Sizing Sun Ray™ Enterprise Servers

Technical White Paper

Printed in the United States of America.
901 San Antonio Road, Palo Alto, California 94303 U.S.A

Contents

Introduction 1

The Sun Ray™ Hot Desk architecture has revolutionized the deployment of workgroup computing resources resulting in greater efficiency and economy through better resource sharing. Users no longer have to contend with inadequate hardware and constant upgrades to their desktop systems — administrators avoid the significant costs of processing power, memory, and I/O that often go unused on idle desktop systems.

Far from being wasted, computing resources in the Sun Ray Hot Desk architecture are shared throughout the workgroup, greatly improving resource utilization over traditional desktop models. Sun Ray 1 enterprise appliance users enjoy server-class performance for a wide range of applications without having to worry about upgrading or replacing their desktop hardware every few years. At the same time, a centralized administration model allows administrators to focus on the server and better understand the characteristics of the applications they support. In this new model, computing resources can be applied where they are needed most to benefit application performance.

To provide a high level of interactivity, Sun Ray enterprise servers must be configured to satisfy the needs of the individuals and workgroups they serve. Going beyond simple ratios of clients to servers or other rules-of-thumb, this document provides an effective methodology for sizing Sun Ray enterprise servers with a particular focus towards meeting individualized workgroup needs.

1

Sun Ray Hot Desk Architecture Basics

The principal goal of sizing Sun Ray 1 enterprise servers is to develop an estimate of an initial configuration that will meet the current needs of the workgroup as well as anticipate short-term growth. Where possible, initial sizing should also anticipate longer-term expansion. Familiarity with the Sun Ray Hot Desk architecture is helpful in understanding how it influences server sizing decisions.

High-level Architecture

The Sun Ray Hot Desk architecture succeeds by combining key advantages of existing architectures with today's inexpensive hardware components and high-speed networking technology. As shown in Figure 1-1, the architecture is comprised of three components: the *Sun Ray 1 enterprise appliance*, a *dedicated interconnect*, and a *Sun™ server* running the *Solaris™ Operating Environment* and *Sun Ray enterprise server software.*

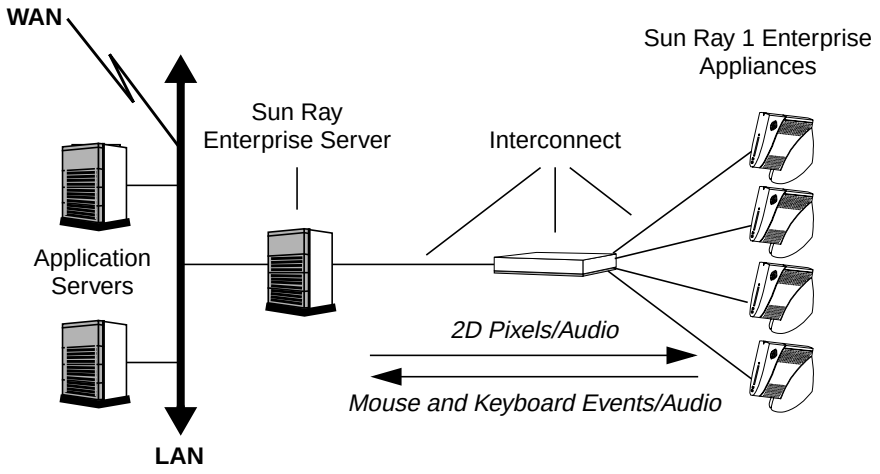

Figure 1-1 High-level perspective of the Hot Desk technology architecture

No client software is stored or executed on the appliance, Unlike X-terminals and similar devices, all user applications run on one or more centralized server systems. The X11 server and GUI for each appliance user run on the Sun Ray

enterprise server. The enterprise appliance contains only the resources necessary for the *human interface* — input devices such as microphone, keyboard and mouse, and output devices such as the display and audio.

All user input (keystrokes, mouse clicks, and audio) is transmitted from the appliance through the interconnect and on to the appropriate client application. 2D pixels and audio output travel back to the appliance across the interconnect. The user sees a fully-functional CDE desktop environment and window system along with all of their active applications.

Virtual Device Drivers

All input and output to the Sun Ray 1 enterprise appliance is accomplished through the use of virtual device drivers on the Sun Ray enterprise server. For example, the X11 server displays to a virtual device driver that translates between the higher-level X11 protocol and the native Hot Desk technology protocol as shown in Figure 1-2. For each appliance, a *virtual display device driver* maintains a copy of the currently active desktop session in a memory-based *virtual framebuffer* on the Sun Ray enterprise server. All display rendering is performed on the server, and pixels are sent by the virtual device driver to the appliance when they need to be updated.

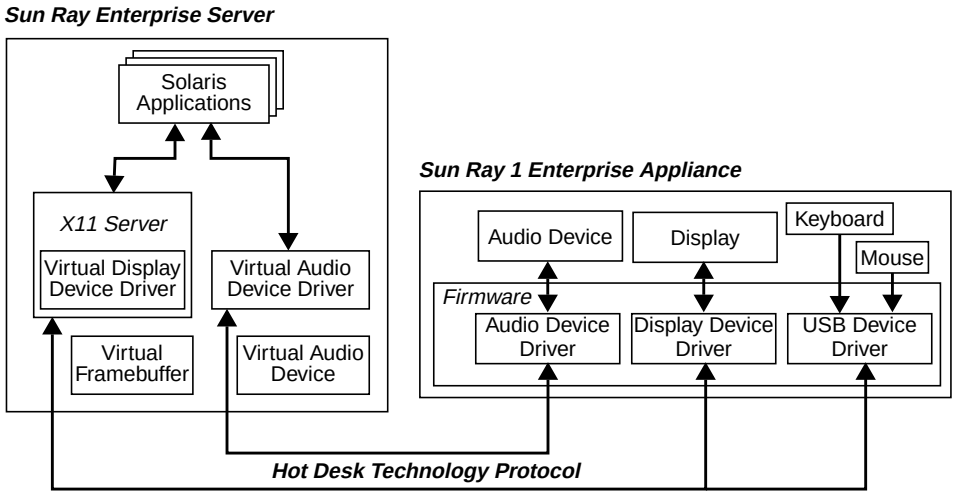

Figure 1-2 Virtual device drivers translate between application-specific protocols (i.e. the X11 protocol) and the native Hot Desk technology protocol

Virtual device drivers also exist for the appliance's audio interface and USB keyboard and mouse. Keyboard and mouse events are sent from the desktop back to the server, converted into X11 events, and then sent on to the appropriate applications.

The Hot Desk technology protocol enables applications to operate independently of the location of their associated input and output devices. Collectively, a user's applications and windowing environment represent a *virtual session* that can be displayed to any appliance on the interconnect. Input and output for the virtual session are automatically redirected to any appliance on the interconnect based on insertion of the users' smart card.

