

The Sun Storage Value Proposition for High Performance Computing

A White Paper



© 2002 Sun Microsystems, Inc. All rights reserved.
901 San Antonio Road, Palo Alto, California 94303 U.S.A.

TRADEMARKS

Sun, Sun Microsystems, the Sun logo, Sun StorEdge, and Solaris are trademarks or registered trademarks of Sun Microsystems, Inc. in the United States and other countries.

Oracle is a registered trademark of Oracle Corporation. UNIX is a registered trademark. in the United States and other countries, exclusively licensed through X/Open Company, Ltd.

THIS PUBLICATION IS PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, OR NON-INFRINGEMENT.

THIS PUBLICATION COULD INCLUDE TECHNICAL INACCURACIES OR TYPOGRAPHICAL ERRORS. CHANGES ARE PERIODICALLY ADDED TO THE INFORMATION HEREIN; THESE CHANGES WILL BE INCORPORATED IN NEW EDITIONS OF THE PUBLICATION. SUN MICROSYSTEMS, INC. MAY MAKE IMPROVEMENTS AND/OR CHANGES IN THE PRODUCT(S) AND/OR THE PROGRAM(S) DESCRIBED IN THIS PUBLICATION AT ANY TIME.



Please
Recycle

The Sun Storage Value Proposition for High Performance Computing



An Analyst's View of High Performance Computing

High Performance Computing (HPC) and SANs have had affinities for each other for years. But, until now, no one's really offered a fully-integrated solution that really exploits those affinities and opportunities.

Performance is the point of HPC — its "middle name" if you will. Everyone can put enough compute resources in a room to execute "billions and billions" of operations per second. Continually feeding that beast enough data to keep it happy and productive is a lot harder. With HPC SAN, Sun claims to have done this-to the tune of two gigabytes per second I/O throughput per file system (and much higher rates than this in the aggregate across multiple file systems and storage arrays). It's an aggressive claim to be sure, but one that we've substantiated with Sun HPC SAN users.

At the foundation of the Sun solution lies QFS. QFS lives up to its billing as the Quick File System. QFS is highly parallelized, intelligently optimized, and it provides many "knobs and dials" that HPC users can employ to tune the file system to the storage environment for maximum performance. Sun also uses SANergy from IBM's Tivoli unit to good effect. SANergy enables deployments across multiple OS environments, which are common in HPC. As a result, applications within the majority of prospective customer sites can be brought under the SAN HPC umbrella with minimal disruption to production environments.

SANs bring data movement, data management, and data availability advantages to HPC that more traditional direct attached storage models just can't match. SANs and HPC are now an attractive combination for both traditional scientific/engineering applications and for an increasing number of commercial applications requiring massive compute power. In HPC SAN, Sun's done a nice job of bringing the required hardware, software, and services components together, qualifying them in concert, and using them to address genuine HPC pain-points.

John Webster,
Senior Analyst and IT Advisor,
Illuminata, Inc.



Introduction

The requirements of HPC and commercial computing markets are converging. In HPC, this convergence is being driven by Web technologies that offer access to advanced computational resources from anywhere in the world, at any time, from any platform. As a result, “commercial” computing requirements, such as high availability and data movement and management, are becoming as important as performance and scalability.

At the same time, business organizations are now using HPC technologies and techniques in a new generation of commercial applications — such as decision support, financial analysis, data mining, and bioinformatics.

To jointly tackle these challenges with customers, Sun is integrating new and existing technologies, products, and services into a new offering for the HPC marketplace.

HPC has long been focused on improving the performance of the compute server — the speed of the processor and how many numbers it can crunch every second. Granted, HPC does — and will always — require an enormous number of compute cycles, a vast number of floating point operations per second (FLOPS). But Moore's Law, Joy's Law, and the rise of highly modular, networked systems annually provide more compute horsepower at a lower cost. The processing complex is no longer where the key HPC problems lie. In fact, the rapid increase in available processing cycles has dramatically increased the load on the I/O, data movement, and storage systems, in effect moving the performance bottleneck from the compute platform to the I/O subsystem. For some time many applications have suffered from the mismatch between the advances in processing power and the relative sluggishness of advances in I/O. In the recent past, the majority of applications have experienced an explosion of data growth that has highlighted this ever-increasing disparity between compute power and getting the data to the processors.

