# Scaling the N-Tier Architecture

*Solaris™ Infrastructure Products and Architecture*

Sun Microsystems, Inc.
901 San Antonio Road
Palo Alto, CA 94303
1 (800) 786.7638
1.512.434.1511

Please
Recycle

# Contents

CHAPTER 1

# Scaling the N-Tier Architecture

The Internet continues to grow in importance, and more businesses are globalizing than ever before. Rapid business growth, increased management costs, implementation complexities, rapid pace of deployment, and frequent application change are elevating the cost of providing a positive customer experience. Organizations must continuously provide high quality of service (QoS) around the world to gain a competitive advantage and foster customer loyalty. After all, the competition is only a click away. As a result, infrastructure scalability, manageability, and availability are paramount if increased service levels are to be achieved. Clearly the pervasiveness and ubiquity of the Internet demands increased flexibility and agility, and enterprise information infrastructures must be retooled for this new competitive economy.

With the use of the Internet and corporate intranets growing at a phenomenal pace, enterprises must position themselves for growth and agility to handle increasing numbers of users, additional services, and more challenging workloads. Rapidly changing business requirements are forcing information systems to interoperate with these corporate and external resources in an interactive, reliable, and secure manner, while maintaining the flexibility to quickly adapt to rapidly changing business environments. The IT infrastructure is critical to enterprise competitiveness, having moved from an internal support function to a business enabler and vehicle for profit. These demands are pushing the limits of existing information infrastructures. To be competitive, organizations must find solutions that will safeguard existing infrastructure investments, yet deploy modern capabilities to provide the flexibility, predictability, and availability needed for success.

This paper provides insight into scaling applications and services in the most popular IT paradigm — the N-tier architecture. The challenges faced by data centers are described, as well as what enterprises must do in order to win in the Net economy. Finally, Sun™ solutions are presented that enable data centers to design and deploy scalable infrastructures that facilitate rapid application deployment, increase service levels, and reduce or contain the risks and costs of providing better customer service.

CHAPTER 2

# Characterization of the N-Tier Architecture

The traditional two-tier, client/server model requires clustering and disaster recovery to ensure resiliency. While the use of fewer nodes simplifies manageability, change management is difficult because it requires servers to be taken offline for repair, upgrades, and new application deployments. In addition, the deployment of new applications and enhancements is complex and time consuming in fat-client environments, resulting in increased risk and reduced availability. Only average resource utilization rates are possible, and the ability to reactively scale resources to meet peak time and seasonal demand is virtually impossible.

The inherent shortcomings of the two-tier model gave rise to N-tier architectures. To mitigate the limitations of traditional client/server environments, the N-tier architecture was designed to enable applications to be distributed as needs dictate. An N-tier application architecture is characterized by the functional decomposition of applications, service components, and their distributed deployment, providing improved scalability, availability, manageability, and resource utilization (Figure 1). A tier is a functionally separated hardware and software component that performs a specific function.

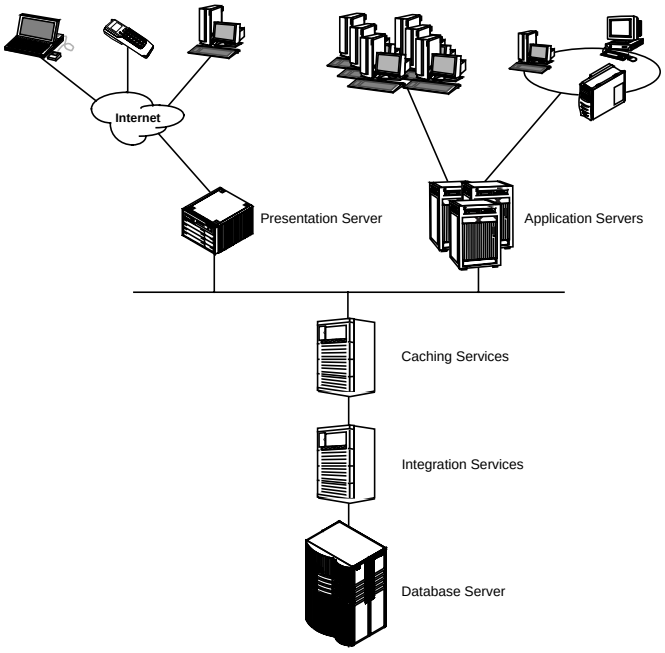


**FIGURE 1** N-tier architectures are characterized by multitiered, server-centric applications.

Because each tier can be managed or scaled independently, flexibility is increased in the IT infrastructure that employs it. Communication between tiers is accomplished through standard protocols such as HTTP, RMI, and XML. All functional components, such as persistent storage, management of clients, and the marshalling of data from many stores, are separated. This componentization is key, with each component ascribed to a tier, providing an abstraction for application architecture, manageability, and flexibility. As a result, individual components can scale and be made highly available with ease.