Selecting an Appropriate Sun Ray Enterprise Server

The Sun Ray Hot Desk architecture can scale to serve from very small to very large workgroups — from single laboratories to several buildings on a campus. For any given workgroup, a range of server solutions is generally appropriate. Minimal configurations should be avoided since interactivity can suffer as soon as initial demands are exceeded. Oversized configurations offer better potential performance for individual applications with corresponding lower levels of resource utilization.

Sizing Versus Tuning

Sizing and tuning are different but related activities. *Sizing* estimates the initial configuration of a Sun Ray enterprise server to meet the needs of a known set of tasks. *Tuning* refers to the process of measuring and adjusting the capacities of an existing Sun Ray enterprise server to meet dynamic needs such as changing individual roles in the workgroup, larger numbers of appliance users, or new applications.

A conservative approach is important when sizing Sun Ray enterprise servers due to the limited accuracy of the assumptions that can be made about sizing data. Though workgroup resource usage and user activity data are readily measurable, they are subject to the typical dynamic nature of the workgroup's usage patterns and other variables. Fortunately, incremental server resources (fast processors, large disks, and blocks of memory) are available at a fairly coarse granularity making precise resource usage and user activity data unnecessary.

Initially, server configurations should be sized to exceed the minimum needs of the workgroup and allow for anticipated *growth* within the workgroup for one to two years. Needs for additional appliances should be forecast along with potential requirements for new applications — particularly applications that are computationally-demanding. Where possible, initial server sizing should also anticipate future *expansion* which might include significant new requirements based on supporting new buildings, new workgroups or significant new application processing demands.

Sun Ray enterprise servers which are initially sized to allow for expansion can be easily tuned at a later time. Changes may involve additional CPU or memory upgrades to support greater numbers of appliances or new applications. New network interfaces may need to be configured in order to support an expanded interconnect. If initial sizing is done correctly, total server system replacement can be avoided.

Tuning may also involve evaluating where applications run best and most economically. Often, new compute-intensive applications can be hosted on other network application servers which are specifically sized for their anticipated needs.

Application Resource Measurement Tools

Gauging workgroup application resource needs is key to configuring properly-sized Sun Ray enterprise servers for a given set of applications and users. Server sizing calculations provided later in this document depend on this information. The Solaris Operating Environment provides tools which can be used to measure the resource use levels associated with any particular application.

- `perfmeter`

  The bundled *perfmeter* tool can be used to determine average CPU and network resources used by a given application. By setting *perfmeter* to log frequent, periodic samples of CPU usage levels and network packet data to a file, later analysis on the saved data can distill useful resource statics.

- `pmap`

  The */usr/proc/bin/pmap* utility is provided with version 2.6 and later of the Solaris Operating Environment. *Pmap* can be used to measure an application's memory footprint.

Server Monitoring Tools

Sizing is just an estimate for the initial server configuration based on educated guesses about user profiles and application usage. Administrators must have the ability to monitor and identify critical resources on the Sun Ray enterprise server. The Solaris Operating Environment also provides effective tools for monitoring server resources.

- `vmstat`

  The */bin/vmstat* command provides valuable information on the system's paging statistics which can help determine when memory is constrained causing excessive paging.

  Vmstat also reports CPU activity and the length of the *run queue*. The run queue is the number of processes which are waiting to run (and are not blocked waiting on I/O or other synchronous activities). Run queue length is a key indicator of system load. A Sun Ray enterprise server which shows frequent or long periods of 100% CPU utilization along with many processes in the run queue may need additional CPU resources to handle the load generated by the appliances it serves.

- `perfmeter`

  Perfmeter is also useful for monitoring and logging resource use information in terms of both CPU utilization and network bandwidth, enabling detection of periods of peak loading and contention for resources.

Configuring for Resiliency

The Sun Ray 1 enterprise appliance is an extremely reliable device since it is fundamentally stateless and has no moving parts. A hardware failure in an appliance is easily remedied by plugging in a new appliance or having the user move to another appliance.

Because all of the Sun Ray 1 enterprise appliances in a workgroup depend on a single server, availability for the workgroup as a whole can be affected by problems on the Sun Ray enterprise server. Until high availability session migration is available, a number of steps can be taken to harden Sun Ray 1 servers against single points of failure.

- *RAS Capabilities*

  RAS, or reliability, availability, and serviceability are highly desirable properties in a Sun Ray enterprise server. Where ever possible, servers should be configured with RAS capabilities to enable them to seamlessly recover from single-component failures. For example, providing disk mirroring or other forms of RAID (Redundant Arrays of Independent Disks) can protect Sun Ray enterprise servers against downtime and data loss in case of the failure of an individual disk. Techniques like RAID may have some small impact on disk I/O performance but this will generally not affect latency and will not be apparent to appliance users.

- *Extra Processors and Memory SIMMS*

  Another useful technique to provide resiliency in the event of processor or memory failure is to over-configure the Sun Ray enterprise server by at least one processor module and one memory SIMM. This approach does not protect against server crashes (which terminate all user sessions), but does allow the server to recover gracefully in the event of a processor or memory component failure.

  Sun's Solaris Operating Environment is able to configure around a failed component upon re-boot. By over-configuring by at least one component, the Solaris Operating Environment running on the Sun Ray enterprise server can reboot, map out a failed component, and continue to operate at an expected level of service until the component replacement can be scheduled.

- *Multiple Servers*

  Multiple servers can be configured to divide users into separate workgroups of Sun Ray 1 enterprise appliances. Though this approach decreases the dependency on a single server, it adds the administration of a second server. Additionally, appliance users with smart cards are unable to migrate their virtual session between appliances served by different servers. Finally, modern queuing theory holds that a single server is more effective at serving the random needs of a workgroup than multiple servers. Multiple servers may be effective where different workgroups are not expected or allowed to share appliances.

1

Sun Enterprise™ Servers and the Solaris™ Operating Environment

Application performance in a Sun Ray enterprise system environment is directly dependent on the servers that provide computational resources. Sun Enterprise™ servers lead the industry in offering some of the most powerful and reliable systems available today. Sun's family of servers provide scalable, symmetric multiprocessing capabilities. From one to 64 high-performance UltraSPARC™ processors can be configured along with up to 64 GB of physical memory and up to 20 TB of disk storage, providing the necessary performance for peak demands as well as virtually unlimited growth.

The power of Sun's servers is further enhanced by the Solaris Operating Environment — the premiere environment for enterprise network computing. Designed with the needs of the enterprise in mind, the Solaris Operating Environment features full 64-bit processing, mainframe-class reliability, superior scalability, and unprecedented performance. These features and others greatly enhance multi-user environments making the Solaris Operating Environment uniquely suited to hosting the Sun Ray enterprise system.

