

# Efficiency Gains at the Edge of the Network

---

*Integrating Intelligent Networking and Security Functions  
Into Web Services Infrastructures for Bottom-Line Benefits*



Sun Microsystems, Inc.  
901 San Antonio Road  
Palo Alto, CA 94303  
1.800.786.7636  
1.512.434.1551

Copyright 2003 Sun Microsystems, Inc., 901 San Antonio Road, Palo Alto, California 94303 U.S.A. All rights reserved.

This product or document is protected by copyright and distributed under licenses restricting its use, copying, distribution, and decompilation. No part of this product or document may be reproduced in any form by any means without prior written authorization of Sun and its licensors, if any. Third-party software, including font technology, is copyrighted and licensed from Sun suppliers.

Parts of the product may be derived from Berkeley BSD systems, licensed from the University of California. UNIX is a registered trademark in the U.S. and other countries, exclusively licensed through X/Open Company, Ltd.

Sun, Sun Microsystems, the Sun logo, Solaris, Sun Fire, and Ultra are trademarks, registered trademarks, or service marks of Sun Microsystems, Inc. in the U.S. and other countries.

**RESTRICTED RIGHTS:** Use, duplication, or disclosure by the U.S. Government is subject to restrictions of FAR 52.227-14(g)(2)(6/87) and FAR 52.227-19(6/87), or DFAR 252.227-7015(b)(6/95) and DFAR 227.7202-3(a).

DOCUMENTATION IS PROVIDED "AS IS" AND ALL EXPRESS OR IMPLIED CONDITIONS, REPRESENTATIONS AND WARRANTIES, INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT, ARE DISCLAIMED, EXCEPT TO THE EXTENT THAT SUCH DISCLAIMERS ARE HELD TO BE LEGALLY INVALID.

---

# Contents

Executive Summary .....	4
The Service Point Architecture .....	5
Overview of the Edge .....	8
Primary Roles .....	8
Operational Environment .....	9
Integrating Edge Functions Into Tier 1 Web Servers .....	10
Blade Server Innovations .....	12
Blade Evolution .....	12
Specialty Blades for Edge Functions .....	13
Optimal Resource Use With N1 .....	15
Conclusion .....	17
Additional References .....	18

---

## Executive Summary

During the Internet boom of the late 1990s, companies deployed new computing infrastructures to efficiently deliver a multitude of Web-based services to millions of users around the world. In doing so, a three-tier architecture became a common infrastructure model for data centers in order to deliver the performance and availability users began to expect. By implementing three tiers—a Web tier, an application tier, and a database tier—businesses were able to optimize processing at each tier by scaling and tuning computing resources based on the unique software requirements of Web, application, and database servers.

Seeking to gain additional performance efficiencies, the industry is now eyeing the next computing frontier: the edge of the network. The edge of the network is the gray area between edge routers and the Tier 1 Web servers of the data center. It is here that, between a packet's arrival and its entry into the server room, intelligent networking and security functions can be implemented to optimize the flow of traffic in and out of the data center in order to achieve higher service levels at lower costs. These specialized functions include traffic management activities such as load balancing and priority processing, security measures such as firewalls and cryptography, and availability features such as redundant components and failover support.

This white paper provides an overview of the edge of the network and its functions. It also outlines the benefits of integrating the intelligent networking and security functions conducted at the edge into the Web tier to simplify Web services delivery, optimize data processing, maximize resource utilization, and improve overall system availability.

