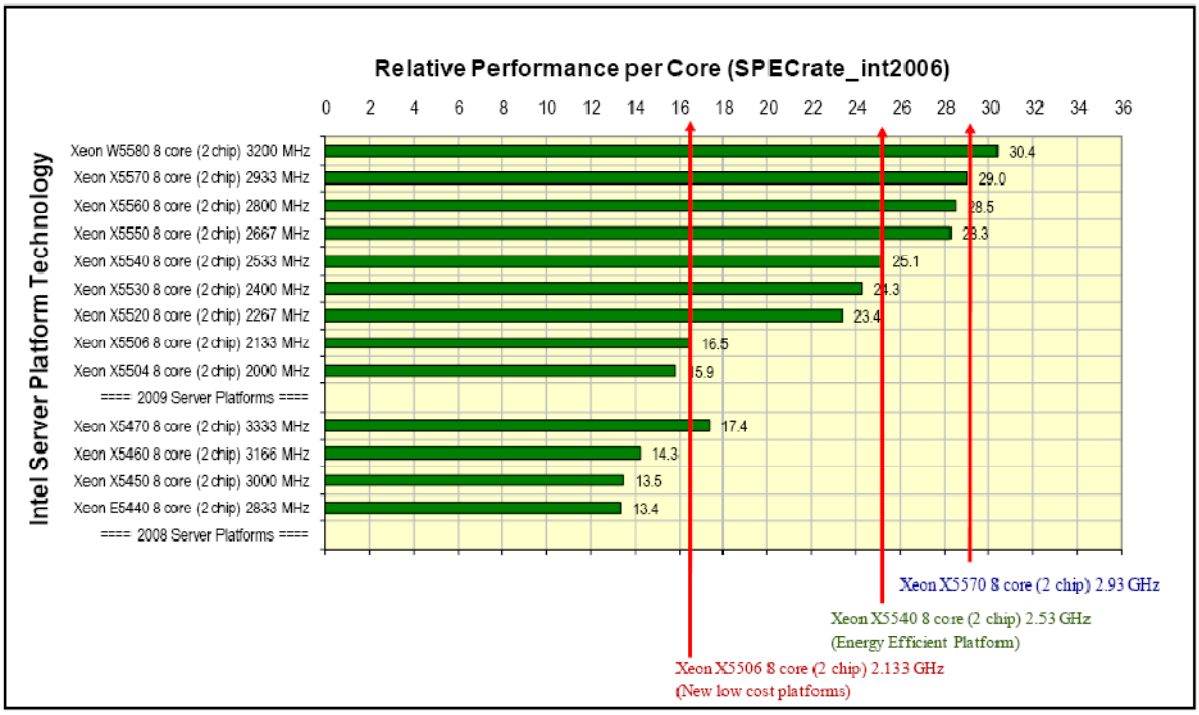
The ideal platform configuration would include the right processor, memory, and hard drive configuration. Configuring a Dell PowerEdge server with two Xeon X5570 quad core 2933 MHz processors, 16GB 1066 MHz memory, 64 Bit standard windows operating system, and dual RAID1 146 GB disk drives cost just over $9,000.

Processors: The X5570 2933 MHz processors provide the best performance/core. Selecting the "Energy Efficient" Xeon X5530 2400 MHz processor reduces the overall cost by just over $1000. What is not shown? Performance of the Xeon X5530 processor is 84 percent of the Intel Xeon X5570 configuration. The reduced cost of the X5504 quad core models attract the budget minded purchaser ($1,450 savings), but the performance is only 55 percent of the X5260 platform.

Selecting the right platform to meet your performance and capacity needs is more challenging than ever before. You really need to know your performance need, and the relative performance of platforms is not identified when you make your purchase - you need to know the model number you are looking for, and do your research before you buy, or you may be very disappointed with a platform that is not designed to meet your performance expectations.

Figure 9-13 provides a graphic overview of the current platforms on the market, and shows the relative performance per core for each. The chart also shows the SPEC baseline performance values for reference.

Figure 9-13
Vendor Published Platform Performance
(Available Dual Core Chip Performance)

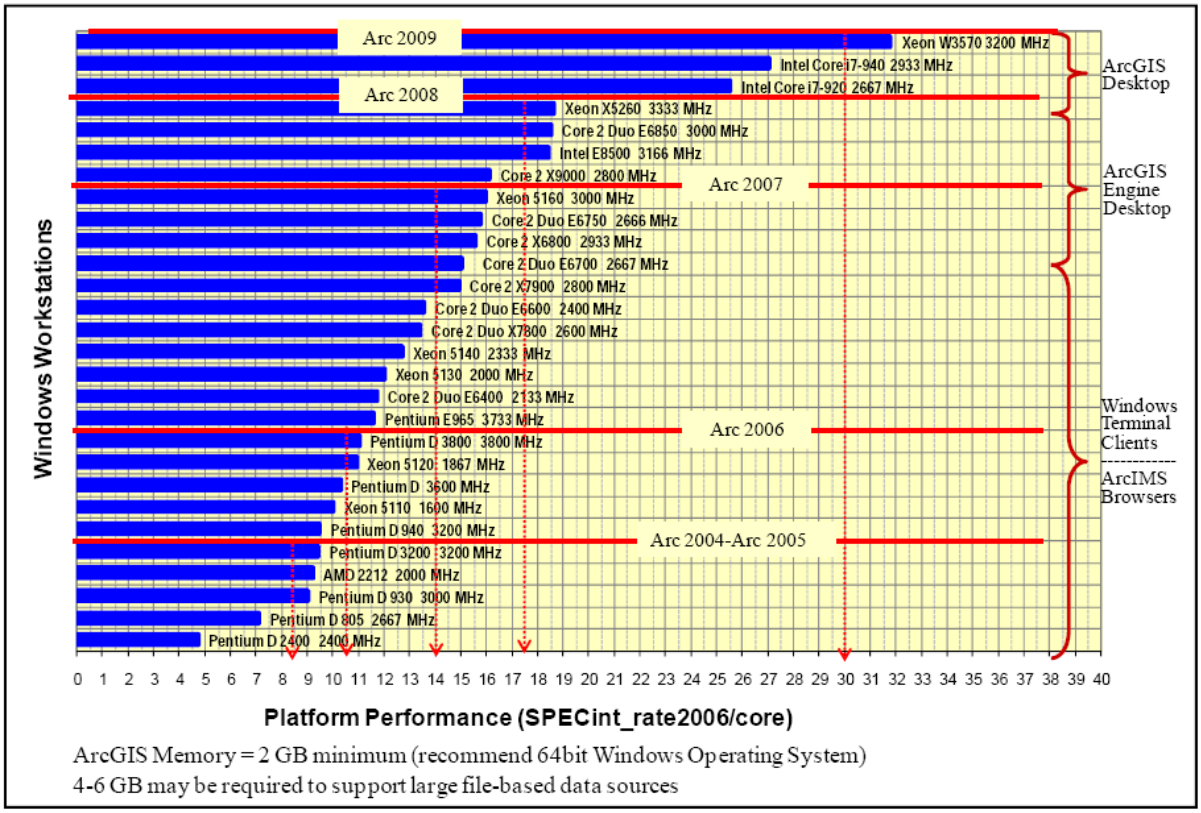


The platforms that run with reduced power are slower than the full power configurations (reduced power means reduced user productivity). Know what you are shopping for before you buy and you will be much happier with the performance of your new platform selection.

9.3 ArcGIS Desktop Platform Sizing

Figure 9-14 provides an overview of supported ArcGIS workstation platform technology. This chart shows the Intel platform performance changes experienced over the past five years. The new Xeon W3570 MHz quad-core processor is more than 6 times faster and almost 24 times the capacity of the Pentium D 2400 MHz platform that supported ARC/INFO workstation users in 2004. The advance of GIS technology is enriched by the remarkable contributions provided by ESRI's hardware partners.

Figure 9-14
Workstation Platform Recommendations



Full release and support for Windows 64-bit operating systems provide performance enhancement opportunities for ArcGIS Desktop workstation environments. The increasing size of the operating system executables and the number of concurrent operations supporting GIS operations makes more memory and improved memory access an advantage for ArcGIS Desktop users. Recommended ArcGIS Desktop workstation physical memory with an ArcSDE data source is 2 GB, and 4-6 GB may be required to support large image and/or file-based data sources.

Most GIS users are quite comfortable with the performance provided by current Windows desktop technology. Power users and heavier GIS user workflows will see big performance improvements with the new Xeon i7 or W5570 quad-core technology. Quad-core technology is now the standard for desktop platforms, and although a single process will see little performance gain in a multi-core environment there will be significant user productivity gains by enabling concurrent processing of multiple executables. Parallel processing environments such a 3D image streaming with ArcGIS Explorer 900 and future enhancements with 3D simulation and geoprocessing will leverage the increased capacity of multi-core workstation environments.

9.4 Server Platform Sizing Models

Figure 9-15 provides a view of the platform configuration module available with the configuration planning tool (overview presented in Chapter 10). This configuration framework will be used in chapter 11 to share a system architecture design methodology for selecting the right hardware solution to support specific operational performance and scalability needs.

Figure 9-15
Server Platform Sizing Models

Platform sizing models have been developed and maintained over the years to support ESRI customers with system architecture design planning and proper selection of supported vendor hardware. A fundamental discussion of these models was presented in chapter 7 Performance Fundamentals.

This chapter applies these capacity planning models to current hardware technology and provides some simple engineering charts that can be used for proper hardware selection. The charts presented in this chapter are consistent with the models described in chapter 7, and the sizing charts can be used to validate standard ArcGIS Desktop and Web mapping service platform recommendations generated by the excel based capacity planning tool.

The ESRI software technology patterns have expanded significantly with the ArcGIS 9.3 release, and the charts included in this section address a small range of the fundamental GIS dynamic deployment strategies. The capacity planning tool includes a much broader range of workflow options, and should be used as the primary capacity planning tool for selecting the most optimum technology solutions.

9.5 Windows Terminal Server Platform Sizing

Windows Terminal Server supports centralized deployment of ArcGIS Desktop applications for use by remote terminal clients. Figure 9-16 identifies three standard Windows Terminal Server software configurations. A separate platform sizing chart will be provided to address each solution architecture.

Figure 9-16
Windows Terminal Server Architecture

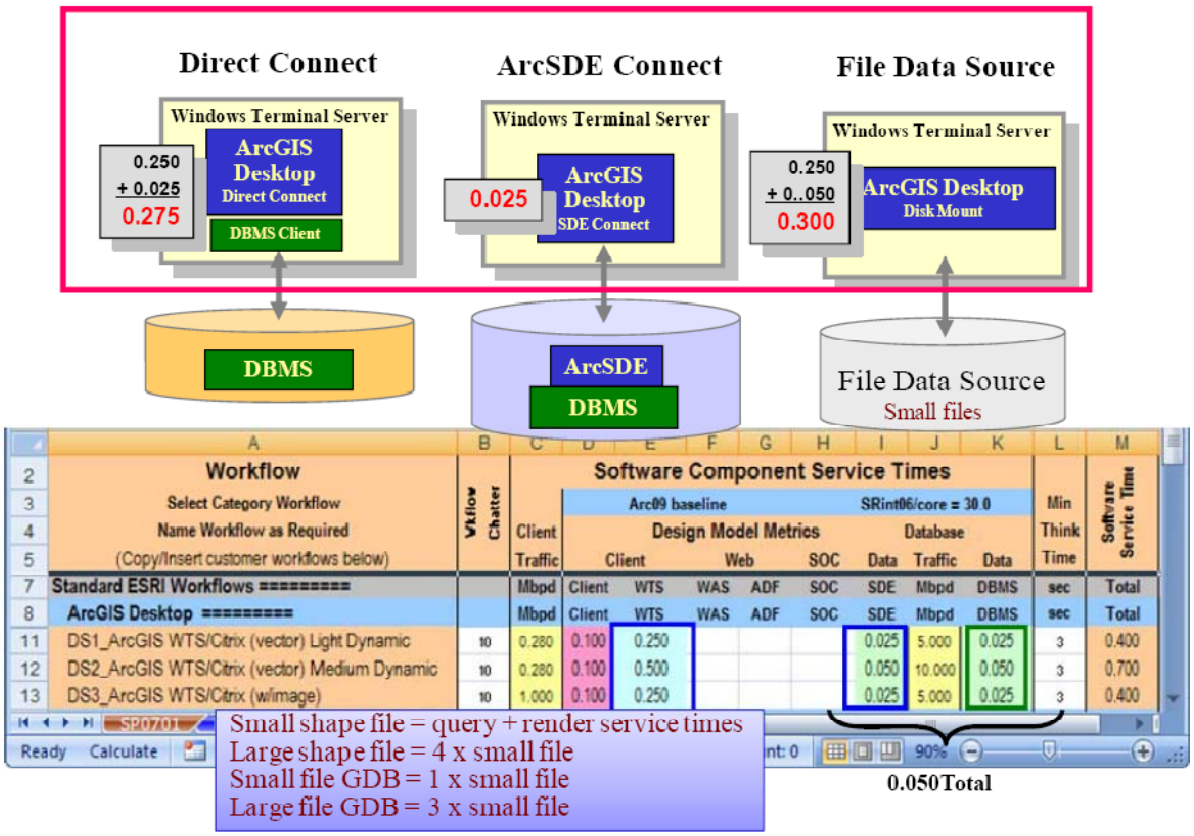
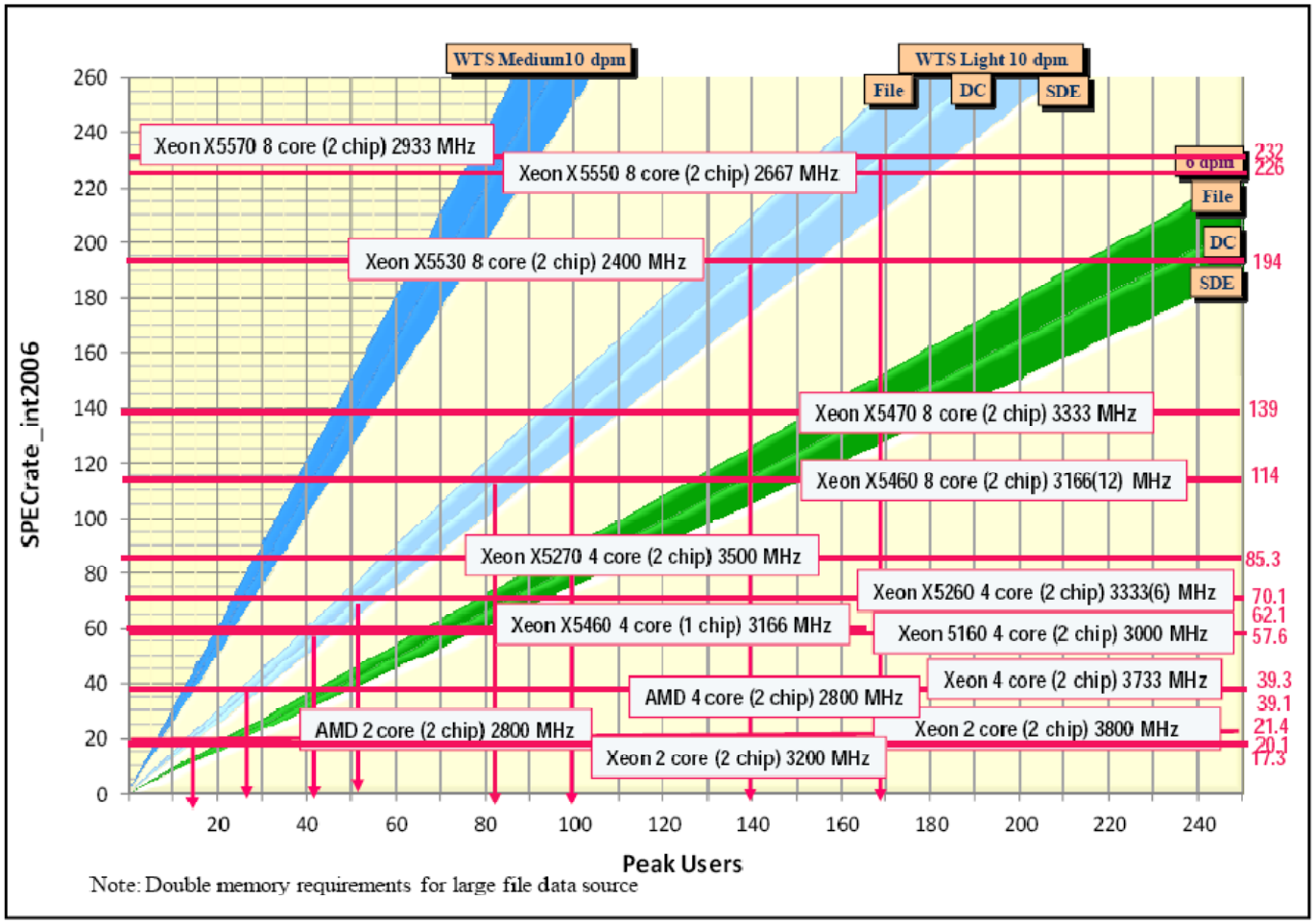


Figure 9-17 introduces a standard platform sizing chart that will be used throughout this chapter as a tool to identify peak concurrent users that can be supported with a selected vendor platform configuration. Hardware platforms are represented on sizing charts as a horizontal line. The location of the platform on the chart is determined by the vendor -published SPECrate_int2006 benchmark results for the represented platform configuration.

The capacity planning models introduced in chapter 7 are represented on these charts. A platform configuration graphic is included showing the software component installation represented by the sizing chart. The peak displays per minute introduced in chapter 7 are used with the Arc09 component service times to identify specifications required to support each specific GIS workflow. The platform performance specifications are represented by the vendor-published SPECrate_int2006 benchmark on the vertical axis of the sizing chart.

There are three diagonal fans on the sizing chart (WTS Medium 10 DPM, WTS Light 10 DPM, and WTS Light 6 DPM). Each fan includes performance based on three different data sources (Small Shape File, ArcSDE Direct Connect, and ArcSDE Application Server connect). Peak user capacity is determined by dropping down from the intersection of the selected platform configuration (horizontal lines) with the data source configuration on the associated user productivity fan (diagonal lines on the productivity fans).

Figure 9-17
Windows Terminal Server Sizing



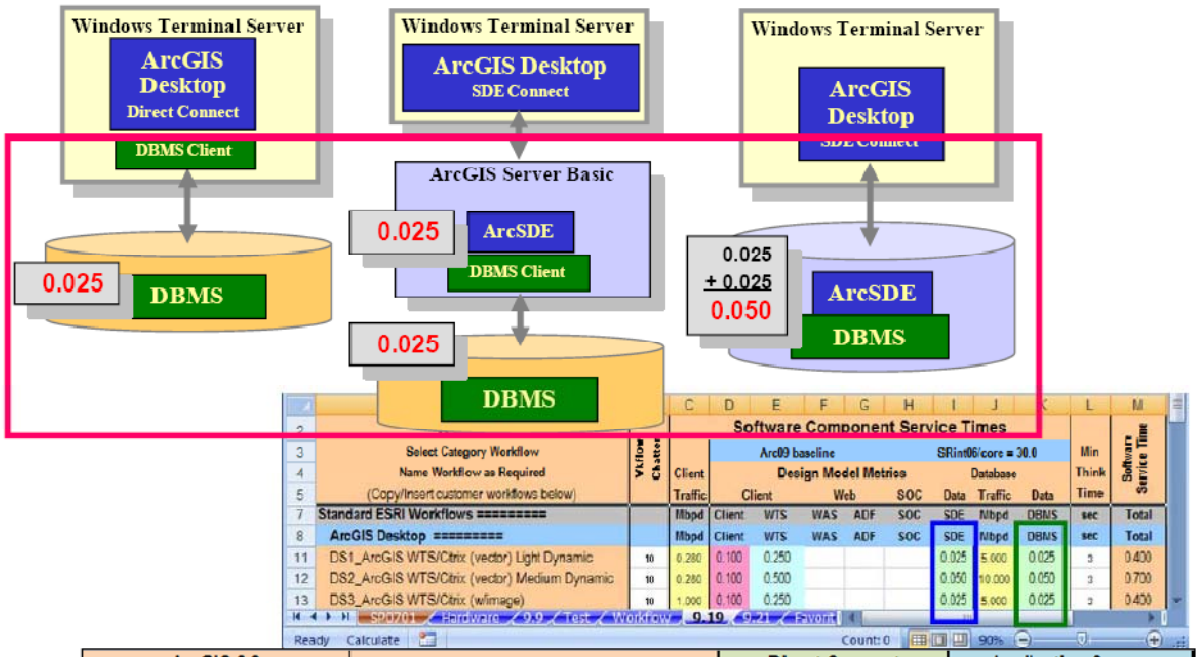
Note: Double memory requirements for large file data source

A WTS Medium workflow is representative of GIS power users with a medium complexity map display environment (many display layers, heavier display queries, heavy map display complexity). The WTS Light workflow represents higher performance optimized workflow environments. The concept of user productivity is a parameter introduced with the Arc06 sizing models, and as user workflows generate more complex processing and rich map displays customers may find user productivity will reduce accordingly (i.e., a display generated in 1 second may support a workflow with 10 displays per minute, which a richer display generated in 2 seconds could reduce user productivity to 5 displays per minute). It is important to understand user information display requirements, and provide applications that will support display requirements with optimum use of available system resources (i.e., simple map displays).

9.6 ArcSDE Geodatabase Server Sizing

Figure 9-18 identifies recommended software configuration options for the geodatabase server platforms. The geodatabase transaction models apply to both ArcGIS Desktop and Web mapping service transactions. Normally a geodatabase is deployed on a single database server node, and larger capacity servers are required to support scale-up user requirements. ArcGIS Server 9.2+ includes distributed geodatabase replication services that can be used to distribute instances of a single SDE geodatabase over multiple server nodes.

Figure 9-18
ArcSDE Geodatabase Server Architecture Alternatives

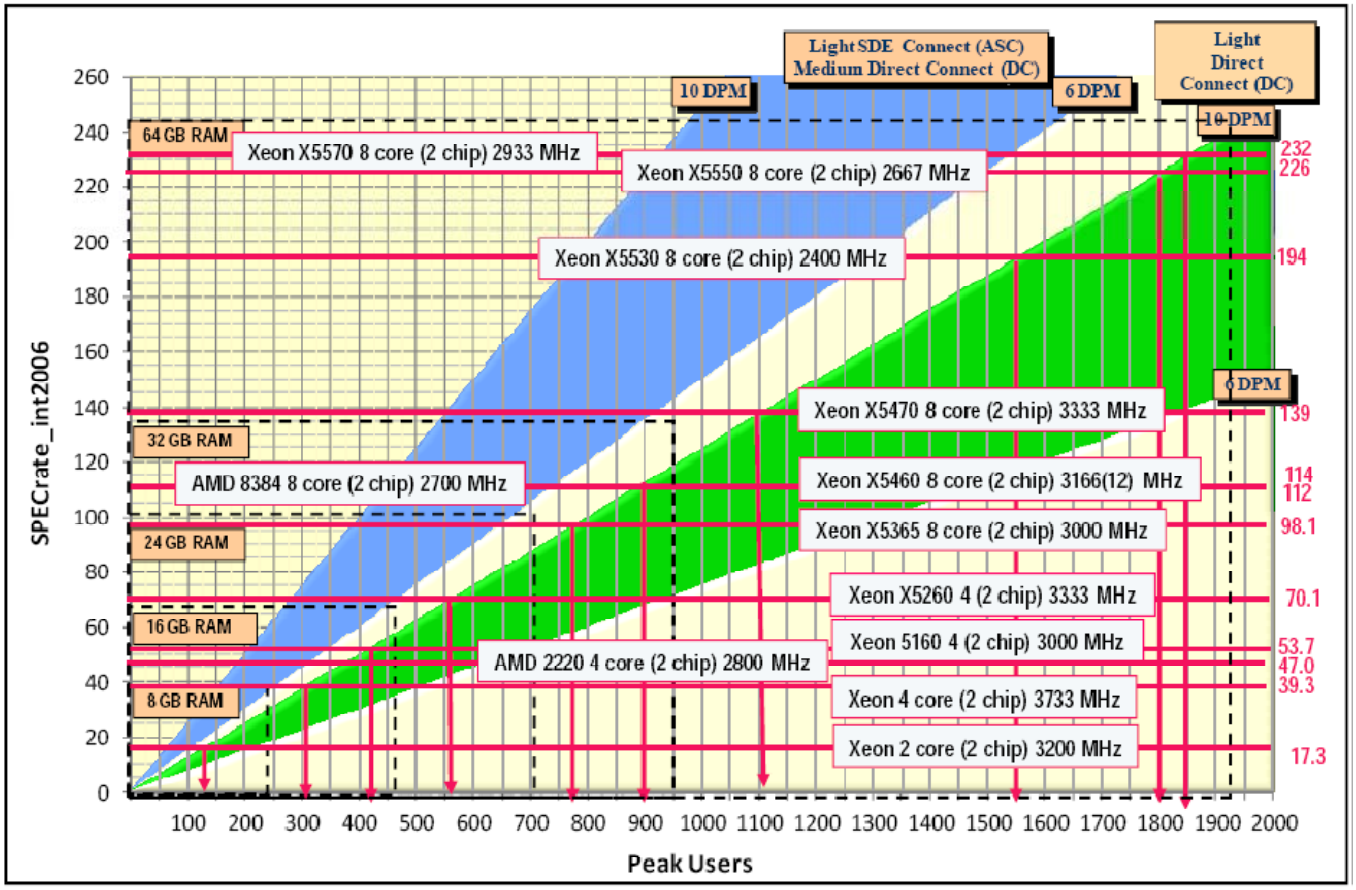


ArcGIS 9.3 Supported Databases	Spatial Data Types	Direct Connect Unversioned	Direct Connect Versioned	Application Server Unversioned	Application Server Versioned
Microsoft SQL Server	SDE Binary, Geography, Geometry	X	X	X	X
Oracle	ST_Geometry, Large Object Binary,	X	X	X	X
	Oracle Spatial (SDO_Geometry), Long Raw	X	X	X	X
IBM DB2	Spatial Extender	X	X	X	X
IBM Informix Datablade	Spatial Datablade	X	X	X	X
Postgre SQL	ST_Geometry, PostGIS	X	X	X	X

The ArcSDE and DBMS display processing times (service times) are roughly the same for capacity sizing purposes, so the DBMS Server and ArcSDE Remote Servers platform sizing models are the same. Three platform sizing charts are provided—two are configured for the more common capacity systems (under 2,000 concurrent users) and the other is configured for high-capacity systems (up to 5,000 users concurrent users).

Figure 9-19 shows a platform sizing chart for the ArcSDE Geodatabase Server showing capacity of the more common commodity Windows server platforms.

Figure 9-19
ArcSDE Geodatabase Windows Server Sizing (up to 8 core platforms)

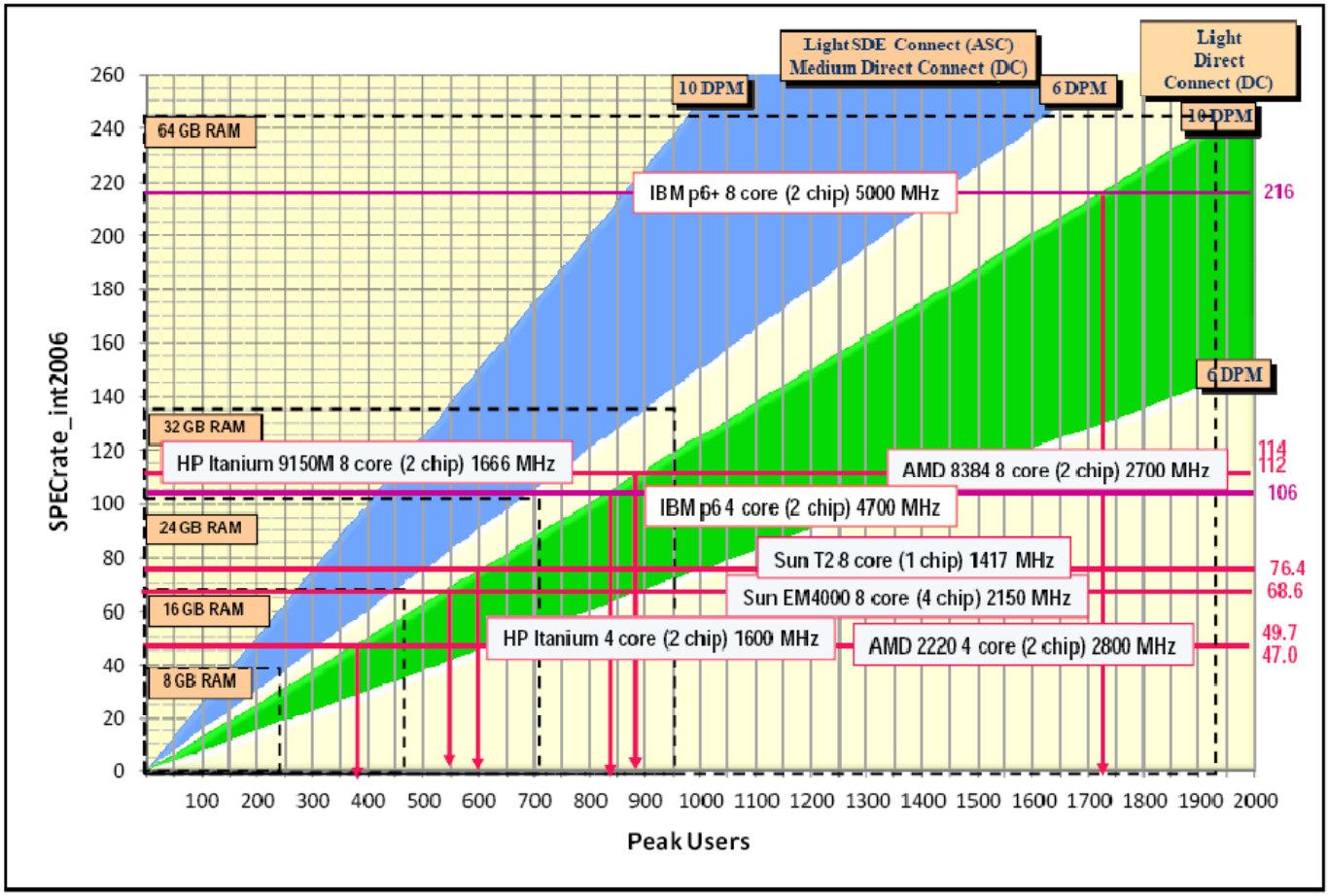


The geodatabase sizing chart above includes two diagonal fans, one showing performance capacity for light complexity geodatabase configurations that include ArcSDE installed on the DBMS server (can also be used for medium complexity geodatabase direct connect configurations) and the other showing performance capacity for configurations where client applications connect to a light SDE Geodatabase server through a direct connect architecture (ArcSDE executable is not installed on the DBMS server). Database server capacity is doubled when client applications use a geodatabase direct connect architecture.

Geodatabase platform capacity has improved dramatically over the past few years. The Xeon 2 core (2 chip) 3200 MHz platforms introduced in CY2004 would support less than 100 concurrent users in an Application Server Connect (ASC) architecture (SDE on the DBMS platform). Hardware technology performance improvements along with introduction of multi-core processors has increased the peak capacity of Intel Xeon two chip commodity server platforms to over 800 concurrent users (Xeon X5460 8 core (2 chip) 3166(12) MHz platform can support up to 18,000 concurrent users with clients using the recommended Geodatabase Direct Connect architecture. Many GIS customers are moving from UNIX to lower cost Windows platforms to support their enterprise GIS DBMS platform requirements. It is important to take advantage of the Windows 64-bit Operating System, since these higher capacity servers will require much more physical memory to handle the high number of current active client connections. Up to 84 GB of memory is required to take full advantage of the 1800 concurrent user capacity available with the Xeon X5570 8 core (2 chip) 2933 MHz platforms.

Figure 9-20 shows an ArcSDE Geodatabase Server platform sizing chart for the smaller capacity UNIX server platforms.

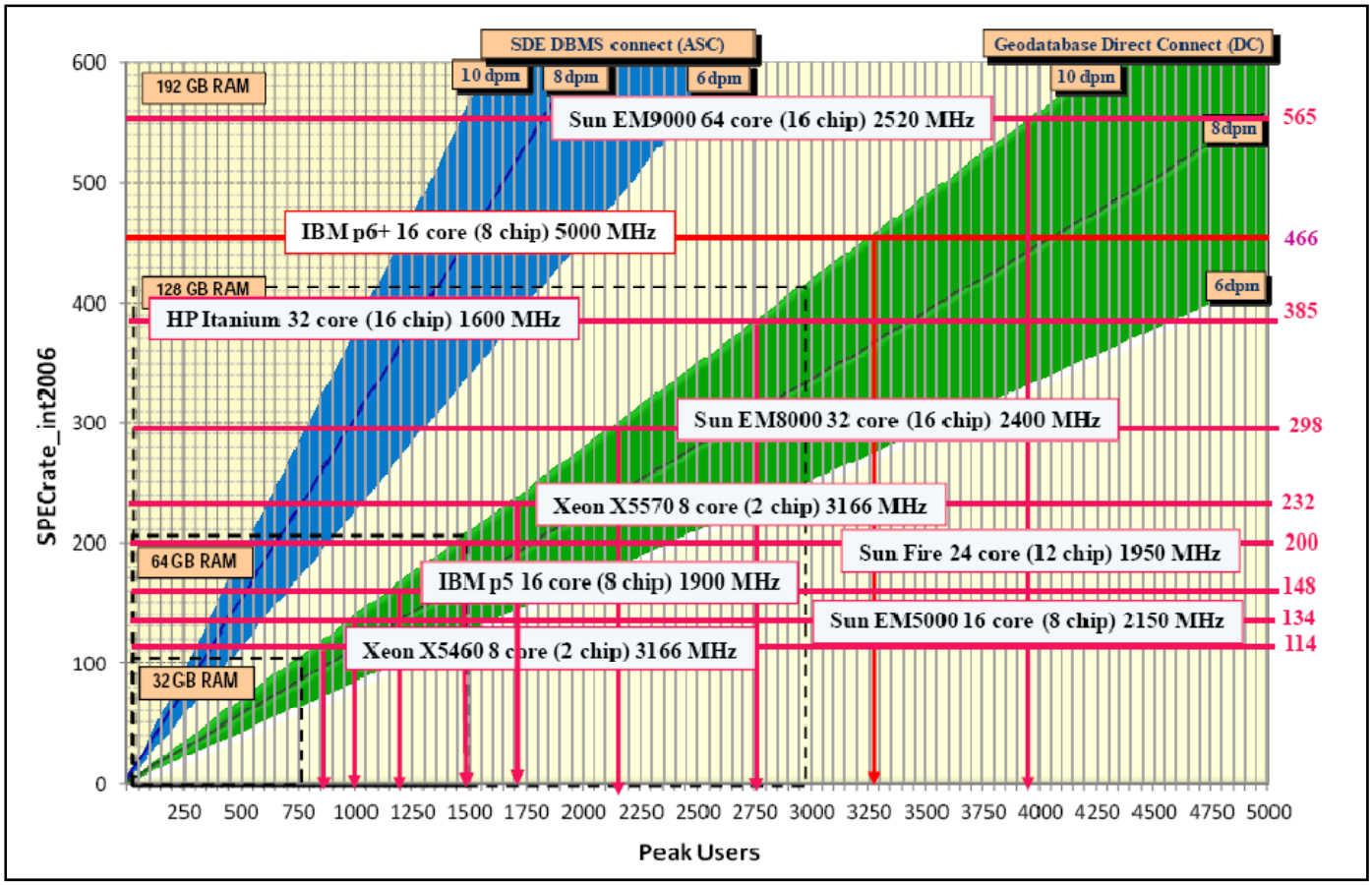
Figure 9-20
ArcSDE Geodatabase UNIX Server Sizing (up to 8 core platforms)



UNIX platforms are having a hard time maintaining their market share within standard GIS enterprise environments. In many cases, platform performance is less than their Intel counterparts and the platform technology cost more. Many critical production database environments continue to be hosted on UNIX platforms, although many IT departments are considering migration to a commodity server architecture to reduce cost and improve adaptability.

Figure 9-21 shows a platform sizing chart for the ArcSDE Geodatabase Server demonstrating the high capacity available with the more scalable DBMS platform configurations (capacity up to 5000 concurrent users).

Figure 9-21
ArcSDE Geodatabase Server Sizing (Large Capacity Platforms)



Larger Unix platforms can provide sufficient processing capacity to support well over 5000 concurrent users. The IBM AIX power6 16 core (8 chip) 4700 MHz platform supports up to 3400 concurrent power users - this platform can provide configurations up to 64 core with sufficient processing capacity for over 12,000 concurrent users. For Enterprise GIS environments, sizing servers with thousands of concurrent users is well beyond demonstrated processing loads and network throughput levels the DBMS software and storage infrastructure has experienced for a single server platform environment. A distributed geodatabase architecture provides a lower risk alternative for handling these peak capacity loads.