Sizing Guidelines and Methodology 2

It is difficult to provide meaningful guidelines for sizing Sun Ray enterprise servers without an understanding of the applications and usage patterns of the appliances they serve. Knowing what users are doing, along with their activity levels and relative time spent in particular applications is key to configuring adequate server resources.

This section describes general server sizing guidelines to help administrators gain a quick perspective on server requirements but also provides detailed server sizing calculations based on real user and application requirements. Information is also provided to assist in sizing servers which provide additional application services (HTTP, file services, mail, multimedia).

Those who wish to size a configuration quickly should proceed to Chapter 3 for information on the Java™ technology Sun Ray server configurator.

Sizing Methodology and Guidelines

The principal goal of correct Sun Ray enterprise server sizing is to provide for robust interactive performance for appliance users under average-use conditions. Sufficient processing capabilities (CPU), memory, I/O, and interconnect bandwidth must be provided to avoid contention between appliance users. The methodology provided herein accomplishes this goal by providing enough CPU, disk, and interconnect resources to support peak load conditions, enough system memory to hold peak active user sessions, and sufficient swap space to hold all user sessions (both active and inactive) within virtual memory.

Note – The guidelines presented here, along with the simple sizing examples that accompany them, are meant as a quick approximation for server sizing. Whenever possible, applications and users should be carefully profiled and measured as a part of a more detailed sizing exercise (see the later section on sizing calculations and the Sun Ray server configurator).

To help illustrate the basic sizing guidelines, an environment with typical office productivity and personal information management (PIM) applications will be used as an example. The environment described here reflects a workgroup of fifty Sun Ray 1 enterprise appliances, of which twenty five are typically active at any one time (a fifty percent activity level).

- *Number of Processors*

  Sufficient processing power must be provided for Sun Ray enterprise servers in order to provide good application performance as well as a high level of interactivity. For most workgroups, Sun Ray enterprise servers should be equipped with at least two processors. Multiple processors enable the Sun Ray enterprise server to provide consistently quick response, even in the presence of high-priority system processes and threads. These high priority tasks can cause somewhat delayed response in uniprocessor servers under peak loading conditions.

  Single processor servers *can* be used for small workgroups — especially if the users have low average activity levels or use applications that require minimal CPU resources (text editing, terminal emulation, etc.). For these users, up to twenty Sun Ray 1 enterprise appliances connected to a powerful (300 or 400 MHz) uniprocessor server is not unreasonable. Configuring a uniprocessor server does require careful attention to peak loading to ensure that good response times can be maintained when all users are active.

  For an example of processor calculations, consider the general-purpose end-user environment under study — fifty users with a fifty percent activity level. Conservatively, office productivity and personal information management applications require two to five percent of a 300 MHz UltraSPARC™ processor. (Computationally-intensive applications like CAD, simulations, and video decompression applications must be profiled individually.) Since one or two applications are typically active at any one time on behalf of a given user, a conservative estimate of an active user's impact is five percent of a 300 MHz processor.

Multiplying five percent by the number of simultaneously active users (twenty five in this example) yields one hundred and twenty five percent of a 300 MHz UltraSPARC CPU for servicing user environments and applications. Adding the roughly ten percent of a 300 MHz CPU required to support the Solaris Operating Environment yields one hundred and thirty five percent of a 300 MHz CPU. Thus, two 300 MHz UltraSPARC processors should be sufficient to serve the average needs of twenty five active users running PIM and general office productivity applications.

- *Sizing System Memory*

  Memory is, perhaps, the most important resource in a Sun Ray enterprise server. A server which has run out of CPU resources will typically degrade gracefully whereas a system that is thrashing due to memory starvation can significantly affect interactive performance for the workgroup. For the Sun Ray enterprise server, memory is sized based on simultaneous *active* users.

  Excluding applications with large memory requirements (imaging, CAD, etc.), 40 MB of system memory should be provided for each active appliance user. This guideline provides memory resources roughly equivalent to those of a single user working on a 64 MB workstation in one or two applications at a time. Enough memory is provided to keep one or two active applications in main memory though the user's inactive applications may be swapped out if necessary.

  Applications with larger memory requirements or work patterns with frequent shifting of focus between applications may require more significant memory resources on the Sun Ray enterprise server. If users are accustomed to workstation environments that require 128 MB or more memory to achieve desired performance levels, at least 100 MB of memory per user should be configured on the Sun Ray enterprise server.

  This per-user amount of memory must be added to the minimum system configuration of 64 MB for the operating system kernel and shared libraries. For the office automation example, multiplying 40 MB by twenty five expected *active* users, plus 64 MB for the operating system yields 1064 MB in system memory requirements.

- *Swap Sizing and Disk Spindles*

  Swap space is used extensively by the Sun Ray enterprise server to share physical memory among appliance users. Sessions belonging to inactive users are paged out as memory is required to support other active users. Sessions which become active are paged back into system memory. Virtual memory must be sized large enough to hold all users' X11 sessions in

addition to providing space for anonymous memory and temporary storage required by the operating system and other applications. Given the importance of swap space and the low cost and high density of today's disk storage, conservative sizing of swap resources for non-restricted desktop environments is strongly recommended.

In addition to having sufficient swap space to support the community of appliance users, sufficient I/O bandwidth to the disk subsystem must be provided to ensure that sessions can be paged in (and out) quickly to support the expected level of user interactivity. In general swap space should be spread across several spindles to avoid bottlenecks resulting from a single disk's I/O constraints. Running swap across one disk spindle for each processor configured in the server is an acceptable minimum.

The virtual memory footprint of each Sun Ray 1 enterprise appliance user in this example is between 40 and 100 MB. Again, virtual memory requirements for the Sun Ray enterprise server are sized based on *all* users, not just active sessions. For example, assuming a 50 MB footprint per user, a fifty user workgroup would require 2.5 GB of virtual memory. To calculate the required swap space, 1064 MB of real memory is subtracted leaving 1.436 GB of swap space. To this number, 500 MB to 1 GB of swap space should be added to provide space for application core dumps and temporary storage yielding a swap requirement of approximately 2 GB. Since the example server configuration will have two processors, the 2 GB of swap space should be spread across at least two disk spindles.

- *Network and Interconnect Interfaces*

  A Sun Ray enterprise server requires at least two network interfaces — one for connection to a LAN, the other for connection to the interconnect. Sun *strongly recommends* a dedicated 100 Mbps interconnect because it guarantees a high level of quality of service for the Sun Ray 1 enterprise appliances. The Hot Desk protocol is UDP/IP based and requires a low-latency environment in order to provide interactive response to the appliance. Shared, routed, general purpose networks are not recommended for use as the interconnect since network events like large file transfers or frequent broadcasts can cause Hot Desk protocol packets to be dropped. Dropped packets or other interference can severely degrade the appliance user's experience.