Typically, N-tier architectural platforms place each service or group of services on a separate server, enabling systems to be divided into easily scalable components. As a result, applications can exploit this modular software architecture approach to increase scalability and availability. Three fundamental aspects must be understood to ensure maximum benefit: horizontal, vertical, and diagonal scaling.

# Horizontal Scaling

Horizontal scaling is characterized by rapid application change and simplified change management. With multiple servers in the horizontal tier, changes can be deployed incrementally, enabling applications or services to be replicated quickly to multiple servers in a controlled manner. In addition, minimal service disruption occurs, even in the face of hardware and software errors. If one server ceases to function or is taken offline, then the remaining servers in the horizontal set of servers continue to provide service. Hence, horizontal scaling provides inherent resiliency and predictable performance. Resource utilization is improved through a load balancer, which strives to spread the load over the horizontally scaled set of servers. Should a single server go offline, the load balancer continues to send traffic to the remaining servers, providing inherent system resiliency.

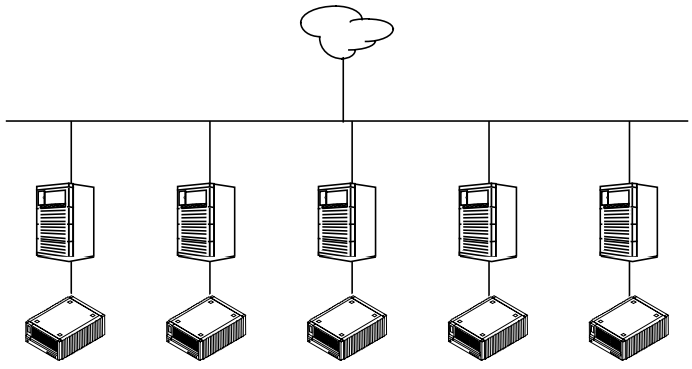

**FIGURE 2** Horizontal scaling employs multiple servers within a tier.

Horizontal scaling can help organizations achieve the right level of resilience. For example, organizations that can handle 10-percent session loss at any given time should scale to ten nodes. This can be done when application workload or data is easily partitioned or abstracted to another tier. When this is impossible, applications should be clustered in the traditional fashion. Horizontal scaling works best in situations where there is minimal state information and replication is easy.

# Vertical Scaling

With vertical scaling, services are scaled within the system — resources such as CPUs, memory, and storage can be incrementally added to the server over time to increase scalability. Vertical scaling is characterized by slowly changing applications and is often necessary for data-intensive services such as databases, video servers, mail stores, and directories. This slow change is a consequence of many things, including database schema definition early in the life of a service, stable metatables, and the practicality of data replication. However, static design stability does not necessarily mean static content — data stores can change in size very rapidly. Stability and simple vertical scaling are driven by uniform access to data and the non-partitioning of data and its core management. As a result, items that cannot be partitioned must be close together to ensure latency and uniform access, resulting in increased scaling.

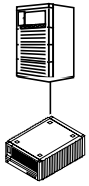

**FIGURE 3** Vertical scaling utilizes systems that can be scaled as needs dictate.

Once horizontal scaling has been deployed, organizations can employ vertical scaling to improve performance. To achieve better resource utilization, organizations can consolidate multiple instances of horizontally scaling applications onto a single server. Indeed, with the right tools, vertical scaling can provide greater benefits than horizontal scaling. Vertical scaling works best when replicating state information proves difficult due to cost, time, performance, or size considerations. While vertical scaling can be implemented at any time, and is simpler to manage in some cases, it may require clustering and disaster recovery to ensure availability. However, if data partitioning is not feasible, vertical scaling is required.