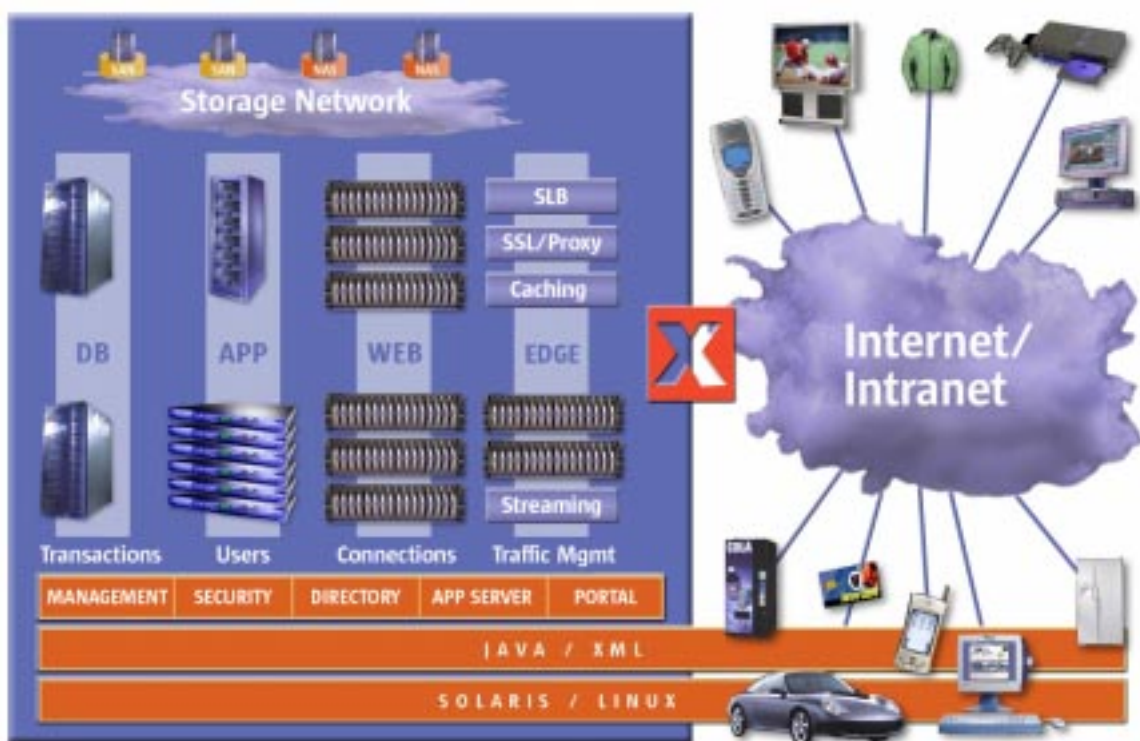
# The Service Point Architecture

One of the largest implications the Internet has brought in terms of change is the way applications are delivered. In the past, IT departments ran applications on hundreds or thousands of PCs to serve individual users. With the rise of the Internet, however, companies now need to serve millions or tens of millions of users around the globe. To achieve this, businesses have had to reinvent the way they develop and deploy their information systems.

Today's data centers must provide information and services to customers, partners, and employees around the clock, with users employing a wide range of connection devices to access data over the Internet, including PCs, laptops, cell phones, and PDAs. This new service architecture must be able to scale massively and provide continuous availability to handle ever-growing user demands. Sun refers to these Web-based applications as network services and calls the infrastructure that supports them the Sun™ Service Point Architecture.

The Sun Service Point Architecture is based on three primary tiers, as shown below:

- The Tier 1 Web or presentation tier
- The Tier 2 application tier
- The Tier 3 database tier



These three tiers generally refer to the functional roles of computer systems involved in

service delivery:

- *Tier 1*, representing the Web tier, presents the service to clients via protocols such as http
- *Tier 2*, representing the application or business logic tier, provides session persistence to clients for various applications
- *Tier 3*, representing the database tier, safeguards and manages processing of corporate data

By layering functions into three tiers, organizations can optimize scalability, availability, manageability, and security by using techniques appropriate for each tier:

- *Tier 1 Web servers* are generally stateless and are horizontally scaled as a way to rapidly respond to increasing workloads. By deploying many smaller servers, companies can quickly process a multitude of smaller, Web transactions that typically require little processing and run independently. Throughput is key, as well as the ability to dynamically provision systems in response to changing workloads.
- *Tier 2 application servers* must execute complex business rules, broker requests for content from the database, and store user session state. Because of the compute power required in this tier, vertical scaling is used for performance, while horizontal scaling is used for availability.
- *Tier 3 database servers* store an organization's most important information in the form of business-critical data. Because database applications are highly stateful, they are often deployed in asymmetric clustered configurations, with one active and one hot stand-by server. Capacity in terms of transactions per second is key for this tier, with larger servers supporting the process-intensive database transactions where large amounts of data are often manipulated in a single instance.

By implementing three tiers, data centers can disaggregate resources so servers can excel at specific tasks. For example, an average high-volume Web site requires systems to process the millions of small requests made each day for information. These requests include the hundreds of tiny transactions needed to assemble the components of a Web page or transmit e-mail. By assigning these tasks to a roomful of single-processor Tier 1 Web servers, larger back-end database systems can perform business-critical tasks more efficiently, helping to meet stringent service-level agreements. This modular building-block approach also allows software and hardware components to be easily replicated wherever additional capacity or availability is needed, whether through horizontal or vertical scaling, supporting highly efficient performance and availability gains.