The *network (LAN) interface* must be sized adequately to support the combined network traffic from all of the applications running on behalf of the workgroup's active users. Sun Ray enterprise servers may require greater than a 100 Mbps LAN connection depending on the number and nature of the applications they host. For example, many customers use the Sun Ray 1 enterprise appliance to access X sessions on other servers which can drive up LAN bandwidth requirements. In addition, other requirements such as connection of the Sun Ray enterprise server to separate subnets can require additional LAN interfaces.

The *interconnect interface* on the server must be sized to support the flow of pixels and keystrokes to the all of the active appliances at a sufficient rate to sustain good response and interactivity. Other than video, games, and other applications which cause large, frequent screen updates, most typical applications use less than 1 Mbps of interconnect bandwidth. Supplying 1 Mbps on the server's interconnect interface for each of the twenty five active users in the workgroup yields 25 Mbps required for the example workgroup.

Conservatively assuming a twenty five percent protocol overhead, a 100 Mbps network interface card should provide 75 Mbps of throughput. For the fifty-seat workgroup example, a 100 Mbps (100BaseT) adapter would suffice. To provide for future expansion, a 1 Gbps interface could be configured into the server. Details for configuring the interconnect for different physical deployment scenarios can be found in the Sun white paper *Deploying The Sun Ray Hot Desk Architecture*.

In summary, the needs of a fifty appliance workgroup with moderate application needs and a fifty percent activity level could be met by a Sun Enterprise server equipped with two 300 MHz UltraSPARC processors, two 100 Mbps Fast Ethernet interfaces, and 1064 MB of RAM. A minimum of two disks would need to be provided with 2 GB of swap space configured across both drives. Sun produces several servers that support at least two-processor configurations. Server selection should ultimately be based on anticipated workgroup growth and expansion.

2

Estimating Real-world Workloads

Workgroups in the real world are seldom as homogeneous as shown in the example above. A more robust approach to server sizing is required that involves accurately measuring application demands and projecting impact based on expected usage patterns of different groups of users. To aid with this process, Sun provides a Java technology *Sun Ray server configurator* (chapter 3).

This section details calculations similar to those used by the server configurator in order to explain the recommended methodology for sizing Sun Ray enterprise servers. Data for these sample calculations is taken from the Educational (kindergarten through twelfth grade or *K-12*) deployment scenario described in chapter 3.

Profiling Users

In most deployment scenarios, different groups of users typically have distinct profiles based on the applications they use as well the usage patterns they display. In the K-12 deployment scenario, all of the users are students, but their application usage and activity level expectations vary depending on the location of the Sun Ray 1 enterprise appliance. For instance, appliances in a computer lab are expected to be 100% busy while a laboratory is in session. Appliances in classrooms or the library may have more occasional use. User data for the K-12 deployment scenario is shown in Table 2-1.

	Computing Lab	Library	Classroom
Number of Users	30	10	60
Percent Active	100%	20%	20%

Table 2-1 Default user categories for the K-12 deployment scenario

In addition to different activity levels, the students will have slightly differing application profiles depending on their location. Table 2-2 shows how much time users are expected to spend in various applications.

	Computing Lab	Library	Classroom
X Baseline	—	—	—
CDE Utilities	20%	20%	20%
Word Processor	50%	30%	40%
Web Browser	30%	45%	40%
Java Software	—	5%	—

Table 2-2 Default user profiles for the K-12 deployment scenario

X Baseline refers to the basic X11 environment used by all students as a part of the Hot Desk environment. *Java software* refers to a Java technology applet that is only used in the library; for example, a Java technology front-end applet or JavaScript™ software used to search a library database.

Profiling Applications

Once a set of user profiles and application usage patterns has been established, individual applications are profiled to understand the demands that they place on system resources. Resource considerations for applications running on a Sun Ray enterprise server are expressed in terms of memory, processor, and interconnect bandwidth.

- *Memory*

  It is important to understand the memory demands (in megabytes) placed on the Sun Ray enterprise server by each *fully-active* instance of the application. The Solaris Operating Environment features an efficient shared memory system which enables multiple instances of applications and libraries to share certain memory segments (i.e. code segments). As a result, application memory usage is divided into private and shared memory. An application's shared memory is allocated only once and shared between all instances of the application whereas private memory is allocated for each additional instance of the application.

- *Processor*

  Processor resources are typically expressed as the percentage of a given processor needed to run a fully-active instance of the application. By default the configurator expresses processor resources as a percentage of a 300 MHz UltraSPARC processor.

- *Interconnect Bandwidth*

  In order to provide robust, interactive performance, it is essential to supply an adequate network interface on the Sun Ray enterprise server to connect to the dedicated interconnect. In order to calculate interconnect bandwidth needs, the application's impact on the interconnect must be understood in terms of megabits per second.

The configurator applet lets the user choose from a variety of pre-defined profiles for common application types. The profiles for the application types used in the K-12 deployment scenario are shown in Table 2-3.

Application	Private Memory (MB)	Shared Memory (MB)	300 MHz CPU (%)	Interconnect Bandwidth (Mbps)
X Baseline	17	3	—	—
CDE Utilities	3	7	2%	0.2
Word Processor	8	7	3%	0.2
Web Browser	9	11	4%	0.8
Java Software	6	3	2%	0.1

Table 2-3 Application profiles used in the K-12 deployment scenario

Sample Calculations

The Java technology configurator automatically performs all necessary calculations to determine server requirements. Several calculations are shown here to illustrate how the configurator determines resource requirements for different classes of users and applications.

Note – These calculations are provided as an example only. They illustrate, but may not exactly reproduce the calculations made by the current version of the Java technology configurator.

Calculating the Average Resource Consumption of a User

To estimate the average impact of users of on the server, the resource demands of each application must be calculated as a function of usage. Table 2-4 shows that each classroom user will require an *average* of 4.88 MB of private memory, 0.64% of a 300 MHz CPU, and 0.088 Mbps of interconnect bandwidth.

Application	%Active	%Usage	Private Memory (MB)	300 MHz CPU (%)	Inter-connect Bandwidth (Mbps)
X Baseline	20%	100%	3.4	—	—
CDE Utilities	20%	20%	0.12	0.08%	0.008
Word Processor	20%	40%	0.64	0.24%	0.016
Web Browser	20%	40%	0.72	0.32%	0.064
	Totals	**100%**	**4.88**	**0.64%**	**0.088**

Table 2-4 Average impact of a single classroom user on the Sun Ray enterprise server

The *average* results are obtained by multiplying the values for each application's resource category (Table 2-3) by both the percentage that the user is active (20 percent for the classroom users), and the percent usage for that particular application (Table 2-2). To solve for *maximum* user impact, the calculations are repeated with the values in the "*%Active*" column set to 100%.