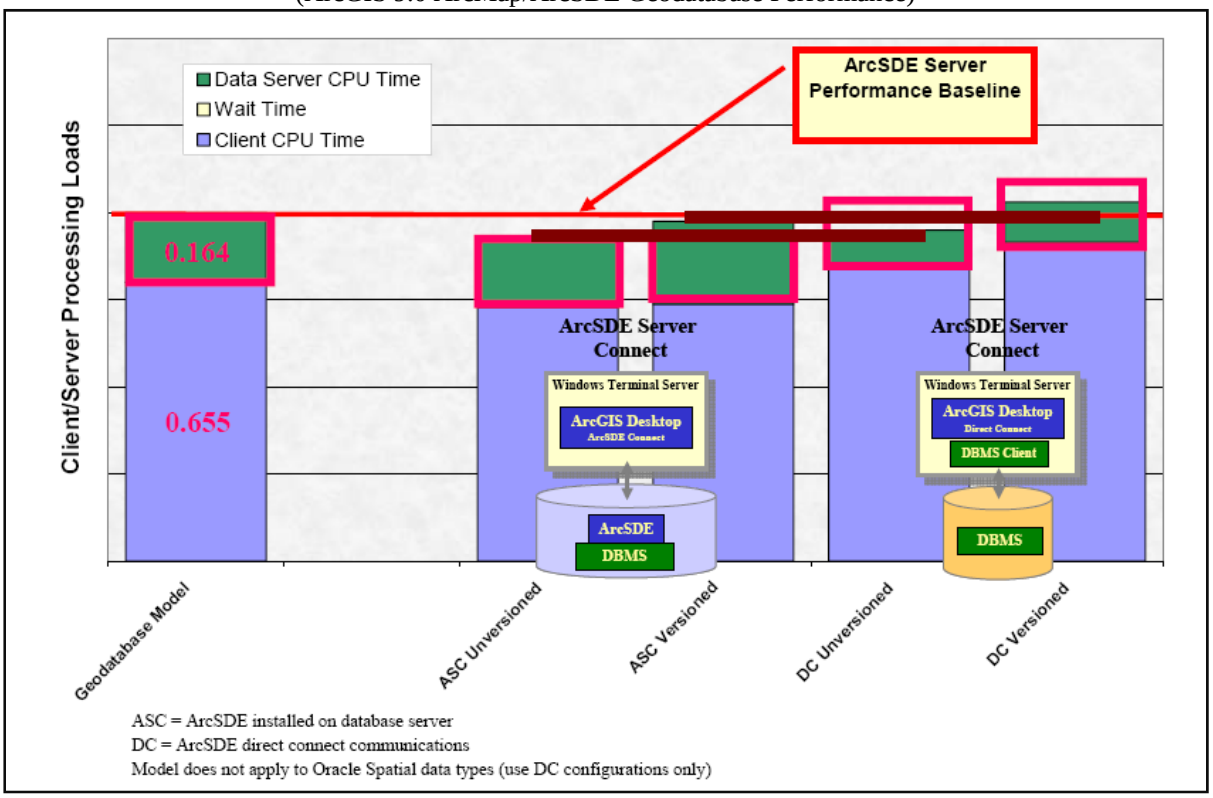
9.7 ArcSDE Application Server Connect vs Direct Connect Architecture

There are three configuration options shown in this section for connecting ArcGIS applications to an ArcSDE Geodatabase. ESRI recommends use of the Direct Connect as the preferred deployment architecture. Software performance improvements introduced in 2004 with the ArcGIS 9.0 release removed user performance as a potential concern in selecting your ArcSDE configuration. This section provides an overview of why ESRI recommends ArcSDE direct connect as the preferred geodatabase architecture.

Figure 9-22 provides a user performance comparison between the ArcSDE Application Server Connect (ASC) and ArcSDE Direct Connect (DC) configuration options. Performance when comparing Unversioned and Versioned database configurations are roughly the same with either configuration choice.

Figure 9-22
Geodatabase Direct Connect Performance Validation Test

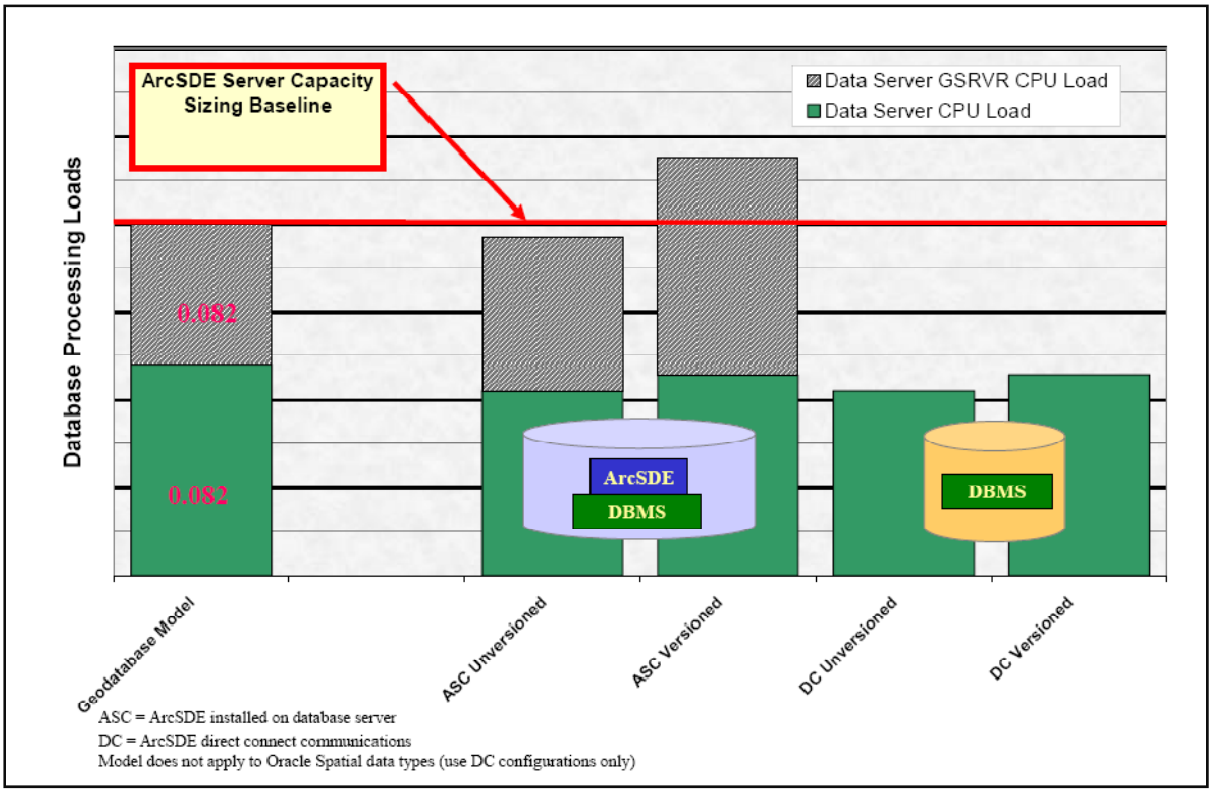
(ArcGIS 9.0 ArcMap/ArcSDE Geodatabase Performance)



A closer look at the database server processing load will highlight the real advantages of the direct connect architecture. When using an ArcSDE Direct Connection to the DBMS, the ArcSDE processing load is supported by the client application and not on the DBMS server. The ArcSDE processing load can be as heavy as the DBMS load, so removing ArcSDE processing from the DBMS server doubles the performance capacity of the data server platform.

Figure 9-23 takes a closer look at comparing the data server loads for the ASC and DC configurations. The Direct Connect options reduce the data server processing load by roughly 50 percent.

Figure 9-23
ArcSDE 9.0 Geodatabase Server Loads

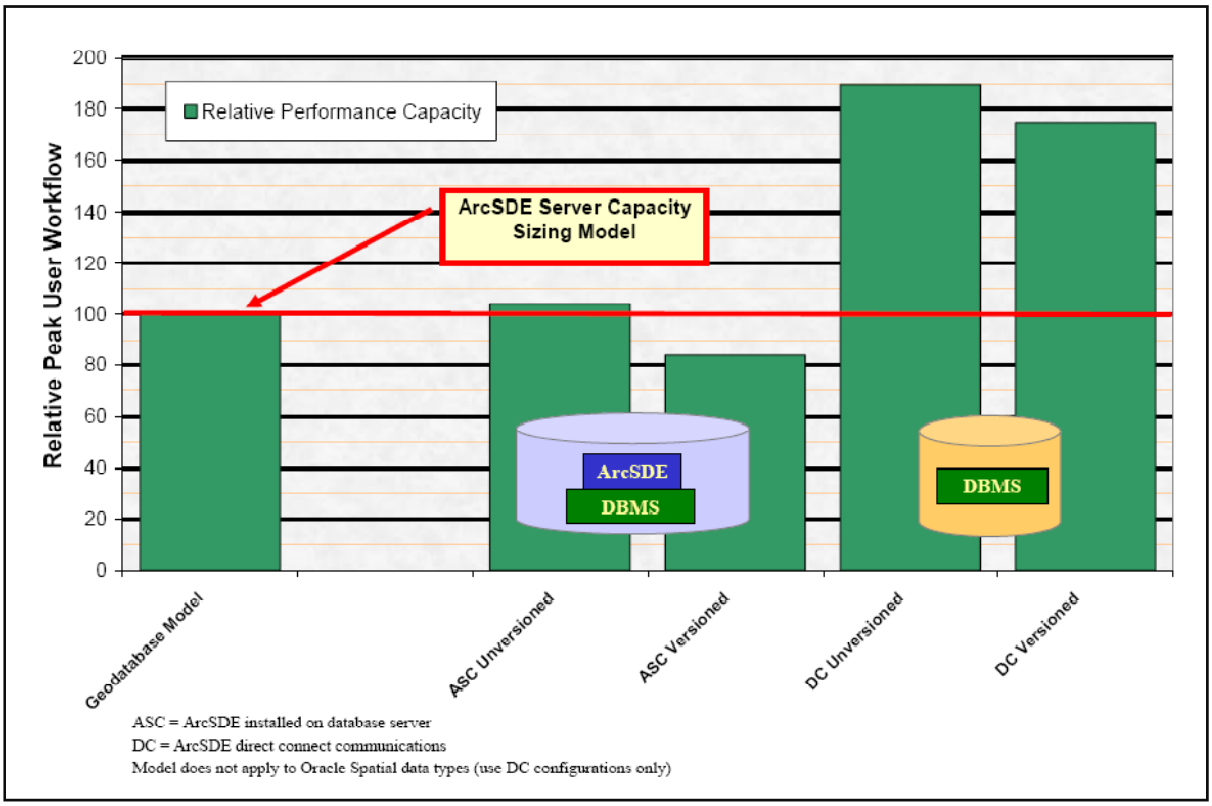


There are several potential advantages for having a lighter DBMS server load. Peak capacity processing loads can be supported on a DBMS server with half the number of core (reduce DBMS licensing by up to 50 percent). Larger capacity Enterprise GIS systems can be supported by lower cost server platform technology. ESRI changed how we license ArcSDE with the ArcGIS 9.2 release - we no longer count the DBMS server core when using an ArcSDE Direct Connect architecture (only when the ArcSDE executables are not installed on the DBMS server).

There were some limitations with the earlier ArcGIS 9 release that led many users to continue supporting ArcSDE executables on the DBMS server. One limitation restricted client applications from connecting to an earlier release geodatabase (ArcSDE executable will connect only to the same version geodatabase configuration). The ArcGIS 9.3 release includes the ability for connecting to previous version geodatabase schema using the client direct connect architecture.

Figure 9-24 shows the capacity change when using a direct connect architecture. The DBMS server capacity increases by up to 50 percent when using the ArcSDE Direct Connect architecture.

Figure 9-24
ArcSDE 9.0 Geodatabase Relative Capacity Sizing

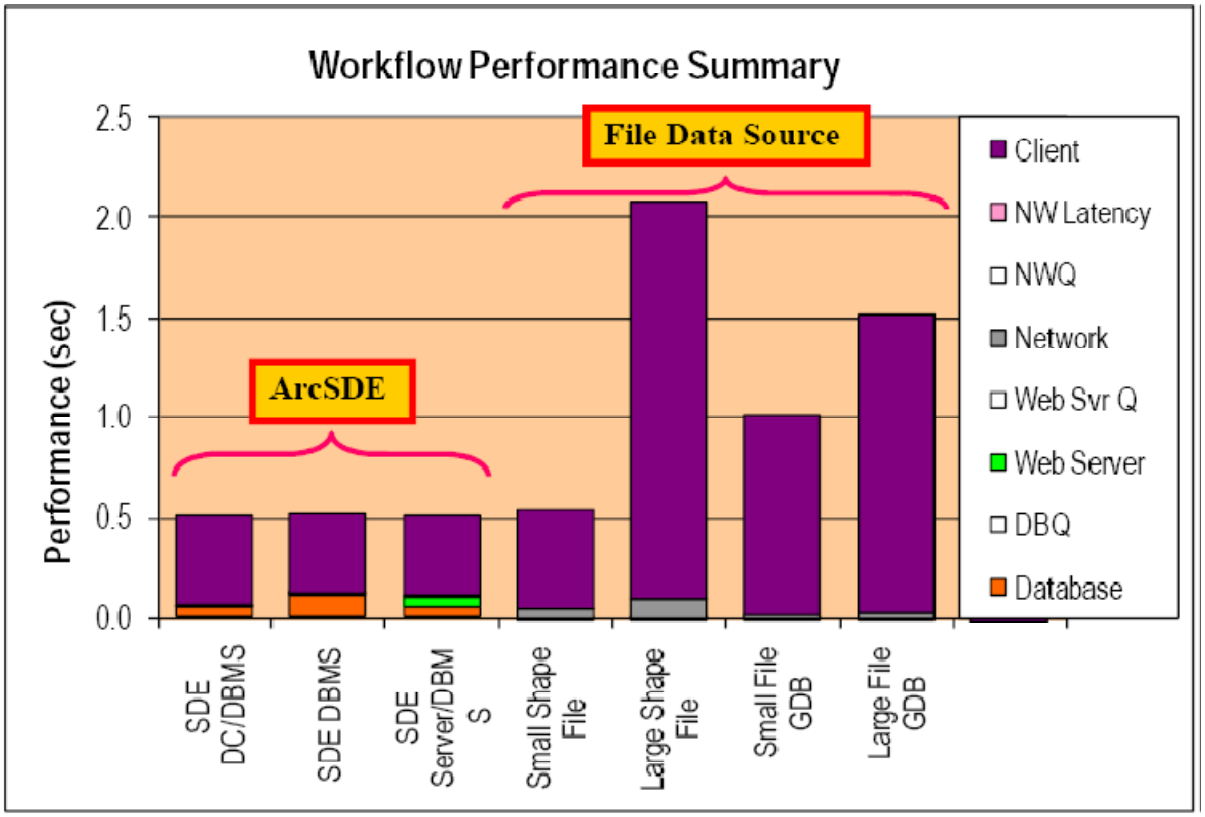


Many ESRI customers continue to configure ArcSDE on the DBMS server. Most customers are moving to a direct connect architecture to simplify administration and reduce licensing costs. ESRI continues to fully support both implementation strategies.

9.8 File Data Server Sizing

Figure 9-25 compares display performance between using an ArcSDE Geodatabase data source and the available File data sources. The file data sources include small and large shape files and file geodatabase. Values shown on the chart are those currently represented in the Capacity Planning Tool data source performance targets. The File Geodatabase was introduced with the ArcGIS 9.2 release.

Figure 9-25
Data Server Relative Performance Test Results



The small Shape File performance is about the same as the ArcSDE Geodatabase. The large Shape File format requires four times the processing required with ArcSDE. The small File Geodatabase loads are about twice the ArcSDE Geodatabase performance values, while the large File Geodatabase is three times the ArcSDE loads.

Figure 9-26 shows the data extent of the San Diego geodatabase used in performance validation testing. Initial evaluation of performance with shape files was conducted using this data set.

Figure 9-26
San Diego Data Set

The performance test series generated common displays for 100 random maps within the neighborhood area identified in the map above. The ArcSDE geodatabase display performance was roughly the same whether using the full San Diego extent or the small neighborhood area as the data source.

The small Shape File testing used vector layers extracted for the neighborhood extent identified above. The large Shape File data set included the complete San Diego area (the Capacity Planning Tool large shape file load would represent twice the full San Diego data extent).

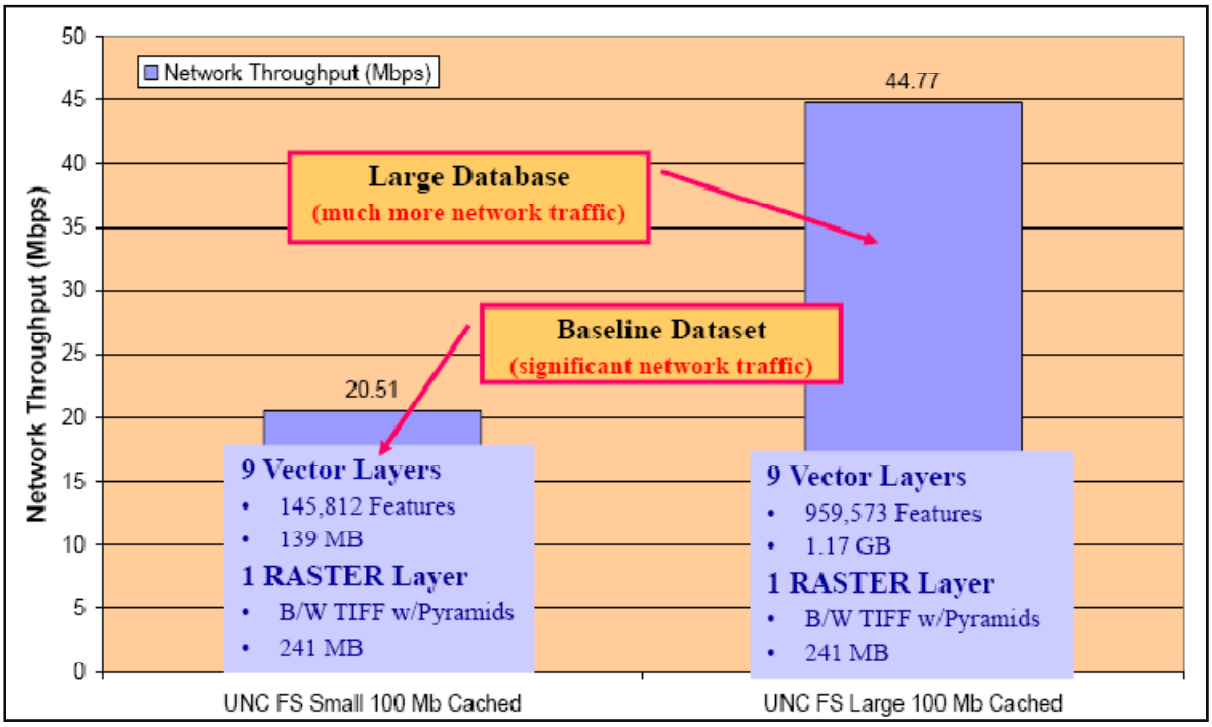
Display performance with the File Geodatabase is much improved over the Shape File format (test results are not shown). Tests have been completed accessing a File Geodatabase with up to 1 TB of data, and performance was quite good. The small File Geodatabase performance targets would likely support a data source representing the full San Diego extent. The large File Geodatabase performance targets would data sources up to 1 TB in size.

It is difficult to be precise when evaluating these performance targets. There are many factors in the database design, number of features per layer, and number of layers in the display that can factor into these performance values. Processing loads on the file data server platform are very light (all query processing loads for a file data source are performed by the client application).

Figure 9-27 provides some additional information on the number of display layers, total number of features, and the size of the San Diego data source used during our initial performance validation testing. The same raster layer was used for both tests.

Figure 9-27
File Server Network Traffic

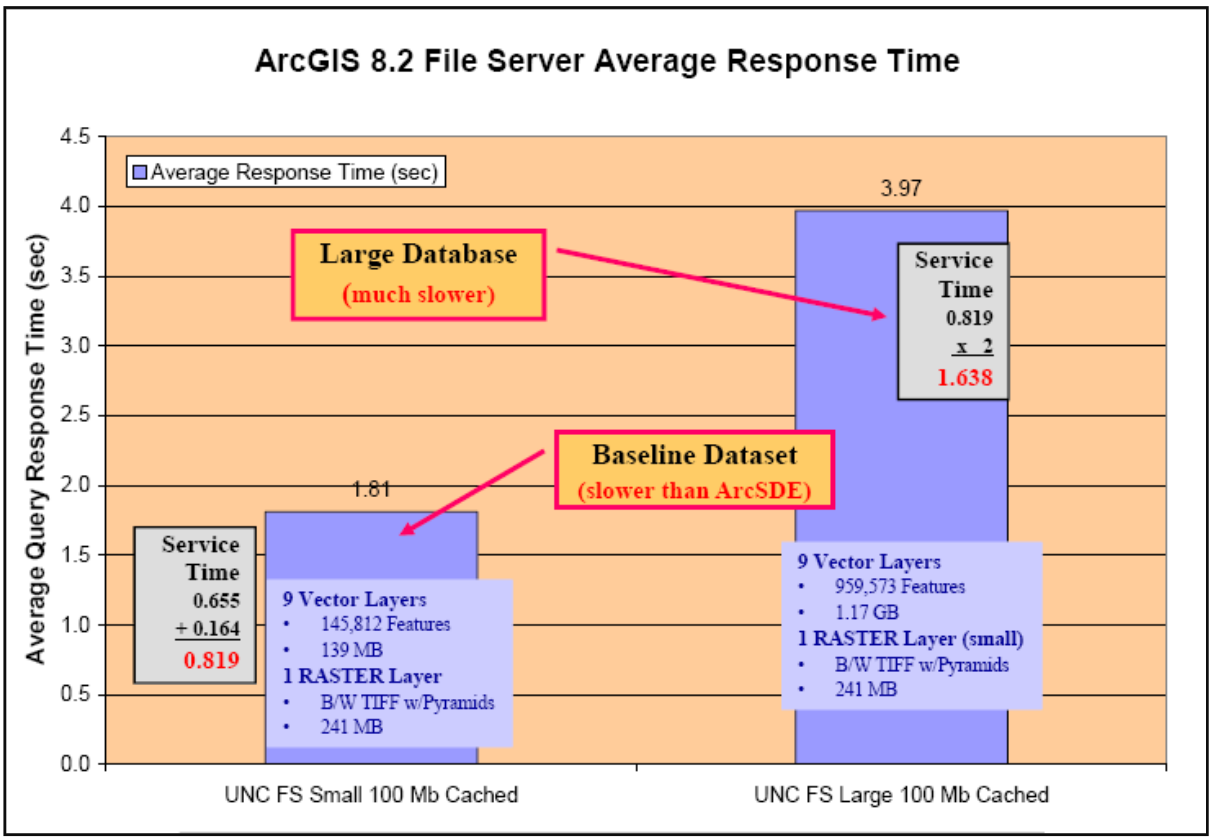
(Shape File Data Source)



Network throughput for the small Shape File data source was half the traffic of the large Shape File data source.

Figure 9-28 shows the different in display performance for the two Shape File data source. The large Shape File required twice the processing of the small Shape File.

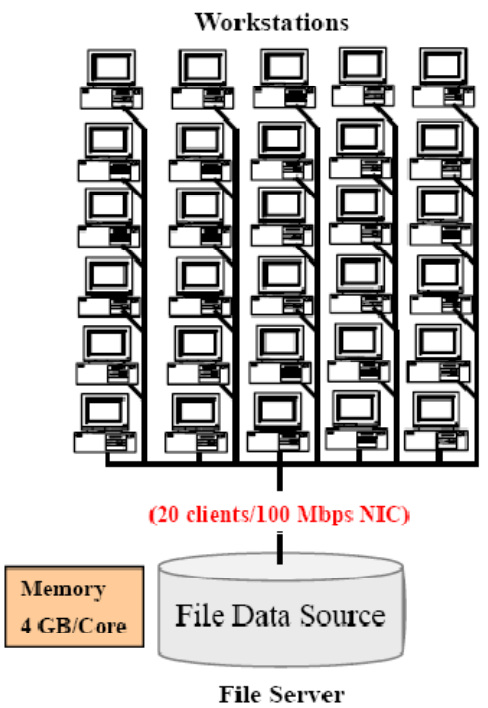
Figure 9-28
File Server Query Performance
(Shape File Data Source)



The large Shape File performance targets included in the Capacity Planning Tool represent 4 times the processing of the small Shape file, thus these performance targets should be adequate to support twice the extent identified in the test above (9 vector layers with up to 2 million features or possibly twice the number of vector layers at the same extent).

Figure 9-29 provides some general guidelines for File Server Platform Sizing. Query processing is not supported by the File Server platform, thus the platform compute loads are quite light. The primary performance factor determining File Server capacity is network capacity (available bandwidth). Network communication guidelines provided in Chapter 3 (Network Communications) identify the factors determining file server capacity. It is important to avoid disk contention, and store data across multiple disk using the standard RAID configurations discussed in Chapter 6 (Data Administration). The file server will need to be configured with network cards that will support peak traffic flow requirements (network interface controller <NIC> cards should have at least twice the capacity of the peak display traffic flow).

Figure 9-29
GIS File Server Platform Sizing



ArcGIS Desktop service time varies with size of the GIS data files

GIS File Servers

Capacity Planning Matrix

Traffic per Map Display (Mb)		File Server NIC Bandwidth (Mbps) 10	100	200	1000
10	Peak concurrent Clients	6	60	120	600
20		3	30	60	300
30		2	20	40	200
40		2	15	30	150
50		1	12	24	120
60		1	10	20	100
70		1	9	17	86
80		1	8	15	75
90		1	7	13	67
100		1	6	12	60

Average 50 Mb Traffic per Map Display (assume 1 display every 6 seconds)

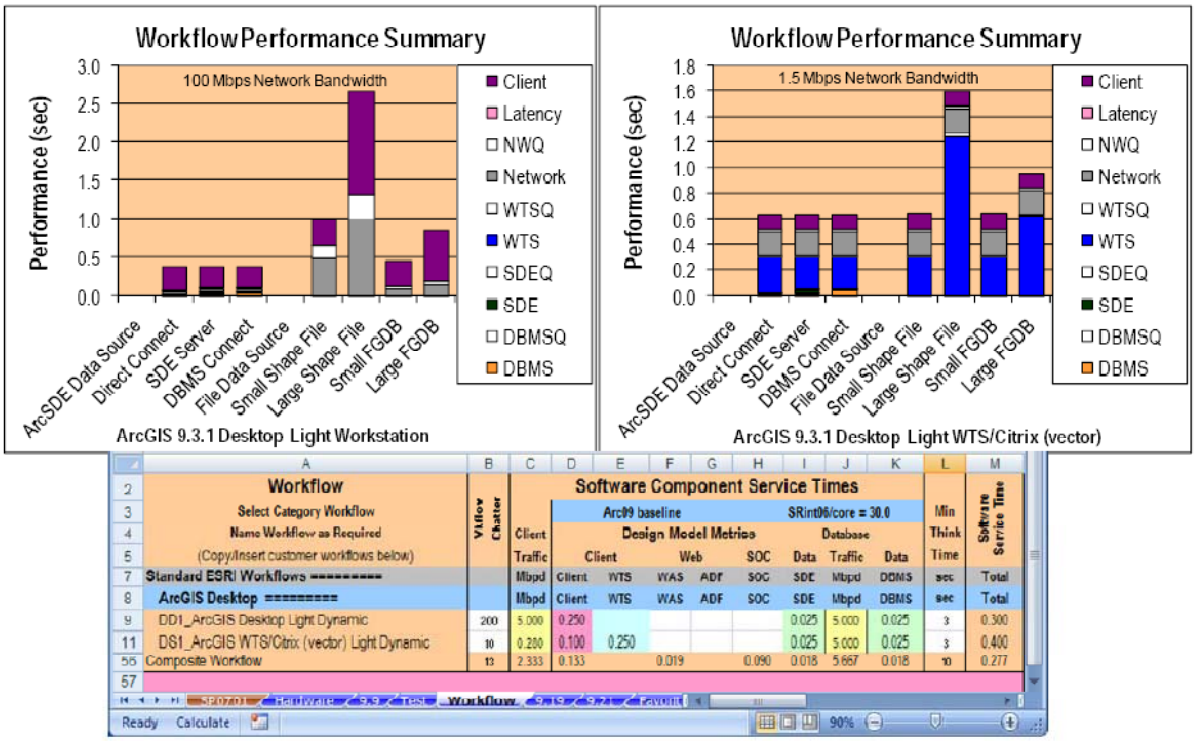
Additional file server memory will be used by the server Operating System as a data cache. More memory improves data access performance. We recommend configuring the file server with 4 GB per core, which would be 16 GB for a four core server and 32 GB for an eight core server.

Network card bandwidth requirements will depend on the display traffic loads. The Capacity Planning Tool uses an estimate of 50 Mbpd for traffic from a Shape File data source. Assuming user productivity of 10 displays per minute, a 100 Mbps NIC card would support up to 12 concurrent GIS users.

9.9 ArcGIS Desktop Standard Workflow Performance

Figure 9-30 provides an overview of the display performance targets currently represented in the Standard ESRI Workflows used in the Capacity Planning Tool.

Figure 9-30
ArcGIS Desktop Performance Summaries
(Standard ESRI Workflows)



The 14 workflow combinations identified above can be generated from just three Standard ESRI Workflows included on the Capacity Planning Tool workflow tab. The first chart shows workflows using the ArcGIS Desktop Workstation workflow, while the second chart shows the same workflows with ArcGIS Desktop application supported on a Windows Terminal Server configuration.

9.10 Web Mapping Servers

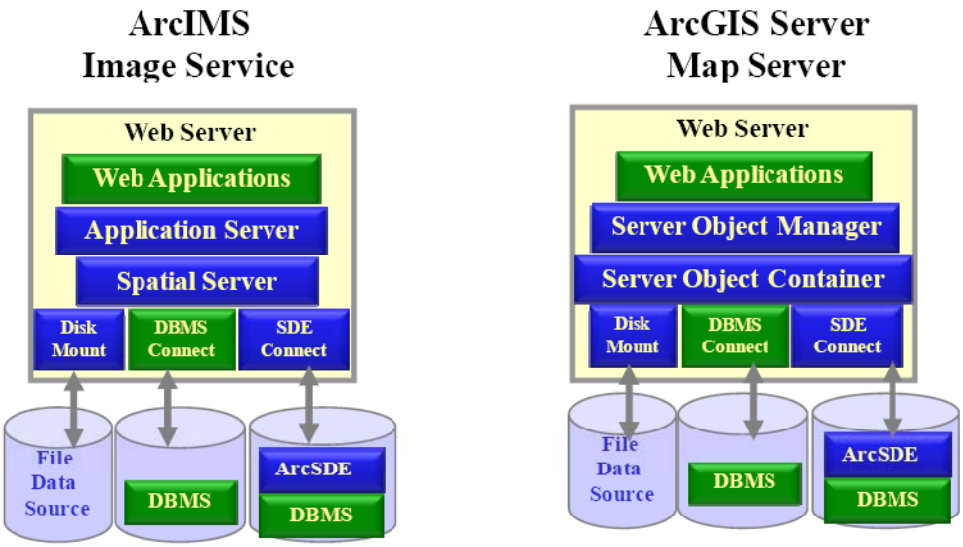
Web mapping services platform sizing guidelines are provided for the ArcIMS and ArcGIS Server software technology. The ArcIMS image service is deployed using the ArcIMS software, and the ArcGIS Server map services are deployed using the ArcGIS Server software. All Web mapping technologies can be deployed in a mixed software environment (they can be deployed on the same server platform together). All mapping services can be configured to access a file data source or a separate ArcSDE database. Geodatabase access can be through direct connect or an ArcSDE server connection.

9.10.1 Web Two-Tier Architecture

Figure 9-31 identifies recommended software configuration options for standard two-tier Web mapping deployments. This configuration option supports the Web server and spatial servers (container machines) on the same platform tier. This configuration is recommended for implementations that can be supported by one- or two-server platforms.

Figure 9-31
Server Performance and Scalability—Two Tier Architecture

Smaller Two-Tier Implementations



	A	B	C	D	E	F	G	H	I	J	K	L	M
2	Workflow	Wkflow Chatter	Software Component Service Times										
3	Select Category Workflow			Arc09 baseline				SRint06/core = 30.0				Min Think Time	Software Service Time
4	Name Workflow as Required		Client	Design Model Metrics					Database				
5	(Copy/Insert customer workflows below)		Traffic	Client		Web		SOC	Data	Traffic	Data		
6	Standard ESRI Workflows =========		Mbpd	Client	WTS	WAS	ADF	SOC	SDE	Mbpd	DBMS	sec	Total
20	SA6_AGS931 ADF_MXD Light Dynamic	10	2.000	0.100		0.053		0.175	0.018	5.000	0.018	3	0.363
21	SA7_AGS931 ADF_MXD Medium Dynamic	10	2.000	0.100		0.105		0.350	0.035	5.000	0.035	3	0.625
27	SS4_AGS93 REST_MXD Light Dynamic Service	10	2.000	0.100		0.018		0.175	0.018	5.000	0.018	3	0.328
28	SS5_AGS931 REST_MSD Light Dynamic Service	10	2.000	0.100		0.018		0.088	0.018	5.000	0.018	3	0.241
42	IMS3_IMS Image Server (Web App)	10	1.000	0.100		0.026		0.088	0.018	5.000	0.018	3	0.250
56													

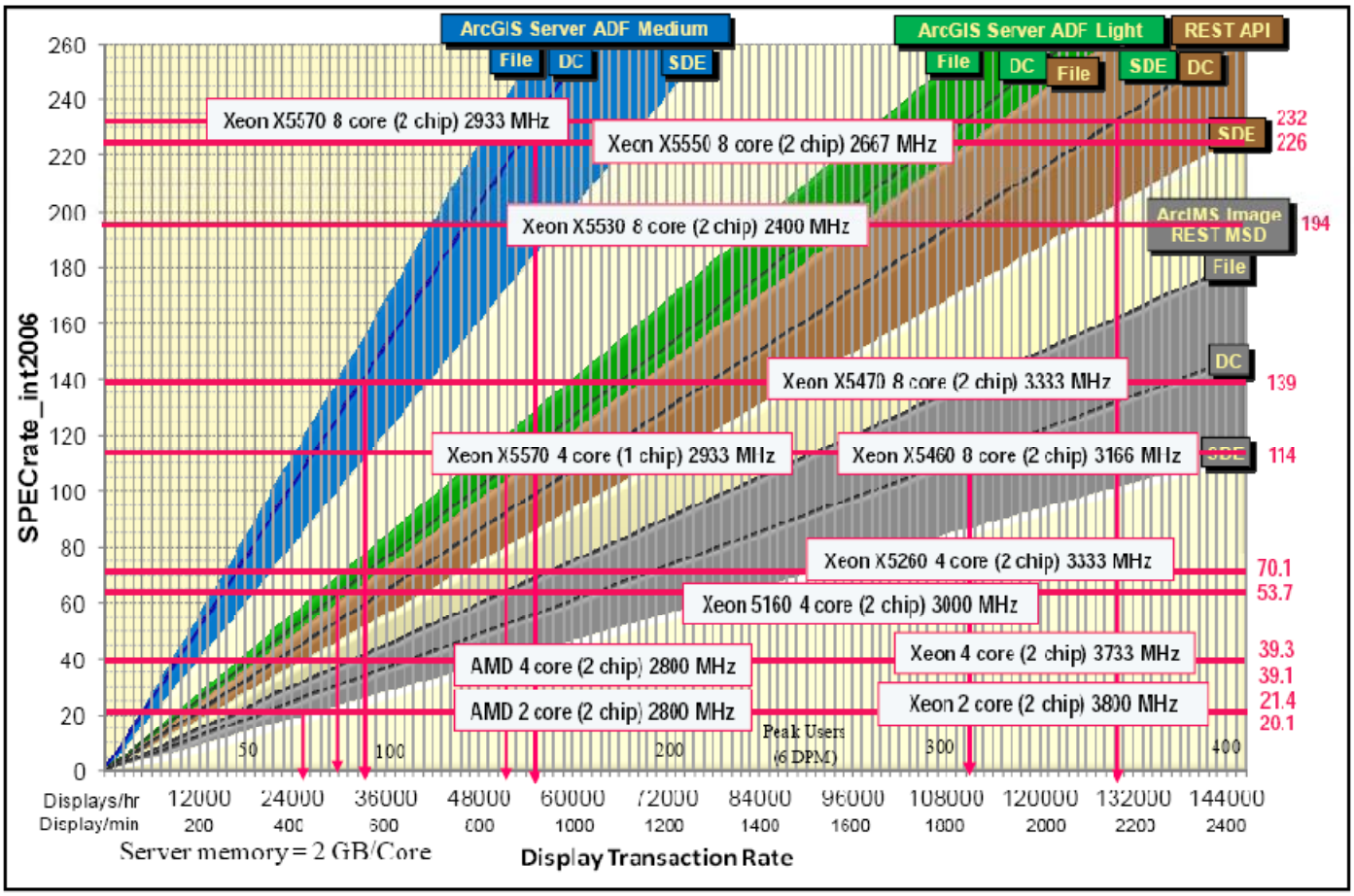
The ArcGIS Server service configurations include performance targets for ArcGIS Server ADF_MXD light and medium dynamic workflows, ArcGIS Server REST MXD and MSD light dynamic services, and ArcIMS Image Service. The ArcGIS Server ADF_MXD light dynamic workflow service times are based on initial performance comparisons between the ArcIMS Image Service and the ArcGIS Server ADF dynamic workflows when generating a simple Image map service. The ArcGIS Server ADF_MXD medium dynamic performance targets represent feedback from customers that are deploying heavier ArcGIS Server map services, where processing loads are twice those required for the simple MXD light dynamic images. Many ArcGIS Server implementations with current technology use the heavier MXD medium dynamic performance targets.

The ArcGIS 9.3 release includes a map service generated with a simple REST API. The ArcGIS Server REST_MXD services do not use the ArcGIS Server Map Editor or Viewer ADF components, and Web processing load is significantly reduced. The ArcGIS Server 9.3.1 REST_MSD light dynamic service uses the new 9.3.1 optimized map document (MSD), using a new graphics rendering engine on the SOC machine. The new graphics rendering engine generates the MSD in roughly the same amount of processing at the legacy

ArcIMS image service. The ArcGIS Server REST_MXD and MSD services are included on the performance sizing charts. The ArcIMS Image Server represents our legacy Web mapping software, and is included on the platform sizing charts for reference purposes.

Figure 9-32 provides a two-tier capacity planning chart for ArcIMS and ArcGIS Server platform selection. This sizing chart identifies peak display transaction rates that can be supported on selected Web server platforms (displays per hour and displays per minute are both include on the chart). Peak users are also identified above these values based on user productivity of six displays per minute.

Figure 9-32
ArcIMS/ArcGIS Server Sizing—Two-Tier Sizing
Two-Tier Map Server/Container Machine



Entry level Xeon X5570 4 core (1 chip) 2933 MHz server platform can support up to 50,000 ArcGIS Server ADF light map displays per hour, 30,000 ArcGIS Server ADF medium map displays per hour, and over 135,000 ArcIMS image map displays per hour. Entry-level ArcGIS Server ADF medium with the Xeon X5570 4 core (2 chip) 2933 MHz platform supports roughly the same peak map transactions per hour that ArcIMS image supported with the Intel Xeon 4 core (2 chip) 3.7 GHz servers just three years ago, ArcGIS Server ADF light supports twice what ArcIMS could do with the same quality display. The new REST MSD light dynamic API extends ArcGIS Server entry level capacity to over 110,000 map displays per hour. Services using pre-cached data sources can support 5-10 times the capacity of the same dynamic services with improved display response times and higher quality maps.