Real-world implementations obviously also need a management plane and storage systems, whether dedicated or serving more than one tier. While this three-tier model is at times incomplete and oversimplified, all data centers can view the model as a useful reference for data center design.

Having fine-tuned this model over the years, data centers are now seeking additional efficiencies to reduce complexities and further lower their total cost of ownership. As a result,

hardware and software vendors are looking closely at the space between the data center and the networking infrastructure. Serving as the boundary between Web servers and networking gear, the edge of the network plays the unique role of traffic processing to support the smooth passage of data through the data center. As such, functions implemented at the edge of the network promise significant opportunities for ongoing innovation in the Sun Service Point Architecture.

---

# Overview of the Edge

The edge of the network is the boundary between the network infrastructure and the data center servers. A data request moves from the client, over the Internet using the networking infrastructure, and then must travel across the gray area of the edge of the network before being passed on to a Tier 1 Web server. It is here, at the edge, where all the preparation for moving data into the server room happens, with these functions focusing on traffic processing rather than actual data processing. The reason this is considered a gray area is because these functions can be performed either by or with Tier 1 systems or by networking equipment such as routers and switches.

## Primary Roles

In determining the best deployment of edge functions, it's essential to review key roles played:

- *Policy enforcement*: Control of what services a given client is allowed to view and access. This is done by setting access rules for individual users or administrative roles using a security product such as a firewall, Virtual Private Network (VPN) gateway, or a proxy that sits between the client and a server. Based on these predetermined rules, users are presented with a limited view of only the services they are authorized to access.
- *Service availability*: Ensure services are readily available to clients. Often this involves contractually specified levels of uptime, called service levels. This is done by deploying a service across multiple servers to protect against failures. Load balancing, including priority request handling for high service-level users, also helps increase service availability.
- *Security*: Protect the data center from system attacks and unauthorized data access. This is achieved through use of products such as firewalls, VPN gateways, proxies, cryptography, intrusion protection, and Denial of Service prevention.
- *Service virtualization*: Enable data centers to present a consistent virtual front to all services while pooling resources to support allocation of only what is needed to different servers and users at any given time. This abstraction hides back-end complexities and supports a lower total cost of ownership through resource sharing. It also supports economies of scale through the efficient provisioning, management, and co-existence of multiple applications for multiple entities on a consolidated platform. For example, service virtualization enables a service provider to host Web sites for two companies on the same Web servers and have the sites share the same load balancer; client requests for Company A content will remain separate from client requests for Company B content. By separating services logically and eliminating the need to deploy new cables or infrastructures, service virtualization simplifies data center management while increasing resource utilization.
- *Performance optimization*: Provide performance boosts without requiring additional system scaling. For example, proxy services such as SSL acceleration—which speeds encryption

and decryption of secure data packets—can deliver performance gains.

The following table shows how various discrete products currently deliver primary functions at the edge of the network. Because each product takes on two or more functions, function management via a sophisticated control plane becomes a complex, critical part of any solution.

<b>Edge Functions</b>					
<b>Products</b> <b>Functions</b>	<i>Load Balancer</i>	<i>Firewall</i>	<i>VPN Gateway</i>	<i>Proxy</i>	<i>Connection Manager</i>
Policy enforcement		√	√	√	
Service availability	√				√
Security		√	√	√	
Service virtualization	√				√
Performance optimization				√	√

## Operational Environment

The operational environment at the edge of the network also has unique attributes:

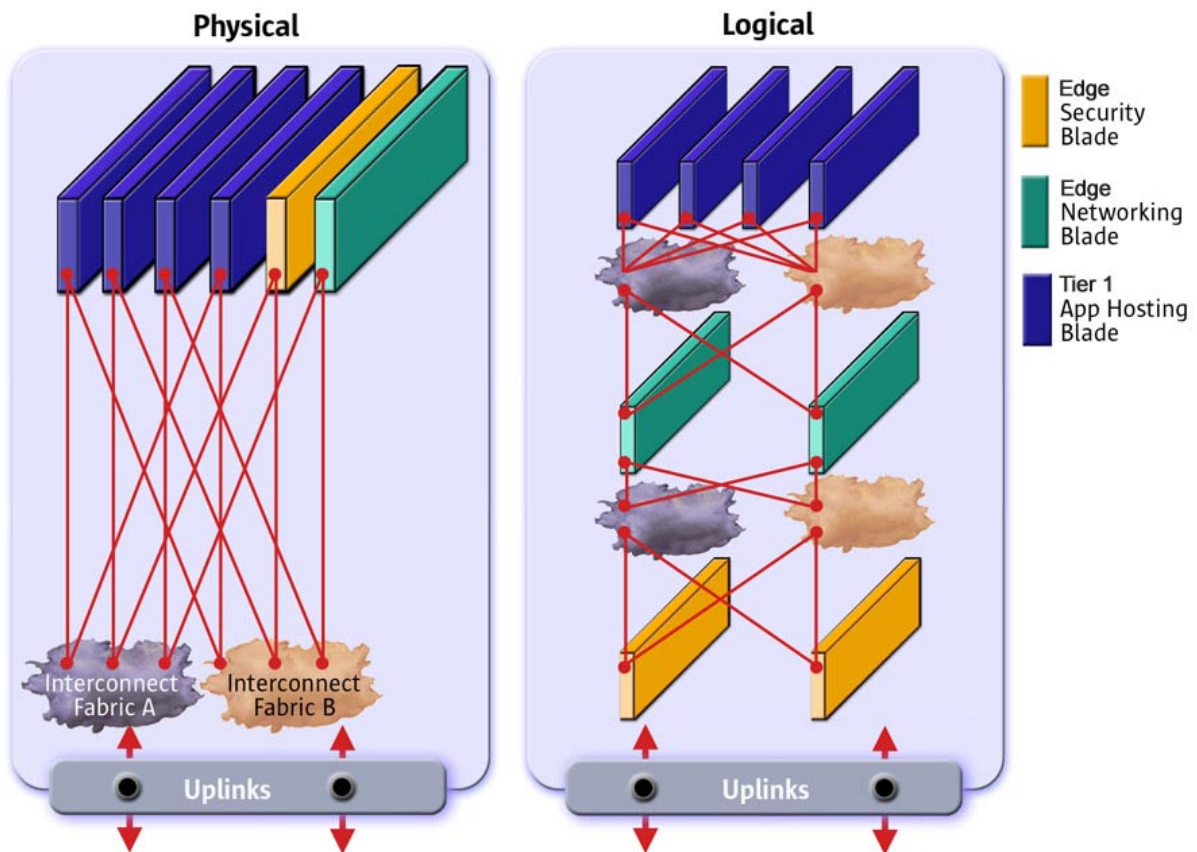
- *Physical deployment location:* Today’s data centers have typically located edge functions as a part of the network infrastructure. By changing this physical deployment and co-locating these functions with Tier 1, data centers can reap significant gains, including reducing cabling and management complexities by an order of magnitude.
- *Coordinated network parameters:* In order to steer and restrict traffic to and from the appropriate places, edge functions need sufficient knowledge about the internal structure of service delivery, such as the Web site tree and file structure. Edge functions also must have knowledge of the content structure, such as secure areas and caching. Currently these address assignments must be configured manually; next-generation edge and data center management products will handle network parameters logically to support automated configuration and modification.
- *Coordinated service and content parameters:* These parameters also need to be configured to define the services and content available at each address. For example, customers with a premium account may have access to a greater amount of services or content. This is another area ripe for innovation and automation. Imagine having an intelligent common management framework where a click of a button automatically configures the network and security parameters to enable premium service for a client.

- *Failover model:* Because specialized functions at the edge of the network affect overall service availability for clients, edge failover models need to complement the service failover model of Tier 1.

## Integrating Edge Functions Into Tier 1 Web Servers

The functions provided at the edge of the network and its required operational environment make it a natural evolution to operationally integrate intelligent networking and security functions into the Web tier instead of rolling these functions into the networking infrastructure. The physical co-location of edge functions in Tier 1 servers supports use of the same chassis and interconnects, minimizing cabling complexities while saving space and power. Such co-location lowers costs even further by supporting optimal performance and resource utilization through tightly integrated system designs. In addition, a single control plane will eventually manage both intelligent edge functions and Tier 1 activities, simplifying service management. For example, instead of having to configure both a load balancer and a Web server to point to a specific address for a service, the local configuration at each tier will be done automatically from a single virtualized system view.

Physical co-location of edge and Tier 1 functions provides management and performance benefits while still enabling functions to logically remain separate, as shown in this diagram.