Application	%Active	%Usage	Private Memory (MB)	300 MHz CPU (%)	Inter-connect Bandwidth (Mbps)
X Baseline	20%	100%	3.4	—	—
CDE Utilities	20%	20%	0.12	0.08%	0.008
Word Processor	20%	30%	0.48	0.18%	0.012
Browser	20%	45%	0.81	0.36%	0.072
Java Software	20%	5%	0.06	0.02%	0.001
	Totals	**100%**	**4.87**	**0.64**	**0.093**

Table 2-5 Average impact of a single library user on the Sun Ray enterprise server

Note that shared memory is not represented in any of these calculation — it will be added accounted for later. Table 2-5 and Table 2-6 list the calculations for the library and computer lab users respectively.

Application	%Active	%Usage	Private Memory (MB)	300 MHz CPU (%)	Inter-connect Bandwidth (Mbps)
X Baseline	100%	100%	17	—	—
CDE Utilities	100%	20%	0.6	0.4%	0.04
Word Processor	100%	50%	4	1.5%	0.10
Web Browser	100%	30%	2.7	1.2%	0.24
	Totals	**100%**	**24.32**	**3.1%**	**0.38**

Table 2-6 Average impact of a single computer lab user on the Sun Ray enterprise server

Calculating Average Resource Consumption for Each User Type

Once the average consumption of resources for an individual user of each type has been calculated, the overall resources needed for each user group can be obtained by simply multiplying by the number of expected users (Table 2-1).

User Group	Number of Users	Private Memory (MB)	300 MHz CPU (%)	Interconnect Bandwidth (Mbps)
Classroom User	60	292.80	38.4	5.28
Library Users	10	48.70	6.4	0.93
Computer Lab Users	30	729.60	93.0	11.40
	Total	**1071.10**	**137.8**	**17.61**

Table 2-7 Resource needs for each application group

From Table 2-7, a minimal Sun Ray enterprise server for the example workgroup would have at least 1166 MB of RAM (1071 MB plus 31 MB of shared memory plus 64 MB for the kernel), at least two processors, and could be served by a single 100BaseT Ethernet interconnect in addition to its LAN interface(s). The configurator does not perform LAN bandwidth calculations.

Calculating SWAP Space Requirements

As mentioned previously, adequate virtual memory is essential to operation of the Sun Ray enterprise server — allowing inactive sessions to be paged out and quickly retrieved when they become active again. Unlike memory which is sized for active users, virtual memory, and therefore SWAP space must be sized to hold *all* user sessions, both active and inactive. Table 2-8 shows the calculations for determining total private memory.

	# Users	Private Memory/User (MB)	Total Private Memory (MB)
Classroom	60	24.4	1,464
Library	10	24.35	243.5
Computer Lab	30	24.32	729.6
		Total	**2437.1**

Table 2-8 Estimated total private memory

31 MB of shared memory and 64 MB reserved for the kernel is added to 2437 MB of private memory yielding a minimum virtual memory requirement of 2532 MB. This amount reflects a minimum value for two principal reasons:

- *Application Memory Bloat*

  The measured memory allocation for the sample applications may not account for all of the memory that can potentially be allocated by the application. This value reflects a snapshot of the application's memory needs which may change over time. Applications like Web browsers can start off small and grow to use significant amounts of virtual memory as plug-ins and other ancillary objects are loaded. In reality an application may have allocated significantly more storage that is currently represented in system memory — unused pages may have been reclaimed by the virtual memory system.

- *Temporary Storage Space*

  Since most user applications run on the Sun Ray enterprise server, sufficient temporary space must be provided to accommodate the needs of active applications. The memory-based *tmpfs* file system in the Solaris Operating Environment means that this memory must come from the virtual memory budget.

Fortunately, configuring additional swap space in today's disk market is extremely cost-effective. *It is always worth erring on the side of configuring too much swap space.* Although applications vary, configuring fifty to one hundred percent more virtual memory above the minimum required is strongly recommended.

In the workgroup example, multiplying the minimal virtual memory figure (2532 MB) by 1.5 yields 3798 MB. Assuming 20 MB of temporary space per user, and 44 active users, 880 MB is added to the total yielding 4678 MB. Subtracting the 1166 MB of RAM memory yields a recommended value of 3512 MB of swap space. This amount of swap space should be distributed across at least two disk spindles.

Accounting for Other Services

Sun Ray enterprise servers need not be dedicated to serving only appliance users' desktop applications. Other applications and network services can be run on Sun Ray enterprise servers including Web, file, and database services as long as the server is configured to meet the demands placed upon it. Sizing for additional applications or services is simply a matter of understanding resource loads for those applications and entering the information into the Sun Ray server configurator.

Web Services

Recent SPECweb results for an Ultra Enterprise™ 250 with one 400 MHz CPU achieved 2625 operations per second. A typical, active browser user is capable of generating one or two operations per second.

For a relevant example, the user profiles in the K-12 deployment scenario would provide for fifteen users active in the browser at any one time under average load and thirty eight users under peak loading conditions. These loads equate to 0.6% to 1.5% of a 400 MHz CPU. Adjusting for a 300 MHz CPU yields a range from 0.8% to 2% of processor capacity.

The memory requirement of the particular Web server would also need to be added into the calculation of memory requirements.

File Services

The SPECsfs benchmark can be used to estimate the impact of file service on a Sun Ray enterprise server. Recent benchmarks achieved 2562 file system operations per second for a single processor 300 MHz server.

Most applications use some level of file system service. Even browsers typically use file services to cache Web page data and to store cookies. For conservative sizing, it is reasonable to assume that most users generate one to two file system operations per second for general application use. This results in a file service CPU load of .15 to .25 percent for each file services user. The total number of file service users relying on the server would need to be accounted for in any calculations.

Mail Services

Using Sun Internet Mail Server (SIMS) 2.0, an Ultra™ 1/140 server with 448 MB of memory can support up to 1800 simultaneous mail clients. Assuming a range of from twelve to twenty two active mail users, this equates to a CPU load of between 0.3% and 0.6% and an additional memory requirement of 3 MB to 6 MB plus the mail server memory needs.