# Diagonal Scaling

Diagonal scaling is a combination of horizontal and vertical scaling. Each server in the group of horizontally scaled servers can be grown within the system, providing increased flexibility.

CHAPTER 3

# N-Tier Architecture Components

While there is no prescribed structure or set of tiers for an N-tier architecture, Figure 4illustrates a common approach. The Presentation, Application Server, and Data Tiers are standard in these environments. An additional tier, the Caching Service Tier, illustrates that tiers can be constructed to accommodate unique environments. This architecture is called an N-tier architecture because the Application Server Tier of one application can call upon the Application Server Tier of another application when the two share data. There is no limit to the amount of inter-application calling that can occur.

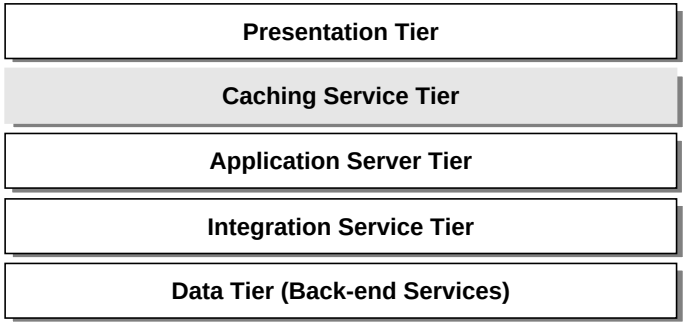


**FIGURE 1** A common N-tier architecture layered approach.

The Presentation Tier is typically a graphical user interface that displays data to users regardless of device type or location and enables information manipulation. In such environments, business logic can be moved to the Application Tier, drastically reducing the number of locations that implement the logic, and thereby simplifying management and control. As a result, only one Web interface is exposed to the client, eliminating the need for clients to run a given application, and ensuring interoperability. The use of thin clients simplifies deployment, enables rapid application change in the middle tiers, reduces deployment risks, and enables businesses to quickly respond to market opportunities.

Sun's Caching Service Tier helps provide fast access to server resources. It runs application proxies that forward requests from the client to the application server, as well as a service locator that finds needed services for clients. The use of one or more Caching Service Tiers that are local in the WAN enables Java™ applets to be cached locally, improving download performance. In addition, more efficient WAN bandwidth usage as well as improved access to printers, files, authentication, and access control is possible.

The Application Server Tier determines which data is to be stored and retrieved, and manipulates that data on behalf of the business logic.

The Integration Service Tier coordinates the passage of synchronous and asynchronous messages and transactions to and from heterogeneous database services.

The Data Tier stores application data in a persistent store, such as a relational database (RDB) or an object-oriented database (OODB). The Data Tier also stores commonly used business logic procedures to reduce the network traffic associated with repeated operations. This tier typically deploys vertical scaling. Consequently, the Application and Data Tiers are responsible for implementation of business rules, validation, and persistent storage, thereby reducing network query traffic and significantly improving performance. This performance increase is accomplished by caching SQL at the Data Tier rather than at the Application Tier.

Data Tier services generally provide system availability through a dedicated system where components can failover. This extra system may actively deliver additional services or passively wait for failover to occur.

CHAPTER 4

# N-Tier Architecture Challenges

While an N-tier application architecture provides virtually limitless scalability, the need to change or add new functionality presents challenges in several key areas. Large growth (from 10 to 100 times) makes capacity planning difficult. When applications exhaust available resources, provision must be made to borrow resources to support unexpected workloads. Manageability requires resource sharing, centralized management, and simplicity. Complexity forces organizations to maintain competitive service levels through a flexible architecture that enables reactive scalability to positively affect both service level and cost. These challenges have exposed the inability of traditional architectures to efficiently utilize existing IT infrastructures.

## Quality of Service

Every day, an increasing number of users depend on the Internet to conduct business and personal transactions. Moreover, organizations need to differentiate classes of users, account for user and group usage, and maximize service-level provisioning, including service availability, performance, and predictability. Consequently, the ability to provide predictable and differentiated quality of service must be designed into the platform infrastructure. What is needed is an infrastructure that supports a service-based application approach, featuring:

- An N-tier architecture
- Dynamic resource allocation
- Accounting and management
- Infrastructure management
- Heterogeneous legacy integration
- Cluster support
- Multiplatform Java technology
- Multilevel security model.