9.10.2 Web Three-Tier Architecture

Map Server (container machine) Sizing: Figure 9-33 identifies recommended software configuration options for the larger three-tier Web mapping deployments. This configuration option supports the Web server and spatial servers (container machines) on separate platform tiers. This configuration is recommended for implementations requiring a large number of concurrent Web mapping clients. Additional spatial servers (container machines) can be included with the configuration to meet peak system capacity needs.

Figure 9-33
Server Performance and Scalability—Three-Tier Architecture

Larger Three-Tier Implementations

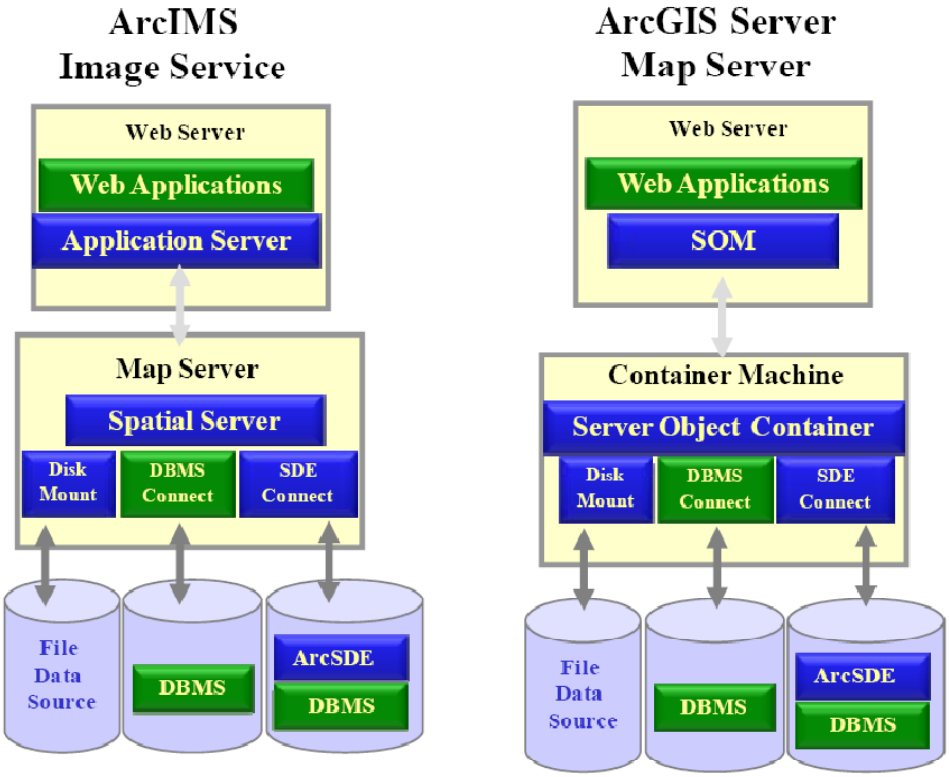
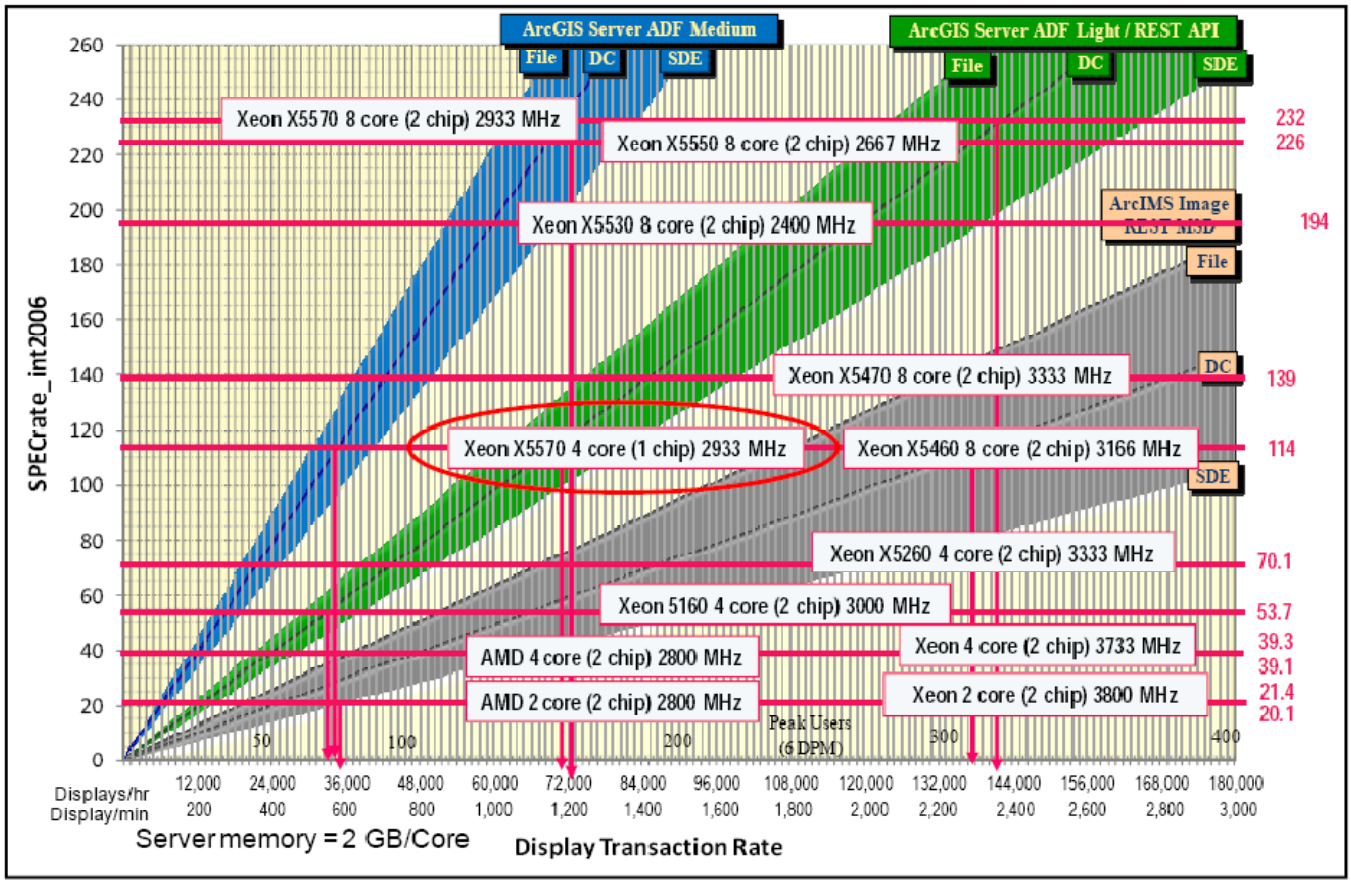


Figure 9-34 provides the three-tier capacity planning chart for ArcIMS and ArcGIS Server deployments. ArcGIS Server ADF light and medium dynamic performance targets are included for planning purposes. The ArcGIS Server 9.3 REST light dynamic API processing targets are the same as ADF light on the container machine. The ArcGIS Server 9.3.1 REST MSD light dynamic processing targets are the same as the legacy ArcIMS processing loads. This sizing chart identifies peak display rates for the map server/container machine platforms. Peak users are also identified based on productivity of six displays per minute.

Figure 9-34
ArcIMS/ArcGIS Server Sizing—Three-Tier Sizing

Three-Tier Map Server/Container Machine

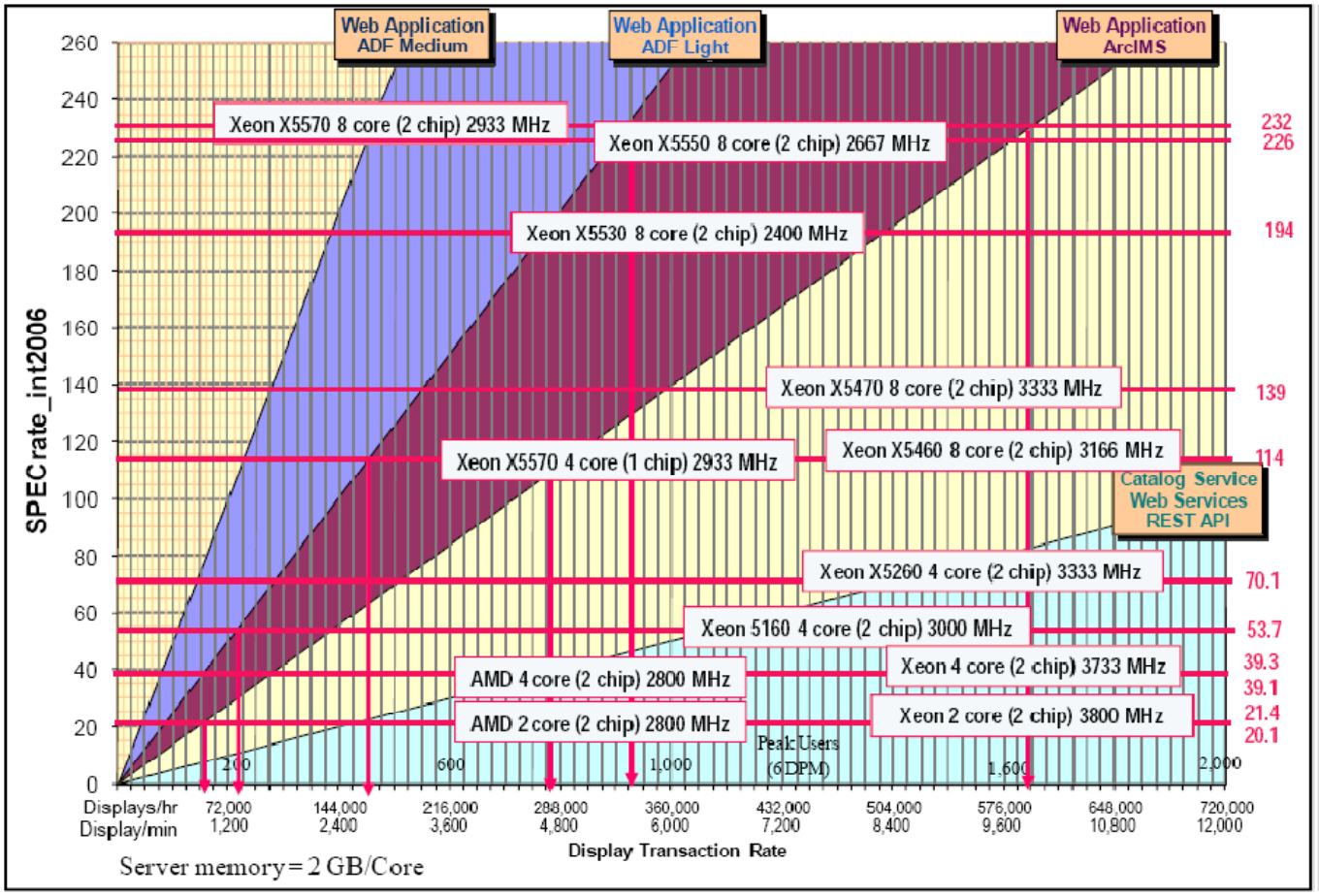


Entry level Intel Xeon 4 core (1 chip) 2933 MHz server platform can support up to 70,000 ArcGIS Server ADF light dynamic map displays per hour, 30,000 ArcGIS Server ADF medium map displays per hour, and over 140,000 ArcGIS Server REST MSD light dynamic or ArcIMS image map displays per hour. Entry-level ArcGIS Server ADF medium dynamic with the Intel Xeon 4 core (1 chip) 2933 MHz platform supports 5,000 more peak map transactions per hour than ArcIMS image supported with the Intel Xeon 2 core (2 chip) 3.8 GHz servers just three years ago. The ArcGIS Server REST API supports over 140,000 map displays per hour with the entry level baseline server platform. Services using pre-cached data sources can support 5-10 times the capacity of the same dynamic services with improved display response times and higher quality maps.

Figure 9-35 provides a three-tier capacity planning chart for ArcGIS Server Web applications (ArcGIS Server ADF medium, ADF light, and REST API) and the standard ArcIMS Web servers. ArcGIS Server ADF web services require roughly the same resources as the REST services. This sizing chart identifies peak display transaction rates that can be supported on the Web server platform. Peak users are also identified based on productivity of six displays per minute.

Figure 9-35
ArcIMS/ArcGIS Server Sizing—Web Server Sizing

Three-Tier Direct Connect



Intel Xeon X5570 4 core (1 chip) 2933 MHz server platform can support up to 150,000 ArcGIS Server ADF light map displays per hour, or over 720,000 map services per hour. Most enterprise solutions would deploy a high available configuration with two Web servers supporting double these numbers. Much higher capacity can be achieved with the larger 8 core (2 chip) platforms. Two Web servers should be adequate to support most large enterprise production environments.

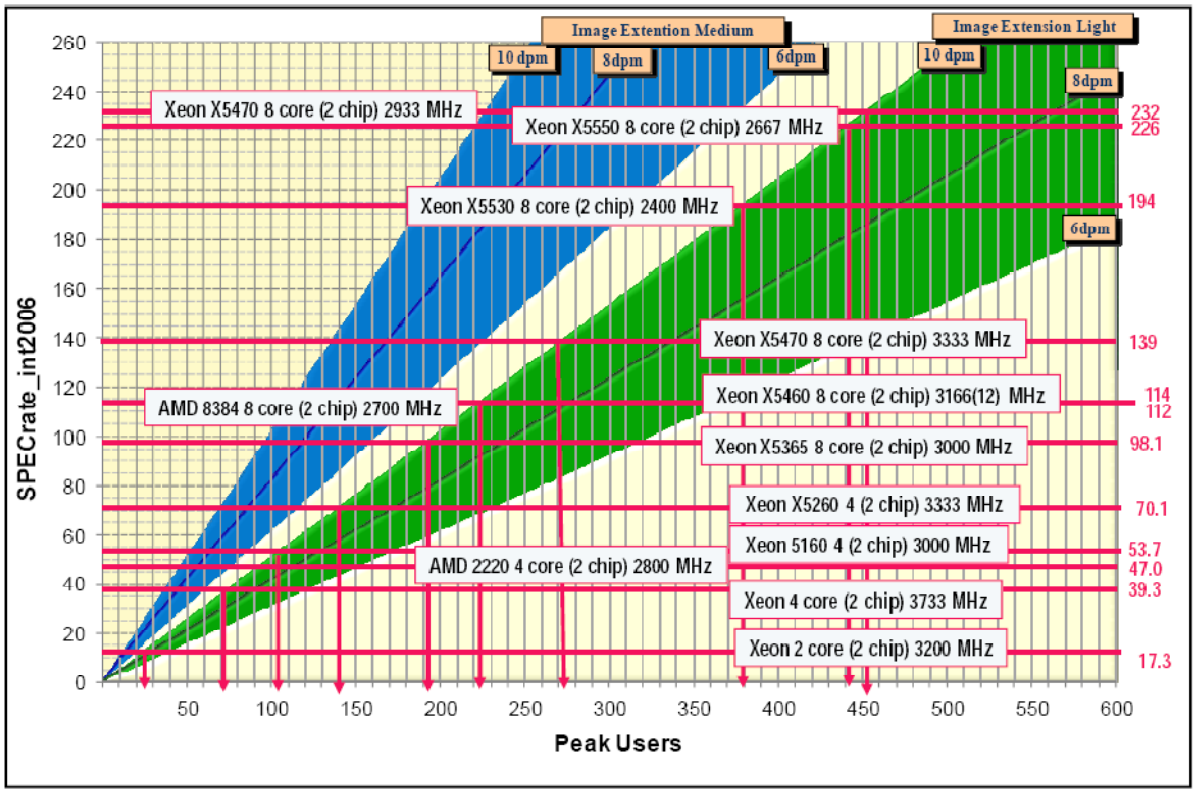
9.11 ArcGIS Server Image Extension Sizing

ArcGIS Server Image Extension provides an alternative data source for image files. The Image Extension uses a standard image file data source (native images stored on a local or network file share), and completes required image processing in response to a specific image service request. The image service providers (service executables) provide a variety of image processing options, with image output available to the full range of ArcGIS Server clients.

With ArcGIS Server 9.3.1, the Image Extension is part of the ArcGIS Server product line. ArcGIS Server 9.3.1 provides an Image Service that can use ArcGIS Server Image Extension as an image data source. The ArcGIS Server Image Extension provides customers with a full range of image processing capabilities delivered to a broad range of clients in a variety of communication formats.

Figure 9-36 provides sizing guidelines that can be used for ArcGIS Server Image Extension platform selection. Two types of service profiles are shown on the chart (Image Extension light and Image Extension medium).

Figure 9-36
ArcGIS Server Image Extension Sizing



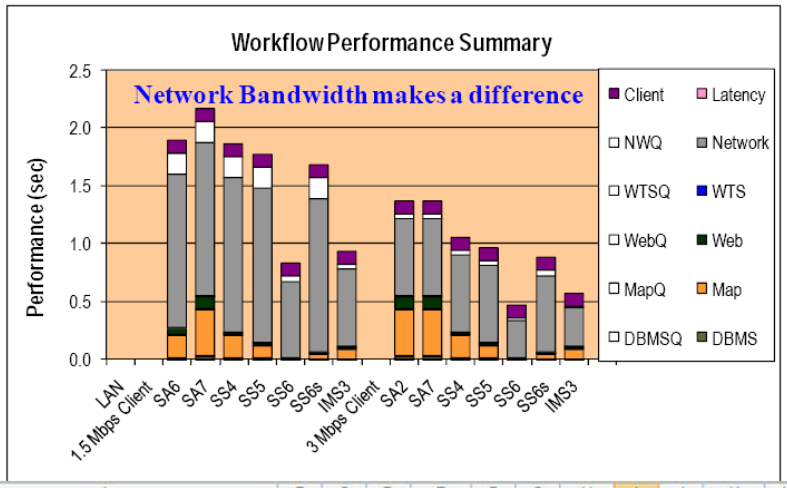
The Image Extension medium performance targets were established from initial customer performance capacity testing. Processing times for a simple image service were doubled to provide a conservative performance target for capacity planning.

The Image Extension light results are extracted from the same test series, and should be used only when publishing a simple image service. These are the service times included with in the current Capacity Planning Tool. These service times do not include any processing required by ArcGIS Server when supporting a composite configuration (ArcGIS Server 9.3.1 Image Service using an Image Extension data source should use the Image Extension medium performance sizing profile).

9.12 ArcGIS Server Standard Workflow Performance

Figure 9-37 provides an overview of the display performance targets currently represented in the Standard ESRI Workflows used in the Capacity Planning Tool.

Figure 9-37
ArcGIS Server Performance Summary
(Standard ESRI Workflows)



	A	B	C	D	E	F	G	H	I	J	K	L	M
2	Workflow	Wkflow Chatter	Software Component Service Times									Min Think Time	Software Service Time
3	Select Category Workflow			Arc09 baseline				SRint06/core = 30.0					
4	Name Workflow as Required		Client	Design Model Metrics					Database				
5	(Copy/Insert customer workflows below)		Traffic	Client		Web		SOC	Data	Traffic	Data		
6	Standard ESRI Workflows =========		Mbpd	Client	WTS	WAS	ADF	SOC	SDE	Mbpd	DBMS	sec	Total
20	SA6_AGS931 ADF_MXD Light Dynamic	10	2.000	0.100		0.053		0.175	0.018	5.000	0.018	3	0.363
21	SA7_AGS931 ADF_MXD Medium Dynamic	10	2.000	0.100		0.105		0.350	0.035	5.000	0.035	3	0.625
27	SS4_AGS93 REST_MXD Light Dynamic Service	10	2.000	0.100		0.018		0.175	0.018	5.000	0.018	3	0.328
28	SS5_AGS931 REST_MSD Light Dynamic Service	10	2.000	0.100		0.018		0.088	0.018	5.000	0.018	3	0.241
29	SS6_AGS93 REST Map Cache Service	10	1.000	0.100		0.005		0.005		2.000		3	0.110
30	SS6s_AGS93 REST 1-5 layer Dynamic Service	10	2.000	0.100		0.005		0.045	0.005	5.000	0.005	3	0.159
42	IMS3_IMS Image Server (Web App)	10	1.000	0.100		0.026		0.088	0.018	5.000	0.018	3	0.250
56													

The 14 workflow combinations above provide a representative subset of the Standard ESRI Workflows for Web Mapping Services included on the Capacity Planning Tool workflow tab. The first chart shows workflows using ArcGIS Server workflows deployed to clients with 1.5 Mbps Web connections, and the second set shows the same workflows deployed to clients with a 3 Mbps Web connection. Network bandwidth is one of the primary factors impacting Web client display performance. Server processing load variations of the different ArcGIS Server deployment patterns have a secondary impact on client display performance and the primary impact on platform sizing and server throughput capacity.

9.13 Platform Selection Criteria

Several factors must be considered in supporting proper hardware selection. These factors include the following:

Platform Performance: Platform must be configured properly to support user performance requirements. Identifying proper platform configurations based on user performance needs and the ESRI design models establishes a solid foundation for proper hardware platform selection.

Purchase Price: Cost of hardware will vary depending on the vendor selection and platform configuration. Pricing should be based on the evaluation of hardware platforms with equal performance capacity.

System Supportability: Customers must evaluate system supportability based on vendor claims and previous experience with supporting vendor technology.

Vendor Relationships: Relationships with the hardware vendor may be an important consideration when supporting complex system deployments.

Total Life Cycle Costs: Total cost of the system may depend on many factors including existing customer administration of similar hardware environments, hardware reliability, and maintainability. Customers must assess these factors based on previous experience with the vendor technology and evaluation of vendor total cost of ownership claims.

Establishing specific hardware performance targets during hardware source selection significantly improves the quality of the hardware selection process. Proper system architecture design and hardware selection provide a basis for successful system deployment.

10 Capacity Planning Tool Introduction

Chapter 7 (Performance Fundamentals) provided an introduction to capacity planning, and described the terms and relationships that determine system performance and scalability. Chapter 8 (Software Performance) identified the software functions that generate the processing loads that must be supported by the selected hardware infrastructure. Chapter 9 (Platform Performance) identified the hardware technology available to handle the required software processing workload, including some core platform sizing charts that can help you identify the right platform selection. This Chapter will introduce the Capacity Planning Tool, which puts the performance fundamentals, software processing requirements, and platform processing capacity together in an excel workbook that can be used to model your system performance needs.

This Chapter provides an overview of the primary CPT views, and is not intended to provide all the information needed to describe how to configure and use the tool to facilitate your system design needs. The System Architecture Design Strategies training class provides a three day workshop, teaching basic performance fundamentals and training attendees on how to use the Capacity Planning Tool. The Capacity Planning Tool is provided to students that complete the training class, for their continued use in supporting their system architecture design and performance tuning needs. The Capacity Planning Tool is also provided with the ESRI Press release of *Building a GIS, System Architecture Design Strategies for GIS Managers.* This book is available at www.esri.com/esripress/buildingagis. Updates to the Capacity Planning Tool, along with a series of flash videos, are available on the *Building a GIS* Online Resource Center.

10.1 Defining User Workflow Requirements

User needs must be identified before completing the system architecture design. A simple user requirements template can be used to review and collect peak user workflow requirements. GIS user workflows are best defined during a formal user needs assessment, where existing user workflows are reviewed to identify technology enhancements that will improve business operations. Roger Tomlinson provides excellent guidance for GIS planning in his ESRI Press release *Thinking about GIS Third Edition, GIS Planning for Managers.*

Figure 10-1 provides a sample user workflow summary, including user locations and peak user loads for planned ArcGIS Desktop and ArcGIS Server workflows. These user workflow requirements will be used in this Chapter to introduce the Capacity Planning Tool.

Figure 10-1
Sample User Workflow Needs

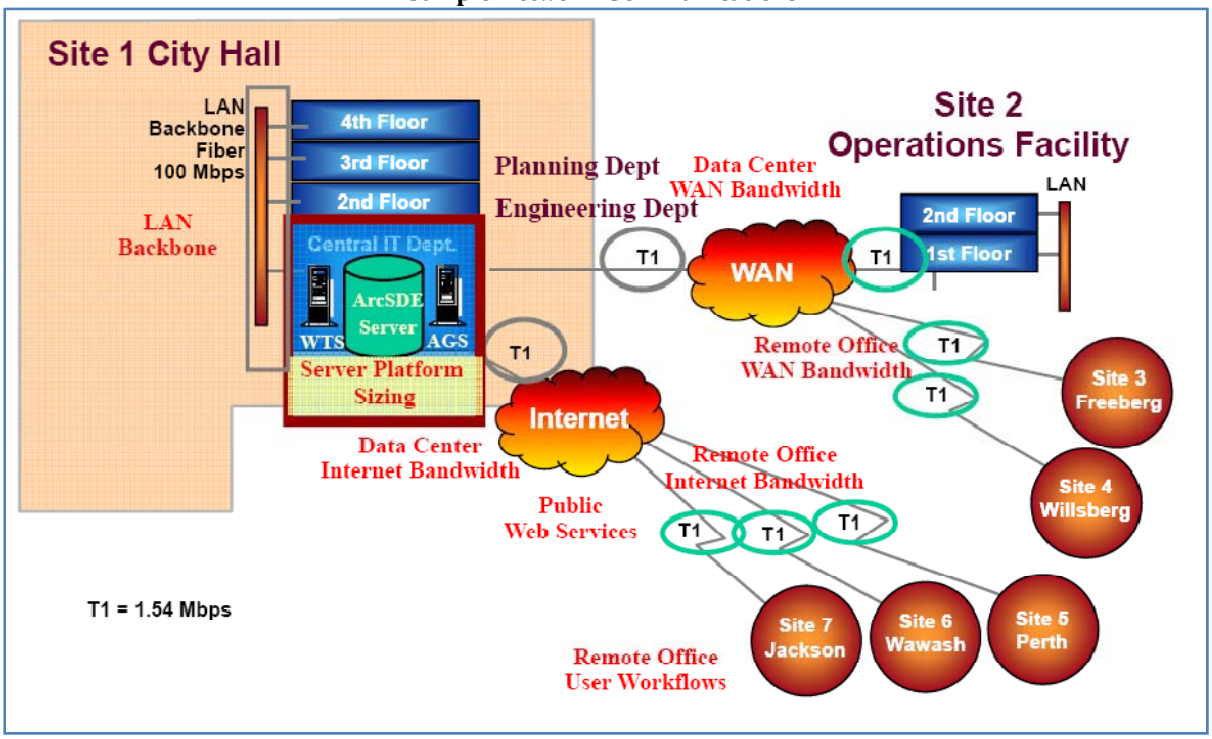
Location	Desktop		Server	
	Editor	Viewer	Batch	Web
LAN				
City Hall	19	44	1	
WAN				
Site 1 - Operations		32		
Site 2 - Freeberg		30		
Site 3 - Willsberg		40		
Internet				
Public				12,000
Site 4 - Perth		2		
Site 5 - Wawash		40		
Site 6 - Jackson		20		

The user workflows above show a peak user load of 19 Desktop Editors (ArcEditor) and 44 ArcGIS Desktop viewers (ArcView) located on the Local Area Network (LAN). Three remote sites are located on the Wide Area Network (WAN) and an additional three remote sites connect through the central data center Internet (ISP) connection. ArcGIS Desktop viewers (ArcView) support the remote site workflows, with 102 WAN users (32 at Operations, 30 at Freeberg, and 40 at Willsberg) and 62 Internet users (2 at Perth, 40 at Wawash, and 20 at Jackson). ArcGIS Server will be used to support public Web services, with estimated peak loads of 12,000 map requests per hour. A batch process is included in the data center to support data maintenance operations during peak operations.

10.2 Identifying user workflow locations and network communications

Figure 10-2 shows a typical IT drawing that provides an overview of the user locations and network communications. The central data center is supported at City Hall, with T1 (1.54 Mbps) bandwidth for the data center WAN and Internet connections. The LAN is supported by a 100 Mbps backbone.

Figure 10-2
Sample Network Communications



Definition of the peak user workflow requirements, user locations, and the associated network communications is information that must be understood in order to complete a system architecture design and select an appropriate platform solution.

10.3 Establishing User Workflow Performance Targets

Once the user requirements are defined and understood, user requirements can be used to configure the Capacity Planning Tool. The first step in the system design process is to define the user workflows and identify appropriate workflow performance targets (software service times). A list of ESRI Standard Workflow software service times are included on the CPT workflow tab to be used as a reference in establishing appropriate workflow performance targets.

Figure 10-3 shows five workflows that were selected from the Standard ESRI Workflows based on the user needs identified in Figure 10-1. The project workflows include the following:

DeskEdit_ArcGIS Desktop Medium Dynamic: Support local ArcEditor workflows installed on user workstations located on the City Hall LAN. Standard ESRI ArcGIS Desktop software service times for medium display complexity were selected for the ArcEditor workflow performance targets.

DeskView_ArcGIS Desktop Medium Dynamic: Support local ArcView workflows installed on user workstations located on the City Hall LAN. Standard ESRI ArcGIS Desktop software service times for medium display complexity were selected for the ArcView workflow performance targets.

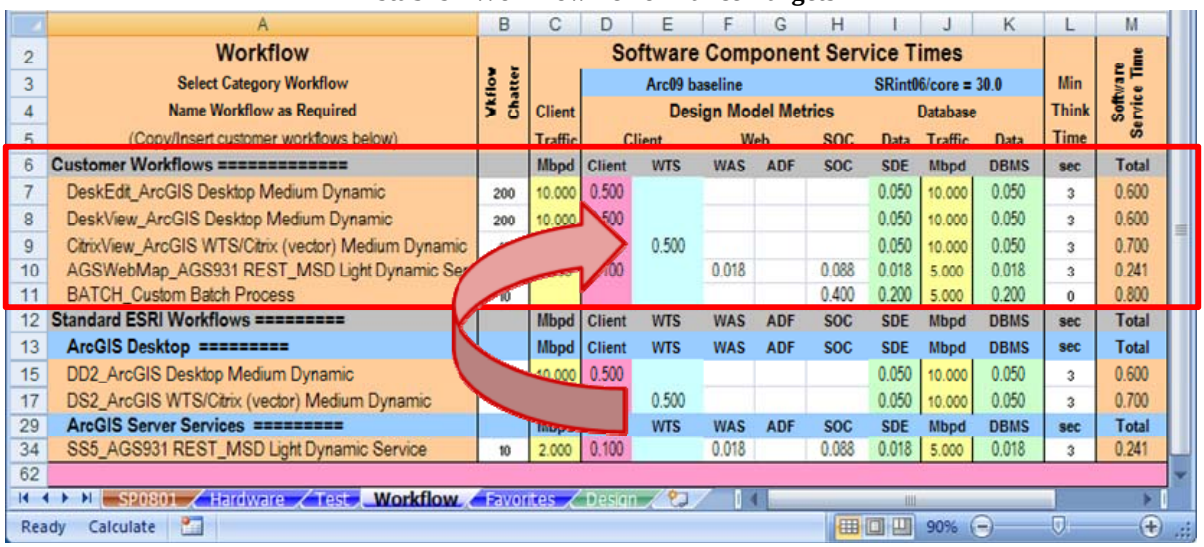
CitrixView_ArcGIS WTS/Citrix (vector) Medium Dynamic: Support remote ArcGIS Desktop (ArcView) clients with applications hosted on Windows Terminal Server farm in data center. Standard ESRI ArcGIS

WTS/Citrix (vector) Medium Dynamic software service times were selected for the remote ArcGIS Desktop viewer workflow performance targets.

AGSWebMap_AGS931 REST_MSD Light Dynamic Service: Support ArcGIS Server Web mapping services. Standard ESRI ArcGIS Server 9.3.1 REST_MSD Medium Dynamic software service times were selected for the Web mapping services performance targets.

BATCH_Custom Batch Process: Data Center batch process reserved for use during peak workflow loads. ArcGIS Server used to host batch process, with target performance service times of 0.4 sec on the SOC machine and 0.2 seconds for the SDE and DBMS query processing. CPT Batch process is specified by establishing a zero (0) minimum think time in column L.

Figure 10-3
Establish Workflow Performance Targets



	A	B	C	D	E	F	G	H	I	J	K	L	M
2	Workflow		Software Component Service Times										
3	Select Category Workflow				Arc09 baseline				SRint06/core = 30.0				
4	Name Workflow as Required				Design Model Metrics				Database				
5	(Copy/Insert customer workflows below)												
6	Customer Workflows ==========		Mbpd	Client	WTS	WAS	ADF	SOC	SDE	Mbpd	DBMS	sec	Total
7	DeskEdit_ArcGIS Desktop Medium Dynamic	200	10.000	0.500					0.050	10.000	0.050	3	0.600
8	DeskView_ArcGIS Desktop Medium Dynamic	200	10.000	0.500					0.050	10.000	0.050	3	0.600
9	CitrixView_ArcGIS WTS/Citrix (vector) Medium Dynamic				0.500				0.050	10.000	0.050	3	0.700
10	AGSWebMap_AGS931 REST_MSD Light Dynamic Se					0.018		0.088	0.018	5.000	0.018	3	0.241
11	BATCH_Custom Batch Process							0.400	0.200	5.000	0.200	0	0.800
12	Standard ESRI Workflows =========		Mbpd	Client	WTS	WAS	ADF	SOC	SDE	Mbpd	DBMS	sec	Total
13	ArcGIS Desktop =========		Mbpd	Client	WTS	WAS	ADF	SOC	SDE	Mbpd	DBMS	sec	Total
15	DD2_ArcGIS Desktop Medium Dynamic		10.000	0.500					0.050	10.000	0.050	3	0.600
17	DS2_ArcGIS WTS/Citrix (vector) Medium Dynamic				0.500				0.050	10.000	0.050	3	0.700
29	ArcGIS Server Services =========				WTS	WAS	ADF	SOC	SDE	Mbpd	DBMS	sec	Total
34	SS5_AGS931 REST_MSD Light Dynamic Service	10	2.000	0.100		0.018		0.088	0.018	5.000	0.018	3	0.241
62													

Workflows are defined on the CPT workflow tab and assigned a unique name (short nickname is provided before the underscore for CPT graphic display purposes). These project level workflows, along with the software service time performance targets, will be used by Excel to complete the User Requirements Analyses.

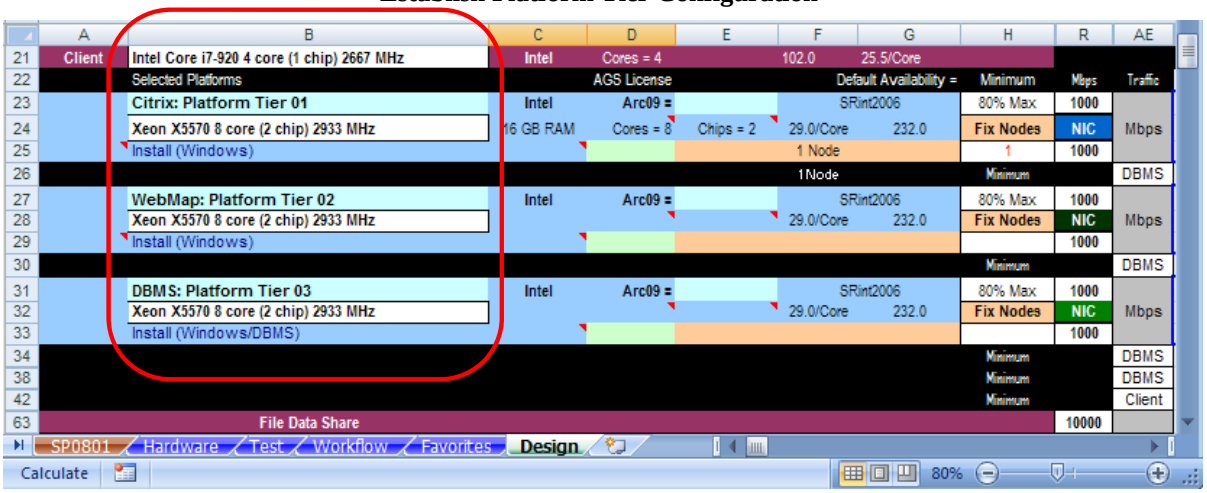
The custom workflows are identified in the Capacity Planning Tool workflow tab by first copying an ESRI Standard Workflow row, and then insert copied cells as a new row for each specific project workflow. Once you establish the project workflows, the software service times can be adjusted to represent your own custom performance targets. The custom workflows are now available for use for configuring the requirements analysis module.

10.4 Data Center Platform Configuration

The Capacity Planning Tool (DPT) provides a variety of platform configuration options. Platform architecture strategies were discussed in Chapter 4. The CPT includes ten separate platform tier, with each tier available to accommodate as many servers as needed to meet peak workflow requirements.

Figure 10-4 shows how we configure the server tier to support our design solution. For the purpose of this design, we will identify three platform tier (Citrix windows terminal server farm, ArcGIS Server map servers, and the SDE Geodatabase). We will also identify the desktop workstation and server platforms that will be provided for our production configuration.

Figure 10-3
Establish Platform Tier Configuration



The first three tier were selected for our configuration. Intel Core i7-920 4 core (1 chip) 2667 MHz workstation was selected to represent our desktop clients. Platform tier nicknames were established for each tier (Citrix: Tier 01, WebMap: Tier 02, and DBMS: Tier 03).

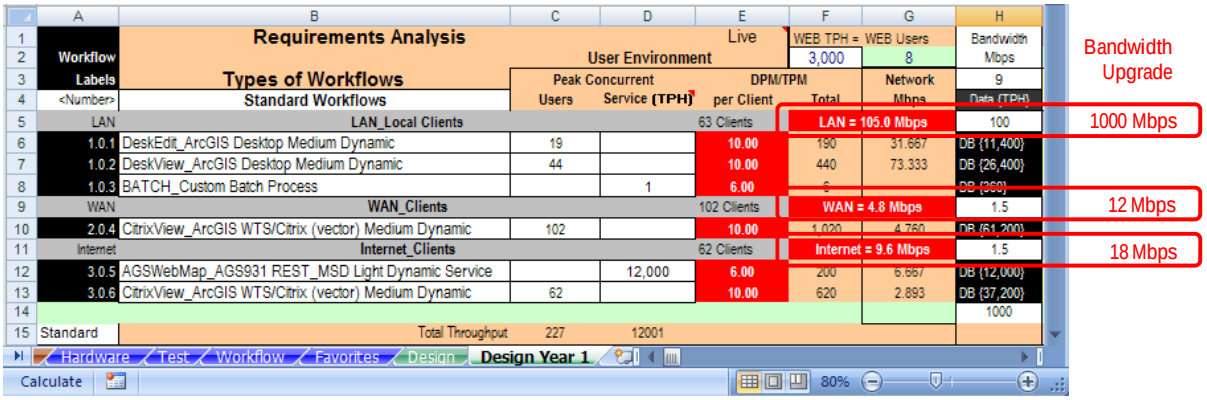
10.5 Data Center User Requirements Analysis

There is more than one way to approach capacity planning. For some organizations, the departments or supported agencies are responsible for their own network connections. The hosting data center may only be interested in completing a design with proper platform selection and adequate data center bandwidth capacity. In other cases, the system architecture design will consider network bandwidth requirements for the data center and the remote site connections.