Video Clients

In the Hot Desk environment, all of the workgroup's video clients (i.e. ShowMe TV™ Receiver) run on the same Sun Ray enterprise server, whether decoding video from a file or a multicast broadcast. Because decoding of video files or streams is a computationally-intensive process, sizing for multimedia must be considered in addition to standard desktop resources. More information on the Sun Ray 1 enterprise appliance in multimedia environment can be found in the Sun white paper *Digital Media on the Sun Ray 1 Enterprise Appliance.*

CPU

The CPU resources needed for decode and display of video vary widely with the type and quality of the video sources being decoded. For instance, MPEG files can be encoded at widely differing rates. Decoding a 1.5 Mbps MPEG-1 file needs very different resources than those required for an 8 Mbps MPEG-2 file. To decode and display a 1.5 Mbps MPEG-1 file at thirty frames per second

typically requires one third of the resources of a 300 Mhz UltraSPARC processor. The resources of an entire 300 Mhz processor are required to decode and display a 6 Mbps MPEG-2 file at twenty four frames per second.

As a real-world example, assume that a maximum of eight appliances in a given workgroup were expected to be decoding and displaying 1.5 Mbps MPEG-1 video files simultaneously. Since each MPEG-1 stream is expected to require one third of one CPU, three additional 300 Mhz UltraSPARC processors should be added to the basic configuration needed to support basic appliance functionality and applications. If instead the eight appliance users were viewing 6 Mbps MPEG-2 video files, as many eight additional 300 Mhz processors would be required. These numbers are offset by the degree to which the server is under- or over-configured for the other workgroup services it provides.

It should also be noted that multicast video works somewhat differently in the Hot Desk environment than in traditional networked environments. Applications like ShowMe TV Transmitter use multicasting to send network video and audio streams only to decoders on networked systems that either request a signal or are permitted to see it. In the current implementation of the Hot Desk environment, though the Sun Ray enterprise server receives a single media stream, a separate client decoder application (i.e. ShowMe TV Receiver) runs on the server for each active appliance user who wishes to view the media stream. Administrators must account for this additional processor load when sizing servers.

Memory

The ShowMe TV 1.3 Receiver requires approximately 30 MB of virtual memory in order to run and presents a resident set size of 12 MB. Of the resident set, 8 MB is required in private memory and 4 MB is shared. The first instance of ShowMe TV Receiver allocates 12 MB of memory (8 MB private, 4 MB shared). Subsequent instances only need allocate their own 8 MB private memory segment. A good estimate is that eight simultaneous ShowMe TV users would require at least 68 MB of additional memory in the Sun Ray enterprise server.

Interconnect Requirements for Multimedia

The Hot Desk protocol works by sending only pixels than change over the interconnect to the appliance. Because displaying video causes a large number of changed pixels, it can demand significant interconnect resources.

Displaying a 320 x 240 MPEG-1 video window typically generates 10 to 12 Mbps of interconnect traffic. Its easy to see how fewer than ten simultaneous video users could easily saturate a server's 100 Mbps interconnect connection, causing performance problems for other users. A Gigabit Ethernet, or *QuadFastEthernet* interface in the server would be required to serve this number of video users and allow for future expandability. Gigabit Ethernet or multiple Fast Ethernet interfaces would also be required to serve appliance users who wish to decode and display MPEG-2 video — a single 640 x 240 MPEG-2 window generates approximately 45 Mbps of interconnect traffic.

2

Server Sizing Tools and Examples 3

In the real world, most workgroups are comprised of different types of users, who in turn run different groups of applications. To assist with the process of sizing Sun Ray enterprise servers, Sun provides StarOffice™ spreadsheets and a Java technology *configurator applet*. Sun and reseller engineers can access these tools at (*http://sunray.corp/sunray1/technical/server/ and* http://channel.sun.com/US/pricelist/tools/sunray/).

The Java Technology Server Configurator Applet

The configurator applet is loaded with data for several pre-defined configurations, in line with Sun's target markets for the Sun Ray 1 enterprise appliance, including:

- *Educational (K-12)*
- *Library Automation*
- *General Purpose Office Environment*
- *Call Center*

The inputs to the configurator are given as numbers of appliance users along with their application profiles and usage patterns. The configurator automatically calculates recommendations for server memory capacity, number of processors, and the number of server network interfaces for the interconnect. These numbers can then be used to select an appropriate Sun Enterprise server.

If target workgroup characteristics match one of the deployment profiles, the configurator applet can be used directly to get a ball-park idea of server requirements based on an input number of users. The configurator can also be used to explore “what-if” scenarios for future growth or expansion in the workgroup. In addition, unique user and application data can be input into the configurator to get a much more accurate picture of server sizing for a given workgroup. Figure 3-1 illustrates the main configurator window containing the basic user classification for the K-12 deployment scenario.

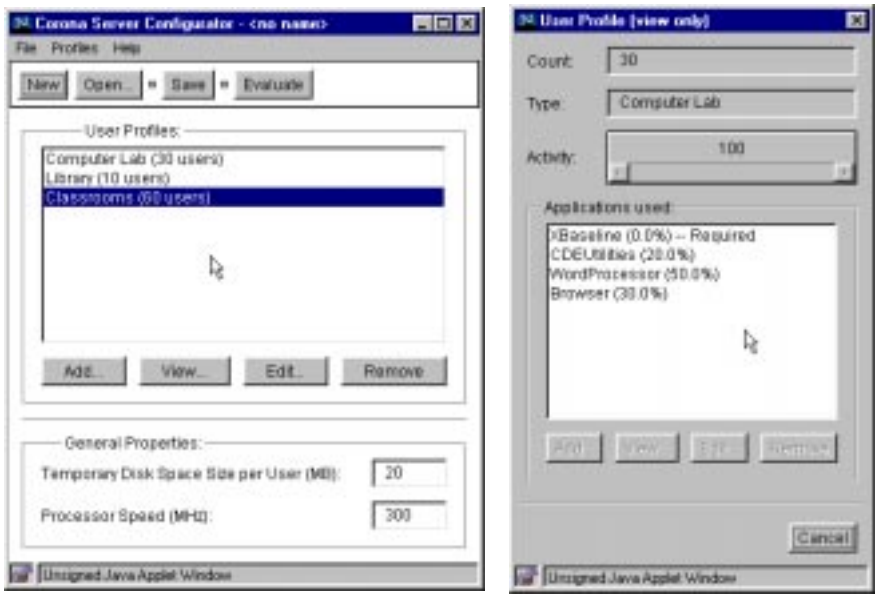

Figure 3-1 The Java technology Sun Ray server configurator and user profiles

Individual user groups are defined by the User Profile window on the right which enables overall activity levels to be set along with application distribution.

The configurator applet allows selection from a pre-defined list of applications as shown in Figure 3-2.

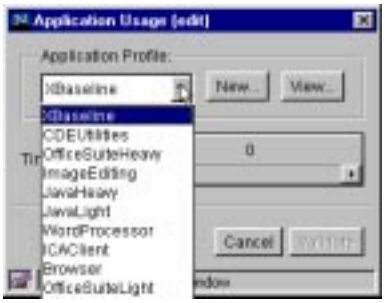

Figure 3-2 Selecting pre-defined applications

For each application defined in the configurator, an application profile defines processor usage, private memory size, shared memory size, and network bandwidth utilization (Figure 3-3).