# Scalability

In the network economy, enterprises are constantly deploying new applications at Internet speed— and it is difficult to predict which ones will need to grow, or how rapidly. Service demand often increases without warning, and enterprises must quickly adapt to keep pace. This explosion of network services is creating larger and more unpredictable demand peaks. As a result, a scalable infrastructure designed to handle these conditions is key to success.

To be effective, organizations must plan for both proactive and reactive scalability. Proactive scaling enables organizations to anticipate increased demand and preallocate system capacity, while reactive scaling ensures that extra resources are available to handle sudden, unanticipated demand. Proactive scalability demands a strong architecture, metric gathering tools, sound data center practices, and a predictable scaling model. Reactive scalability requires an architecture that provides simple horizontal scaling, ensures seamless scaling at the back end, and enables rapid reallocation of resources as business priorities change. Even when it is possible to plan for demand, it may not be possible or cost-effective to proactively purchase additional hardware resources to support new or existing applications. As a result, organizations must find ways to create an infrastructure that enables resources to be applied where they are needed most.

Scalability is critical to companies needing to make the most cost-effective use of their computing resources, gracefully handle peak workloads, and grow their computing environment along with their business. Many applications, such as ERP, data warehousing, and some e-commerce applications, have been developed with scalability in mind. Others, including those in the burgeoning B2B marketplace, face unique application architecture challenges that require new and different strategies to achieve equivalent scalability. Applications must be written to take advantage of the architectural platform on which they run. As a result, organizations are finding they must either choose application software optimized for scalability, or plan for scalability in the development and implementation process. To overcome these and other limitations, applications must be designed to transparently take advantage of the underlying hardware and software architecture to ensure highly scalable services.

## Resource Utilization

The cost of providing increased service levels is reduced through improved resource utilization. Control must be provided through hardware and software to enable more than a single application to run on a single machine. Server consolidation is paramount to achieve increased return on investment from under-utilized resources. While mainframes often run at 80 to 90 percent of capacity, distributed systems typically run at only 15- to 25-percent utilization. Organizations are finding that allocation adjustments must be made to take full advantage of available resources. What is needed is a way to run multiple applications on a single server, giving each application a minimum level of service that is free from security and resource contention concerns. Further control could enable dynamic adjustment through management policies.

## Availability and Predictability

Today's businesses are information driven. The need to access and analyze corporate information in real time, update databases, perform trend analysis, provide high customer satisfaction, and operate in 24x7 environments is changing the demands placed on IT infrastructures. It is no longer sufficient for computing environments to simply provide increasing levels of capacity — they must also be available, reliable, and predictable in order to meet the requirements of both users and applications.

The unpredictable demands of the Internet, as well as traditional operations, are forcing organizations to ensure the data center is available. With competitors just a click away, services must always be available to customers and clients. Service disruption must be minimized, even during system upgrades and routine maintenance. Systems must be capable of being debugged, repaired, or patched on-line. Resources must be automatically, dynamically, and gracefully redirected in the event of a failure to ensure expected service levels are met. To be effective, organizations must deliver availability, predictability, and capacity through well-chosen infrastructure components and a reliable, scalable, manageable operating system platform (Table 1).

| | |
|---|---|
| **Availability** | • High-quality, redundant components with low failure rates<br>• Marshals redundant components to deliver seamless availability<br>• Provides an abstraction which simplifies the view of the components and reduces the risk of operator error<br>• Simple architectural model to reduce risk of unstable deployments |
| **Capacity** | • Single instance of an application – if properly architected and implemented<br>• Across multiple instances of an application through replication<br>• Combines single and multiple instances when a service has many components<br>• Good linear platform scalability |
| **Predictability** | • Predictable response times, scaling, and availability |

**TABLE 1** Infrastructure component strategies that ensure availability, capacity, and predictability.

Three factors that affect system availability: people, process, and product. While people and process typically account for 80 percent of a system's ability to remain available, only 20 percent originates from the system itself. It is important to remember that product manageability — which reduces operator error —affects both the people and process aspects of system availability. Increasing availability, however, requires disciplined procedures and processes that are consistently maintained. To further impact availability, infrastructure platforms must simplify management, deployment, and maintenance operations.