For illustration purposes, the Capacity Planning Tool will be configured initially to address only the data center platform and network capacity requirements. Once these requirements are complete, we can configure the CPT requirements analysis to include the remote office connections.

Figure 10-5 shows the CPT Design tab Requirements Analysis Module configured to address data center requirements.

Figure 10-5
Data Center Requirements Analysis



The CPT design requirements analysis module is configured to represent the user requirements (user requirements were identified in Figure 10-1). A total of 19 ArcGIS Desktop Editors and 44 ArcGIS Desktop Viewers were located on the LAN along with the BATCH process in the data center, 102 ArcGIS Desktop Viewers hosted on Citrix were located on the WAN, and 62 ArcGIS Desktop Viewers were located on the

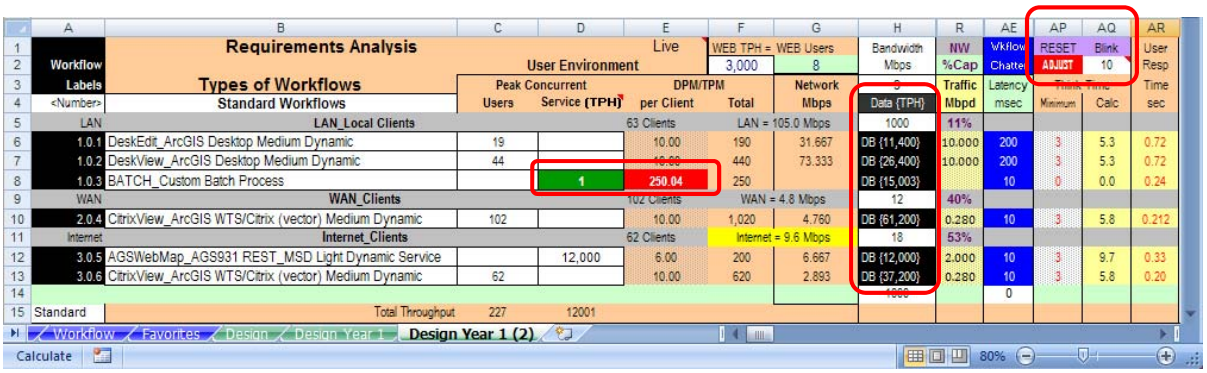
Internet. The Web mapping clients (12,000 transactions per hour) were located on the Internet. Standard Excel row copy and insert copied cells commands are used to configure the requirements analysis module. You can add additional rows or delete rows to configure the appropriate workflows for each network segment. Workflow selection (column B) is a dropdown menu that selects the appropriate target performance values from the project workflow previously defined on the CPT workflow tab.

Peak user workflow requirements and services are entered in columns C and D . Data center network bandwidth connections must be identified on the network rows in column H (LAN - 100 Mbps, WAN - 1.5 Mbps, Internet - 1.5 Mbps). As the user requirements are entered Excel completes the Requirements Analysis, including a network suitability assessment. WAN and Internet traffic blocks will turn RED if the projected network traffic exceeds the existing network bandwidth. The LAN traffic cell will turn YELLOW if that the projected traffic is more than 50 percent of the existing bandwidth. User productivity cells (column E) will turn RED if calculated think time is less than zero (YELLOW if calculated think time is positive but less than minimum think time). RED cells indicate that system adjustments must be made to support the identified user workflow requirements. In this case, the peak network traffic is well above the available network bandwidth, and the data center bandwidth connections will need to be increased to satisfy peak processing loads.

10.6 Data Center Network Performance Tuning

Planned Network bandwidth should be at least twice the projected traffic flow. Network bandwidth can be updated in the CPT to provide the recommended capacity (LAN - 1000 Mbps, WAN - 12 Mbps, Internet - 18 Mbps). Figure 10-6 shows the requirements analysis view following the bandwidth upgrades.

Figure 10-6
Updated Data Center Bandwidth



	A	B	C	D	E	F	G	H	R	AE	AP	AQ	AR
1		Requirements Analysis			Live	WEB TPH =	WEB Users	Bandwidth	NW	Wkflow	RESET	Blink	User
2	Workflow			User Environment		3,000	8	Mbps	%Cap	Chatter	ADJUST	10	Resp
3	Labels	Types of Workflows	Peak Concurrent		DPM/TPM		Network		Traffic	Latency			Time
4	<Number>	Standard Workflows	Users	Service (TPH)	per Client	Total	Mbps	Data (TPH)	Mbpd	msec	Minimum	Calc	sec
5	LAN	LAN_Local Clients			63 Clients	LAN = 105.0 Mbps		1000	11%				
6	1.0.1	DeskEdit_ArcGIS Desktop Medium Dynamic	19		10.00	190	31.667	DB {11,400}	10.000	200	3	5.3	0.72
7	1.0.2	DeskView_ArcGIS Desktop Medium Dynamic	44		10.00	440	73.333	DB {26,400}	10.000	200	3	5.3	0.72
8	1.0.3	BATCH_Custom Batch Process		1	250.04	250		DB {15,003}		10	0	0.0	0.24
9	WAN	WAN_Clients			102 Clients	WAN = 4.8 Mbps		12	40%				
10	2.0.4	CitrixView_ArcGIS WTS/Citrix (vector) Medium Dynamic	102		10.00	1,020	4.760	DB {61,200}	0.280	10	3	5.8	0.212
11	Internet	Internet_Clients			62 Clients	Internet = 9.6 Mbps		18	53%				
12	3.0.5	AGSWebMap_AGS931 REST_MSD Light Dynamic Service		12,000	6.00	200	6.667	DB {12,000}	2.000	10	3	9.7	0.33
13	3.0.6	CitrixView_ArcGIS WTS/Citrix (vector) Medium Dynamic	62		10.00	620	2.893	DB {37,200}	0.280	10	3	5.8	0.20
14										0			
15	Standard	Total Throughput	227	12001									

Excel updates the network suitability analysis as the changes are made, and as a result the identified performance issues are resolved. The CPT includes an automatic adjust function that is used to adjust workflow productivity when calculated think time (column AQ) is less that minimum think time (column AP). The adjust function is also used to compute batch process productivity when minimum think time is zero (0). For this analysis, the adjust function was used to calculate batch process productivity in cell E8. The ADJUST is an iterative Excel calculation, and the calculation will continue until the minimum and calculated think times are the same. When complete, the peak concurrent user or service cell will turn green (cell D8).

The network bandwidth updates were included in column H, and this resolved the network traffic issues.

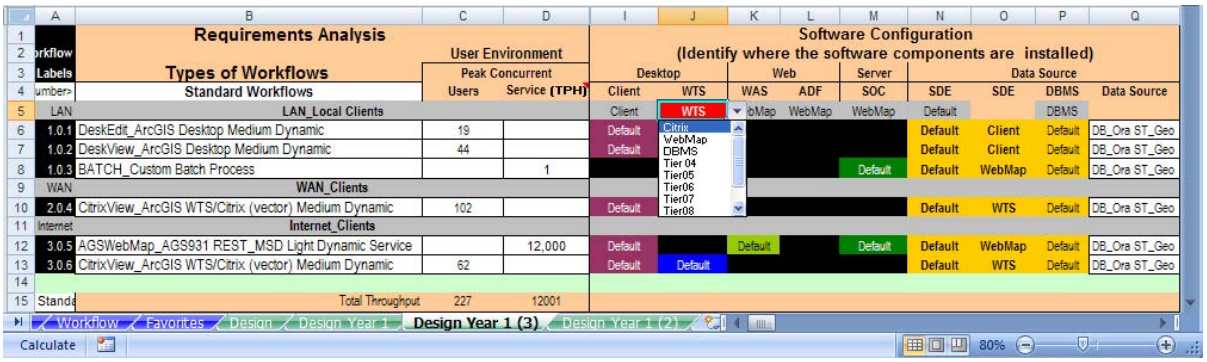
The Workflow Performance Summary shows all workflow response times are now less than a second, verifying that the network upgrades resolved the performance issues.

10.7 Workflow Software Install

Once the user workflows are included in the CPT Requirements Module, the workflow software can be installed on the appropriate platform tier using the Software Configuration module located in columns I through Q. Figure 10-7 shows the workflow software installation module. The default installation tier can be identified in the LAN row, or each workflow can be configured separately. In this example, the WTS software is installed on the Citrix tier, the Web, ADF, and SOC software is installed on the WebMap tier, and the DBMS software is installed on the DBMS tier. The nicknames we identified earlier for each platform tier show up in the

dropdown software assignment list. The Oracle ST_Geometry geodatabase (column Q) is identified data source as the data source for all workflows. The default SDE install (column N) is direct connect, and column O identifies the host platform for the SDE executables.

Figure 10-7
Workflow Software Installation

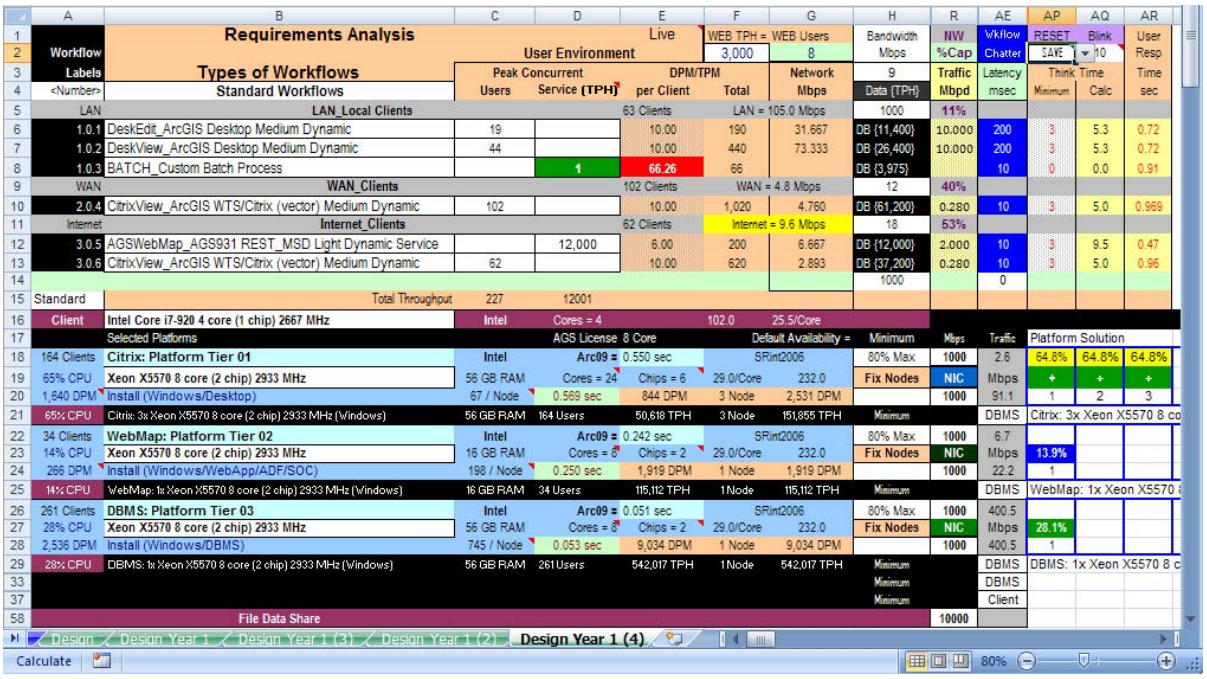


10.8 Data Center Platform Solution

Once user requirements are included in the CPT requirements module and the software is installed on the selected platform tier, the CPT translates user workflow loads to the selected platforms identified in the CPT platform selection module and completes the system design analysis.

Figure 10-8 provides an overview of the design solution. User requirements are identified at the top half of the design tab by user location, identifying the peak user workflow and workflow display response times (column AR).

Figure 10-8
Data Center Platform Selection

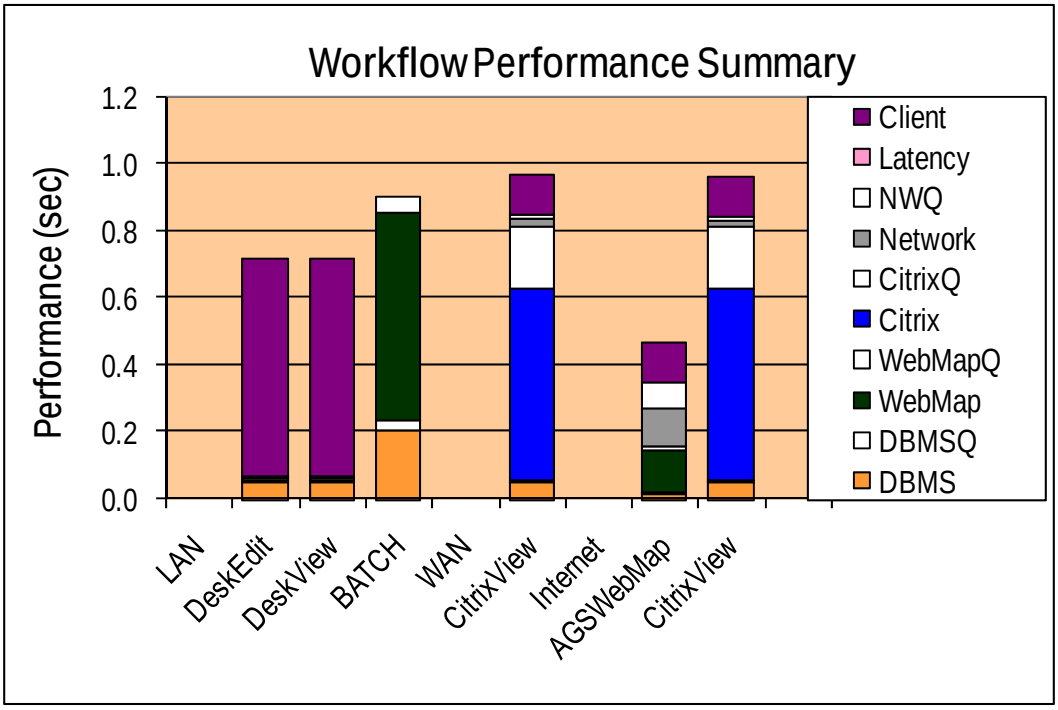


The platform solution is displayed on the bottom half of the design tab by platform tier. Selected platform and platform performance metrics are displayed for each tier, along with a graphic display of the final platform configuration to the right.

10.9 Workflow Performance Summary

Figure 10-7 provides a view of the Workflow Performance Summary chart included with the Requirements Analyses Module. Workflow nicknames are displayed at the bottom of the display for simple recognition.

Figure 10-7
Server Platform Utilization Profile



The Workflow Performance Summary identifies the display processing time for each server platform and network segment along with the associated queue times and over all display response time. Component queue times shown above each platform and network component are based on calculated component utilization metrics. Queue time assumes random arrival time distributions, and the relationship with service time and existing component utilization measurements is discussed in Chapter 7. User display response time is the sum of all component service and queue times included in each workflow; slow display response times can impact user productivity.

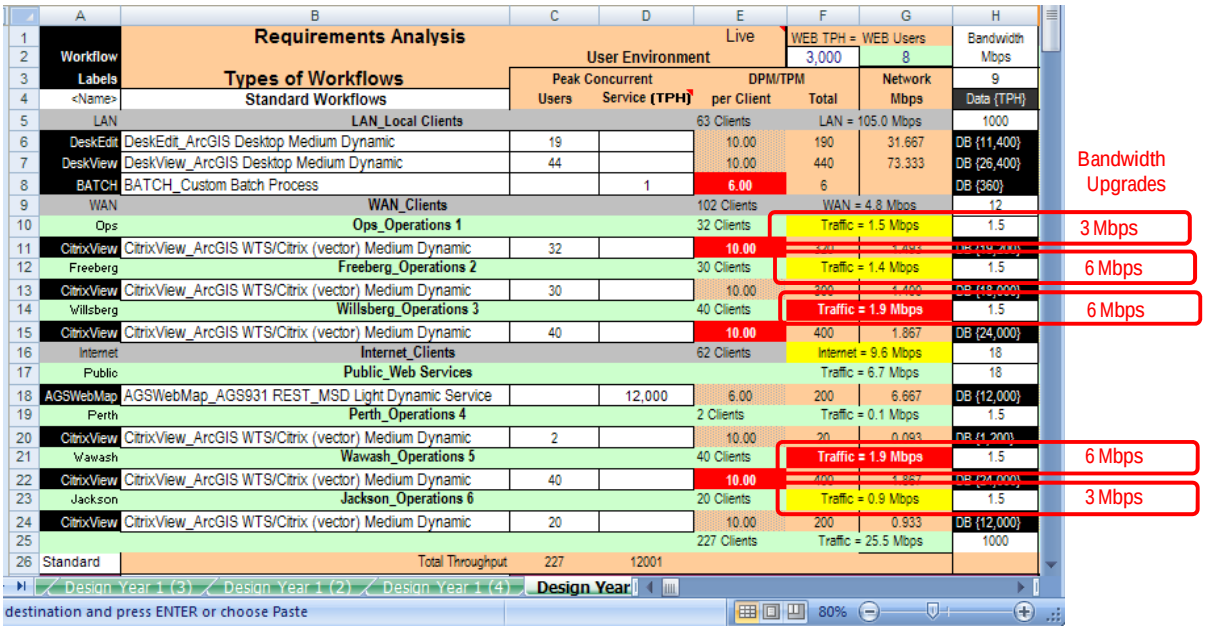
The Workflow Performance Summary is a powerful information product generated by the Capacity Planning Tool. Each of the component platform and network services times are identified in the key, with the associated queue time for each platform and network component identified by a WHITE bar above each. Each workflow is identified by a workflow number or nickname from in the CPT (column A) located left of each workflow in the User Requirements module. For each workflow, you have a visual display of the service times and the overall response time (sum of all service times and queue times). The Workflow Performance Summary is dynamically updated (based on the selected hardware in the hardware selection module) as you complete the user requirements configuration.

10.10 Including Remote Site Locations in the User Requirements Analysis

For most design studies, the remote site locations are also included in the system design analysis. Remote site bandwidth connections can introduce performance bottlenecks and should be considered any time this information is available.

Figure 10-8 shows the same user requirements module configured to include the remote site locations. Remote site locations are included in the design analysis using the GREEN row located at the bottom of the CPT requirements module section.

Figure 10-8
Remote Site Locations included in the Requirements Analysis



	A	B	C	D	E	F	G	H
1		Requirements Analysis			Live	WEB TPH =	WEB Users	Bandwidth
2	Workflow			User Environment		3,000	8	Mbps
3	Labels	Types of Workflows	Peak Concurrent		DPM/TPM		Network	9
4	<Name>	Standard Workflows	Users	Service (TPH)	per Client	Total	Mbps	Data {TPH}
5	LAN	LAN_Local Clients			63 Clients	LAN = 105.0 Mbps		1000
6	DeskEdit	DeskEdit_ArcGIS Desktop Medium Dynamic	19		10.00	190	31.667	DB {11,400}
7	DeskView	DeskView_ArcGIS Desktop Medium Dynamic	44		10.00	440	73.333	DB {26,400}
8	BATCH	BATCH_Custom Batch Process		1	6.00	6		DB {360}
9	WAN	WAN_Clients			102 Clients	WAN = 4.8 Mbps		12
10	Ops	Ops_Operations 1			32 Clients	Traffic = 1.5 Mbps		1.5
11	CitrixView	CitrixView_ArcGIS WTS/Citrix (vector) Medium Dynamic	32		10.00	[illegible]	[illegible]	[illegible]
12	Freeberg	Freeberg_Operations 2			30 Clients	Traffic = 1.4 Mbps		1.5
13	CitrixView	CitrixView_ArcGIS WTS/Citrix (vector) Medium Dynamic	30		10.00	[illegible]	[illegible]	[illegible]
14	Willsberg	Willsberg_Operations 3			40 Clients	Traffic = 1.9 Mbps		1.5
15	CitrixView	CitrixView_ArcGIS WTS/Citrix (vector) Medium Dynamic	40		10.00	400	1.867	DB {24,000}
16	Internet	Internet_Clients			62 Clients	Internet = 9.6 Mbps		18
17	Public	Public_Web Services				Traffic = 6.7 Mbps		18
18	AGSWebMap	AGSWebMap_AGS931 REST_MSD Light Dynamic Service		12,000	6.00	200	6.667	DB {12,000}
19	Perth	Perth_Operations 4			2 Clients	Traffic = 0.1 Mbps		1.5
20	CitrixView	CitrixView_ArcGIS WTS/Citrix (vector) Medium Dynamic	2		10.00	20	0.093	DB {1,200}
21	Wawash	Wawash_Operations 5			40 Clients	Traffic = 1.9 Mbps		1.5
22	CitrixView	CitrixView_ArcGIS WTS/Citrix (vector) Medium Dynamic	40		10.00	400	1.867	DB {24,000}
23	Jackson	Jackson_Operations 6			20 Clients	Traffic = 0.9 Mbps		1.5
24	CitrixView	CitrixView_ArcGIS WTS/Citrix (vector) Medium Dynamic	20		10.00	200	0.933	DB {12,000}
25					227 Clients	Traffic = 25.5 Mbps		1000
26	Standard	Total Throughput	227	12001				

To include a remote site location in the network suitability analysis, you first use Excel to copy the GREEN row shown on line 25 above and then 'insert copied cells' within the data center networks to represent the remote site locations. The display shows the remote WAN sites (Operations, Freeberg, and Willsberg) and the remote Internet sites (Perth, Wawash, and Jackson) inserted within their respective network environments. A Public site is included to represent Web services published for Public Internet use. Once the site locations are inserted into the requirements module, the site name can be identified in column B. Workflows can be inserted for each site location by copying any existing workflow row, and using the insert copied cells function to insert the workflow rows that represent clients at each site location. Workflow location is determined by the client location and must be located on the proper network segments.

Once the workflow rows are in place, the proper workflow can be selected in column B and the peak users for each workflow can be typed in column C. Each remote site network bandwidth must be identified in the WHITE cells in column I.

There are two additional Summation functions that you will need to update for each new site. Site clients (formula in column E on each GREEN row) and site traffic (formula in column F for each GREEN row). Range for each of these summation formulas must be extended to include all workflows within the respective remote site (i.e. traffic range for site 1 - Operations located in G10 must be extended to include H10:H12). Once each of the network traffic ranges is extended to include workflows within the respective remote site, Excel will complete the network suitability analysis.

The results above shows that several of the remote site traffic flows are more than 50 percent of the available network bandwidth, and the network bandwidth should be increased to provide sufficient capacity to support the peak traffic. The CPT adjust function can be used to reduce the workflow productivity to identify the impact if

the remote site network bandwidth recommendations were not implemented.

Figure 10-9 shows the results of the user productivity adjustments if the remote site network bandwidth upgrades were not made.

Figure 10-9
Adjusted User Productivity (no remote site network bandwidth upgrades)

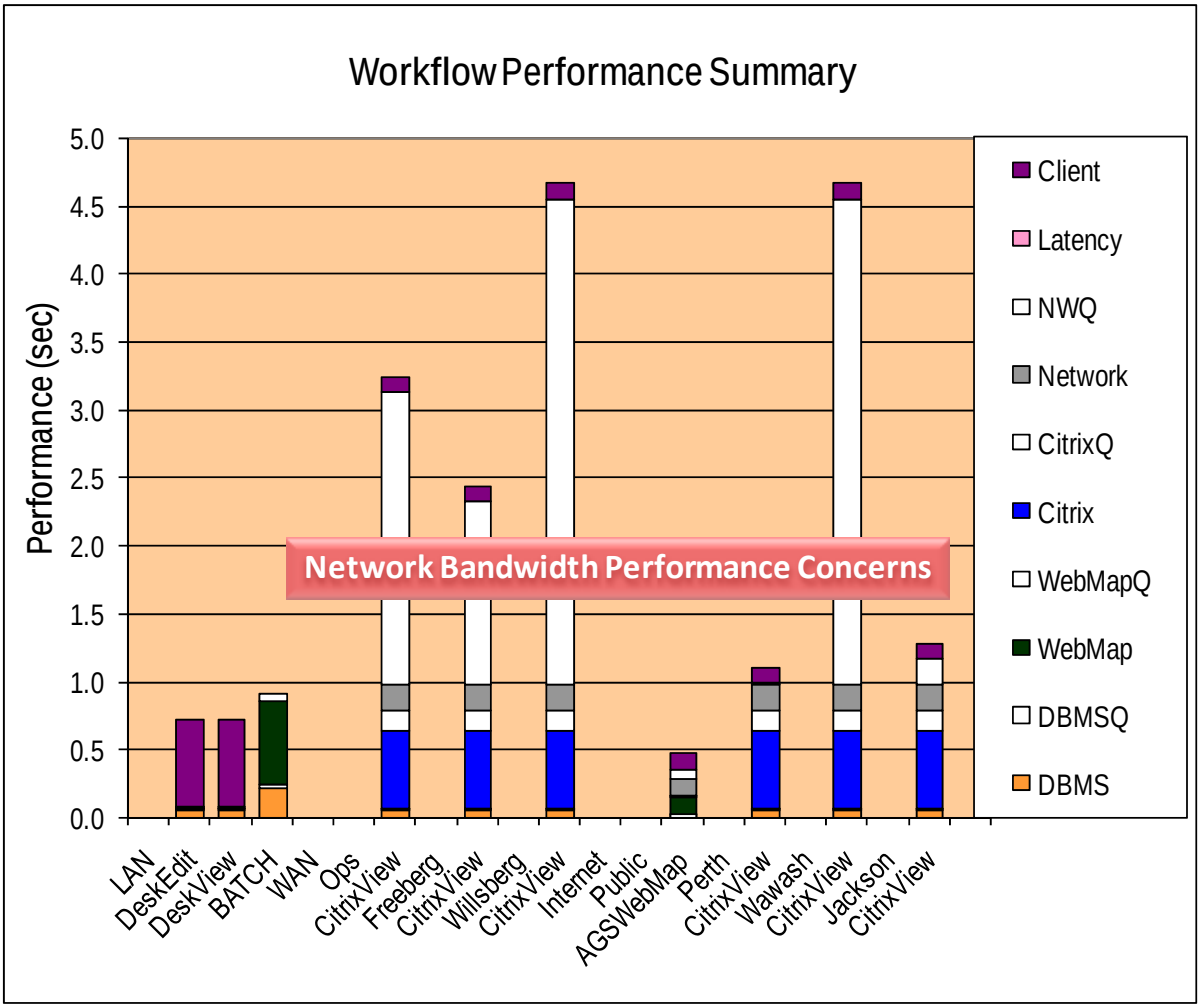
	A	B	C	D	E	F	G	H	R	AE	AP	AQ	AR
1		Requirements Analysis			Live	WEB TPH =	WEB Users	Bandwidth	NW	Wkflow	RESET	Blink	User
2	Workflow			User Environment		3,000	8	Mbps	%Cap	Chatter	ADJUST	1	Resp
3	Labels	Types of Workflows	Peak Concurrent		DPM/TPM		Network	9	Traffic	Latency	Think Time		Time
4	<Name>	Standard Workflows	Users	Service (TPH)	per Client	Total	Mbps	Data {TPH}	Mbpd	msec	Minimum	Calc	sec
5	LAN	LAN_Local Clients			63 Clients	LAN = 105.0 Mbps		1000	11%				
6	DeskEdit	DeskEdit_ArcGIS Desktop Medium Dynamic	19		10.00	190	31.667	DB {11,400}	10.000	200	3	5.3	0.72
7	DeskView	DeskView_ArcGIS Desktop Medium Dynamic	44		10.00	440	73.333	DB {26,400}	10.000	200	3	5.3	0.72
8	BATCH	BATCH_Custom Batch Process		1	66.36	66		DB {3,981}		10	0	0.0	0.90
9	WAN	WAN_Clients			102 Clients	WAN = 4.3 Mbps		12	36%				
10	Ops	Ops_Operations 1			32 Clients	Traffic = 1.4 Mbps		1.5	96%	0			
11	CitrixView	CitrixView_ArcGIS WTS/Citrix (vector) Medium Dynamic	32		9.62	308	1.436	DB {18,468}	0.280	10	3	3.0	3.238
12	Freeberg	Freeberg_Operations 2			30 Clients	Traffic = 1.4 Mbps		1.5	93%	0			
13	CitrixView	CitrixView_ArcGIS WTS/Citrix (vector) Medium Dynamic	30		10.00	300	1.400	DB {18,000}	0.280	10	3	3.6	2.436
14	Willsberg	Willsberg_Operations 3			40 Clients	Traffic = 1.5 Mbps		1.5	97%	0			
15	CitrixView	CitrixView_ArcGIS WTS/Citrix (vector) Medium Dynamic	40		7.83	313	1.461	DB {18,789}	0.280	10	3	3.0	4.664
16	Internet	Internet_Clients			62 Clients	Internet = 9.2 Mbps		18	51%				
17	Public	Public_Web Services				Traffic = 6.7 Mbps		18	37%	0			
18	AGSWebMap	AGSWebMap_AGS931 REST_MSD Light Dynamic Service		12,000	6.00	200	6.667	DB {12,000}	2.000	10	3	9.5	0.46
19	Perth	Perth_Operations 4			2 Clients	Traffic = 0.1 Mbps		1.5	6%	0			
20	CitrixView	CitrixView_ArcGIS WTS/Citrix (vector) Medium Dynamic	2		10.00	20	0.093	DB {1,200}	0.280	10	3	4.9	1.10
21	Wawash	Wawash_Operations 5			40 Clients	Traffic = 1.5 Mbps		1.5	97%	0			
22	CitrixView	CitrixView_ArcGIS WTS/Citrix (vector) Medium Dynamic	40		7.83	313	1.461	DB {18,789}	0.280	10	3	3.0	4.66
23	Jackson	Jackson_Operations 6			20 Clients	Traffic = 0.9 Mbps		1.5	62%	0			
24	CitrixView	CitrixView_ArcGIS WTS/Citrix (vector) Medium Dynamic	20		10.00	200	0.933	DB {12,000}	0.280	10	3	4.7	1.27
25					227 Clients	Traffic = 25.5 Mbps		1000	3%	0			
26	Standard	Total Throughput	227	12001									
27	Client	Intel Core i7-920 4 core (1 chip) 2667 MHz	Intel	Cores = 4		102.0	25.5/Core						
28		Selected Platforms		AGS License	8 Core		Default Availability =	Minimum	Mbps	Traffic	Platform Solution		
29	164 Clients	Citrix: Platform Tier 01	Intel	Arc09 =	0.550 sec		SRint2006	80% Max	1000	2.3	57.5%	57.5%	57.5%
30	57% CPU	Xeon X5570 8 core (2 chip) 2933 MHz	56 GB RAM	Cores = 24	Chips = 6	29.0/Core	232.0	Fix Nodes	NIC	Mbps	+	+	+
31	1,454 DPM	Install (Windows/Desktop)	76 / Node	0.569 sec	844 DPM	3 Node	2,531 DPM		1000	80.8	1	2	3
32	57% CPU	Citrix: 3x Xeon X5570 8 core (2 chip) 2933 MHz (Windows)	56 GB RAM	164 Users	50,618 TPH	3 Node	151,855 TPH	Minimum		DBMS	Citrix: 3x Xeon X5570 8 c		
33	34 Clients	WebMap: Platform Tier 02	Intel	Arc09 =	0.242 sec		SRint2006	80% Max	1000	6.7			
34	14% CPU	Xeon X5570 8 core (2 chip) 2933 MHz	16 GB RAM	Cores = 8	Chips = 2	29.0/Core	232.0	Fix Nodes	NIC	Mbps	13.9%		
35	266 DPM	Install (Windows/WebApp/ADF/SOC)	198 / Node	0.250 sec	1,917 DPM	1 Node	1,917 DPM		1000	22.2	1		
36	14% CPU	WebMap: 1x Xeon X5570 8 core (2 chip) 2933 MHz (Windows)	16 GB RAM	34 Users	115,048 TPH	1 Node	115,048 TPH	Minimum		DBMS	WebMap: 1x Xeon X5570		
37	261 Clients	DBMS: Platform Tier 03	Intel	Arc09 =	0.051 sec		SRint2006	80% Max	1000	369.5			
38	26% CPU	Xeon X5570 8 core (2 chip) 2933 MHz	56 GB RAM	Cores = 8	Chips = 2	29.0/Core	232.0	Fix Nodes	NIC	Mbps	26.1%		
39	2,350 DPM	Install (Windows/DBMS)	802 / Node	0.053 sec	9,014 DPM	1 Node	9,014 DPM		1000	369.5	1		
40	26% CPU	DBMS: 1x Xeon X5570 8 core (2 chip) 2933 MHz (Windows)	56 GB RAM	261 Users	540,815 TPH	1 Node	540,815 TPH	Minimum		DBMS	DBMS: 1x Xeon X5570 8 c		
44								Minimum		DBMS			
48								Minimum		Client			
69		File Data Share							10000				

Several of the remote user site user workflows are performing below the default productivity levels of 10 displays per minute for ArcGIS Desktop clients (green boxes in column C identify the adjusted workflows). Column AQ shows the display response time for these workflows. Display response time for the Operations site is over 3 seconds, and over 4.5 seconds for Willsberg and Wawash clients.

10.11 Workflow Performance Summary

Figure 10-10 shows the workflow performance summary once the remote sites are included in the requirements analysis module and before the remote site network bandwidth adjustments are made. The workflow performance summary is an excellent tool for addressing user productivity concerns. Each user workflow is identified on the chart based on the labels located in column A left of the workflows. The workflow label numbers are automatically generated based on network location, remote site location, and workflow number or you can select to have the workflow nicknames displayed (nicknames are displayed in the chart below). Option for displaying workflow nicknames makes the Workflow Performance Chart easier to read.

Figure 10-10
Workflow Performance Summary

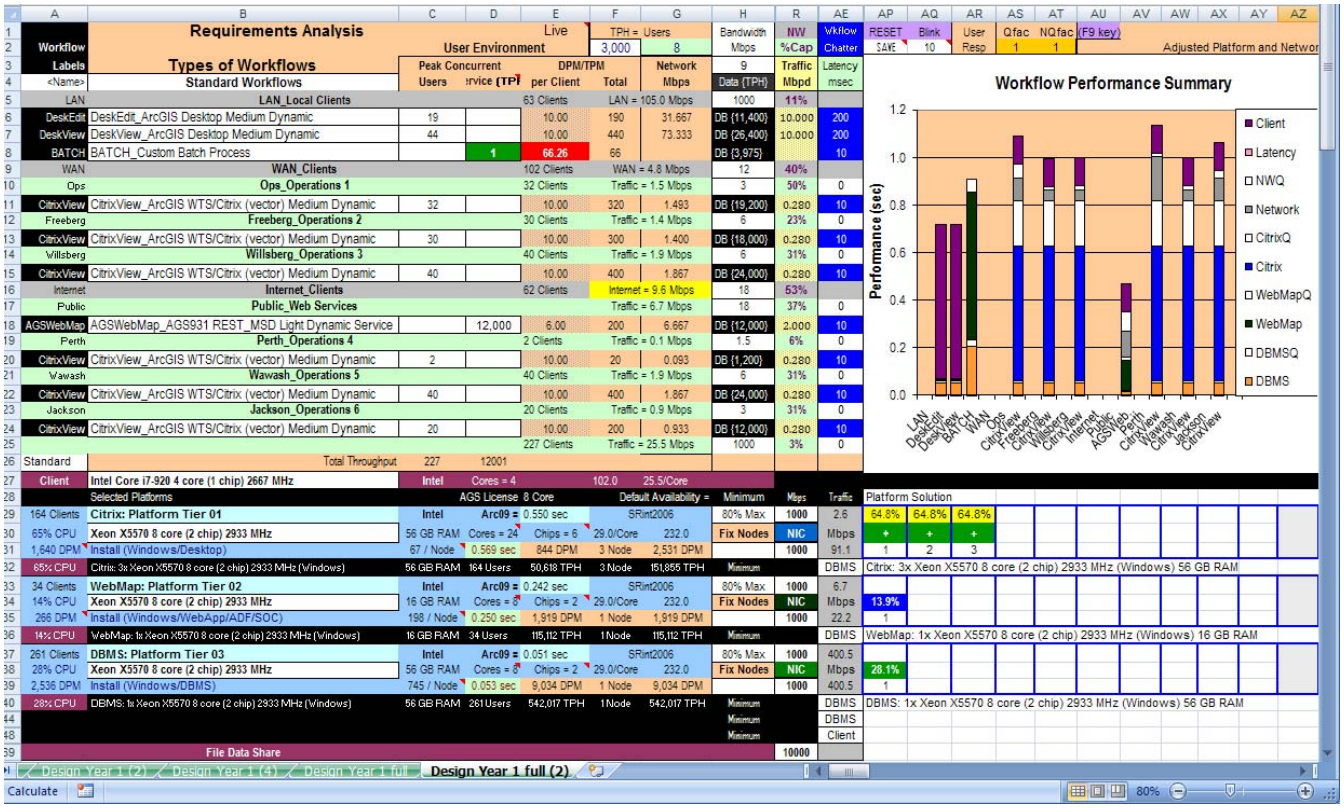


The performance concerns identified by the network traffic queue times are consistent with the YELLOW and RED traffic performance flags identified in the requirements analysis module (Figure 10-9). Adjusting the remote office network bandwidth to accommodate the projected traffic flow will resolve the performance issues.