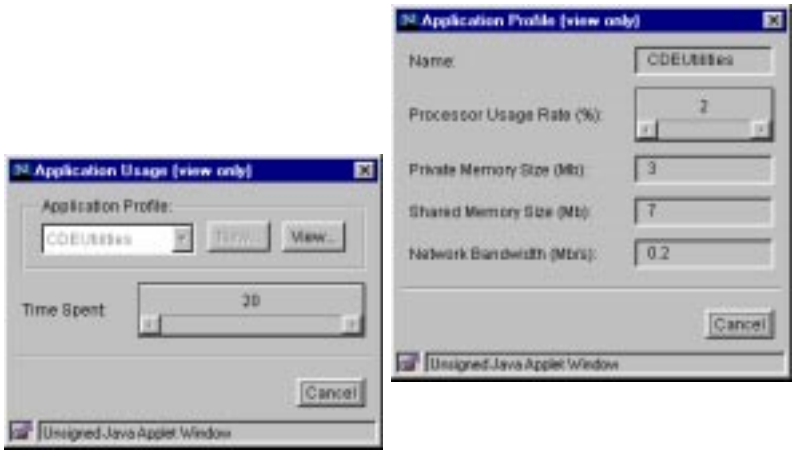

Figure 3-3 Profiling an individual application in the configurator

3

Example Sun Ray Enterprise Server Configurations

The configurator provides several starter profiles for common Sun Ray enterprise appliance deployment scenarios.

Educational (K-12) Environment

The Kindergarten through 12th grade (K-12) educational deployment scenario illustrates typical user and application profiles for three deployment locations in a K-12 school environment that includes a computing laboratory, library, and classroom environments. All appliances are to be connected to a single Sun Ray enterprise server and native applications will provide all functionality including Netscape™ Communicator for browsing, native CDE utilities (file, e-mail, text editing, etc.), and StarOffice 5.1 software for word processing and other office automation applications.

Figure 3-4 illustrates the main GUI of the configurator applet as configured with data for an Educational (K-12) scenario.

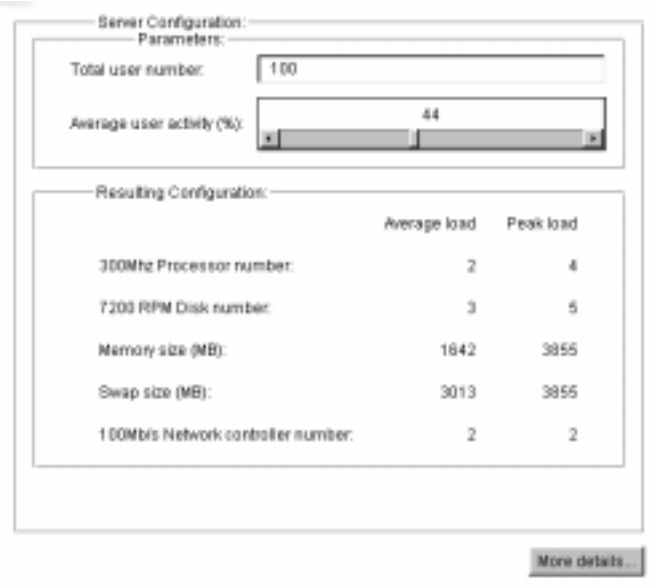

Figure 3-4 Configurator results for a K-12 environment

The configurator presents sizing estimate results in two columns: *Average Load* and *Peak Load*. Average load represents the server configuration required to maintain robust interactivity during average loading conditions. Peak load numbers represent the server configuration required if all of the users were to be active simultaneously.

Together the two columns define a range of appropriate server configurations which can be used to size an initial server. Factors like price and future expansion requirements should also be considered. One useful approach to sizing is to *select* a server platform based on the peak-load values from the configurator or spreadsheet and then *configure* the system with some extra capacity above the average-use level. This approach provides a server sized to meet current needs but with the capacity for future expansion.

From the main configurator applet GUI, the user can adjust both the total number of users, and the average user activity level to explore "what-if" scenarios. Average Load and Peak Load numbers are automatically updated. Users have full access to the data that drives the configurator applet and can explore specific target configurations by selecting the *More details* button which brings up the configurator GUI (Figure 3-1).

Library Automation

The library automation deployment scenario includes user and application profiles for a typical library. In the deployment scenario, library patrons use a Netscape browser for general Internet access and also for accessing most library automation systems. Telnet sessions are used to access older library automation services. Two user profiles are defined, one for librarians, and one for patrons.

Librarian profiles include limited use of a Citrix's ICA(R) client to access Microsoft Office applications on a separate server running Windows NT 4.0 TSE and Citrix's Metaframe^{TM} server software. CDE utilities are used for direct file management, editing, e-mail, etc. Both types of users primarily use the Netscape browser for accessing the Web interfaces to most library automation systems (such as *SIRSI*, and *Endeavor*). These applications provide catalog research and other library services through a combination of HTML, Java technology, and JavaScript software. General purpose Internet browsing for research and entertainment is also expected from both types of users.

The results for the library deployment scenario are shown in Figure 3-5. The default case can be customized to match a desired deployment through the configurator GUI.

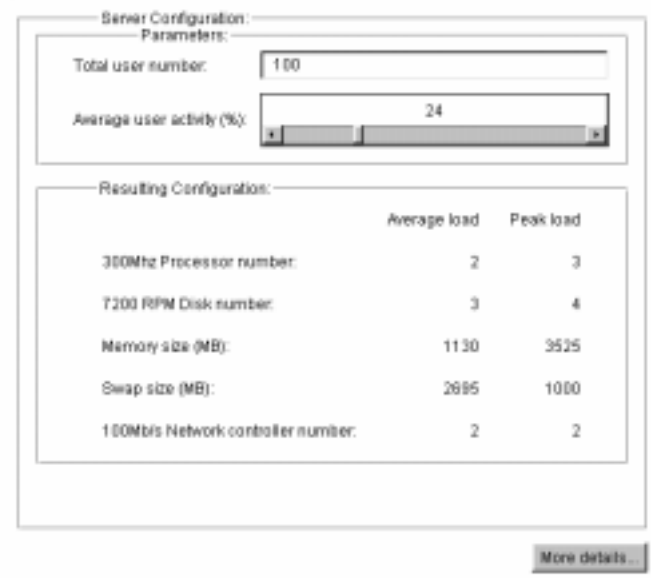

Figure 3-5 Configurator results for a library automation environment

General Office Automation

The general office deployment scenario includes typical user and application profiles for an enterprise office environment with heavy dependence on a company intranet for accessing enterprise applications and services. Three user profiles are defined: one for typical users, one for administrators, and one for power users.