## Manageability

As the enterprise IT infrastructure scales, it inevitably becomes more complex. Unfortunately, increased complexity often renders the data center environment less capable of coping with rapid application change and business demand for services. Indeed, the effort required to manage resources tends to grow faster than the resources themselves. As a result, manageability has a direct impact on scalability and availability. To be effective, organizations must improve management efforts and simplify data center architecture. Toward this end, small and large businesses alike must centralize, simplify, and automate processes, as well as incorporate a management framework that improves platform architecture manageability.

CHAPTER 5

# Sun and the N-Tier Architecture

To date, N-tier software architecture components have been mapped to individual systems. This correlation between hardware and software must be removed if increasing levels of platform utilization and manageability are to be achieved. Sun believes an adaptive, service-driven architecture is emerging — application services will no longer be mapped one-to-one to a specific set of hardware components. Service requirements will be specified through a simplified management framework, and infrastructures will automatically obtain the resources needed to accomplish service goals.

An adaptive, service-driven architecture fosters:

- The right number of nodes for resiliency and manageability
- The right number of service component instances for scalability, mapped appropriately onto the nodes, resulting in proper component resiliency
- The right number of consolidated components, load balancing, and resource management tools to ensure good resource utilization

Today, Sun provides a host of hardware and software solutions that deliver high-performance, highly scalable, and predictable N-tier environments.

## Scalable Servers from Sun

Organizations are migrating from complex, static architectures to a new generation of agility and flexibility. Companies that originally relied on small, uniprocessor servers are now running on large, multiprocessing systems. With massive scalability, organizations are turning their attention to ensuring the architecture is predictable, serviceable, and maximizes uptime. The business-critical information stored in very large enterprise systems, as well as the people who utilize those systems and data, require processing power and I/O bandwidth that work in concert. To be effective, N-tier architectures must employ high-performance scalable servers, high-capacity and high-performance storage subsystems, and an operating environment that supports a wide variety of applications.

Sun Enterprise™ servers provide symmetric multiprocessing, scaling from one to 64 high-performance UltraSPARC™ processors, up to 64 gigabytes of memory, and supporting up to 20 terabytes of disk storage. This allows database management systems to be configured with the optimal balance of processing power and I/O bandwidth. Sun systems are ideally suited to deliver the scalability, robustness, and performance needed for the most demanding enterprise applications, as well as virtually unlimited future growth (Figure 5). Furthermore, Sun servers support advanced clustering and dynamic system domains, increasing availability and facilitating online vertical scaling.

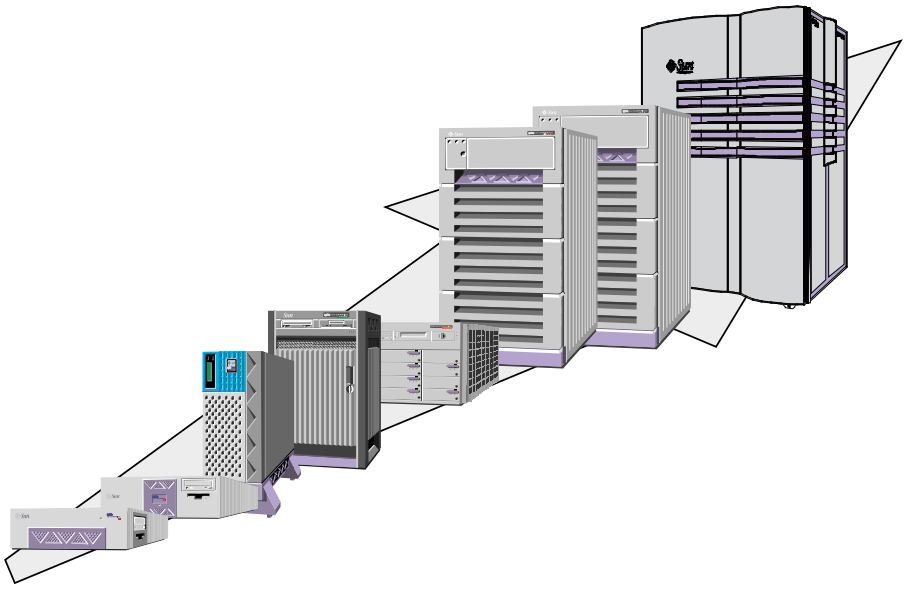

**FIGURE 1** Sun's server family provides unprecedented levels of scalability.

Storage systems are a key element in providing balanced performance, and must be tuned to complement the performance and scalability of powerful multiprocessor systems like those offered from Sun. Utilizing second-generation fibre channel technology, the Sun StorEdge™ product family is a breakthrough offering in high-performance, high reliability, availability, and serviceability (RAS) storage systems that scale from workgroup and departmental computing to data center and mainframe-class environments. With carefully balanced performance, a full complement of advanced hardware and software features, and full compatibility with existing Sun desktop, server, and storage systems, Sun StorEdge subsystems are a powerful expression of next-generation technology that promises to usher in a new era in intelligent network storage.