10.12 Capacity Planning Display Overview

Figure 10-11 shows the final results once the recommended remote site bandwidth updates have been made and the design is adjusted to calculate final batch process productivity. Workflow display response times are now less than 1.2 seconds for all workflows, both in the local network environment and at the remote locations. User productivity for all desktop workflows are the standard 10 displays per minute, and 6 displays per minute for the public Web services. The ArcGIS Server 9.3.1 REST_MSD Light Dynamic services display in less than 0.5 seconds for client locations with minimum Internet access of 18 Mbps (lower bandwidth connection performance would be more due to increased network transport times).

Figure 10-11
Capacity Planning Tool Summary



10.13 How to get the Capacity Planning Tool

The Capacity Planning Tool is provided to students who complete the System Architecture Design Strategies training class or customers who receive ESRI System Architecture Design consulting services. The Capacity Planning tool is also included on a CD with the ESRI Press *Building a GIS* publication. Updates to the Capacity Planning Tool are available for download on the *Building a GIS* Online Resource Center.

This chapter shared a high level overview of the Capacity Planning Tool. This is not a complete review, and the Capacity Planning Tool alone will not address your system performance problems. Understanding the technology and the ways that you can leverage the technology to support your workflow needs is the road to success.

Chapters 1 through 6 shared an overview of GIS technology; Chapter 7 through 9 shared the performance fundamentals. The next Chapter will use a case study to demonstrate how we can apply what we have learned to address the System Architecture Design needs of the City of Rome.

11 Completing the System Design

System architecture design provides a methodology for establishing hardware and network requirements that support the performance and communication needs of GIS application users. Hardware requirements should be established based on identified business needs. A fundamental understanding of user workflow requirements and the supporting GIS technology is required before one can identify the appropriate hardware and network requirements for supporting effective enterprise GIS operations.

City of Rome is the name of the case study provided to demonstrate the planning process presented in Roger Tomlinson's book *Thinking about GIS.* Both his book's chapter 10 and this chapter show standard templates that can be used for most enterprise design studies. In this book we will use the Capacity Planning Tool to model the user requirements for three planned years of expansion and growth.

An application needs assessment should be completed by the business units that benefit from the GIS information products. The needs assessment should be led by in-house GIS staff with support from an executive sponsor. You may have used professional GIS consultants to facilitate the planning process, but all the decision making will come from your end—from the organization's managers and decision makers. Planning is critical in justifying the required GIS investments, in providing a framework for enterprise GIS implementation, and in ensuring upper management support throughout the process.

Most GIS deployments evolve over several years of incremental technology improvements, and implementation plans normally address a two- or three-year schedule, to ensure that the budget is in place for the anticipated deployment needs. So in this case study, we will use the CPT to prepare for year 1, year 2, and year 3.

11.1 GIS user needs assessment

There are a few basic user requirements that must be understood before launching into the system design process. In order to design an effective system for GIS, you must come to grips with the three basic factors that are the focus of the system architecture needs assessment. These requirements compose the system architecture needs assessment: identifying where the GIS users are located in relation to the associated data resources (site locations); what network communications are available to connect user sites with the GIS data sources; and what are the peak user workflow requirements for each user location. Figure 11-1 provides an overview of the system architecture needs assessment.

Figure 11-1
System Architecture Needs Assessment

- **GIS User Locations**
  - User Departments
  - Site Locations
- **Network Communications**
  - Local Area Network Standards
  - Wide Area Network Bandwidth
- **GIS User Types**
  - **ArcGIS** (ArcInfo, ArcEditor, ArcView)
  - **Web Services** (Web information products, project reporting)
  - **Concurrent Batch Processes** (Examples: Reconcile/Post, on-line compress, data loading, on-line backup, replication, and other heavy geoprocessing during peak production workflow)

Capacity Planning Tool
- Opportunity for more granular workflow models.
- Opportunity to validate workflow metrics during deployment.

Caution: Keep planning workflows simple!

GIS user locations. All user locations requiring access to GIS applications and data resources must be identified. You want to include everyone who might need access to the system during peak work periods. The term "user locations" encompasses local users, remote users on the Wide Area Network (WAN), and Internet users (Internal and Public).

The enterprise infrastructure must be able to support the peak user workflows. Knowing where users are located, understanding what applications they will need to do their work, and identifying the location of the required data resources provide the basis for system design analysis.

Network Communications. In the system design assessment, you must identify the network communication bandwidth between the different user locations and the data center. Network bandwidth may establish communication constraints to consider when developing the software technology solution. The final technology solution may require upgrades to the network communication infrastructure.

User Workflows. The peak system workflow loads are the peak users and workflow platform service times discussed in Chapter 10. These will determine computer processing and network capacity requirements. Traditional capacity planning was done with simple workflow models, using separate models for GIS desktop workflows, Web Services, and batch processing. Platform sizing models were established to help select the right server platform technology. General network design guidelines were used to address communication suitability. These simple platform sizing models and network design guidelines demonstrated the value of a proper design in ensuring successful enterprise GIS implementations.

The capacity planning tool introduced in CY2006 (chapter 10) now provides an opportunity for more granular workflow models. The CPT incorporates all of the platform sizing models we have developed over the past 16 years; only what we have added to it is new. The tool's ability to dynamically model the design alternatives provides an adaptive, integrated, management view of the complete technical solution.

There is a tradeoff between simplicity and complexity in building a system performance model. Simplicity is easier to understand, simpler to quantify, enables broader validation, helps quantify business risk, and provides valuable information on which to base business decisions. Complexity may be more accurate, may provide a closer representation of the final implementation, and may lead to more detailed results. Yet complex models can be much more difficult to understand, harder to quantify, more difficult to validate, and may include hidden risk. During the planning phase, it is best to provide a simple model to represent the technical solution and lead to the right business decisions. A simple model is best: it highlights the relationship between what the organization wants to do with GIS (what you want out of the system) and the technology you need to do it (software and hardware procurement decisions).

Planning should establish performance targets that you can manage throughout the implementation phase. The capacity planning tool models used during planning are based on what other people are able to do with the technology. You can use the CPT models to establish your own system performance targets, based on your specific infrastructure limitations and operational needs. This chapter will show you how to build your own solution—one that can that can achieve your performance goals--according to the guidelines described in chapter 8. Afterwards, you must monitor and validate that these goals are met throughout the GIS Implementation (see Chapter 12).

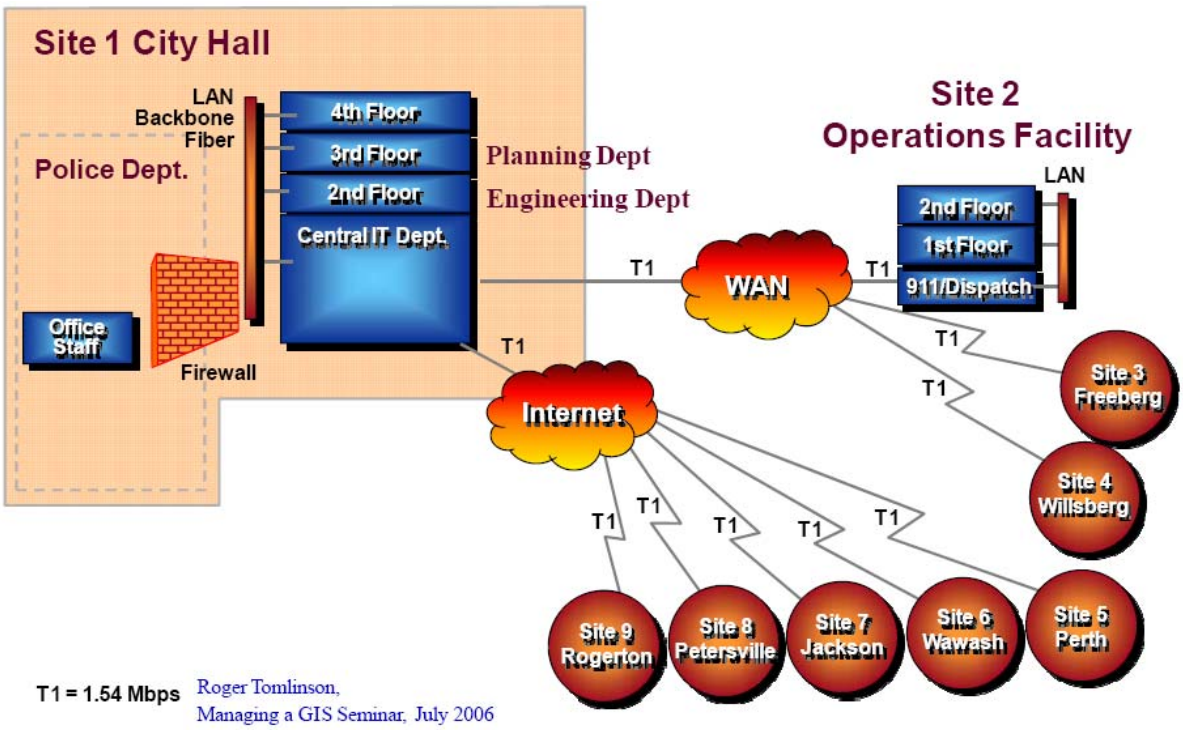
11.2 City of Rome user requirements analysis

The fictional City of Rome represents a typical organization, just right as a case study to demonstrate how you can use the capacity planning tool in your system design process. In planning a GIS for this city, we're going to look ahead to year 2 and year 3 even as we plan for the year 1 implementation.

Let's begin by taking stock of the city government's current situation and how exactly the organization and its employees are looking forward to using GIS. This city has more than 580 employees who require GIS information to help in their normal work processes (information products). These employees are located in the Planning, Engineering, Police, and Operations departments throughout the city. The public will also benefit from deployment of standard GIS information products through published Web applications (services). Each department provides a set of these Web services, which it shares with the public on the city's Web site.

Figure 11-2 provides a sample format for identifying user locations throughout the operations environment. This is an overview of the facility locations and network communications to be addressed in the system design study. The point is to show how each user location is connected to the data center (LAN, WAN, Internet) along with the available network capacity (T1 here, a 1.54 Mbps bandwidth).

Figure 11-2
User Locations and Network Communications



11.2.1 GIS user types (GIS use cases)

The types of GIS users can be divided into three basic categories. The ArcGIS Desktop user type will require desktop applications for GIS processing. Web users will be supported by ArcGIS Server Web applications. An additional batch process to support replication services and standard administrative batch processing (reconcile and post, online backups, etc.) is the third user type.

ArcGIS Desktop: This category includes ArcGIS Desktop specialists doing general spatial query and analysis studies, simple map production, and general-purpose query and analysis operations, including all ArcEditor and ArcView clients. GIS applications for custom business solutions and any custom ArcGIS Engine clients that support specific business needs should also be included in this category.

For this study, separate workflows are identified for ArcGIS Desktop Editors and Viewers. A separate workflow is also included for ArcGIS Desktop Business Analyst. The initial Desktop workflows use the published Standard ESRI Workflow models. The Editors and Business Analyst workflows will use the medium dynamic performance targets, and the viewers will use the light dynamic performance targets. These workflows can be modified as required during system design and implementation as capacity planning performance targets are validated.

Web services: ESRI ArcGIS Server software provides transaction-based map products to Internet browser clients and supports synchronization with remote mobile clients. The City of Rome will use the ArcGIS Server REST API with dynamic vector operational layer mashup with a pre-processed map cache base layer. Mobile clients will use the ArcGIS Server Mobile ADF (ArcGIS Mobile) workflows. The Standard ESRI Workflow performance targets are used for initial planning purposes (AGS93 REST 1-5 layer Dynamic Service for Web mapping and the AGS93 Mobile ADF Client and AGS93 Mobile ADF Service for ArcGIS Mobile). These workflows can be modified as required during system design and implementation as capacity planning performance targets are validated.

Batch processing: A batch process load profile will be included in the design to account for the administrative batch process workflows planned during peak processing periods. These may include online backups, replication services, reconcile-and-post services, and so forth. A platform utilization profile can be established to represent the batch processing model loads (this profile would depend on potential batch processing needs). It will be a system administration responsibility to make sure the number of concurrent batch processes are managed within these limits during peak processing loads.

The capacity planning framework introduced in the last chapter provides the flexibility to include a variety of desktop and server workflow models in a single system design assessment. Workflow models are defined in terms of software component service times. Component service time target performance loads are provided for ESRI standard COTS (off the shelf) software based on internal performance validation testing and customer feedback. These Standard ESRI Workflow models, which are used as the ESRI baseline for system architecture design consulting, have proven to help customers identify successful design solutions. New technology options introduced with each ArcGIS software release may require different workflow models, and these new models can be represented in the future by appropriate component service times identified for these workflows in the capacity planning tool. Careful selection of appropriate target service times is an important part of the system architecture design.

11.2.2 User workflow requirements

The user-needs template of figure 11-3, used here to document the year 1 user application requirements for the City of Rome, was designed to integrate the requirements analysis provided in Tomlinson's *Thinking about GIS* with what is needed to complete the system architecture design. The spreadsheet identifies user workflow requirements (peak workflow loads) at the department level for each user location.

Figure 11-3
City of Rome User Needs - Year 1

City of Rome - Year 1				Total Users	Peak Workflow Loads		
					Desktop (users)		Server (req/hr)
Department	Workflow	IPD	User type		ArcInfo	ArcView	AGS
Site 1 - City Hall							
Planning	Zoning	1.0	Planner	20		8	
		1.1	Web services				2,600
	Permits	1.2	Inspector	20		10	
		1.3	Appraiser	15	8		
		1.4	Supervisor	2		2	
		1.5	Web services				600
Engineering	Sewer Backup	2.1	Engineer	4		3	
		2.2	Web services				800
	Electrical Breaks	2.3	Electrician	13	6		
		2.4	Supervisor	2	1		
		2.5	Web services				600
	Hwy. repair	2.6	Field engineer	10		4	
		2.7	Contracts	4		4	
City hall totals				90	15	31	
IT Department	Public		Web Services				**5,800**
Site 2 - Operations							
Operations	Clean-up prog.	3.1	Ops. Staff	4		2	
Operations totals				4	0	2	0
Remote field offices (WAN)							
Site 3 - Freeberg	Inspection	4.1	Field engineer	40		30	
Site 4 - Willsberg	Inspection	4.1	Field engineer	30		20	
Field Offices	Inspection	4.2	Web Services				1,200
Remote totals				70	0	50	
City Totals				164	15	83	0

Note: Peak Web Users = Peak Requests per minute / 6 displays per minute

Public Internet Site

Peak workflow loads (identified during the requirements definition—or user needs assessment—stage) establish processing and traffic requirements used by the capacity planning tool to generate hardware specifications. Each department manager should be called upon during the design process to validate that these peak user loads are accurate estimates. Final negotiations on hardware selection should focus on these peak user workflow requirements. Peak user requirements are used by the Capacity Planning Tool to generate system loads (the processing and traffic requirements referred to above) and determine the final hardware and network solution.

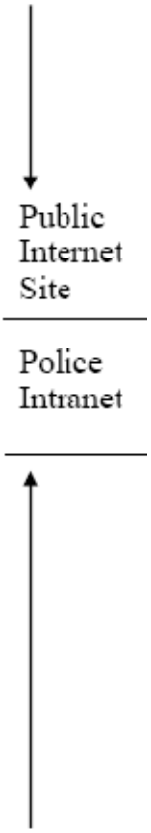
The Web services are published from the IT Data Center to the public Internet site (in this case study, departments did not access Web services from internal locations).

Again, each department manager is responsible for validating the final workflow requirements. Workflow requirements are identified for each implementation phase. Figure 11-4 identifies the City of Rome year 2 workflow requirements, and figure 11-5 identifies workflow requirements for year 3.

Figure 11-4
City of Rome User Needs - Year 2

City of Rome - Year 2				Total Users	Peak Workflow Loads			
					Desktop (Users)		Server (req/hr)	
Department	Workflow	IPD	User type		ArcInfo	ArcView	Map	Mobile
Site 1 - City Hall								
Planning	Zoning	1.0	Planner	25		15		
		1.1	Web services				2,600	
	Permits	1.2	Inspector	25		15		
		1.3	Appraiser	20	10			
		1.4	Supervisor	5	2			
		1.5	Web services				900	
Engineering	Sewer Backup	2.1	Engineer	5		3		
		2.2	Web services				1,000	
	Electrical Breaks	2.3	Electrician	13	6			
		2.4	Supervisor	2	1			
		2.5	Web services				1,900	
	Hwy. repair	2.6	Field engineer	11		7		
		2.7	Contracts	4		4		
City Hall LAN Totals				110	19	44		
IT Department	Public		Web Services				12,900	
Police (Firewall)	Patrol sched.	5.1	Admin.	10		3		
		5.4	Web services					100
	Crime analysis	5.2	Detectives	10	5			
	Spec. events	5.3	Traffic	10		3		100
Remote Patrols	Patrols	5.5	Patrol officers	20				
Police Network Totals				30	5	6		100
Site 2 - Operations								
Operations	Clean-up prog.	3.1	Ops. staff	4		2		
911	Response	3.2	Call takers	30		30		
		3.3	Web services				4,000	
Remote vehicles	Dispatch	3.4	Drivers	30				
Operations totals				34	0	32		
Remote field offices (WAN)								
Site 3 - Freeberg	Inspection	4.1	Field engineer	45		30		
Site 4 - Willsberg	Inspection	4.1	Field engineer	60		40		
WAN Field Offices	Inspection	4.2	Web Services				1,200	
Remote field office (WAN) totals				105	0	70		
Remote field offices (Internet)								
Site 5 - Perth	Inspection	4.3	Field engineer	10		2		
Site 6 - Wawash	Inspection	4.3	Field engineer	60		40		
Site 7 - Jackson	Inspection	4.3	Field engineer	60		20		
WAN Field Offices	Inspection	4.2	Web Services				1,300	
Remote field office (Internet) totals				120	0	62		
City totals (excluding Police private network)				419	19	200	12,900	

Note: Peak Web Users = Peak Requests per minute / 6 displays per minute



Year 2 includes implementation of a separate secure network to support Police Operations. The Police network will be a separate design, and geodatabase replication services will provide communication between the City Network and the Police Network. A new Mobile ADF application will support the Police Patrols, using background communications over dial-up connections to synchronize with the ArcGIS Server mobile application.

Year 2 City Network deployment adds 911 services within the Operations department along with a new dispatch operation and implementation of three additional field offices (Perth, Wawash, and Jackson).

Figure 11-5
City of Rome User Needs - Year 3

City of Rome - Year 3				Total Users	Peak User Workflow				
					Desktop (Users)			Server (req/hr)	
Department	Workflow	IPD	User type		ArcInfo	ArcView	Bus Anal	Map	Mobile
Site 1 - City Hall									
Planning	Zoning	1.0	Planner	25		15			
		1.1	Web services					2,600	
	Permits	1.2	Inspector	25		15			
		1.3	Appraiser	20	10				
		1.4	Supervisor	5	2				
		1.5	Web services					900	
Engineering	Sewer Backup	2.1	Engineer	6		3			
		2.2	Web services					1,000	
	Electrical Breaks	2.3	Electrician	13	6				
		2.4	Supervisor	2	1				
		2.5	Web services					1,500	
	Hwy. repair	2.6	Field engineer	11		7			
		2.7	Contracts	4		4			
	Work Orders	2.8	Managers	4		3			
	Work Reports	2.9	Field Units	10		2			
Business Development	Site Sel. and Natureserve	6.1	Planners	10			8		
	FMA Flood Zone, Serviced Land	6.2	Planners	10			8		
	Emergency Response, Time, etc.	6.3	Planners	2			2		
	Web Services	6.4						2,000	
City Hall LAN Totals				146	19	49	18		
IT Department	Public		Web Services					17,600	
Police (Firewall)	Patrol sched.	5.1	Admin.	10		3			
		5.5	Web services						200
	Crime analysis	5.2	Detectives	20	15				
	Spec. events	5.3	Traffic	10		3			
Remote Patrols	Police Dispatch	5.4	Traffic	10		3			
	Patrols/Routing	5.6	Patrol officers	20					200
Police Network Totals				70	15	9			200
Site 2 - Operations									
Operations	Clean-up prog.	3.1	Ops. staff	4		2			
911	Response	3.2	Call takers	50		30			
		3.3	Web services					4,000	
Remote vehicles	Fire and Ambulance Dispatch	3.4	Schedulers	30		30			
	Routing	3.5	Drivers	30					
Snow Cleaning	Scheduling	3.6	Engineers	4		4			
	Snow Plows	3.7	Drivers	100					
Operations totals				218	0	66			
Remote field offices (WAN)									
Site 3 - Freeberg	Inspection	4.1	Field engineer	45		30			
Site 4 - Willsberg	Inspection	4.1	Field engineer	60		40			
WAN Field Offices	Inspection	4.2	Web Services					1,200	
Remote field office (WAN) totals				105	0	70			
Remote field offices (Internet)									
Site 5 - Perth	Inspection	4.3	Field engineer	10		2			
Site 6 - Wawash	Inspection	4.3	Field engineer	50		40			
Site 7 - Jackson	Inspection	4.3	Field engineer	60		20			
Site 8 - Petersville	Inspection	4.3	Field engineer	80		60			
Site 9 - Rogerton	Inspection	4.3	Field engineer	80		60			
Internet Field Offices	Inspection	4.2	Web Services					4,000	
Remote field office (Internet) totals				120	0	182			
City totals (excluding Police private network)				**589**	**19**	**367**	**18**	**0**	

Note: Peak Web Users = Peak Requests per minute / 6 displays per minute

Public Internet Site

Police Intranet

Year 3 includes deployment of new Business Analyst and Joint Tracking (JTX) work order management applications. A Tracking Server implementation is deployed to facilitate snowplow scheduling. Two new remote field offices are added to the system. The police department adds a Police Dispatch and implements Tracking Server solution with the 20 Police Patrols.

Now that you've got it all down on the template, your overall task is to figure out what infrastructure you'll need to support all these new workflows. In other words, to design the system you must do a GIS user requirements analysis. Performing a proper one is hard work. The organization staff must work together to identify workflow processes and agree on business needs. Estimating the peak number of users for each user workflow is a fundamental part of doing business; once identified, this number will affect decisions on staffing and software licensing, as well as the hardware and infrastructure cost to support these workflows. Understanding and getting this right will make a big difference in user productivity and the success of the system.

11.3 City of Rome system architecture design

People skills and experience in maintaining distributed computer system solutions are important considerations when selecting a system design. Maintenance of the distributed computer environment is a critical factor in selecting appropriate vendor solutions. What experience and training in maintaining specific computer environments already exists in the organization? The answer to this may in itself identify a particular design solution as the best fit for an organization. This and the other considerations listed in figure 11-6 must be analyzed and understood before you can develop a proper design solution.

Figure 11-6
General System Design Considerations

- **Platform and Network Environments**
  - Hardware Experience
  - Maintenance Relationships
  - Staff Training
- **Hardware Policies and Standards**
  - Management Preferences
  - Established Vendor Relationships
- **Operational Constraints and Priorities**
  - System Availability Requirements
  - System Security Requirements
  - Application Performance Needs
- **System Administration Experience**
  - System Administration Support
  - Network Administration Support
  - Support Staff Administration Strategies
- **Financial Considerations**
  - Available Financial Resources
  - Performance/Cost Considerations

Platform and Network Environments: Whether on your own or in concert with a design consultant, you should review the vendor platforms and network environments currently maintained by the organization. Hardware experience, maintenance relationships, and staff training represent a considerable amount of investment for any organization. Proposed GIS design solutions should take advantage of corporate experience gained from working with the established platform and network environment.

Hardware Policies and Standards: Organizations develop policies and standards that support their hardware investment decisions. Understanding management preferences and associated vendor relationships will provide insight into a design solution that can be supported best by the organization.

Operational Constraints and Priorities: Understanding the type of operations supported by the GIS solution will identify requirements for fault tolerance, security, application performance, and the type of client/server architecture that would be appropriate to support these operations.

System Administration Experience: The skills and experience of the system support staff provide a foundation for the final design solution. Understanding network administration and hardware support experience, in conjunction with support staff preference for future administration strategies, will help guide the consultant to compatible hardware vendor solutions.

Financial Considerations: The final design must be affordable. An organization will not implement a solution that is beyond its financial resources. With system design, cost is a function of performance and reliability. If cost is an issue, the system design must facilitate a compromise between user application performance, system reliability, and cost. The design consultant must identify a hardware solution that provides optimum

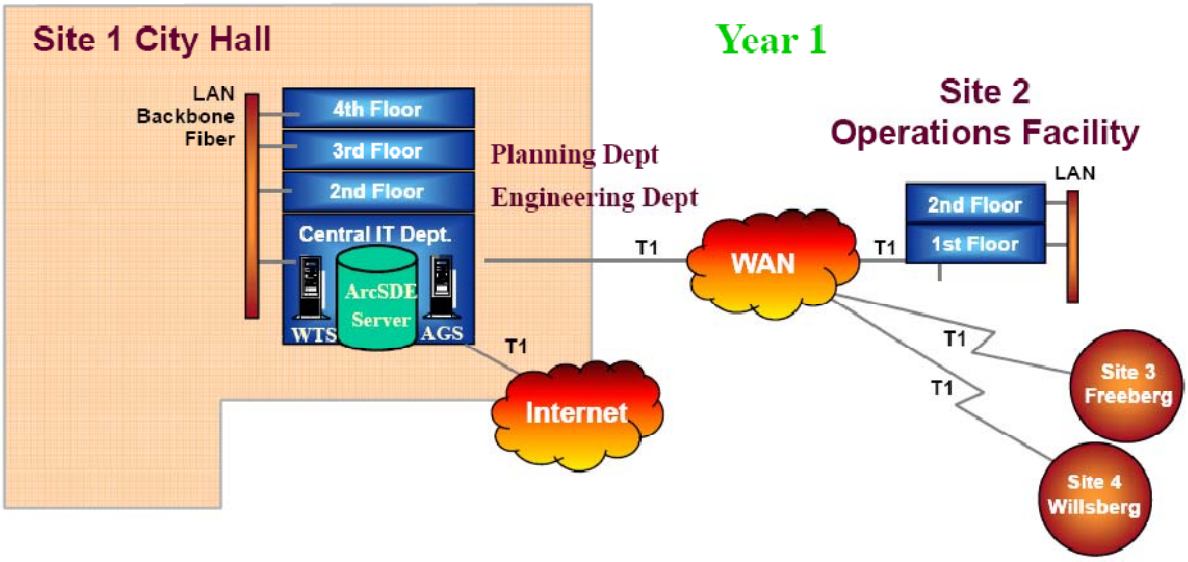
performance and reliability within identified budget constraints.

Current technology enables distribution of GIS solutions to clients throughout an enterprise environment, but there are limitations that apply to any distributed computer system design. It is important to clearly understand real GIS user needs and discuss alternative options for meeting those needs with system support staff to identify the most cost-effective solution. It may be necessary to review multiple software workflows and a variety of system deployment alternatives to identify and establish the best implementation strategy.

11.4 Year 1 capacity planning

Figure 11-7 provides an overview of City of Rome year 1 implementation strategy. A server-based architecture will be deployed from the central IT data center for the year 1 implementation. Server platforms will include a Windows Terminal Server farm to support the remote ArcGIS Desktop users, ArcGIS Server to support the public Web services, and a central GIS data server to support the enterprise geodatabase.

Figure 11-7
User Locations and Network Communications - Year 1



Roger Tomlinson,
Managing a GIS Seminar, July 2006

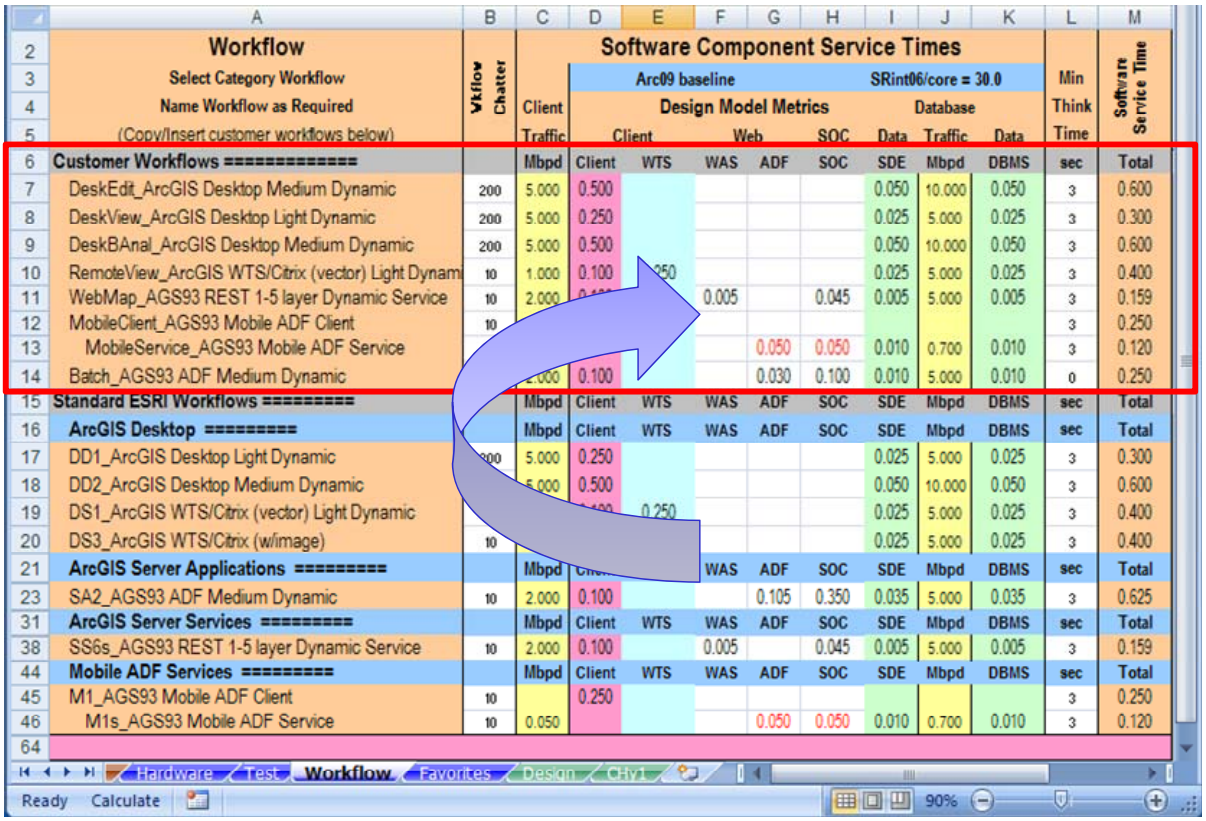
11.4.1 Year 1 Workflow Requirements Analysis

The system architecture design process is supported by the capacity planning tool. The planning workflow requirements identified earlier in figure 11-5 are used for the workflow analysis. For some organizations, the capacity planning tool may be used to collect workflow requirements and complete the analysis without using separate workflow requirements templates. Organizations should use the methodology appropriate to their design needs. The templates pictured earlier, with planning workflow requirements inserted as on a spreadsheet, provide the most complete user needs representation and establish appropriate documentation for the simpler workflow representations displayed in the CPT, as follows.

The first step in the system design process is to create custom GIS workflows for City of Rome in the CPT workflow tab. Workflow consolidation simplifies the look of the requirements analysis, but user workflows can only be consolidated when they are supported by the same software technology and have the same software service times. The current Capacity Planning Tool includes a tool for building a composite workflow from multiple individual workflows, and this should be used in establishing appropriate workflow targets for use cases that include multiple workflows and services. In most cases, the Standard ESRI Workflows should do fine for planning purposes.

You can use a large number of separate workflows in the capacity planning tool, yet keeping workflows to a reasonable number will clarify the presentation. The City of Rome example simplifies the analysis display by assuming all ArcGIS Desktop workflows are similar and can be represented by common productivity (divided into three categories - Editors, Viewers, and Business Analysts). If a more detailed analysis is needed, this can be completed on more detailed tabs with a summary provided for presentation. Figure 11-8 shows the custom workflows established for the City of Rome system design.

Figure 11-8
City of Rome Workflow Performance Targets

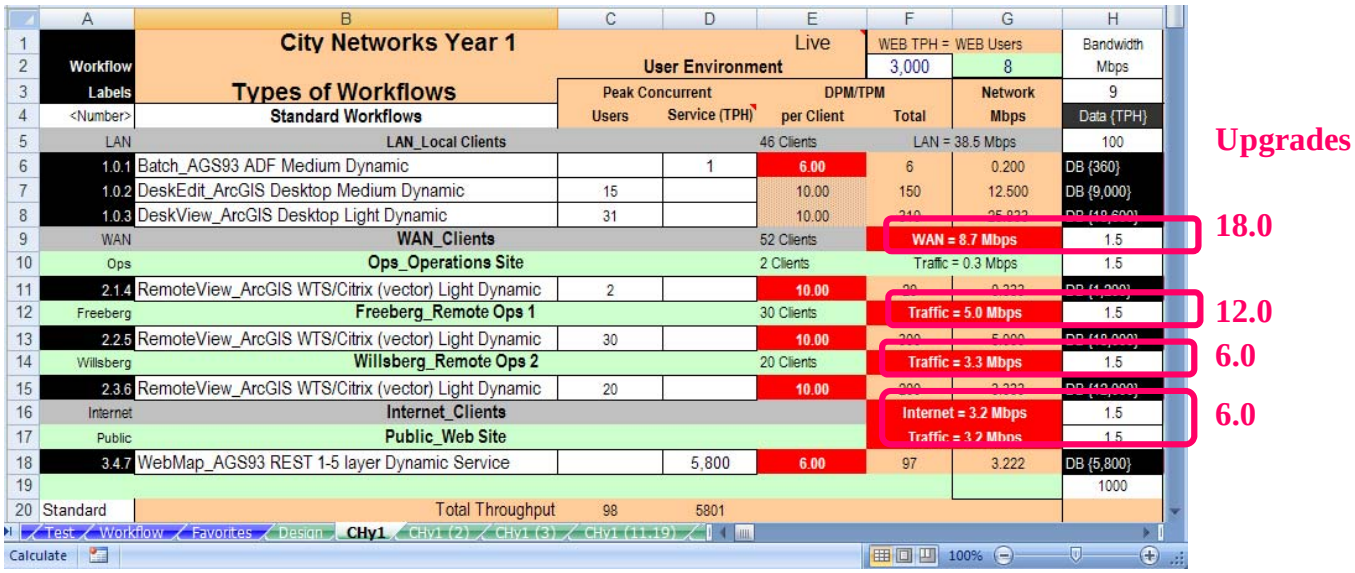


Workflow	Wkflow Chatter	Client Traffic	Software Component Service Times: Arc09 baseline, Design Model Metrics: Client	WTS	WAS	ADF	SOC	SRint06/core = 30.0, Database: SDE	Mbpd	DBMS	Min Think Time	Software Service Time
Select Category Workflow / Name Workflow as Required / (Copy/Insert customer workflows below)		Mbpd	Client	WTS	WAS	ADF	SOC	SDE	Mbpd	DBMS	sec	Total
Customer Workflows ==========												
DeskEdit_ArcGIS Desktop Medium Dynamic	200	5.000	0.500					0.050	10.000	0.050	3	0.600
DeskView_ArcGIS Desktop Light Dynamic	200	5.000	0.250					0.025	5.000	0.025	3	0.300
DeskBAnal_ArcGIS Desktop Medium Dynamic	200	5.000	0.500					0.050	10.000	0.050	3	0.600
RemoteView_ArcGIS WTS/Citrix (vector) Light Dynami	10	1.000	0.100	250				0.025	5.000	0.025	3	0.400
WebMap_AGS93 REST 1-5 layer Dynamic Service	10	2.000			0.005		0.045	0.005	5.000	0.005	3	0.159
MobileClient_AGS93 Mobile ADF Client	10										3	0.250
MobileService_AGS93 Mobile ADF Service						0.050	0.050	0.010	0.700	0.010	3	0.120
Batch_AGS93 ADF Medium Dynamic		2.000	0.100			0.030	0.100	0.010	5.000	0.010	0	0.250
Standard ESRI Workflows =========		Mbpd	Client	WTS	WAS	ADF	SOC	SDE	Mbpd	DBMS	sec	Total
ArcGIS Desktop =========		Mbpd	Client	WTS	WAS	ADF	SOC	SDE	Mbpd	DBMS	sec	Total
DD1_ArcGIS Desktop Light Dynamic	200	5.000	0.250					0.025	5.000	0.025	3	0.300
DD2_ArcGIS Desktop Medium Dynamic		5.000	0.500					0.050	10.000	0.050	3	0.600
DS1_ArcGIS WTS/Citrix (vector) Light Dynamic				0.250				0.025	5.000	0.025	3	0.400
DS3_ArcGIS WTS/Citrix (w/image)	10							0.025	5.000	0.025	3	0.400
ArcGIS Server Applications =========		Mbpd			WAS	ADF	SOC	SDE	Mbpd	DBMS	sec	Total
SA2_AGS93 ADF Medium Dynamic	10	2.000	0.100			0.105	0.350	0.035	5.000	0.035	3	0.625
ArcGIS Server Services =========		Mbpd	Client	WTS	WAS	ADF	SOC	SDE	Mbpd	DBMS	sec	Total
SS6s_AGS93 REST 1-5 layer Dynamic Service	10	2.000	0.100		0.005		0.045	0.005	5.000	0.005	3	0.159
Mobile ADF Services =========		Mbpd	Client	WTS	WAS	ADF	SOC	SDE	Mbpd	DBMS	sec	Total
M1_AGS93 Mobile ADF Client	10		0.250								3	0.250
M1s_AGS93 Mobile ADF Service	10	0.050				0.050	0.050	0.010	0.700	0.010	3	0.120

Each organization's solution will be different. Several decisions must be made during the design process before a final representation is collected in the capacity planning tool. The process and discussion leading up to the final design should be documented as a record of decisions made during the design process. Design documentation should clearly define the basis for the final workflow representation.

Once the custom workflows are defined, we are ready to complete the Phase 1 workflow requirements analysis. Figure 11.9 shows a CPT representation of the year 1 workflow requirements analysis for City of Rome.