All three user profiles include a mix of applications with primary use of Netscape Communicator and native applications for e-mail, calendar, name directories, etc. Native office automation applications are also part of the user profiles with sizing based on the server resources required for StarOffice 5.1 software.

The results for the general office deployment scenario are shown in Figure 3-6. The default case can be quickly adjusted to match an intended deployment by changing the number of users (both active and inactive) and altering other data through the configurator GUI.

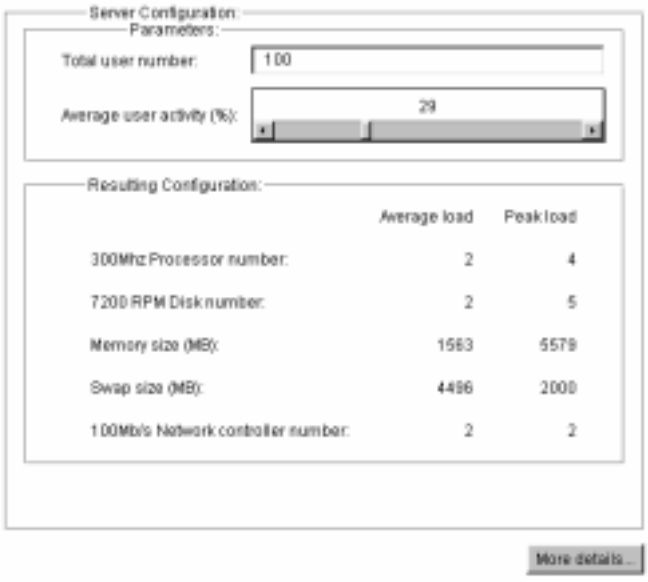

Figure 3-6 Configurator results for an office automation environment

Call Center

The call center deployment scenario includes typical user and application profiles for a Java technology and browser-based customer management system environment. Only one type of user is defined for sizing purposes. The most unique characteristic of the call center user profile is that users are expected to be active all of the time. As a result, average load is equivalent to peak load in the final analysis as shown in Figure 3-7.

The default case can be quickly and easily adjusted to match an intended deployment by changing the total number of users. In the case of the call center, the activity level is the amount of time a user actively uses the appliance or the percentage of appliances being actively used at any given time.

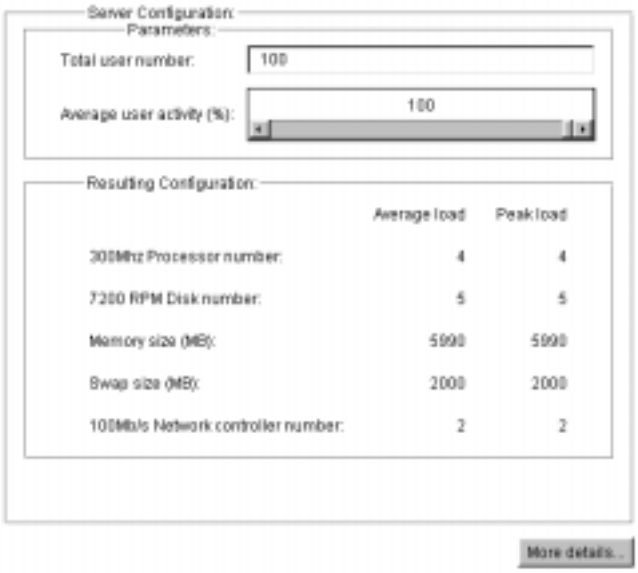

Figure 3-7 Configurator results for a call center environment

References

Sun Microsystems posts product information in the form of data sheets, specifications, and white papers on its Internet World Wide Web page at: *http://www.sun.com*.

Look for abstracts on these and other Sun technology white papers:

Sun Ray 1 Enterprise Appliance Overview and Technical Brief, White Paper, Sun Microsystems.

Assessing Scalability of the Sun Ray 1 Enterprise Appliance, White Paper, Sun Microsystems.

Deploying the Sun Ray Hot Desk Architecture, White Paper, Sun Microsystems.

Digital Media on the Sun Ray 1 Enterprise Appliance, White Paper, Sun Microsystems.

Sun Microsystems, Inc.
901 San Antonio Road
Palo Alto, CA 94303 USA
650 960-1300
FAX 650 969-9131
http://www.sun.com

Sales Offices

Africa (North, West and Central):
+33 1 30674680
Argentina: +54-11-4317-5600
Australia: +61-2-9844-5000
Austria: +43-1-60563-0
Belgium: +32-2-716 79 11
Brazil: +55-11-5181-8988
Canada: +905-477-6745
Chile: +56-2-3724500
Colombia: +571-629-2323
Commonwealth of Independent States:
+7-502-935-8411
Czech Republic: +420-2-33 00 93 11
Denmark: +45 4556 5000
Estonia: +372-6-308-900
Finland: +358-9-525-561
France: +33-01-30-67-50-00
Germany: +49-89-46008-0
Greece: +30-1-6188111
Hungary: +36-1-202-4415
Iceland: +354-563-3010
India: +91-80-5599595
Ireland: +353-1-8055-666
Israel: +972-9-9513465
Italy: +39-039-60551
Japan: +81-3-5717-5000
Kazakhstan: +7-3272-466774
Korea: +822-3469-0114
Latvia: +371-750-3700
Lithuania: +370-729-8468
Luxembourg: +352-49 11 33 1
Malaysia: +603-264-9988
Mexico: +52-5-258-6100
The Netherlands: +31-33-450-1234
New Zealand: +64-4-499-2344
Norway: +47-2202-3900
People's Republic of China:
Beijing: +86-10-6803-5588
Chengdu: +86-28-619-9333
Guangzhou: +86-20-8777-9913
Shanghai: +86-21-6466-1228
Hong Kong: +852-2802-4188
Poland: +48-22-8747800
Portugal: +351-1-412-7710
Russia: +7-502-935-8411
Singapore: +65-438-1888
Slovak Republic: +421-7-522 94 85
South Africa: +2711-805-4305
Spain: +34-91-596-9900
Sweden: +46-8-623-90-00
Switzerland: +41-1-825-7111
Taiwan: +886-2-2514-0567
Thailand: +662-636-1555
Turkey: +90-212-236 3300
United Arab Emirates: +971-4-366-333
United Kingdom: +44-1-276-20444
United States: +1-800-555-9SUN OR +1-650-960-1300
Venezuela: +58-2-905-3800
Worldwide Headquarters:
650-960-1300 or 800-555-9SUN
Internet: www.sun.com

Printed in USA