In addition to the wide scalability of the product line, all Sun systems run the proven Solaris™ Operating Environment. Solaris software provides unlimited, transparent access to systems, servers, printers, remote databases, and other resources, with the scalability to support virtually any mix of applications and peripherals. As the world's premier UNIX® environment, the Solaris Operating Environment commands an installed base of two million users and supports over 12,000 applications. The binary compatibility of Solaris software helps ensure that customers can migrate in-house and third-party applications from smaller development and test environments to large-scale production servers.

Additional features of Sun systems that support complex N-tier environments include:

- *A solution to the challenge of vertical scaling*

  Traditional vertical scaling requires organizations to buy larger systems to increase capacity, resulting in higher costs, increased down time, and parallel testing efforts that often fall short. Sun servers mitigate these problems with support for dynamic system domains, dynamic reconfiguration, alternate pathing, resource and cluster management options, live upgrades, and hot patching technologies.

- *Server flexibility*

  Sun technologies enable administrators to borrow resources from under-utilized machines and domains to meet planned and unplanned demand. Administrators can redeploy CPU, memory, or storage resources on a temporary basis, and return them after peak periods have passed.

- *Reduced training and support costs*

  Sun systems employ consistent, standardized user interfaces. Role-based access control (RBAC) in Solaris 8 software extends management by enabling profiles to be defined to control access, thereby minimizing operator error and improving overall data center security and management. These and other features significantly reduce training and support costs. As a result, less-experienced staff can perform tasks traditionally handled by senior personnel, freeing them to focus on tasks that require senior-level expertise.

- *Strong relationships with third-party application vendors*

  Sun understands that high-performance hardware alone is insufficient for today’s demanding environments. Sun continues to work with third-party application developers to deliver software programs that scale optimally on Sun systems. Organizations can evaluate application scalability by running benchmark baselines, identifying bottlenecks that limit scalability and performance, and tuning systems appropriately. Further tests provide valuable information for analysis and capacity planning efforts.

# Resource Utilization

Patterns of network use are changing, and demands for bandwidth are increasingly unpredictable. Global users access the network 24 hours a day, 7 days a week. People may stay at a Web site for extended periods and download large amounts of data. As information appears and disappears on Web sites, saturation moves around the network. Emerging Internet applications are both bandwidth intensive and time sensitive. They often require support for voice, video, and data, which consumes increasing amounts of bandwidth. Yet users expect instant access, especially if the information is critical to their work. As a result, organizations must find ways to ensure system predictability and proper resource utilization.

## Solaris Resource Manager™ Software

Sun's Solaris Resource Manager™ software is an effective tool for creating and managing shared service environments. Beyond simple time-sharing schemes, it provides fine-grained, hierarchical control of system resources for users, groups, and applications, enabling an equitable distribution of computational resources within a given Solaris system and promoting server consolidation. Solaris Resource Manager software is particularly effective for use in enterprise servers since it can prevent server resources from being usurped by rogue processes, abusive users, and large computational loads.

Solaris Resource Manager software continuously evaluates application utilization while considering organizational policies. Toward that end, it continuously monitors applications to ensure expected resource allocation levels are achieved. Service levels can be modified instantaneously by changing policies. Online vertical scalability is facilitated through several key features:

- Reconfiguration of hardware components without service disruption
- Online expansion and maintenance
- Dynamic resource configuration
- Control of system resources, including CPU, virtual memory, number of processes, number of logins, connect time, fare resource allocation
- Resource sharing between multiple applications running on the same machine
- Semi-automatic loop control
- Online proactive and reactive capacity adjustments