Figure 11-9
Workflow Requirements Analysis - Year 1



	A	B	C	D	E	F	G	H
1		City Networks Year 1			Live	WEB TPH =	WEB Users	Bandwidth
2	Workflow			User Environment		3,000	8	Mbps
3	Labels	Types of Workflows	Peak Concurrent		DPM/TPM		Network	9
4	<Number>	Standard Workflows	Users	Service (TPH)	per Client	Total	Mbps	Data {TPH}
5	LAN	LAN_Local Clients			46 Clients	LAN = 38.5 Mbps		100
6	1.0.1	Batch_AGS93 ADF Medium Dynamic		1	6.00	6	0.200	DB {360}
7	1.0.2	DeskEdit_ArcGIS Desktop Medium Dynamic	15		10.00	150	12.500	DB {9,000}
8	1.0.3	DeskView_ArcGIS Desktop Light Dynamic	31		10.00			
9	WAN	WAN_Clients			52 Clients	WAN = 8.7 Mbps		1.5
10	Ops	Ops_Operations Site			2 Clients	Traffic = 0.3 Mbps		1.5
11	2.1.4	RemoteView_ArcGIS WTS/Citrix (vector) Light Dynamic	2		10.00			
12	Freeberg	Freeberg_Remote Ops 1			30 Clients	Traffic = 5.0 Mbps		1.5
13	2.2.5	RemoteView_ArcGIS WTS/Citrix (vector) Light Dynamic	30		10.00			
14	Willsberg	Willsberg_Remote Ops 2			20 Clients	Traffic = 3.3 Mbps		1.5
15	2.3.6	RemoteView_ArcGIS WTS/Citrix (vector) Light Dynamic	20		10.00			
16	Internet	Internet_Clients				Internet = 3.2 Mbps		1.5
17	Public	Public_Web Site				Traffic = 3.2 Mbps		1.5
18	3.4.7	WebMap_AGS93 REST 1-5 layer Dynamic Service		5,800	6.00	97	3.222	DB {5,800}
19								1000
20	Standard	Total Throughput	98	5801				

Upgrades
18.0
12.0
6.0
6.0

The CPT workflow requirements analysis includes all of the workflows identified during planning. Common workflows can be consolidated at each site location to simplify the display. The CPT workflows must track back to represent the individual Planning Workflow Requirements.

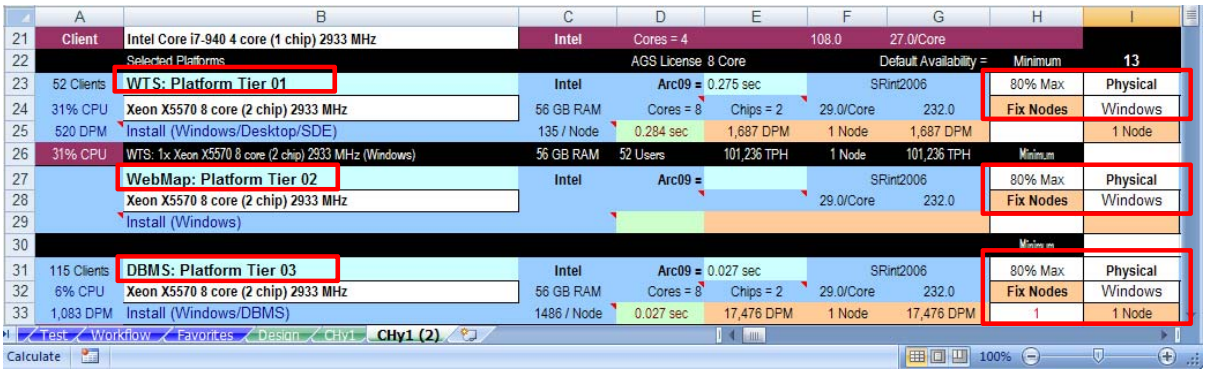
While you configure the user requirements and site locations in the CPT, and update the site traffic summation ranges to include all site workflow traffic going over the network connections (these CPT configuration procedures were discussed in chapter 10), the CPT is completing the performance analysis. Workflow network traffic is computed from total workflow requirements (DPM x Mbpd / 60 sec.). Network traffic cells show RED when there is insufficient bandwidth to support the required traffic.

System design planning should include discussions with the Network Administrator to make sure budgets are in place to accommodate required network upgrades. Network bandwidth should be at least twice the traffic to avoid workflow performance problems.

11.4.2 Platform Configuration Strategy and Software Installation

Several upgrades were made to the Capacity Planning Tool since August 2008. The system can be configured with up to 10 separate platform tier, you can change the tier name and include a nickname, you can identify whether it is a physical or virtual server deployment, and you can identify when you want to add additional platform nodes (utilization rollover percentage). Figure 11-10 shows the platform configuration strategy (3 tier architecture), platform tier naming assignment (WTS, WebMap, DBMS), minimum platform rollover setting (80 percent), and operating system environment (Windows physical servers) selected for the City of Rome design.

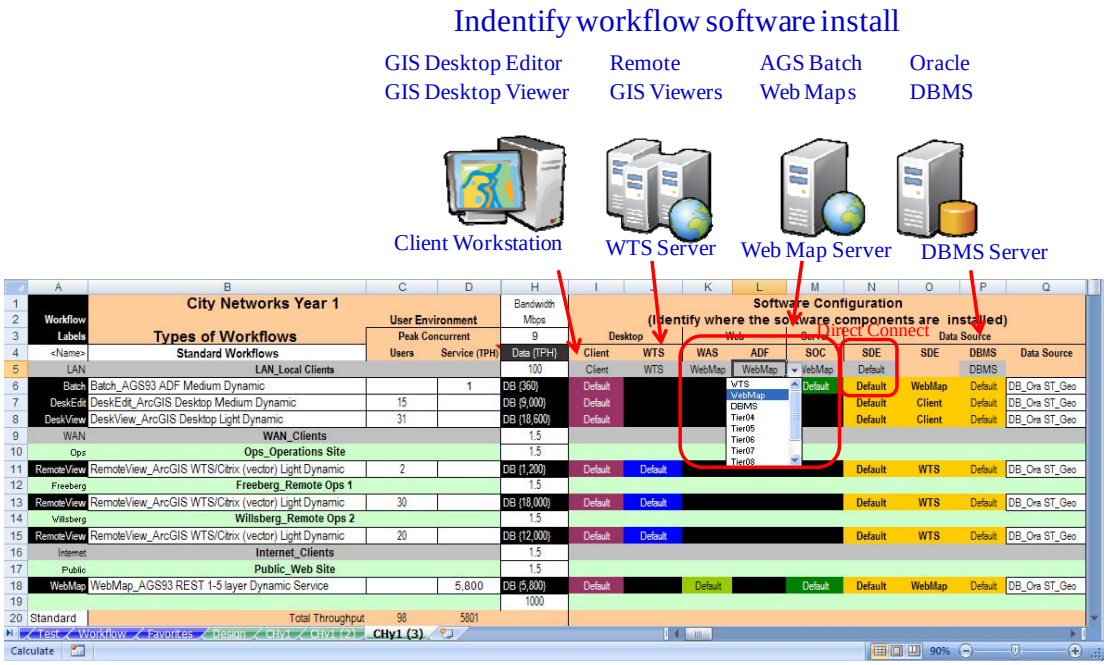
Figure 11-10
Platform Tier Configuration



	A	B	C	D	E	F	G	H	I
21	Client	Intel Core i7-940 4 core (1 chip) 2933 MHz	Intel	Cores = 4		108.0	27.0/Core		
22		Selected Platforms		AGS License 8 Core			Default Availability =	Minimum	13
23	52 Clients	WTS: Platform Tier 01	Intel	Arc09 =	0.275 sec	SRint2006		80% Max	Physical
24	31% CPU	Xeon X5570 8 core (2 chip) 2933 MHz	56 GB RAM	Cores = 8	Chips = 2	29.0/Core	232.0	Fix Nodes	Windows
25	520 DPM	Install (Windows/Desktop/SDE)	135 / Node	0.284 sec	1,687 DPM	1 Node	1,687 DPM		1 Node
26	31% CPU	WTS: 1x Xeon X5570 8 core (2 chip) 2933 MHz (Windows)	56 GB RAM	52 Users	101,236 TPH	1 Node	101,236 TPH	Minimum	
27		WebMap: Platform Tier 02	Intel	Arc09 =		SRint2006		80% Max	Physical
28		Xeon X5570 8 core (2 chip) 2933 MHz				29.0/Core	232.0	Fix Nodes	Windows
29		Install (Windows)							
30								Minimum	
31	115 Clients	DBMS: Platform Tier 03	Intel	Arc09 =	0.027 sec	SRint2006		80% Max	Physical
32	6% CPU	Xeon X5570 8 core (2 chip) 2933 MHz	56 GB RAM	Cores = 8	Chips = 2	29.0/Core	232.0	Fix Nodes	Windows
33	1,083 DPM	Install (Windows/DBMS)	1486 / Node	0.027 sec	17,476 DPM	1 Node	17,476 DPM	1	1 Node

Software can be installed on any platform tier, and the selected platform tier must be identified in the software configuration module in design columns I through Q. For City of Rome, the WTS software will be hosted on the WTS platform tier, the Web and SOC software will be installed on the WebMap platform tier, SDE will use the default SDE direct connect installation, and the DBMS software will be installed on the DBMS platform. Oracle will be the DBMS software, using the ST_Geometry data type for the geodatabase.

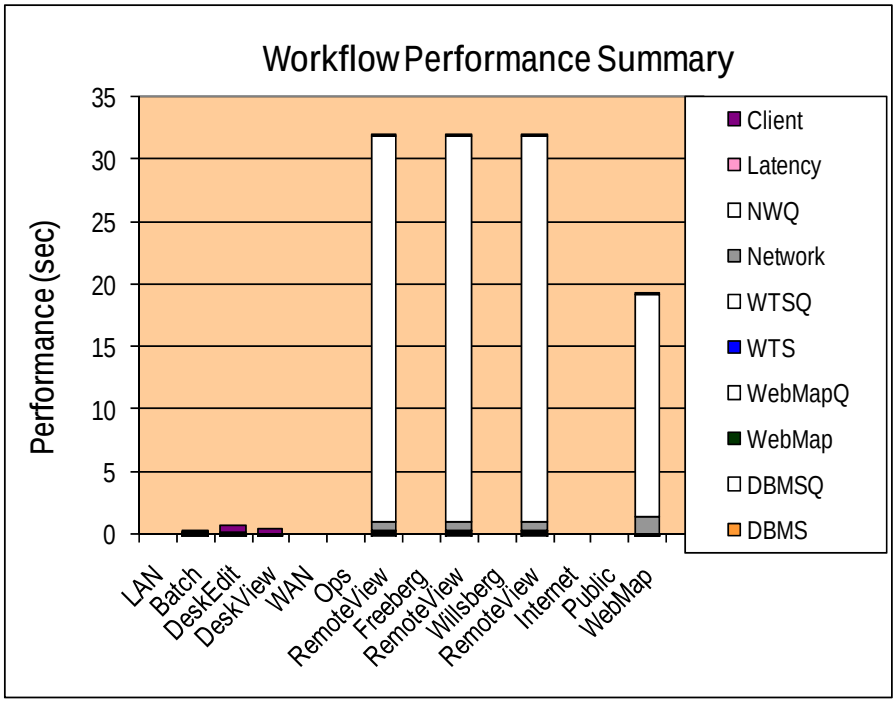
Figure 11-11
Workflow Software Installation



11.4.3 Year 1 network bandwidth suitability

Once the network connections are defined and the summary ranges validated, the capacity planning tool will identify potential traffic bottlenecks. The bottlenecks show up as colored network traffic summary cells and as network queue time on the Workflow Performance Summary. Once you identify the platform configuration strategy and identify host platform technology, the CPT can provide a workflow performance summary. Figure 11-12 shows the performance problems identified by the CPT during the network suitability analysis in figure 11-9.

Figure 11-12
Network Bandwidth Suitability - Year 1



In figure 11-9, the network traffic summary cells turn YELLOW when traffic is more than 50 percent of existing network bandwidth, and turn RED when traffic exceeds existing network bandwidth. The network queue times (NWQ) are identified on the Workflow Performance Summary above the network service time for each workflow.

For year 1, both the WAN and Internet traffic are well above the current bandwidth capacity. The Workflow Performance Summary shows the expected display response time due to network contention – the Internet application display response time (WebMap) is almost 20 seconds during peak loads. The Workflow Performance Summary clearly shows what could happen if the network bandwidth issue were not addressed during the design process.

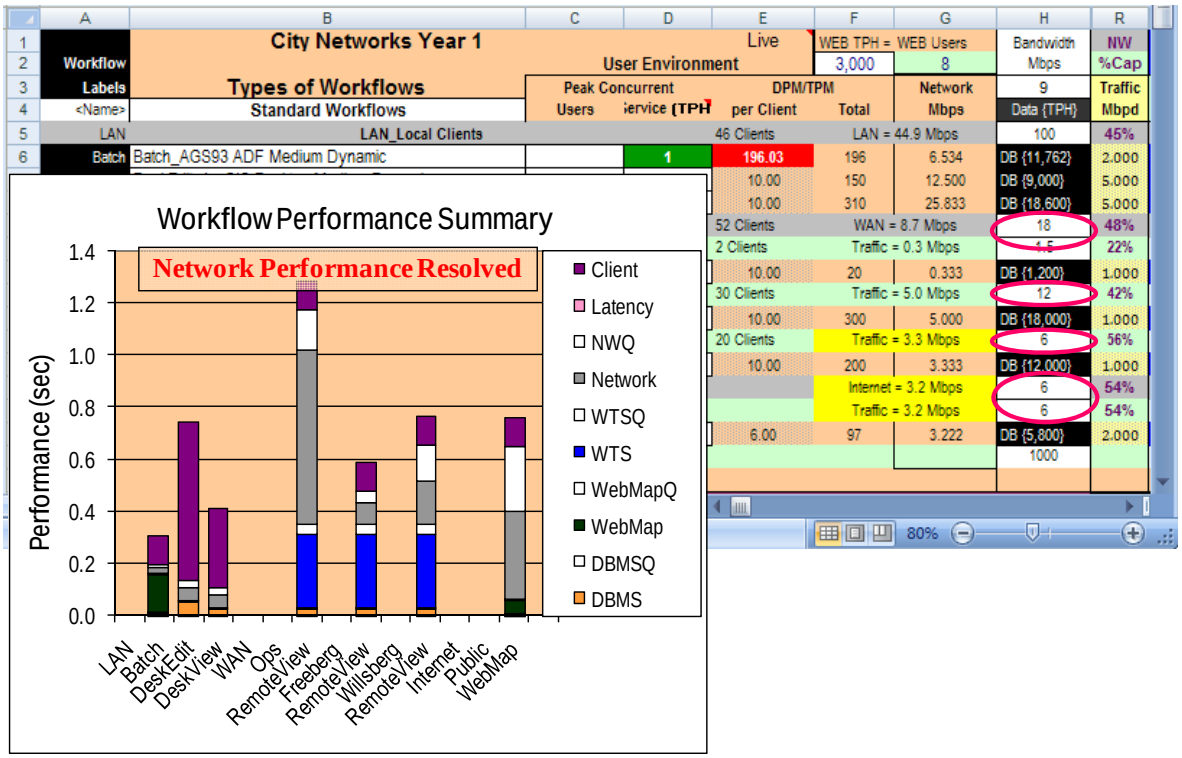
Year 1 network performance tuning

After identifying the network performance bottlenecks, the Capacity Planning Tool can be adjusted to resolve the particular performance problems. The standard recommendation is to configure network bandwidth to at least double the projected peak traffic requirements.

Remember, we only show the GIS traffic in this analysis; the WAN and Internet connections are shared by other users throughout the organization and their needs should also be represented in the analysis. It is possible to include an additional workflow representing other projected network traffic – this would certainly be advisable when performance is important, which it usually is.

City of Rome decided to upgrade the Data Center WAN and Internet connections 18 Mbps, the Freeberg site to 12 Mbps, and the Willsberg and Data Center Internet connections to 6 Mbps. Figure 11-13 shows the Workflow Performance Summary following the recommended network bandwidth upgrades. The Public WebMap workflow display response time is now less than 0.8 seconds.

Figure 11-13
Network Performance Tuning - Year 1



Appropriate network bandwidth upgrades can be represented in the CPT, and the network traffic cell colors and Workflow Performance Summary will respond with the proper adjustments. This is a discussion you should have with the network administrator; s/he can identify other traffic requirements which can easily be included in the analysis and confirm what network upgrades are possible.

Once the capacity planning tool is properly configured, you can compare the peak network bandwidth (provided by the workflow analysis tool) with existing bandwidth connections (represented on the earlier network diagram). Then you can identify the required upgrades, if any, and recommend them in your system design.

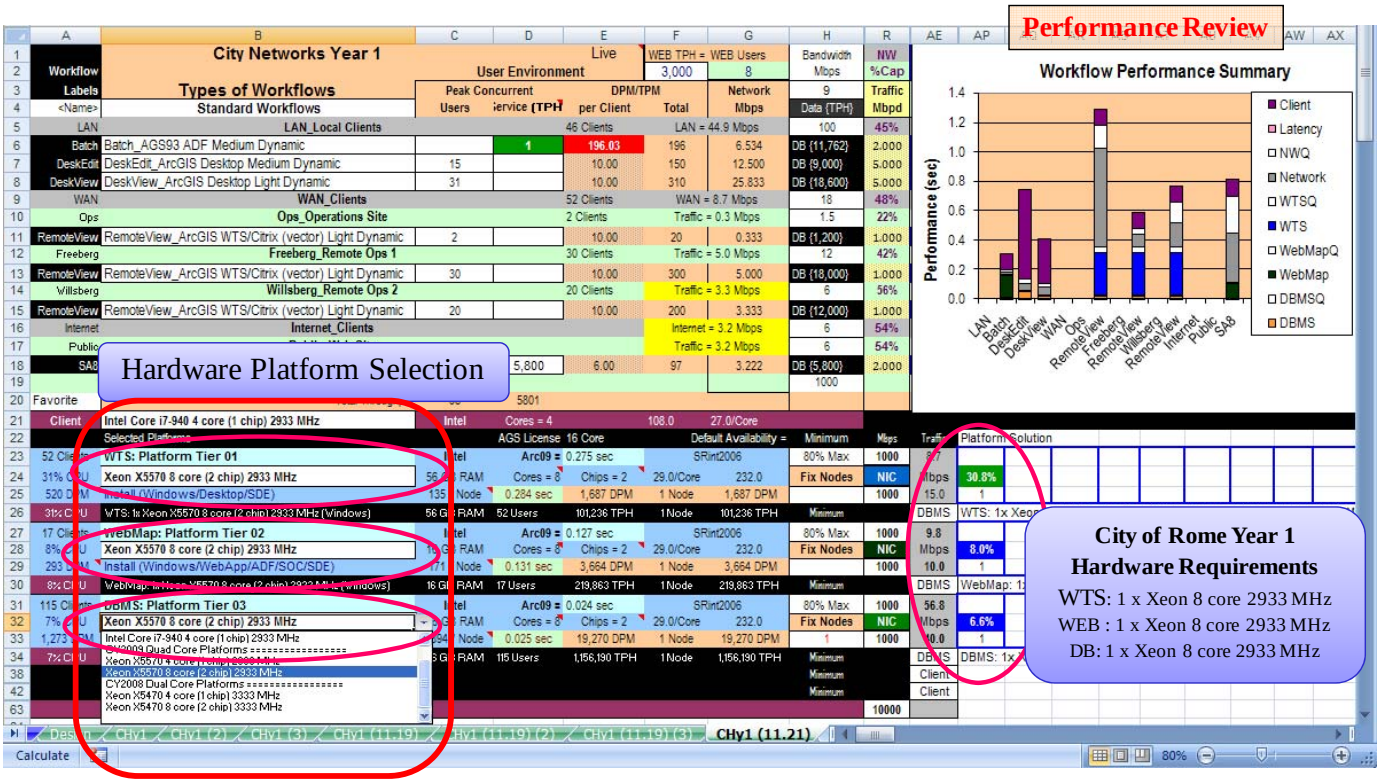
The City of Rome year 1 implementation encompasses users at four locations. City Hall includes network connections to enable WAN communications with the remote offices and an intranet connection over which the published public Web mapping services can be transported. Each remote office includes a router connection to the city WAN. Traffic requirements for each network connection are represented in the workflow analysis network traffic (column H).

11.4.4 Year 1 hardware platform configuration

After your have input the year 1 peak user workflow requirements in the CPT User Requirements Module and completed the software platform configuration, the CPT Platform Configuration Module can be used to make the final platform selection and complete the system configuration recommendations. The CPT incorporates the performance models (chapter 7) and the vendor platform performance benchmarks (chapter 9) discussed earlier. The CPT uses the established peak user workflow requirements, along with vendor-relative performance benchmarks, to generate the number of platform nodes required to handle the peak workflow loads. (The system architect can select from a variety of vendor platforms for the final configuration.)

The system design configuration in figure 11-14 is supported by Xeon X5570 8 core (2 chip) 2933 MHz hardware platforms (as of mid-2009, our favorite application server platform).

Figure 11-14
Hardware Platform Selection - Year 1

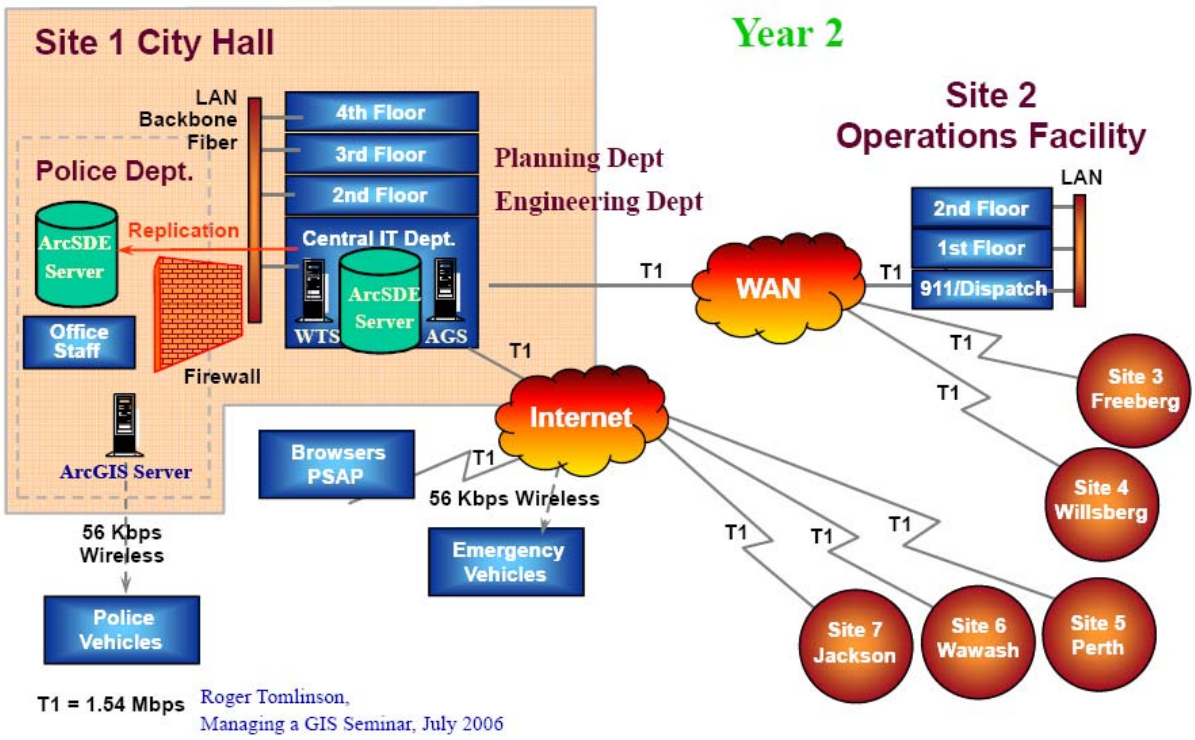


Peak capacity workflows projected for the year 1 deployment can be supported with one Windows Terminal Server; one Web mapping server (hosting the Web and ArcGIS Server software); and one database server. All hardware platforms are supported with the Xeon X5570 8 core (2 chip) 2933 MHz platforms configured with minimum of 56 GB physical memory (RAM) for the Windows Terminal Server and Database Tier and 16 GB RAM for the Web mapping. Projected CPU utilization rates are shown for each tier of the recommended platform solution server (less than 10 percent utilization for the Web mapping and Database platforms - these platforms could be hosted on two virtual servers supported by one of the physical machines).

11.5 Year 2 capacity planning

A similar analysis can be completed for the year 2 City of Rome implementation. Most GIS deployments evolve over several years of incremental technology improvements, and implementation plans normally address a two- or three-year schedule, to ensure that the budget is in place for the anticipated deployment needs. Figure 11-15 identifies user locations for the year 2 implementation.

Figure 11-15
User Locations and Network Communications - Year 2



Several additional Internet remote sites will be included in year 2, along with deployment of the 911 dispatch and the initial police network. The police network will be supported by a separate server environment (ArcSDE DBMS server for the geodatabase and an ArcGIS Server for the Mobile ADF Police Patrol application. Geodatabase replication will be used for data updates to the Police geodatabase). City network operations will continue to be administered from the IT Data Center.

11.5.1 Year 2 workflow analysis

See figure 11-16 for results of the year 2 city network workflow requirements analysis, as they appear in the CPT. You will need to transfer the workflow requirements identified in the user needs assessment (figure 11-4) to the workflow analysis module in the configuration tool. Two separate CPT tabs must be completed, one for the city network and a separate CPT tab for the police network. The Year 2 worksheet for the city network was updated in the capacity planning tool by copying the year 1 worksheet to a separate tab and inserting the additional locations and user workflows, to complete the user requirements. Year 1 network bandwidth upgrades were included as a starting point for the year 2 analysis.

Figure 11-16
Workflow Requirements Analysis - Year 2

	A	B	C	D	E	F	G	H	
1		City Networks Year 2			Live	WEB TPH =	WEB Users	Bandwidth	
2	Workflow			User Environment		3,000	8	Mbps	
3	Labels	Types of Workflows	Peak Concurrent		DPM/TPM		Network	9	Upgrade
4	<Name>	Standard Workflows	Users	Service (TPH	per Client	Total	Mbps	Data (TPH)	
5	LAN	LAN_Local Clients			63 Clients	LAN = 52.7 Mbps		100	1000
6	Batch	Batch_AGS93 ADF Medium Dynamic		1	6.00	6	0.200	DB {360}	
7	DeskEdit	DeskEdit_ArcGIS Desktop Medium Dynamic	19		10.00	190	15.833	DB {11,400}	
8	DeskView	DeskView_ArcGIS Desktop Light Dynamic	44		10.00	440	36.667	DB {26,400}	
9	WAN	WAN_Clients			102 Clients	WAN = 17.0 Mbps		18	45.0
10	Ops	Ops_Operations Site			32 Clients	Traffic = 5.3 Mbps		1.5	12.0
11	RemoteView	RemoteView_ArcGIS WTS/Citrix (vector) Light Dynamic	32		10.00	320	5.333	DB {19,200}	
12	Freeberg	Freeberg_Remote Ops 1			30 Clients	Traffic = 5.0 Mbps		12	
13	RemoteView	RemoteView_ArcGIS WTS/Citrix (vector) Light Dynamic	30		10.00	300	5.000	DB {18,000}	
14	Willsberg	Willsberg_Remote Ops 2			40 Clients	Traffic = 6.7 Mbps		6	12.0
15	RemoteView	RemoteView_ArcGIS WTS/Citrix (vector) Light Dynamic	40		10.00	400	6.667	DB {24,000}	
16	Internet	Internet_Clients			62 Clients	Internet = 17.5 Mbps		6	45.0
17	Public	Public_Web Site			62 Clients	Traffic = 17.5 Mbps		6	
18	WebMap	WebMap_AGS93 REST 1-5 layer Dynamic Service		12,900	6.00	215	7.167	DB {12,900}	
19	Perth	Perth_Remote Ops 3			2 Clients	Traffic = 0.3 Mbps		1.5	
20	RemoteView	RemoteView_ArcGIS WTS/Citrix (vector) Light Dynamic	2		10.00	20	0.333	DB {1,200}	
21	WaWash	WaWash_Remote Ops 4			40 Clients	Traffic = 6.7 Mbps		1.5	12.0
22	RemoteView	RemoteView_ArcGIS WTS/Citrix (vector) Light Dynamic	40		10.00	400	6.667	DB {24,000}	
23	Jackson	Jackson_Remote Ops 5			20 Clients	Traffic = 3.3 Mbps		1.5	6.0
24	RemoteView	RemoteView_ArcGIS WTS/Citrix (vector) Light Dynamic	20		10.00	200	3.333	DB {12,000}	
25								1000	
26	Favorite	Total Throughput	227	12901					

The year 2 implementation includes three additional remote sites with access over the Data Center Internet network. The new remote site network bandwidth connections represented on the network diagram (figure 11-14) are identified on the CPT workflow requirements analysis display. The new remote site bandwidth connections identified from figure 11-14 are as follows:

New Remote Site Network Connections

Perth: 1.5 Mbps (Cell H19)

Wawash: 1.5 Mbps (Cell H21)

Jackson: 1.5 Mbps (Cell H23)

Internet Web browser PSAP and emergency vehicles were included with the public Web services.

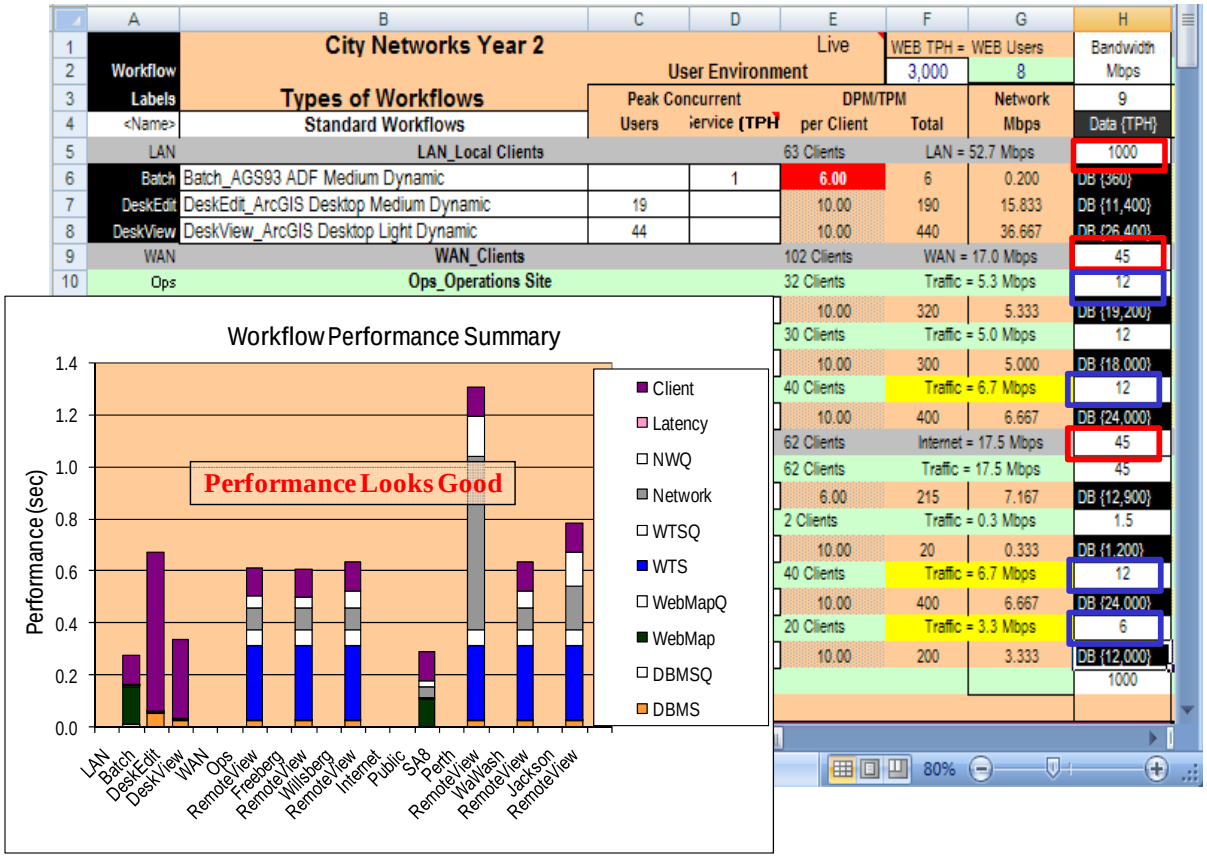
The site level network traffic totals should be checked to make sure the SUM ranges are updated (remote site ranges should include upper and lower network line bounding the site workflows, and the central data center ranges should include the total range of each gray network (LAN,WAN, Internet). Validating that these ranges are correct will ensure the proper calculations are supported by the Capacity Planning Tool.

11.5.2 Year 2 network suitability

With the network connections defined and the summary ranges validated, the capacity planning tool will identify potential traffic bottlenecks. The bottlenecks show up both as colored network traffic summary cells. The City Network Workflow Requirement Analysis (figure 11-15) shows seven site connections over 50 percent capacity. Recommended network upgrades should be at least twice the peak traffic. The Data Center network bandwidth upgrades include the LAN connection to 1000 Mbps and the WAN and Internet connections to 45 Mbps. Remote site upgrades include Ops, Willsberg, and Wawash connections to12 Mbps and the Jackson connection to 6 Mbps. Cost for these upgrades must be included in the network administrator's infrastructure design budget.

City of Rome decided to upgrade network connections based on the design recommendations. Figure 11-17 provides an overview of the City Network Year 2 Workflow Performance Summary once the recommended network upgrades are included in the CPT design.

Figure 11-17
Network Bandwidth Suitability Upgrades - Year 2

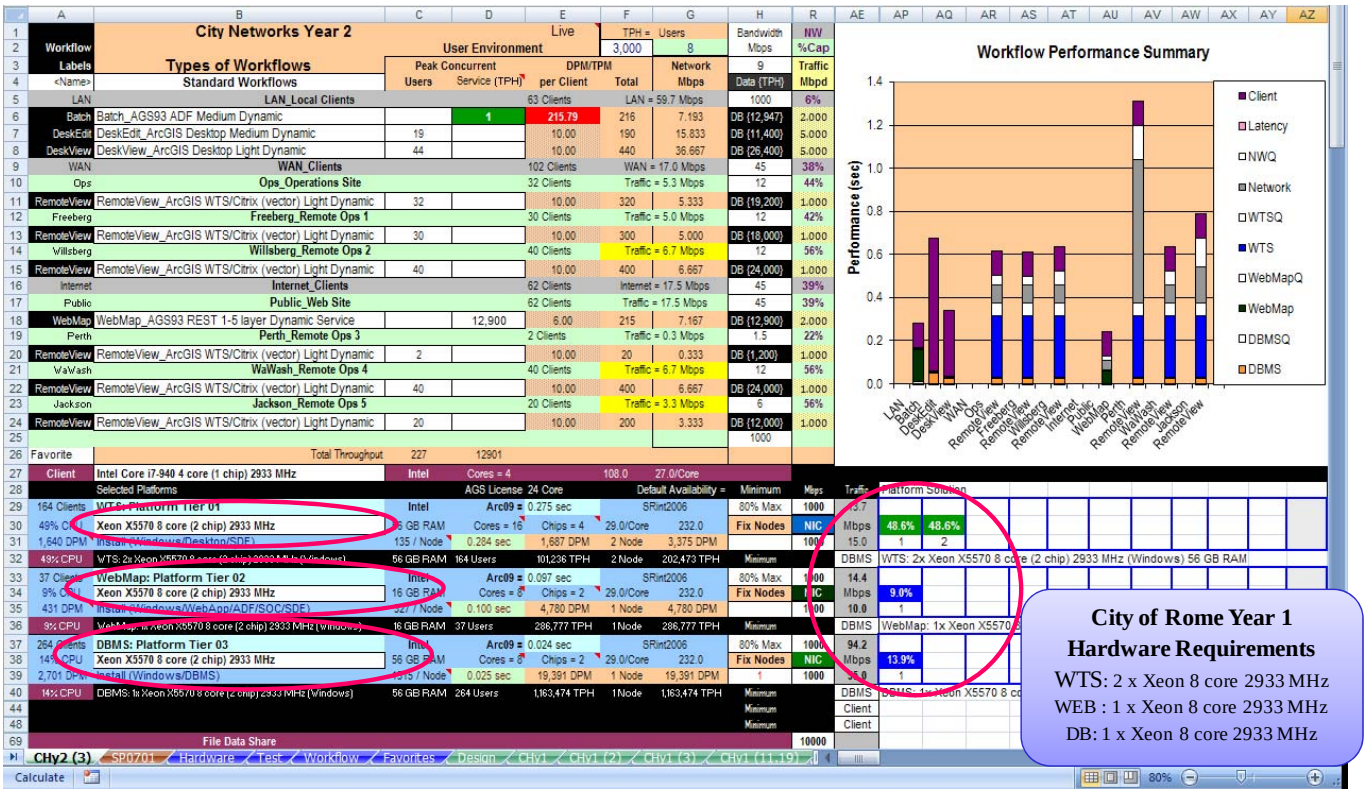


Appropriate network bandwidth upgrades can be entered in the Capacity Planning Tool, and the network traffic cell colors and Workflow Performance Summary will respond with the proper adjustments. Other traffic requirements can easily be included in the analysis, so have a discussion with the network administrator to see what they are and confirm that your recommended network upgrades can be supported

11.5.3 Year 2 hardware selection and performance review

The capacity planning tool is now ready to identify the year 2 platform solution. The CPT display (figure 11-18) shows the minimum recommended platform solution for the year 2 city network workflows when using the Xeon X5570 8 core (2 chip) 2933 MHz platforms. Platform configuration requirements are provided by the capacity planning tool after the user requirements and platform selections are complete. Figure 11-18 also provides an overview of the workflow peak loads and the Workflow Performance Summary.

Figure 11-18
Hardware Platform Selection - Year 2



The Phase 2 City network workflow requirements can be supported with two (2) Windows Terminal Servers, one (1) Web mapping server, and one (1) Database Server with the selected platform configurations.

The Workflow Performance Summary can be used to review user response times for the selected configuration. Process queue times located above each hardware component represent platform processing wait times (queue-time models are discussed in chapter 7). User performance will decrease as platform capacity increases due to process instructions waiting for access to processor core for execution.

11.5.5 Police network year 2 platform solution

Figure 11-19 provides the results of the year 2 Police network workflow requirements analysis. The workflow requirements identified in the user needs assessment (figure 11-4) were transferred to the configuration tool workflow analysis module. You will need to complete the Year 2 Police Network user requirements on a new CPT spreadsheet, configured as inputs to the Police network workflow requirements analysis.