HPC storage has been typically characterized by hard linkages between servers and storage subsystems, resulting in poor reuse of resources, limited scalability, and cumbersome “islands” of physical storage. In addition, HPC environments, like most others, have become increasingly decentralized over time. Systems producing massive amounts of data need to move it — both to other systems within what is often a worldwide network, as well as to long-term tape storage. Whether performing a geographical analysis of satellite telemetry or moving pharmaceutical results around a grid-computing



framework, in HPC, time-to-analysis is critical. The Sun HPC storage area network (SAN) solution breaks through traditional limitations. It augments data availability as well as solving compute pipeline challenges, either for large single files requiring a few huge pipes or for thousands of smaller files requiring a multitude of large pipes.

Addressing the Problem

HPC systems traditionally use tightly coupled architectures in which the storage for each operation is attached directly to the associated computation platform. These tightly-coupled architectures prove cumbersome when data must be moved from one system to another. Over time they become extremely expensive to grow and manage.

Until recently, a storage area network (SAN) that can decouple the one-to-one linkage between computation engine and associated data storage has been available only to commercial computing environments. In those environments, SANs typically deliver:

- Enhanced utilization of storage resources
- Greater data availability
- Centralization of data and storage management functions
- The potential for data sharing among applications residing on disparate OS platforms

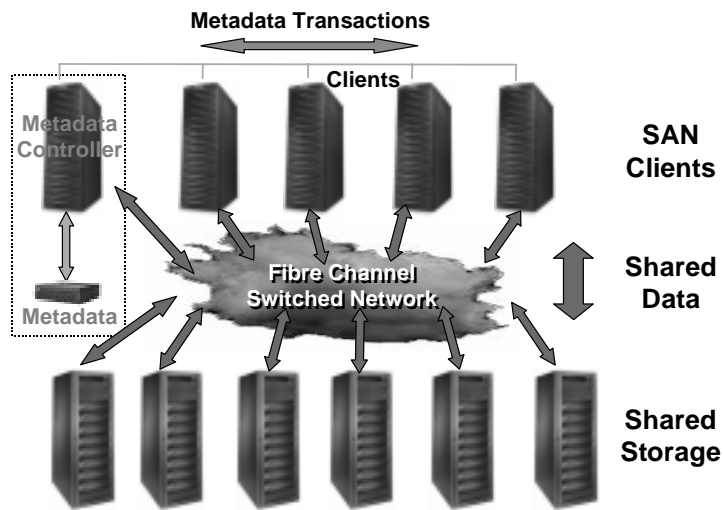
As with other advances in technical computing, HPC users have been early adopters of SANs as well as in pioneering the use of shared file systems. Indeed, SANs address exactly the data movement and management problem that HPC has ordinarily faced. However, HPC users need performance that would be considered in the “lunatic fringe” category by most commercial SAN users. Integrated SAN solutions have not until now focused on extremely high data rates between individual files, file systems, and applications.

At the same time, the do-it-yourself SAN alternative is fraught with integration nightmares. Now, however, the wait is over. The Sun HPC SAN is a leader in providing a fully integrated SAN solution which combines the SAN value propositions — efficiency and manageability — with the high level of performance and data sharing that HPC users require.



In an HPC environment, a SAN can enable a more efficient and cost-effective solution by providing a single, logical view of data and a single point of storage management. In addition, a Sun HPC SAN enables the sharing of storage devices and data between heterogeneous platforms such as UNIX[®] and Windows, allowing HPC data center managers to eliminate redundant islands of data attached to each of its computational hosts. A SAN can also free up compute resources that were previously performing I/O management, allowing them to focus on what they do best — computation — while SAN-based components handle data management and movement.