## Solaris™ Bandwidth Manager Software

Solaris™ Bandwidth Manager enables organizations to control the bandwidth assigned to particular applications, users, and departments that share the same network link. By installing the software on major network links and application servers and setting consistent policies, bandwidth can be distributed evenly. In addition, traffic can be prioritized to prevent a small number of applications or users from consuming all available bandwidth. Solaris Bandwidth Manager software enables organizations to:

- Provide differentiated classes of service to users, and bill accordingly
- Guarantee bandwidth to priority users, applications, or servers
- Reduce traffic congestion and increase network efficiency
- Control user and application access to network resources
- Gather detailed network use statistics and accounting data for usage-based billing and trend analysis

## Processor Sets

Processor sets enable a server's processors to be divided into dedicated groups. Processes may be bound to a processor set, preventing processor contention and eliminating processor starvation. Processor sets also enable server consolidation — multiple applications run within their own processor sets on dedicated processors.

# Highly Available Cluster Technology and Management

In the rapidly changing Net economy, businesses must deploy their information infrastructure to remain available around the clock. With an increasing number of business-critical applications servicing partners, suppliers, employees, and customers, companies must ensure systems remain online — even for routine maintenance or capacity expansion — to survive in a highly competitive economy. Adding complexity is the need to plan for highly available e-business and e-commerce applications and services. While the dot-com economy represents an opportunity for explosive growth, organizations will only succeed if IT systems scale quickly and seamlessly, and the environment remains manageable.

To address the expanding role of the data center, organizations are deploying pools of computing resources, including CPUs, disks, and network interfaces, in cluster computing architectures. Traditionally expensive to buy and operate, cluster products focus on high availability and host specialized applications. Today's computing requirements, however, are forcing data centers to move toward general-purpose computing environments to host virtually any application or service with minimal or no modification.

Cluster computing solves many data center issues, including:

- *Continuity.* By duplicating server resources, a failover mechanism can be created to increase availability. Service continuity is significantly enhanced.
- *Capacity.* More servers can be used to increase the overall processing power. This is particularly useful in specialized applications, such as data mining and trend analysis, where extremely large amounts of data must be analyzed.
- *Resource Management*. By running multiple services on the same managed environment, resource utilization can be increased.
- *Improved Manageability.* When properly implemented, cluster administration can provide a centralized management interface and simplify management, thereby reducing complexity and risk.

Sun™ Cluster 3.0 software is a new approach to creating a cluster computing environment for the networked data center. Based on abstracting applications and services, such as data storage and network connectivity, from the physical hardware. Sun Cluster 3.0 extends a high-availability (HA) environment to provide a single, logical view of a commercial computing environment from both an application services and administrative perspective.

Sun Cluster 3.0 software is a strong, capable, and agile platform for delivering a highly reliable, available, scalable, and manageable solution for the network economy.

- *Resiliency*

  The Sun Cluster 3.0 environment is designed and implemented on an efficient, high-availability framework, including improved fault isolation and recovery capabilities within the hardware, operating environment, network, and service environments.

- *Performance and scalability*

  Sun Cluster 3.0 software scales in multiple dimensions. Sun Enterprise servers offer the best “in-system” scalability — up to 64 processors in a machine. With Sun Cluster 3.0, up to eight servers may be configured into a cluster. For both horizontal and vertical scalability, Sun Cluster 3.0 software offers unparalleled scalability and performance.

- *Manageability*

  A centralized management view of all service delivery components lowers administrative costs, improves response, and reduces risk.

- *Resource utilization*

  Sun Cluster 3.0 software not only provides the capability to consolidate services, it also features intelligent load distribution within a single framework. Service levels are increased while costs are contained and predictability is improved.

# Enterprise Systems Management

An architecture that scales horizontally works well — until it reaches the point where the number of servers becomes unmanageable. Conventional tools cannot make the move from managing small pockets of disparate systems and services to managing an entire enterprise. IT organizations need an integrated set of systems management tools — a platform that offers common services for all enterprise management applications.

## Sun™ Management Center

The most advanced systems management tool from Sun to date, Sun™ Management Center offers a single point of management for all Sun systems, the Solaris Operating Environment, applications, and services for the data center and highly distributed computing environments. With Sun Management Center, IT organizations can efficiently manage and arbitrate between users, applications, and resources. Designed to support Sun systems, Sun Management Center provides a platform upon which the enterprise can base its administrative and management operations to ensure the availability of all systems and the services they provide. A powerful tool for managing the enterprise network, Sun Management Center enables system administrators to configure remote systems, monitor performance, and isolate hardware and software faults, all through an easy-to-use graphical user interface. In addition, Sun Management Center integrates with heterogeneous environments through several popular tools from companies including Tivoli, OpenView, UniCenter, and BMC Patrol.