Figure 11-19
Police Workflow Requirements Analysis - Year 2

	A	B	C	D	E	F	G	H
1		Police Network Year 2			Live	WEB TPH =	WEB Users	Bandwidth
2	Workflow			User Environment		3,000	8	Mbps
3	Labels	Types of Workflows	Peak Concurrent		DPM/TPM		Network	9
4	<Name>	Standard Workflows	Users	Service (TPH)	per Client	Total	Mbps	Data {TPH}
5	LAN	LAN_Local Clients			11 Clients	LAN = 9.2 Mbps		100
6	DeskEdit	DeskEdit_ArcGIS Desktop Medium Dynamic	5		10.00	50	4.167	DB {3,000}
7	DeskView	DeskView_ArcGIS Desktop Light Dynamic	6		10.00	60	5.000	DB {3,600}
8	WAN	WAN_Clients			20 Clients	WAN = 0.1 Mbps		1.5
9	MobileClient	MobileClient_AGS93 Mobile ADF Client	20		10.00	200		DB {12,000}
10	MobileService	MobileService_AGS93 Mobile ADF Service		20	6.00	120	0.100	DB {7,200}
15								1000
16	Standard	Total Throughput	31	20				

The year 2 Police implementation includes local network clients and remote mobile police patrol cars. Each police patrol communications unit is supported over wireless service routing mobile communications through the data center WAN service connection.

Figure 11-20 shows the Police hardware configuration and software install. The Web, ArcGIS Server, and DBMS software are installed on a single workgroup server.

Figure 11-20
Police Workflow Requirements Analysis - Year 2

Single Server Install

	A	B	H	I	J	K	L	M	N	O	P	Q
1		Police Network Year 2	Bandwidth	Software Configuration								
2	Workflow		Mbps	(Identify where the software components are installed)								
3	Labels	Types of Workflows	9	Desktop		Web		Server	Data Source			
4	<Name>	Standard Workflows	Data {TPH}	Client	WTS	WAS	ADF	SOC	SDE	SDE	DBMS	Data Source
5	LAN	LAN_Local Clients	100	Client	Workgroup	Workgroup	Workgroup	Workgroup	Default		Workgroup	
6	DeskEdit	DeskEdit_ArcGIS Desktop Medium Dynamic	DB {3,000}	Default					Default	Client	Default	DB_Ora ST_Geo
7	DeskView	DeskView_ArcGIS Desktop Light Dynamic	DB {3,600}	Default					Default	Client	Default	DB_Ora ST_Geo
8	WAN	WAN_Clients	1.5									
9	MobileClient	MobileClient_AGS93 Mobile ADF Client	DB {12,000}	Default								DB_Ora ST_Geo
10	MobileService	MobileService_AGS93 Mobile ADF Service	DB {7,200}				Default	Default	Default	Workgroup	Default	DB_Ora ST_Geo
15			1000									
16	Standard	Total Throughput										
17	Client	Intel Core i7-940 4 core (1 chip) 2933 MHz								WTS/DBMS		
18		Selected Platforms	High Avail	13	53			Software Install		RAM/User		
19	31 Clients	Workgroup: Platform Tier 01	80% Max	Physical	1.00	100%	/WebApp	/ADF	/SOC	45 MB		
20	2% CPU	Xeon X5570 8 core (2 chip) 2933 MHz	Fix Nodes	Windows	8 core/node	Workgroup		/Desktop	16			
21	230 DPM	Install (Windows/WebApp/ADF/SOC/Desktop/SDE/DBMS)		1 Node	1,563 Total	4	/SDE	/DBMS	7 /client			
22	2% CPU	Workgroup: 2x Xeon X5570 8 core (2 chip) 2933 MHz (Windows)	High Avail									

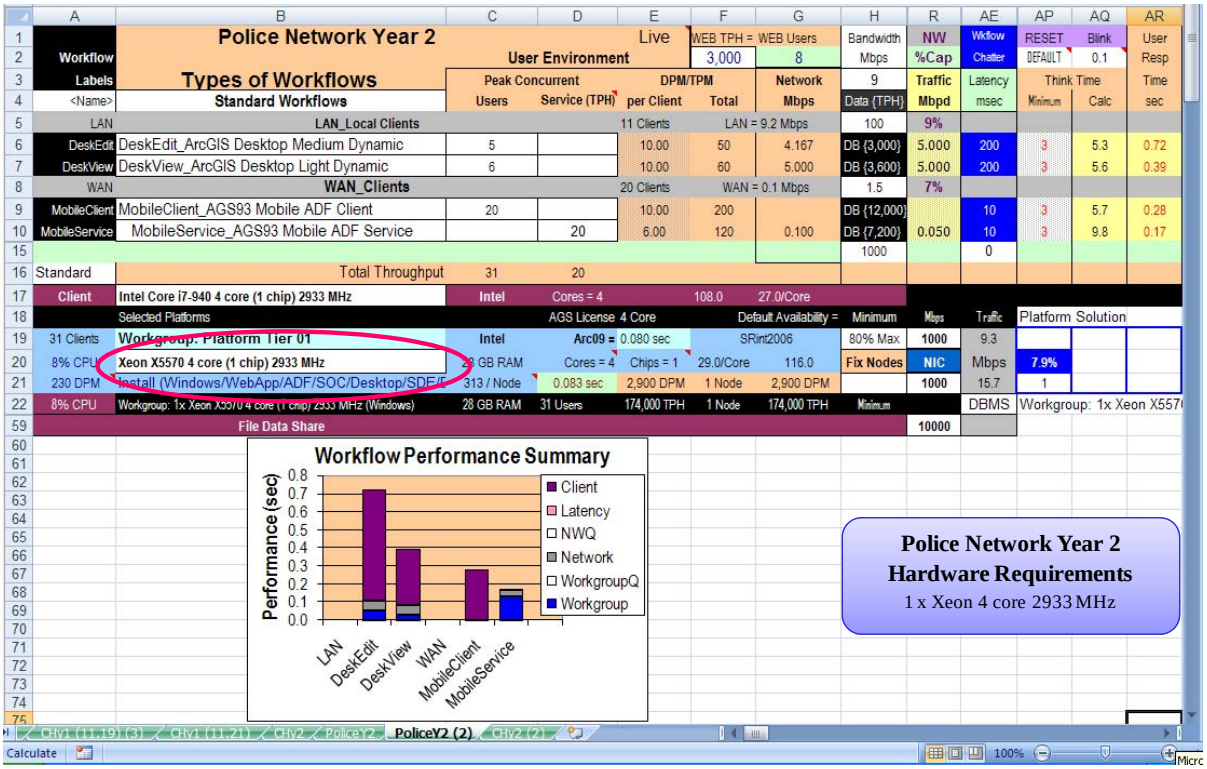
The Police Workgroup server is a Xeon X5570 with a Windows operating system. Software is installed in a Physical platform environment.

The police network platform final design is shown in figure 11-21, which includes the Police user workflows and design solution at the top and the Workflow Performance Summary at the bottom (you can drag the graphs located in the CPT to locations that are convenient for display purposes).

The peak numbers of ArcGIS Desktop users were identified as 11 in City of Rome year 2. The ESRI workgroup server license can support up to 10 concurrent ArcGIS Desktop connections and could be used to support the police network needs, if the peak workflow requirements could be reduced by one desktop user. This should be discussed during the design evaluation as a potential cost reduction—it is important that these peak loads were not overestimated.

In reviewing the final design analysis, it was clear that the Xeon X5570 platform could support the peak workflow loads with a single quad core chip. The selected police network workgroup server is the Xeon X5260 4 core (1 chip) 2933 MHz platform satisfying peak processing loads at less than 10 percent utilization. Figure 11-21 shows the selected Police design solution.

Figure 11-21
Police Hardware Platform Requirements - Year 2

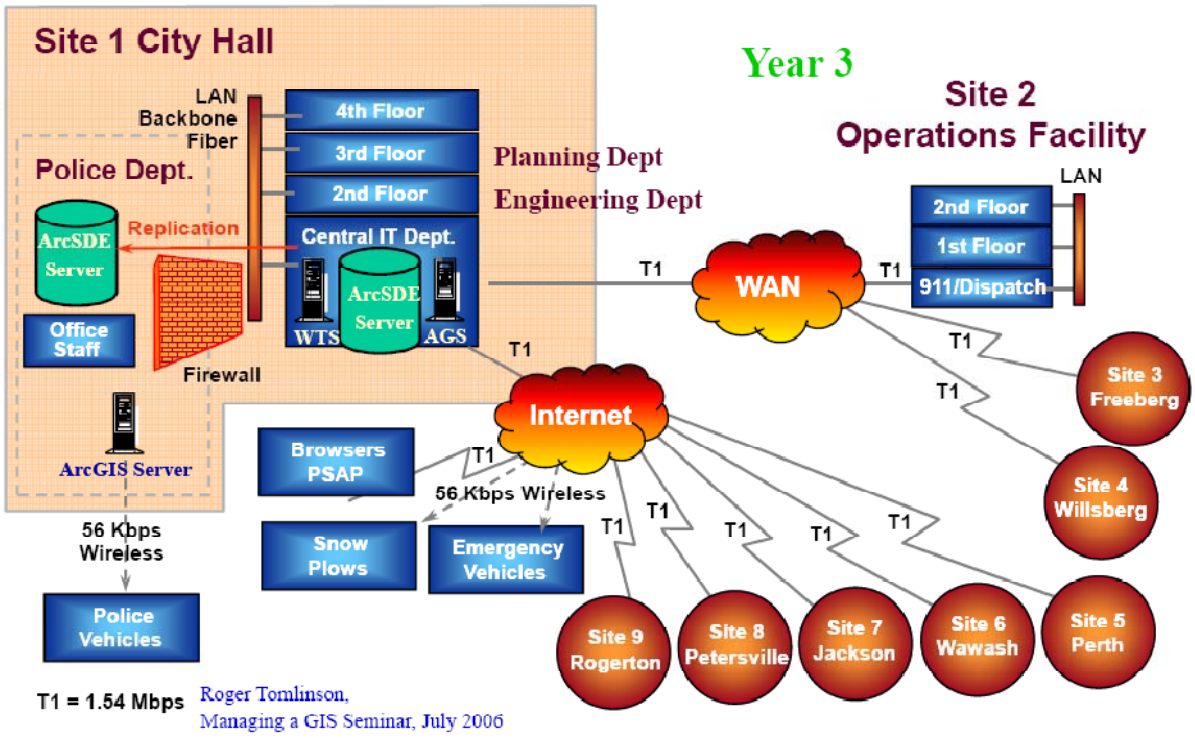


A review of the Workflow Performance Summary suggests there are no identified performance bottlenecks.

11.6 Year 3 capacity planning

Year 3 includes deployment of two new regional office sites, increase in the number of City Hall departments and workflows using GIS technology, and introduction of tracking analysis to monitor mobile vehicle operations, including up to 100 snowplows during winter storm operations. Here are the user locations for year 3 (figure 11-22).

Figure 11-22
User Locations and Network Communications - Year 3



11.6.1 Year 3 workflow analysis

Figure 11-23 provides the results of the year 3 workflow analysis. The workflow requirements identified in the user needs assessment (figure 11-4) were transferred to the configuration tool workflow requirements module. The workflow templates for the city network and the police network were generated from the capacity planning workbook in Excel, to complete the site configurations: by copying the year 2 worksheets to separate tabs and inserting the additional remote site locations and workflow upgrades.

Figure 11-23
Workflow Requirements Analysis - Year 3

	A	B	C	D	E	F	G	H
1		City Networks Year 3			Live	WEB TPH =	WEB Users	Bandwidth
2	Workflow			User Environment		3,000	8	Mbps
3	Labels	Types of Workflows	Peak Concurrent		DPM/TPM		Network	9
4	<Name>	Standard Workflows	Users	Service (TPH)	per Client	Total	Mbps	Data (TPH)
5	LAN	LAN_Local Clients			86 Clients	LAN = 71.9 Mbps		1000
6	Batch	Batch_AGS93 ADF Medium Dynamic		1	6.00	6	0.200	DB {360}
7	DeskEdit	DeskEdit_ArcGIS Desktop Medium Dynamic	19		10.00	190	15.833	DB {11,400}
8	DeskView	DeskView_ArcGIS Desktop Light Dynamic	49		10.00	490	40.833	DB {29,400}
9	DeskBAnal	DeskBAnal_ArcGIS Desktop Medium Dynamic	18		10.00	180	15.000	DB {10,800}
10	WAN	WAN_Clients			136 Clients	WAN = 22.7 Mbps		45
11	Ops	Ops_Operations Site			66 Clients	Traffic = 11.0 Mbps		12
12	RemoteView	RemoteView_ArcGIS WTS/Citrix (vector) Light Dynamic	66		10.00	660	11.000	DB {39,600}
13	Freeberg	Freeberg_Remote Ops 1			30 Clients	Traffic = 5.0 Mbps		12
14	RemoteView	RemoteView_ArcGIS WTS/Citrix (vector) Light Dynamic	30		10.00	300	5.000	DB {18,000}
15	Willsberg	Willsberg_Remote Ops 2			40 Clients	Traffic = 6.7 Mbps		12
16	RemoteView	RemoteView_ArcGIS WTS/Citrix (vector) Light Dynamic	40		10.00	400	6.667	DB {24,000}
17	Internet	Internet_Clients			182 Clients	Internet = 40.1 Mbps		45
18	Public	Public_Web Site			182 Clients	Traffic = 40.1 Mbps		45
19	WebMap	WebMap_AGS93 REST 1-5 layer Dynamic Service		17,600	6.00	293	9.778	DB {17,600}
20	Perth	Perth_Remote Ops 3			2 Clients	Traffic = 0.3 Mbps		1.5
21	RemoteView	RemoteView_ArcGIS WTS/Citrix (vector) Light Dynamic	2		10.00	20	0.333	DB {1,200}
22	Wawash	Wawash_Remote Ops 4			40 Clients	Traffic = 6.7 Mbps		12
23	RemoteView	RemoteView_ArcGIS WTS/Citrix (vector) Light Dynamic	40		10.00	400	6.667	DB {24,000}
24	Jackson	Jackson_Remote Ops 5			20 Clients	Traffic = 3.3 Mbps		6
25	RemoteView	RemoteView_ArcGIS WTS/Citrix (vector) Light Dynamic	20		10.00	200	3.333	DB {12,000}
26	Petersville	Petersville_Remote Ops 6			60 Clients	Traffic = 10.0 Mbps		1.5
27	RemoteView	RemoteView_ArcGIS WTS/Citrix (vector) Light Dynamic	60		10.00	600	10.000	DB {36,000}
28	Rogerton	Rogerton_Remote Ops 7			60 Clients	Traffic = 10.0 Mbps		1.5
29	RemoteView	RemoteView_ArcGIS WTS/Citrix (vector) Light Dynamic	60		10.00	600	10.000	DB {36,000}
30								1000
31	Favorite	Total Throughput	404	17601				

Upgrade

24.0

90.0

24.0

24.0

Year 3 includes adding a new Business Analyst workflow to the City LAN, and adding two new remote sites (Petersville and Rogerton) to the Data Center Internet connections. The remote site network bandwidth connections are shown on the far right of the CPT display (column H):

New Remote Site Network Connections

Site 8 – Petersville: 1.5 Mbps (Cell H26)

Site 9 – Rogerton: 1.5 Mbps (Cell H28)

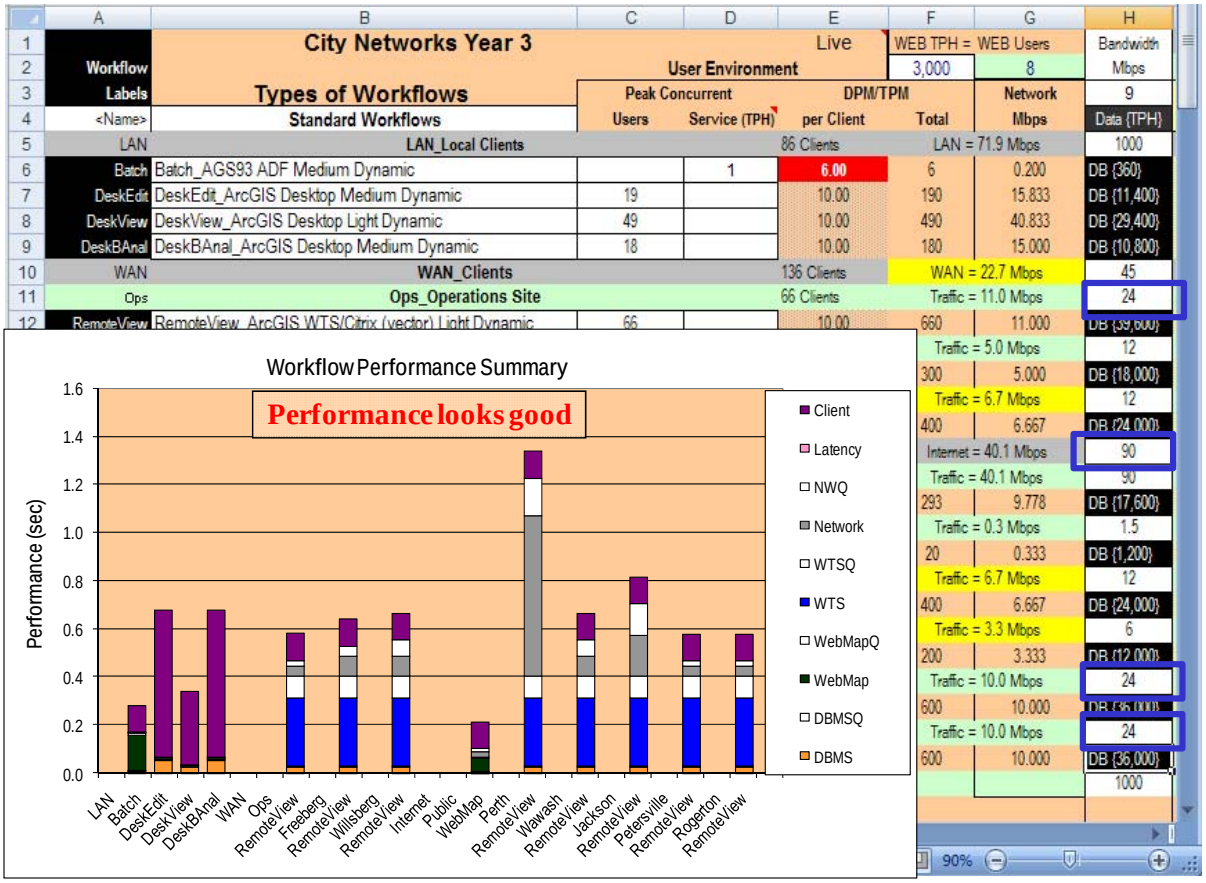
The network traffic totals should be checked to make sure the SUM ranges are correct (remote site ranges should include upper and lower green line bounding the site workflows, and the central data center ranges should include the total range of each network (LAN, WAN, Internet). Validating that these ranges are correct will ensure that the Capacity Planning Tool makes the proper calculations.

11.6.2 Year 3 network suitability

Once the network connections are defined and the summary ranges are validated, the CPT will identify potential traffic bottlenecks. The bottlenecks show up both as colored network traffic summary cells and as network queue time on the Workflow Performance Summary. The CPT (figure 11-23) shows network traffic over 50 percent bandwidth capacity for Ops, Willsberg, Wawash, Jackson and for the City WAN and Internet connections. Petersville and Rogerton show traffic over 100 percent capacity.

The City of Rome decides to plan for upgrading the City Internet connection to 90 Mbps, and connections for Ops, Petersville, and Rogerton to 24 Mbps. The remaining sites were only slightly over 50 percent utilization and display performance should remain acceptable during these peak loads. Figure 11-24 provides an overview of the upgraded City Network Year 3 design including the Workflow Performance Summary.

Figure 11-24
Network Bandwidth Suitability - Year 3



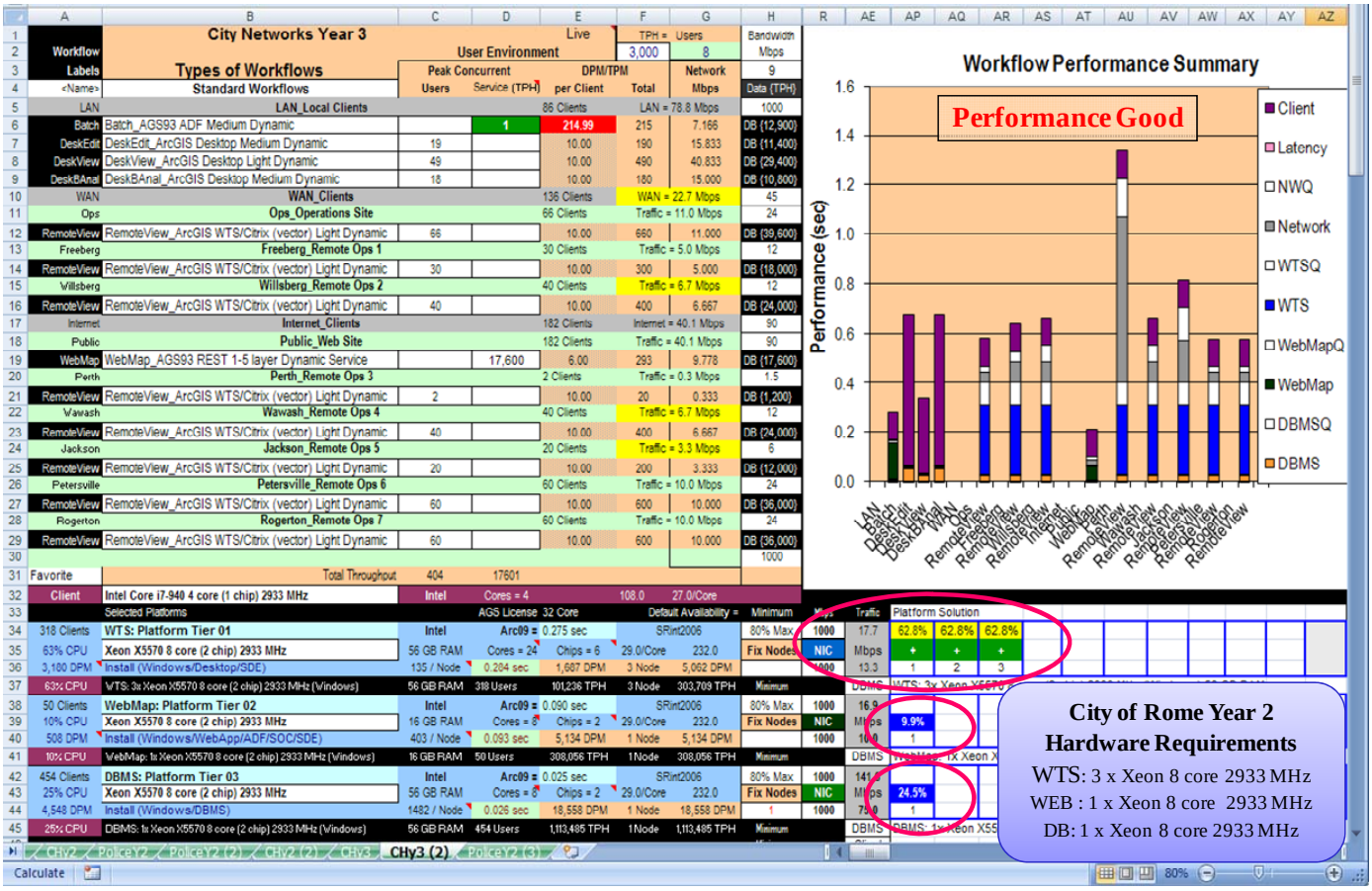
Appropriate network bandwidth upgrades can be represented in the Capacity Planning Tool, and the Workflow Performance Summary will respond with the proper adjustments. Most display response times are less than 1 second, with the remote ArcGIS Desktop viewers at Perth slowing to almost 1.4 sec per display (Perth has the lowest bandwidth at 1.5 Mbps). This analysis considers only the GIS traffic, so there will come a time to conduct the discussion with the network administrator about including the other traffic requirements. They can be estimated using a custom workflow representing existing traffic loads, and thereby easily included in the analysis.

11.6.3 Year 3 hardware selection and performance review

The capacity planning tool is now ready to validate the final year 3 platform configuration. The CPT interface below (figure 11-25) shows the minimum recommended platform solution for the year 3 city network workflows when using the Xeon X5570 8 core (2 chip) 2933 MHz platforms. Platform configuration requirements are provided by Excel once the user requirements and platform selections are complete.

The Workflow Performance Summary shows the user workflow display response times for the selected configuration. Process queue times are represented above each hardware component to identify platform processing wait times (queue time models are discussed in chapter 7). For this solution, Windows Terminal Server platforms are at 63 percent capacity (Cell A35) and, slightly above 50 percent utilization.

Figure 11-25
Hardware Platform Selection - Year 3

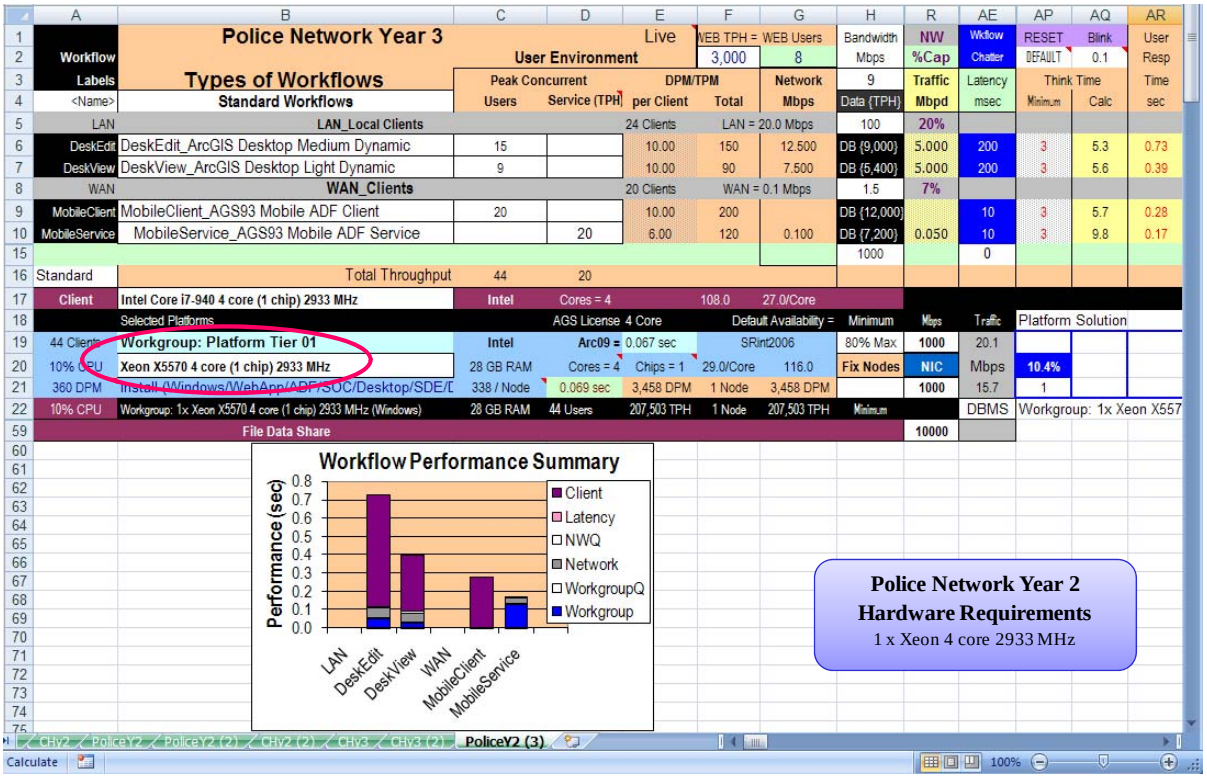


Final platform selection includes three (3) Xeon X5570 8 core (2 chip) 2933 MHz platforms for the WTS tier, one (1) Xeon X5570 8 core (2 chip) 2933 MHz platform for the Web mapping tier, and one (1) Xeon X5570 8 core (2 chip) 2933 MHz platform for the DBMS tier. Additional platforms can be provided to support high availability requirements. Memory recommendations include 56 MB RAM for the WTS and DBMS servers and 16 GB RAM for the Web mapping server. Display performance for all workflows is well under 2 seconds.

11.6.4 Police network year 3 platform solution

The police network year 3 design is provided in Figure 11-26. Capacity requirements continue to be satisfied with a single-socket server configuration. Peak desktop user connections have increased to a total of 24 ArcGIS Desktop users, which would require an enterprise ArcGIS Server license, although peak loads could continue to be supported by the SQL Express database.

Figure 11-26
Police Hardware Platform Requirements - Year 3



The Workflow Performance Summary shows excellent performance with the proposed system design solution.

11.7 Choosing a system configuration

The best solution for a given organization depends on the distribution of the user community and the type of operational data in use. User requirements determine the number of machines necessary (to support the operational environment), the amount of memory required (to support the applications), and the amount of disk space needed (to support the system solution). The system design models provide target performance metrics to aid in capacity planning. The capacity planning tool incorporates standard templates representing the sizing models and provides a manageable interface to help in enterprise-level capacity planning efforts. The CPT can be a big help in applying the results of the user needs assessment.

User needs change as organizations change, so this assessment not only identifies platform and infrastructure specifications and sets performance targets for the initial implementation, it is also part of the process going forward. System upgrades, new technology solutions, tuning and optimizing performance--every implementation or change is like a new launch, insofar as you need to plan for it. Planning provides an opportunity to establish performance milestones that can be used to manage a successful GIS implementation. Performance targets used in capacity planning can provide target milestones to validate performance and scalability throughput deployment of the system. The next and last chapter is about what to do with a system design once you have it: we will discuss the fundamentals of systems integration and some best practices that lead you to a successful implementation.

12 System Implementation

Successful system implementation requires good leadership and careful planning. A good understanding of every component of the system is critical in putting together an implementation strategy. Enterprise IT environments involve integration of a variety of vendor technologies. Interoperability standards within commercial software environments are voluntary, and even the most simple system upgrade must be validated at each step of the integration process.

Enterprise GIS environments include a broad spectrum of technology integration. Most environments today include a variety of hardware vendor technologies including database servers, storage area networks, Windows Terminal Servers, Web servers, map servers, and desktop clients,—all connected by a broad range of local area networks, wide area networks, and Internet communications. All these technologies must function together properly to support a balanced computing environment. A host of software vendor technologies including database management systems, ArcGIS Desktop and ArcGIS Server software, Web services, and hardware operating systems—all integrated with existing legacy applications. Data and user applications are added to the integrated infrastructure environment to support the final implementation. The result is a very large mixed bag of technology that must work together properly and efficiently to support user workflow requirements.

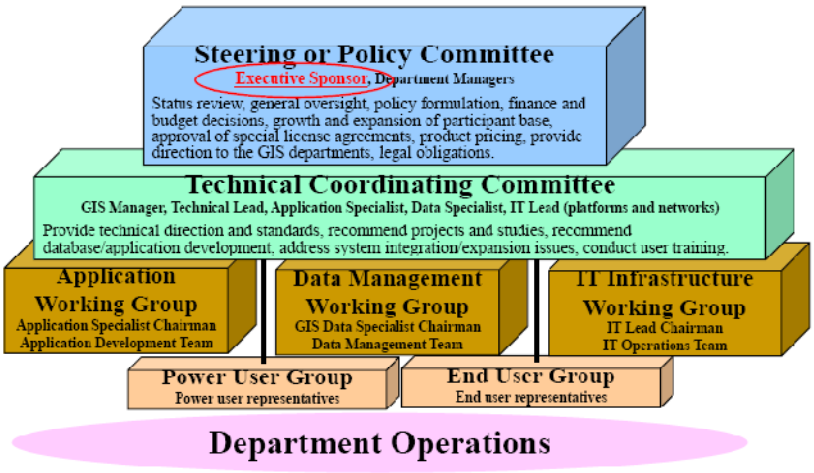
The integration and implementation of distributed computer technology have become easier over the years as interface standards have matured. At the same time, enterprise environments have become larger and more complex. The complexity and risk associated with an enterprise system deployment are directly related to the variety of vendor components required to support the final integrated solution. Centralized computing solutions with a single database environment are the easiest environments to implement and support. Distributed computer systems with multiple distributed database environments can be very complex and difficult to deploy and support. Many organizations are consolidating their data resources and application processing environments to reduce implementation risk and improve administrative support for enterprise business environments.

12.1 GIS Staffing

Good leadership begins with proper staffing. Successful GIS enterprise deployments are normally supported by an executive business sponsor, and the GIS manager should report to senior management.

Figure 12-1 shows an overview of a traditional GIS organization structure. Enterprise GIS operations are supported by an executive committee with influence and power to make financial and policy decisions for the GIS user community. A technical coordinating committee is responsible for providing technical direction and leadership. Working groups are assigned and normally aligned with each technical discipline to address organizational issues and report on system status. The user community should be represented throughout the review process.

Figure 12-1
Traditional GIS Organizational Structure



A formal organizational structure provides a framework for establishing and maintaining long-term support required for successful enterprise GIS operations. This basic organization structure can be useful in managing small to large organizations, and the same type of organizational structure can be effective in managing community GIS operations.

Several technical disciplines are required to support successful GIS operations. Figure 12-2 provides an overview of functional responsibilities required to support enterprise GIS operations.

Figure 12-2
GIS Functional Responsibilities

STAFF	ROLE	RESPONSIBILITIES
GIS Technician	GIS technical lead	Technical leadership to GIS users Executive secretary of the Technical Coordinating Committee Project implementation services Interdepartment technical coordination GIS technical support services Coordinate project work for departments and outside agencies Provide user training services Perform troubleshooting on custom application problems
GIS Manager	Enterprise GIS Operations Management	Provide planning and direction for GIS growth to serve multiple departments Chairman of the Technical Coordinating Committee Provide overall management for all GIS implementation tasks Manage setting of priorities for database and applications development Act as liaison to other departments and outside agencies Provide overall management for all contracted work
Geodatabase Administrator	Data and geodatabase manager	ArcSDE installation and software upgrades Ownership and management of ArcSDE and geodatabase schema objects Manages the ArcSDE service Compresses a geodatabase Data model design Data validation and quality - topology, domains Version management Spatial data management - data loading, spatial index tuning, DBMS statistics
System administrator	System Manager	Hardware procurement, installation and configuration Network infrastructure - installation and configuration System performance tuning / troubleshooting Operating System System backup and recovery Related software (upgrades and service packs)
Database administrator (DBA)	DBMS manager	Database configuration Data model implementation Database security Performance tuning Data backup and recovery Data replication DBMS software upgrades and service packs
Web Administrator	Web Services Administration	Web services development and maintenance Web services test and system performance tuning Web Security and Configuration Strategies
Web Developer	Web Services Development	Web services applicatin development JAVA or .NET Web Application Programming Experience Web services application and system performance tuning
Application Programmer	Desktop application development and support	Develop and enforce programming standards Desktop application development and code maintenance Desktop client application support Desktop application performance tuning

The complexity of these responsibilities will vary with the size and extent of each individual GIS implementation, although every organization will need some level of support and expertise in each of these areas.

12.2 Building Qualified Staff

Training is available to help develop qualified staff and support GIS user productivity. Organizations should make sure their teams receive required training. Figure 12-3 provides an outline of recommended ESRI training courses established to support GIS staff.

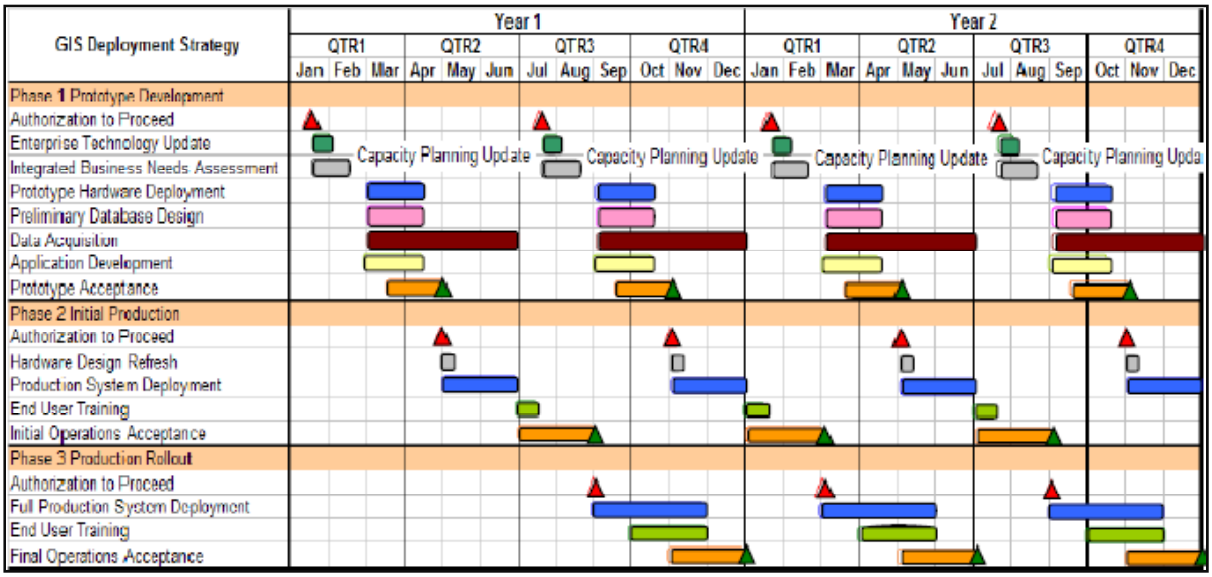
Figure 12-3
Training Opportunities

GIS Training Courses	GIS Technician	GIS Manager	Geodatabase Administrator	System Administrator	Web Administrator	Web Developer	Application Programmer
Advanced Analysis with ArcGIS	●						
ArcGIS Server Enterprise Configuration and Tuning for Oracle			●				
ArcGIS Server Enterprise Configuration and Tuning for SQL Server			●				
Building Geodatabases	●	○	●				○
Cartography with ArcGIS	○						
Creating and Editing Parcels with ArcGIS							
Data Management in the Multiuser Geodatabase			●				
Data Production and Editing Techniques	●						
Data Production with PLTS							
Developing Applications with ArcGIS Server in the Microsoft .NET Framework						●	○
Developing Applications with ArcGIS Server Using the Java Platform						●	○
Extending ArcGIS Desktop Applications							●
Geodatabase Design Concepts		○					
Hydrologic and Hydraulic Analyses Using ArcGIS							
Introduction to ArcGIS Business Analyst							
Introduction to ArcGIS I or Learning ArcGIS Desktop (Virtual Campus)	●	●	●				●
Introduction to ArcGIS II	●	○					○
Introduction to ArcGIS Server		○		○	●	●	
Introduction to Geoprocessing Scripts Using Python	●						●
Introduction to Programming ArcObjects Using the Microsoft .NET Framework						●	●
Introduction to Programming ArcObjects with VBA							●
Introduction to the Multiuser Geodatabase	○	○	○	○			
Learning GIS Using ArcGIS Desktop		●		●	●	●	
Managing Cartographic Data in the Geodatabase							
Managing Editing Workflows in a Multiuser Geodatabase	○		○				
QA/QC for GIS Data	●						
System Architecture Design Strategies	○	●	○	●	○	○	○
Working with ArcGIS Network Analyst	○						
Working with ArcGIS Schematics	○						
Working with ArcGIS Spatial Analyst	○						
Writing Advanced Geoprocessing Scripts Using Python	○						●

12.3 System Architecture Deployment Strategy

Normally planning is the first step in supporting a successful system deployment. A system design team should review current GIS and hardware system technology, review user requirements, and establish a system architecture design based on user workflow needs. A deployment schedule, as shown in Figure 12-4, should be developed to identify overall implementation objectives.