The left side shows four Tier 1 application blade servers housed with a couple of specialized

blades for edge functions, in this case a security blade such as an SSL accelerator and an intelligent networking blade such as a load balancer. Each of these elements are attached to redundant interconnects. This physical co-location delivers good component packaging, increased density, and the flexibility to mix and match components within a chassis.

At the same time, however, this co-location does not force the components to function in a specific fashion. The diagram on the right demonstrates how system administrators can logically create a desired sequence of functions, even if the physical view has no such form. Physically all elements are directly connected to the interconnect fabric, but from the service composition point of view, they appear to function sequentially. As a result, infrastructures become fabrics instead of fixed linked chains. This supports intelligent packet routing, enabling a given packet to skip, for example, unneeded SSL decryption and move directly to load balancing to speed processing.

Due to the tight integration possible between edge and Tier 1 functions, Sun believes that the edge of the network is an ideal place for continued innovation, particularly when innovations can occur transparently for both servers and clients.

---

## Blade Server Innovations

Sun believes that the primary edge functions—including policy management, load balancing, and security services—are more tightly tied to Tier 1 functionality than that of the networking infrastructure. By integrating edge and Tier 1 functions, enterprises and service providers can reduce the cost and complexity of Web services delivery. This integration will eventually allow data centers to manage systems virtually by providing the comprehensive intelligence required to automatically configure functions for optimal performance, maximize resource utilization, and deliver higher availability across all tiers. Due to these affinities, Sun envisions development of blade servers that incorporate specific edge functions.

Blade servers are super-thin servers—not much more than a CPU with memory and a system controller—that can fit very efficiently into a large data center environment. Administrators can pack a dozen or more blades vertically into a single, small, rack-mounted chassis, making blades an ideal choice for Tier 1. Millions of requests demanding Web site data can be spread across multiple blade servers, with the server density minimizing use of floor space and power, providing excellent performance, increasing availability, and simplifying management.

Blade platforms have built-in intelligence that allows administrators to easily manage a pool of resources housed in the same chassis. This enables data centers to disaggregate networked compute resources from specific tasks or applications running on them, delivering exceptional flexibility in service provisioning. By adding intelligent edge functions to blade platforms in the form of individual specialty blades, data centers can further increase the flexibility and performance of their Web services.

### Blade Evolution

Today, many companies meet their network throughput needs by deploying dozens or even hundreds of small, relatively inexpensive servers known as 1RU servers because they take up one 1.75-inch, industry-standard rack unit of space in the data center. Blades are the evolution of this architecture and enable customers to once again improve economies of scale.

Blade computing provides exceptional price/performance through an integrated solution design that supports component sharing, optimized processing, and high availability. Blades can be stacked in a chassis, like books in a bookshelf, to pack more CPUs into a single rack, thereby creating very dense compute environments. Because blade CPUs use less power and generate less heat than traditional single-processor servers, cooling and power are shared. This reduces per-unit costs by spreading the cost of power supplies, fans, and other shared resources across multiple blades, while also making high-availability features like redundant power supplies and system controllers economically feasible.



Each chassis also only has one set of network cables and power cords, cutting down the hassles of management. In addition, each blade and chassis component is customer replaceable, making the entire platform easy to deploy and service, reducing both administration costs and errors.

Individual blade servers are hooked together using a common midplane and share common components, such as integrated switches. They are complemented by management software that keeps tabs on all the available system resources and assigns tasks to specific blades. While enterprises can do this today with a group of 1RU servers, blades provide the added advantage of integrated networking, hot-swappable components, and a seamless management framework.

The blade computing architecture is inherently redundant because many blades sit in each chassis. As a result, if a blade fails, managers can provision another one to take over without an interruption in service. Scaling is also easy, requiring a manager to simply slide in additional blades and allocate those resources wherever they are needed most. The end result is a flexible computing platform that is easy to scale, highly available, and very cost efficient.

## Specialty Blades for Edge Functions

As the demand for blade computing grows, hardware and software vendors will begin producing specialty blades for specific computing functions. Many of these specialty blades are expected to deliver optimized functionality for the edge of the network, including:

- VPN and SSL proxy blades with on-board hardware acceleration
- Load-balancing blades to replace costly, stand-alone load-balancing switches or appliances
- Firewall blades with software preloaded on a hardened operating system
- Caching blades that can act as local proxy caching servers

By integrating edge functions into specialty blades, data centers can gain significant benefits:

- Dramatic improvements in overall system density, utilization, performance, and availability
- Simplified management of systems and services
- Faster deployment and uniform provisioning of services
- Accommodation and management of fast growth and changing workloads
- Flexibility in resource provisioning

For more information on the Sun Fire™ Blade Platform and Sun's specialty blades for the

edge of the network, see the Additional References section below.