CHAPTER 6

# Support and Professional Services

Today's businesses are operating global, heterogeneous enterprises comprised of complex and distributed networks, vast amounts of data, and diverse systems. This environment is fundamentally changing the way operations are managed. Processes and technologies must be designed to support flexible business units, yet provide seamless enterprise management. These requirements are placing increased importance on enterprise operations management as a critical step in planning the IT infrastructure foundation.

The Sun Professional Services℠ program provides a comprehensive suite of services that help organizations take a strategic approach to enterprise operations management. Sun consultants help address the people, process, and technology issues related to functional areas of the enterprise. Beginning with a brief assessment of the current operations environment, this enables enterprises to discover the strengths and weaknesses of the organization. Using this assessment, Sun consultants work with organizations to design an operations management strategy that addresses technical requirements and supports short-term as well as long-term business goals. Deployment and architecture consultants may aid organizations with system, network, data, storage, software, asset, security, performance, capacity, and change management strategies that take advantage of Sun systems and employ the right scaling techniques as business needs dictate.

CHAPTER 7

# Summary

The use of the Internet with corporate intranets is challenging organizations. Quality of service, scalability, and availability demands are pushing the limits of existing information infrastructures. As a result, organizations need to retool and take advantage of flexible N-tier architectures. With the ability to scale both horizontally and vertically, N-tier architectures are changing the flexibility of the IT infrastructure.

Sun has provided foundation-level products for mission- and business-critical computing for more than 18 years, and is delivering the solutions needed to advance the effectiveness of N-tier computing environments. The combination of powerful software running on scalable Sun servers and the Solaris Operating Environment gives organizations the ability to configure servers for unprecedented levels of reliability, availability, serviceability, predictability, and agility.

CHAPTER 8

# Glossary

| | |
|---|---|
| **Adaptive Service-Driven Architecture** | A hardware and software architecture in which applications automatically obtain the resources needed to accomplish service goals. |
| **Application Server Tier** | Software that determines which data is to be stored and retrieved, and manipulates that data on behalf of business logic. |
| **Caching Service Tier** | Software that runs application proxies which forward requests from a client to the application server, as well as a service locator that finds needed services for clients. |
| **Data Tier** | Software that stores application data in a persistent store, such as a relational database or an object-oriented database. |
| **Diagonal Scaling** | A combination of horizontal and vertical scaling, enabling each server in a group of horizontally scaled servers to be grown within the system. |
| **Horizontal Scaling** | The ability to use multiple servers within a single tier. |
| **Integration Service Tier** | Software that coordinates the passage of synchronous and asynchronous messages and transactions to and from heterogeneous database services. |
| **N-Tier Architecture** | A computing paradigm in which the application architecture is characterized by the functional decomposition of applications, service components, and their distributed deployment. |
| **Presentation Tier** | A graphical user interface that displays data to users regardless of device type or location, and enables information manipulation. |
| **QoS** | Quality of Service. |
| **Server Consolidation** | The migration of applications from multiple servers onto a single system. |
| **Server Flexibility** | The ability to borrow resources from under-utilized machines and domains to meet planned and unplanned demand. |

**Tier** A functionally separated hardware and software component that performs a specific function.

**Vertical Scaling** The ability to scale services within a system.

**WAN** Wide area network.

**XML** The eXtensible Mark-up Language.

CHAPTER 9

# References

Sun Microsystems posts product information such as data sheets, specifications, and white papers on its Web site at *www.sun.com*. For more information on Sun's Solaris architecture and product family, visit *www.sun.com/solaris*.

For additional information on Sun products and technologies, please visit the following Web sites:

- www.sun.com
- www.sun.com/servers
- www.sun.com/blueprints
- www.sun.com/clusters
- www.sun.com/storage

Sun Microsystems, Inc.
901 San Antonio Road
Palo Alto, CA 94303

1 (800) 786.7638
1.512.434.1511

http://www.sun.com/solaris

September 2000