Figure 12-4
GIS System Deployment Strategy



Phased implementation strategies can significantly reduce implementation risk. Computer technology continues to evolve at a remarkable pace. Integration standards are constantly changing with technology and, at times, may not be ready to support immediate system deployment needs. New ideas are introduced into the market place every day, and a relatively small number of these ideas develop into dependable long-term product solutions. The following best practices are recommended to support a successful enterprise GIS implementation.

Pilot Phase
- Represent all critical hardware components planned for the final system solution.
- Use proven low-risk technical solutions to support full implementation.
- Include test efforts to reduce uncertainty and implementation risk.
- Qualify hardware solutions for initial production phase.

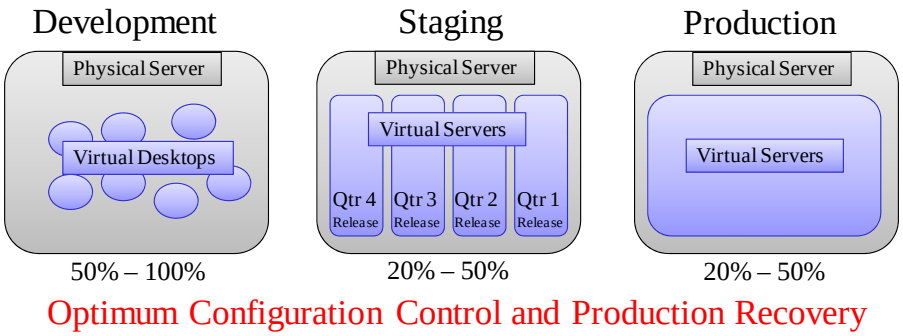
Initial Production Phase
- Do not begin until final acceptance of pilot phase.
- Deploy initial production environment.
- Use technical solutions qualified during the pilot phase.
- Demonstrate early success and payoff of the GIS solution.
- Validate organizational readiness and support capabilities.
- Validate initial training programs and user operations.
- Qualify advanced solutions for final implementation.

Final Implementation Phase
- Do not begin until final acceptance of initial production phase.
- Plan a phased rollout with reasonable slack for resolving problems.
- Use technical solutions qualified during previous phases.
- Prioritize rollout timelines to support early success.

12.3.1 Virtual Desktop and Server Technology

Virtual server technology is continuing to mature and reduce the cost of managing a rapidly changing IT environment. Figure 12-5 identifies a deployment strategy taking advantage of virtual server technology. Many GIS operations are being deployed into virtual server environments. Many ESRI development and testing operations are currently supported in virtual desktop or virtual server environments. Vendors are improving management and performance monitoring of virtual server environments, and it is becoming more practical to manage and deploy production environments in virtual server deployments.

Figure 12-5
Virtual Server Deployment Strategies



Platform virtualization technology provides IT managers with a way to abstract the installed platform software environment from the physical platform hardware. There are two fundamental levels of virtualization, one being virtual desktop environments hosted within a physical platform operating system that interfaces with the physical platform hardware and the other being a virtual server environment hosted on a hypervisor layer that interfaces the virtual servers with the assigned physical platform hardware. In both cases, the virtual desktop or server contains its own dedicated operating system and software install separate from other virtual systems deployed on the same hardware.

There are many recognized benefits and some potential disadvantages with virtual desktop/server deployments. The benefits include faster provisioning times, physical server platform consolidation, fast recovery from system failures, simplified production delivery and recovery, and optimum configuration control. All of these benefits directly contribute to lower overall systems management costs and a more stable operating environment. The disadvantages include additional software cost and some performance overhead. There are also some functional limitations (limited access to hardware graphic cards and performance monitoring software) which in many cases can be managed with the proper deployment selections. The real need for more rapid production update deployment schedules (to keep pace with technology change) which also must be coupled with more stable production deployments (reduced production downtime and more rapid failure recover) drive the need for virtualized platform environments - one solution that addresses some real IT management needs.

The potential disadvantages can be managed by proper deployment strategies. The performance overhead for virtual desktop environments is much higher than for server environments, and for this reason virtual desktops are normally limited to software development operations where performance and scalability is not a critical factor. Server consolidation benefits can be leveraged in a Staging environment, were multiple release candidates can undergo test and validation in preparation for production deployment. Production deployment can benefit from deploying an existing virtual server install (Staging configuration that has completed final test and acceptance) to a higher capacity production physical server by simply moving the Staging server release to the production platform. Also, if there are production failures identified after deployment it is a simple process to move the production environment back to the previous release. Virtual server software is available to accomplish these provisioning tasks during live operations with no production downtime.

Virtual server performance impacts will depend on the workflow environment, and can vary between 20 - 50

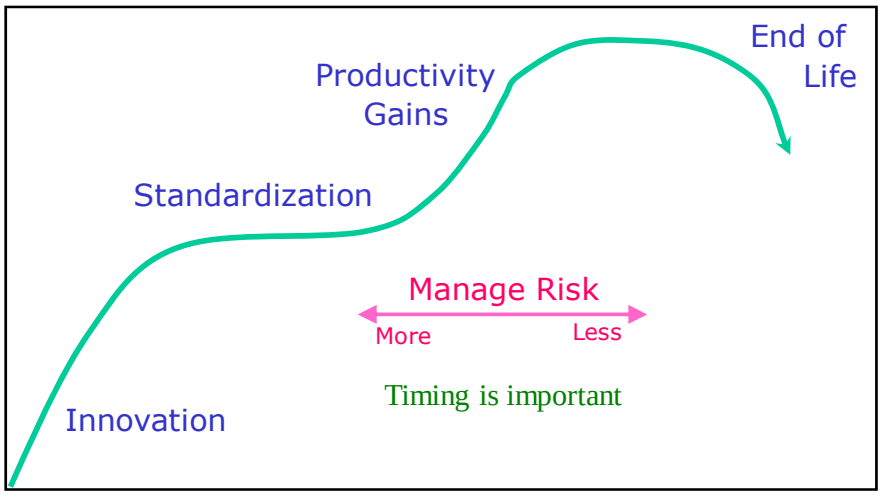
percent (and sometimes higher) in the most efficient virtual server configurations. Hardware platform performance has improved over 70 percent within the past year, which more than overcomes the virtual server processing overhead. The new servers also provide higher capacity (discussed in chapter 9), which opens the door wider for server consolidation benefits.

Virtual server deployments appear to be moving to mainstream IT production environment. The big question is no longer whether it makes sense to deploy on virtual servers, but rather when and which software vendor virtualization solution will provide the highest return on investment.

12.3.2 Technology Life Cycle

Figure 12-6 provides an overview of the technology life cycle, from initial introduction of a new idea (product innovation) through end of life. Technology is changing faster every year, and managing technology change within a production environment is a challenge for GIS managers and IT administrators. We are seeing an increasing number of new ideas introduced into the marketplace, with each idea promising improved user productivity and simplified system administration. These new ideas must integrate with existing systems that are constantly changing, and initial implementation can be painful. Some of these ideas do deliver on their promises, and in time they can provide significant productivity advantages and reduce overall cost of administration. Before long a new idea comes along and organizations move on to new frontiers, leaving their legacy systems behind.

Figure 12-6
Technology Life Cycle



Selecting the right technology at the right time can optimize business performance. Introducing new technology before it is ready for prime time can reduce productivity and increase implementation cost. Delaying too long can result in missed opportunities. Getting the timing right promotes success.

12.4 System Testing

Conducting proper testing at the right time can contribute to implementation success. Functional component and system integration testing should be conducted for new technology during prototype development and before introduction into production. The primary concern during this phase is to make sure everything works. Performance targets established during the initial system design can be evaluated during early testing, paying close attention to map display performance and layer complexity (see Chapter 8 Software Performance). This is the opportunity to evaluate workflow functions and reduce processing overhead.

Enterprise system environments are becoming more complex. Testing should be conducted in a production software environment. Configuration challenges such as firewall access, security, and high availability should be configured and tested before deployment. Figure 12-7 shows a typical data center architecture for emerging enterprise GIS operations. Development and test environments should be established to represent the complete software configuration for each production release cycle.

Figure 12-7
Data Center Architecture

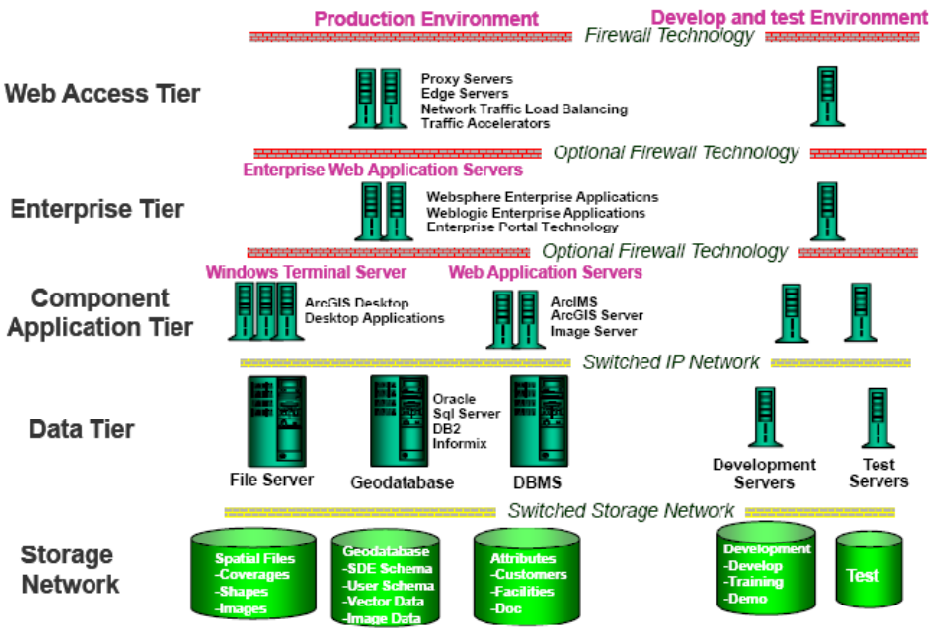


Figure 12-8 identifies best practices for planning and conducting functional system testing.

Figure 12-8
Functional System Testing Best Practices

- **Test Planning**
  - Complete a risk analysis: Identify functionality that requires testing.
  - Identify test objectives and establish configuration control plan.
  - Identify test hardware/software configuration.
  - Develop test procedures.
- **Test Implementation**
  - Identify implementation team and establish implementation schedule.
  - Order hardware and software and publish installation plan.
  - Conduct test plan and validate functional acceptance.
  - Collect test performance parameters (CPU, memory, network traffic, etc.).
- **Test Results and Documentation**
  - Document the results of the testing.
    - Include specific hardware/software/network components that were tested.
    - Include installation and test procedures that were followed, test anomalies, and final resolution.
    - Complete test compliance matrix identifying validation of functional requirements.
  - Publish the test results for reference during system implementation.

Complete prototype integration testing before production deployment.
Test in a production environment (configuration control).
Document functional requirements and test procedures.

A test plan should be developed to identify test requirements, establish configuration control (software versions, operating system environment), and provide test procedures. Testing should be completed before production deployment. Testing should be conducted using the software versions and operating system that will be deployed in the production environment.

Figure 12-9 identifies some cautions and warnings associated with system performance testing.

Figure 12-9
Performance Testing Pitfalls

- **False Sense of Security**
  - People tend to accept test results over analysis.
  - Problem 1: Test results are a function of input parameters often not understood.
  - Problem 2: System bottlenecks in testing can generate false conclusions.
- **Simulated Load Testing May Not Represent Real World**
  - Load generation seldom represents actual user environments.
  - Relationships between load generation and real world are seldom understood.
  - Several system configuration variables can contribute to test anomalies.
- **Performance Testing Best Practices (Scientific Method)**
  - Model system components and response parameters.
  - Models should match real-world user experience.
  - Predict test results from models (hypothesis) before conducting testing.
  - Evaluate test results against models and original hypothesis.
  - Update models and hypothesis, repeat testing until reaching consensus.

TEST ONLY WHEN YOU THINK YOU KNOW THE ANSWER.
TESTING ONLY CONFIRMS WHAT YOU ALREADY KNOW.
TESTING DOES NOT ANSWER WHAT YOU DON'T KNOW.

Performance testing can be expensive and the results misleading. Normally initial system deployments need to be tuned and optimized to achieve final performance goals. Often system performance bottlenecks are identified and resolved during initial deployment. Early application development focuses primarily on functional requirements, and performance tuning is not complete until the final release. Actual user workflow environments are difficult to simulate, and test environments seldom replicate normal enterprise operations.

The scientific method introduced with grade school science fair projects provides the fundamental best practices that directly apply to system performance testing. Performance testing should only be conducted to validate a hypothesis (something you think you know). The primary objective of a performance test is to validate the hypothesis (confirm what you know). The test is a success only if it proves the hypothesis (testing does not teach you what you don't know).

Initial performance testing results often fail to support the test hypothesis. With further analysis and investigation, test bottlenecks and/or improper assumptions are identified that change the test results. Performance testing is only successful if it validates the test hypothesis.

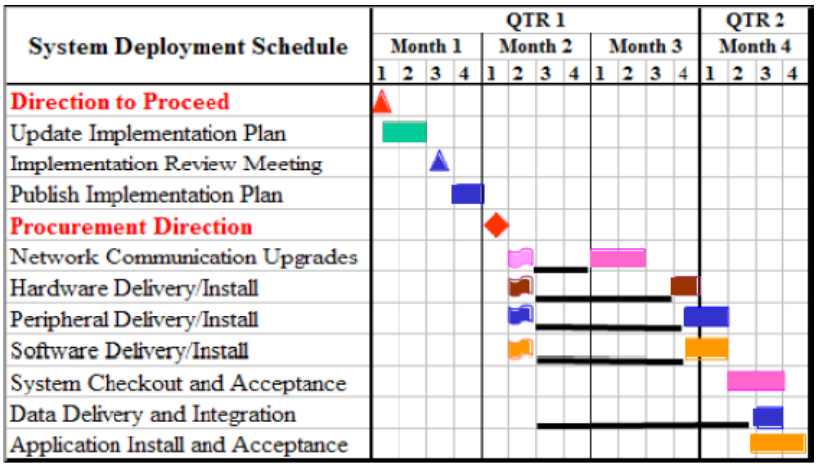
System performance testing is best conducted during the initial production deployment. During this phase, real users doing real workflow can generate a real user environment. Critical system components should be monitored during the initial deployment to identify processing bottlenecks and resolve system conflicts. Compliance with design performance targets can be validated during initial user operations. Initial deployment acceptance should include validation that user workflow performance goals are met.

12.5 Management

Basic project management practices promote implementation success. Project teams should be established, individuals should be assigned specific responsibilities, a task plan should be developed to support implementation planning, a configuration control plan and change control process should be established, and an implementation schedule should be published to support project deployment milestones.

A system architecture design can provide the framework for establishing an implementation plan. The implementation plan should be developed after final selection of the hardware vendor solution. Figure 12-10 provides a typical system deployment schedule. Specific decision milestones should be included in the schedule and each major task effort clearly identified. An implementation project manager should be assigned to make sure all tasks are well-defined, and every participant has a clear understanding of his/her responsibilities. A clear set of acceptance criteria should be developed for each implementation task and a formal acceptance process followed to ensure integration issues are identified and resolved at the earliest opportunity.

Figure 12-10
Systems Integration Management



The capacity planning tool can be used by project managers to validate performance targets are met throughout deployment. Peak users can be identified for different project milestones, and server platform utilization can be reported at each milestone to demonstrate performance goals are met.

Figure 12-11 shows some of the ArcGIS Server and system tools used for performance monitoring. Additional performance validation tools were discussing in Chapter 8 (ArcGIS Server 9.3.1. map optimization tools and the mxdperfstat performance monitoring tools). The performance terms and relationships discussed in Chapter 7, particularly the relationship between throughput (peak users or peak transaction rates) and utilization (server CPU or network bandwidth utilization), can identify if the deployed solution is performing within the initial project performance milestone.

Figure 12-11
Performance Validation Testing

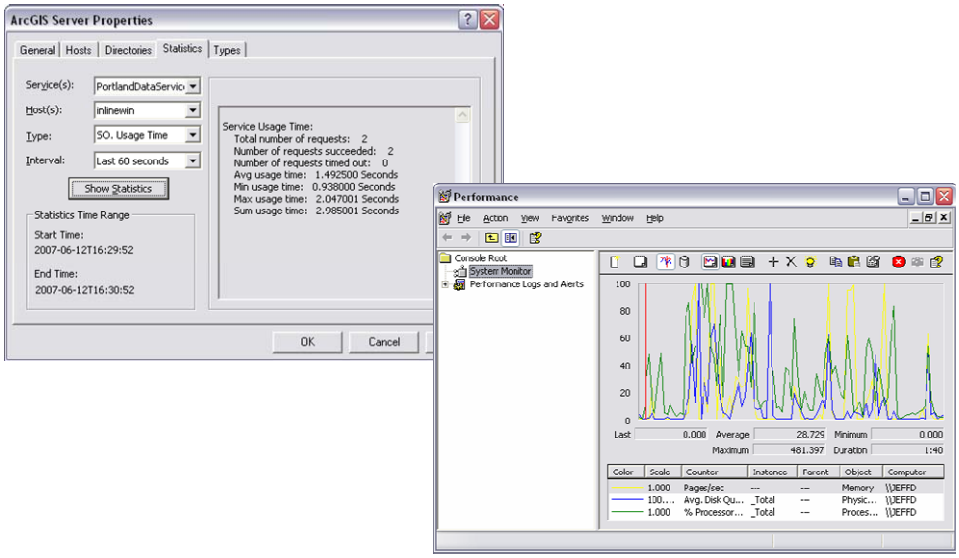
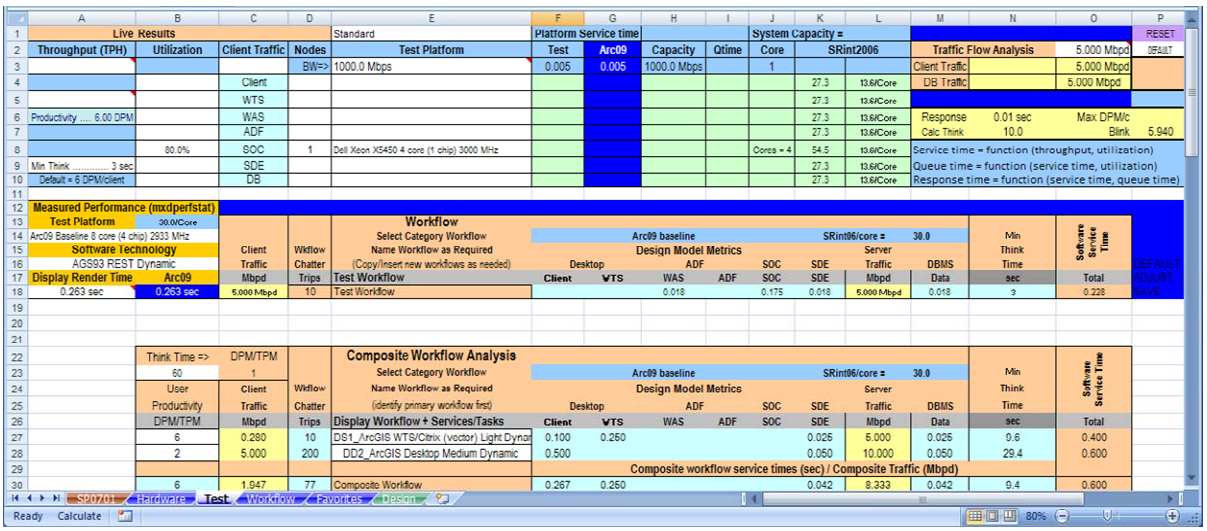


Figure 12-12 shows how the capacity planning tool (CPT) can be used by project managers to establish specific project performance milestones and measure compliance. The CPT test tab is designed to translate throughput and utilization measurements to workflow service times that can be compared directly to the initial project workflow performance targets. Monitoring progress in meeting performance milestones can reduce deployment risk and ensure project delivery success.

Figure 12-12
Performance Validation Testing



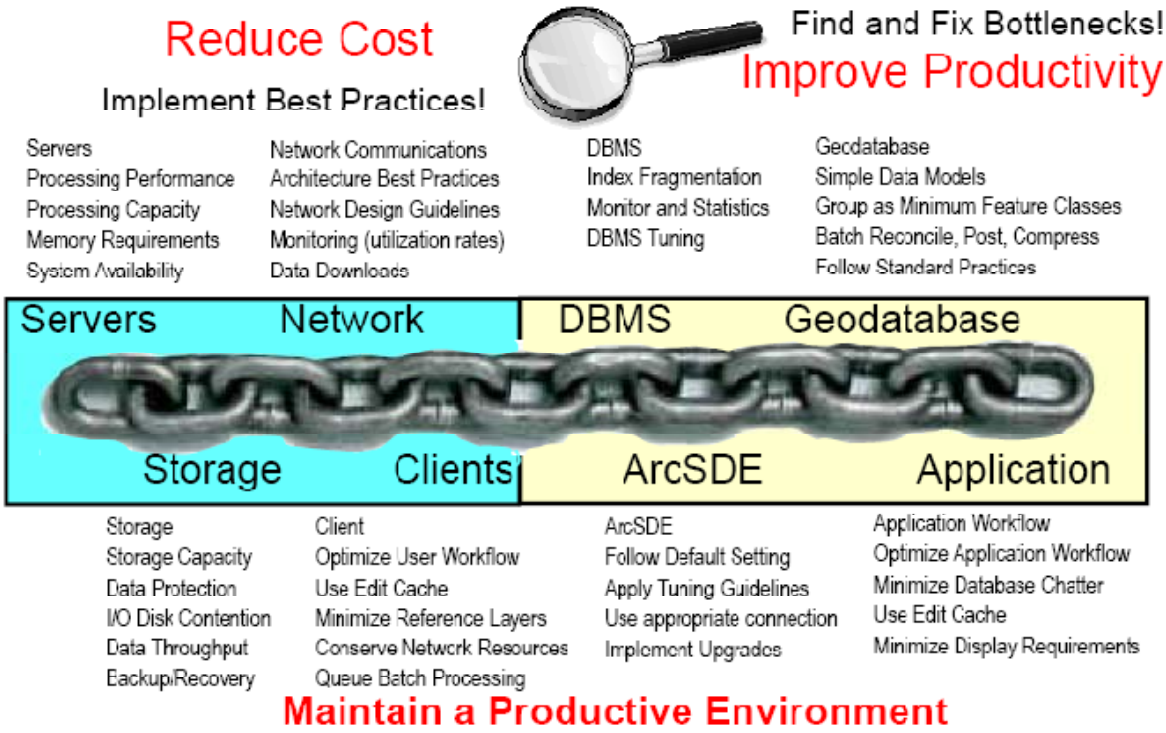
When performance issues are identified early in deployment, proper adjustments can be made before impacting production workflow productivity (simpler map displays - less layers or generalize layers with large number of

features, reduce number of current batch jobs during peak system loads, evaluate preprocessing alternatives (map cache, generalized geodatabase layers, etc). Chapter 8 identified a variety of ways to improve GIS display performance. Identifying and resolving performance issues before they become production level performance problems will promote deployment success.

12.6 System Tuning

System tuning is a critical part of final system integration and deployment. Initial user requirement planning is the first opportunity to begin performance tuning. Heavy batch processing efforts should be separated from interactive user workflows and supported through a separate batch process queue. System backups and heavy processing workloads should be planned during off-peak workflow periods. System component performance metrics should be monitored on a periodic basis particularly during peak workflow periods to identify performance bottlenecks and address system deficiencies. Figure 12-31 provides an overview of the components supporting an enterprise GIS environment. Any component has the potential to introduce a weak link in the overall system performance equation.

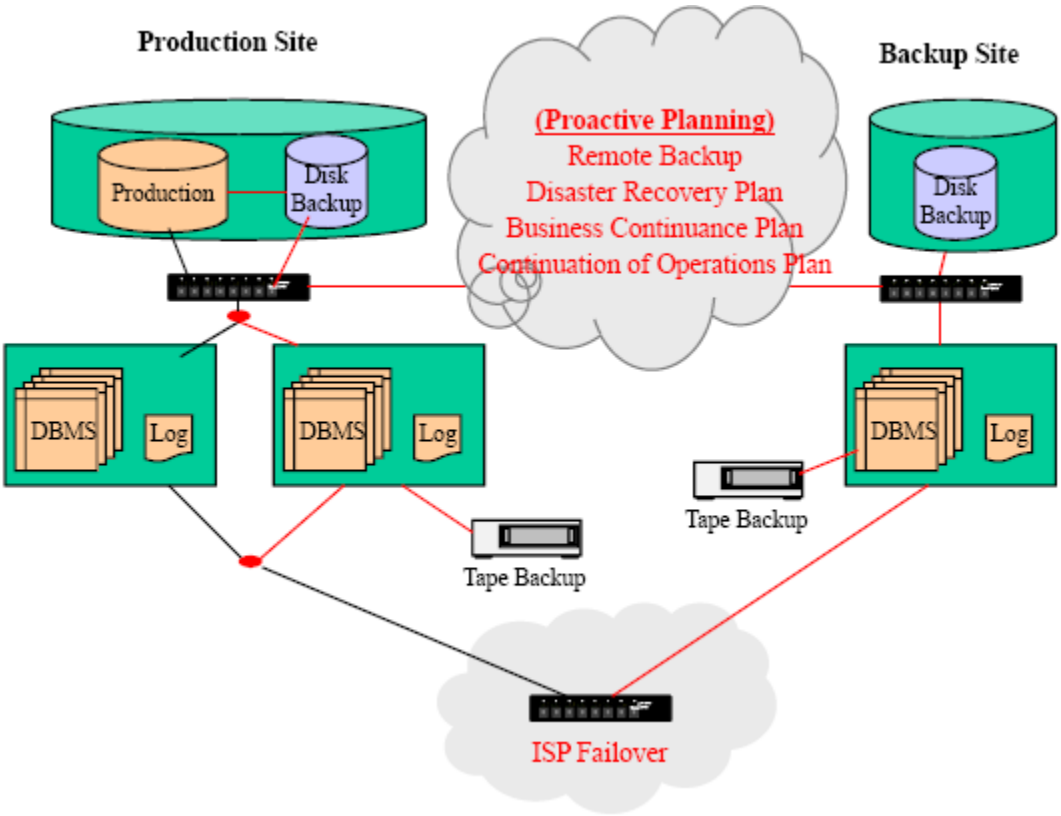
Figure 12-13
System Performance Factors



12.7 Business Continuance Plan

Every organization should carefully assess the potential failure scenarios within its system environment and protect critical business resources against such failures. Enterprise GIS environments require a significant investment in GIS data resources. These data resources must be protected in the event of a system failure or physical disaster. Business recovery plans should be developed to support all potential failure scenarios. Figure 12-14 provides an overview of the different system backup strategies. A business continuance plan should be developed to address specific organizational needs in the event of a system failure or disaster recovery.

Figure 12-14
Plan for Business Continuance

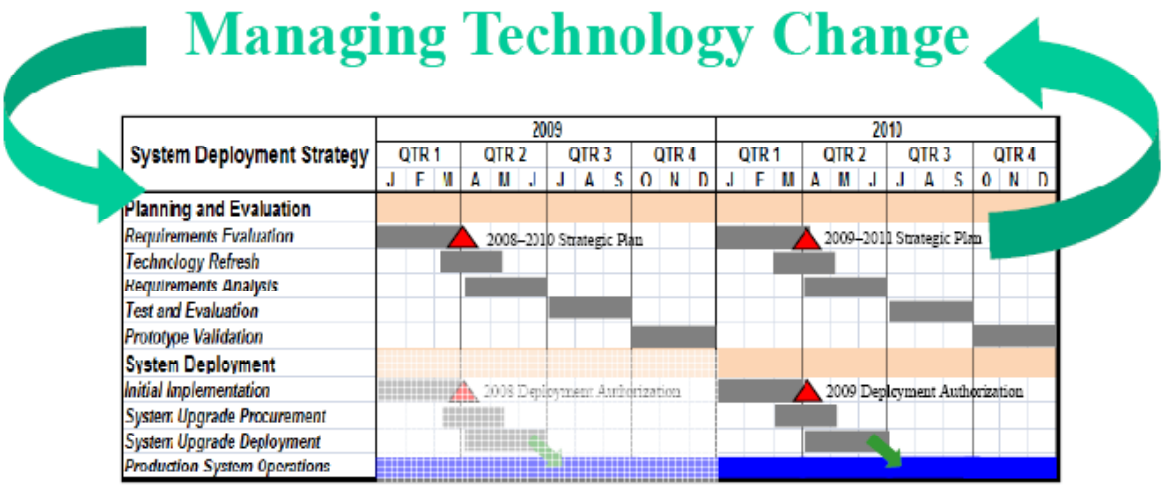


12.8 Managing Technology Change

Enterprise GIS operations require a combination of strategic planning and a continued investment in technology. Technology is changing very rapidly, and organizations that fail to manage this change fall behind in productivity and operational cost management. Managing technology change is a major IT challenge.

Enterprise operations should include a periodic cycle of coordinated system deployment. The planning and technology evaluation should occur one periodic cycle ahead of technology deployment, and these efforts should be coordinated to support operational deployment needs. Figure 12-15 identifies a conceptual system architecture planning and deployment strategy for technology change management.

Figure 12-15
System Architecture Design Strategic Planning



Planning and Evaluation: Planning activities should be established in a periodic cycle, coordinated to support the organization's operational and budget planning needs. Strategic plans should be updated to support a multiyear deployment strategy and published periodically (normally on an annual cycle).

The planning and evaluation process should include a requirements evaluation (strategic plan update), technology refresh (training and research), requirements analysis (process and requirements review), test and evaluation (evaluate new technology alternatives), and prototype validation (pilot test programs). Efforts should be scheduled to support the annual system deployment upgrade cycle.

System Deployment: Operational system upgrades should be planned on a periodic cycle, scheduled to implement validated operational enhancements from the planning and evaluation program. System deployment phases should include initial implementation (implementing changes in an operational test environment) to support deployment authorization. The program should also include planned schedules for new technology procurement and deployment on a periodic schedule (normally on an annual cycle). All production system upgrades should be planned and scheduled with full support for ongoing operations.

12.9 Conclusion

Successful implementation depends on a good solid design, appropriate hardware and software product selection, successful systems integration, and careful incremental evaluation during installation. A phased approach to implementation reduces project risk and promotes success, providing the opportunity for early success and flexibility to incorporate new technology at low risk prior to final system delivery. Guidelines are available to support a successful system design, even for large complex systems. Final purchase decisions are influenced by both operational requirements and budget limitations, introducing unique challenges for system design. Good leadership, qualified staff, and proven standard practices support successful deployments.

Attachment A

System Architecture Design Strategies Course

Attachment A—System Architecture Design Strategies

Overview

This three-day course introduces a proven system architecture design methodology for developing successful GIS design and implementation techniques. This methodology was developed and tested during years of successful ESRI system design consulting efforts. The objective of this course is to help users improve performance of existing and future GIS environments by using this methodology . The system design models, configuration guidelines, and capacity planning tools provide class participants with a proven path to successful GIS solutions. Lectures and hands-on exercises help those responsible for GIS system architecture decisions select a system design that will support GIS user performance requirements.

Audience

System Architecture Design Strategies will appeal to those in charge of developing and maintaining hardware or software systems designs and to those in the business of supporting software or application development and technical marketing for system design, testing, and configuration of client solutions. It also provides an excellent conceptual framework for anyone in the position of supporting and securing GIS hardware or software solutions. Senior architecture consultants will benefit from the GIS design methodology presented, while GIS managers will come away with a better understanding of system architecture and hardware selection criteria.

Goals

Understand the relationships that support successful GIS solutions.
Learn where ESRI software solutions fit in an evolving GIS environment.
Learn how to integrate ESRI software in a distributed enterprise environment.
Learn how to provide high-performance remote user performance.
Understand the prerequisites and recommendations before selecting hardware solutions.
Learn how to summarize user requirements to support system architecture design.
Identify system components that contribute to application performance.
Learn how to select an optimum enterprise design solution.
Learn how to identify relative performance of Windows and UNIX platforms.
Learn how to use capacity planning tools for selecting appropriate hardware technology.
Learn practical design guidelines for network communications.
Understand the process for conducting a system design review and selecting a distributed GIS hardware solution.
Learn how to use the Capacity Planning Tool

Topics covered

System design process
GIS software solutions
Network communications
GIS product architecture
Enterprise security
Data administration
Performance Fundamentals
Software Performance
Platform Performance
Capacity Planning Tool
Completing the System Design
System implementation

Prerequisites and recommendations

Registrants should have an interest in understanding GIS product architecture and how today's computer technologies can support successful GIS solutions.

Class available at ESRI training centers.

Fixed price for on-site training available upon request.

Course Outline *System Architecture Design Strategies*

DAY ONE

System design process

What is system architecture design?
Why is system design important?
Integrated Business Needs Assessment
System design support efforts

GIS software solutions

GIS software / architecture evolution
ESRI product family
Central/Distributed Architecture
Federated / Service-oriented architecture
Software Technology Selection

Network communications

GIS network protocols
Network performance guidelines
System performance latency
Network configuration guidelines
Network design planning factors

GIS product architecture

ArcGIS system architecture overview
ArcSDE technology overview
ArcGIS Desktop workstation configurations
ArcGIS Desktop terminal server configurations
ArcGIS Web services configurations

DAY TWO

Enterprise Security

Levels of Security
Security in-depth strategy
Risk Management Frameworks
Current Security Trends
Firewall Configuration Strategies

Data Administration

Ways to store spatial data
Ways to protect spatial data
Ways to backup spatial data
Ways to move spatial data
Geodatabase transition (ETL migration)
Database / Storage replication services
Geodatabase replication alternatives
Distributed geodatabase architecture

Performance fundamentals

Understanding the technology
Performance baseline history
Performance terms and relationships
Computing platform service times
What is system performance / response time?
Platform capacity planning

Software Performance

Optimum Map Display makes a difference
Number of display layers makes a difference
Use ESRI Optimized lines and polygons
Select the right map image type
Geodatabase / Web services performance tips
ArcGIS Server service configuration guidelines
Select the right software technology
Take advantage of caching

DAY THREE

Platform Performance

User performance expectations
Platform performance history
How to address platform performance differences
2008 supported platform performance
Workstation platform recommendations
Windows terminal server platform sizing
ArcSDE Geodatabase platform sizing
ArcGIS Image Server platform sizing
ArcGIS Server platform sizing (includes ArcIMS)

Configuration Planning Tool

Capacity planning tool overview
Requirements analysis module
Platform configuration Module
Workflow tab overview
Hardware tab overview
Vendor published hardware benchmarks

Completing the System Design

GIS Business Planning
GIS User Needs Assessment
User needs templates
City of Rome case study
- City of Portland Class Exercise

System Implementation

GIS Staffing
Phase system deployment
System Testing Overview
Performance targets / milestone validation
System performance tuning
Managing technology change

Students received a copy of the Excel workbook version of the capacity planning tools used during the class.

Attachment B

List of Figures

Attachment B—List of Figures

Figure **Page**

Attachment C

Acronyms

Attachment C—Acronyms

ADF	**Application Development Framework**
AML	**ARC Macro Language**
API	**application program interface**
ASC	**application server connect**
ASP	**application service provider**
CF	**ColdFusion**
CIFS	**Common Internet File Services**
COTS	**commercial off-the-shelf**
DC	**direct connect**
DCOM	**Distributed Component Object Model**
DMZ	**demilitarized zone**
DPM	**displays per minute**
DS	**data source**
EDN	**ESRI Developer Network**
ETL	**Extract-Translate-Load**
GB	**gigabyte**
Gbps	**gigabits per second**
g.net	**(regional) geography network**
GIS	**geographic information system**
HTTP	**Hypertext Transfer Protocol**
I/O	**input/output**
ICA	**independent computing architecture**
IP	**Internet Protocol**
IPD	**Information Product**
ISP	**Internet service provider**
JBOD	**just a bunch of disks**
Kb	**kilobit**
KB	**kilobyte**
Kbps	**kilobits per second**
LAN	**local area network**
MAC	**Media Access Control**
Mb	**megabit**
MB	**megabyte**
Mbps	**megabits per second**
NFS	**network file server**
NIC	**Network Interface Card**
ODBC	**Open Database Connectivity**
RAID	**redundant array of independent disks**
RDP	**remote desktop protocol**
SAN	**storage area network**
SDE	**Spatial Database Engine**

SM	**service manager**
SMP	**Symmetrical Multiple Processor**
SOA	**service-oriented architecture**
SOAP	**Simple Object Access Protocol**
SOC	**Server Object Containers**
SOM	**Service Object Manager**
SPEC	**Standard Performance Evaluation Corporation**
SS	**spatial services**
SSL	**Secure Socket Layer**
SSO	**single sign on**
TCP	**Transmission Control Protocol**
TPH	**Transactions per hour**
WA	**Web applications**
WAN	**wide area network**
WSE	**Web services extensions**
WTS	**Windows Terminal Server**
XML	**Extensible Markup Language**