---

# Optimal Resource Use With N1

As part of the shift from application delivery to network service delivery, data management becomes considerably more complex. Companies need to be able to consolidate computing resources, easily manage them from a centralized view, and optimize resource utilization to maximize their return on investment.

Sun solutions first disaggregate resources by delivering component solutions that are tuned for specific functions. At the lowest end, blade servers will provide tightly integrated edge and Tier 1 functionality, enabling intelligent blades to achieve optimal processing for a specific function, such as SSL encryption or load balancing.

Sun plans to then reaggregate all of a company's optimized computing resources with N1—Sun's vision, architecture, and products for making entire data centers appear as one system. Instead of requiring system management for each server, N1 will provide a single virtual view of a company's complete computing infrastructure. This will enable a roomful of servers, blades, storage devices, applications, and networking components to appear as a single entity. As a result, system administrators will be able to manage all of these infrastructure components from a single, central view instead of sending out a team of engineers to reconfigure compute resources as workloads change. This virtual view will also support automated configuration, enabling a manager to state a parameter that will then be implemented intelligently across all affected tiers, converting centralized policy into distributed local policy.

N1 data center management promises revolutionary benefits, including the following:

- Increases business agility by supporting dynamic reallocation of resources as processing demands and business needs change
- Eliminates the need for individual systems to maintain excess capacity for peak processing demands by allowing excess capacity to be shared
- Boosts server utilization from industry norms of 15 to 30 percent up to 80 percent or higher
- Significantly reduces resource management complexity and the need for manual intervention
- Simplifies deployment of new services
- Protects technology investments by integrating existing equipment
- Increases availability by leveraging the N1 pool of resources to reassign services
- Provides a Web-based single point of control that delivers anywhere, anytime administration

Ultimately, these benefits can result in significantly lower operations costs by eliminating manual management tasks and simplifying resource allocation. These cost savings are critical, as today's companies spend more than 70 percent of their information technology budget on managing data center complexity. In fact, in order to afford the continued scaling of network services, companies must adopt a new approach to systems management.

By integrating edge functions into intelligent blade servers that support centralized N1 management, Sun will help to optimize resource use, simplify installation and administration, and deliver exceptional price/performance both in Tier 1 and at the edge of the network.

For more information on Sun's N1 vision, see the Additional References section below.

---

## Conclusion

Sun's vision for the future is data centers that support a flexible pool of compute, storage, and networking resources that can be configured and commissioned as rapidly as business requirements change. By integrating edge functionality into Tier 1 systems, Sun will enable data centers to disaggregate functions for optimized performance while supporting management innovations such as automated configuration. As a result, specialty blades that support next-generation N1 data center management will support more efficient, higher performing IT operations by radically simplifying business computing, reducing costs, and improving overall business agility.

---

## Additional References

### **For more information about the Sun Fire Blade Platform:**

Sun Fire Blade Platform Web page  
([www.sun.com/servers/entry/blade](http://www.sun.com/servers/entry/blade))

Sun Fire Blade Platform Technical White Paper  
([www.sun.com/servers/entry/b100s/b1600\\_wp.pdf](http://www.sun.com/servers/entry/b100s/b1600_wp.pdf))

### **For more information about Sun's specialty blades for the edge of the network:**

Sun Fire SSL Proxy Blade Web page  
([www.sun.com/networking/blades/ssl/](http://www.sun.com/networking/blades/ssl/))

Sun Fire Content Load Balancing Blade Web page  
([www.sun.com/networking/blades/lb/](http://www.sun.com/networking/blades/lb/))

### **For more information about Sun's N1 vision for next-generation data management:**

N1 Web page  
([www.sun.com/software/solutions/n1](http://www.sun.com/software/solutions/n1))

N1 Executive Brief  
([www.sun.com/software/solutions/n1/wp-n1.pdf](http://www.sun.com/software/solutions/n1/wp-n1.pdf))

N1 Executive Essays  
([www.sun.com/software/solutions/n1/essays](http://www.sun.com/software/solutions/n1/essays))



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

1.800.786.7638  
1.512.434.1511

<http://www.sun.com/>