Physical SAN Architecture



The Sun HPC SAN

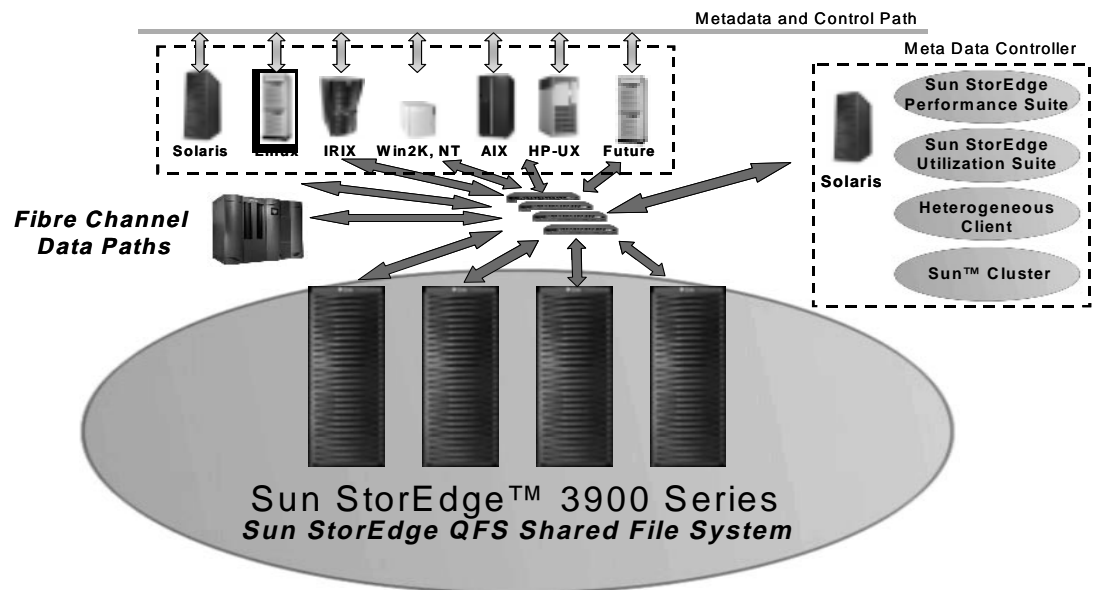
Sun integrates all of the hardware and software components required to deliver a robust SAN for HPC environments. The Sun HPC SAN is capable of scaling to meet the capacity requirements of the largest technical computing sites while preserving essential I/O performance characteristics. I/O performance is maintained by combining a shareable file system optimized for HPC environments (Sun StorEdge[™] QFS) with a Fibre Channel SAN in which data movement is managed by Tivoli SANergy. The combination is capable of delivering throughput at channel speeds; in multi-channel configurations, the combination can drive in excess of 2 gigabytes per second per single file or aggregate within the file system (tested), and throughput rates much higher are possible.



The components of the Sun HPC SAN include:

- *Appropriate host bus adaptors (HBAs) and switches* for high-bandwidth data transfers using widely available, industry-standard Fibre Channel (FC) connections
- *The Sun StorEdge™ Performance Suite* shared file system service (Sun StorEdge QFS), a file system optimized for HPC applications
- *Tivoli SANergy*, a SAN management software package that allows heterogeneous, multi-system data sharing and minimizes data movement among applications and user groups
- *The Sun StorEdge™ Utilization Suite* recovery and archive service (Sun StorEdge SAM-FS), a hierarchical storage manager which automatically archives data, including very large files, without administrator intervention
- *Storage equipment* — the Sun StorEdge™ 3900 series storage arrays
- *Support services* for equipment sizing, installation, integration, testing and support

The following pages describe some of these key components.





Sun StorEdge Performance Suite Shared File System Service (Sun StorEdge QFS)

The Sun StorEdge QFS file system is a shared file system service in the Sun StorEdge Performance Suite software was originally developed and marketed by LSC. It is designed to deliver maximum performance in environments where very large files — in the gigabyte range — are often encountered. Sun has alleviated many potential performance bottlenecks along the I/O chain between the application and its associated storage media. The developers of this file system are extremely “hardware aware,” and are focused on wringing out every ounce of performance.

There are several file system features that are specifically interesting to HPC users:

- *Data Sharing.* Rather than sending files from one application to another, or from one group of users to another, Sun StorEdge QFS allows users to share files in place. This “in situ” sharing saves network bandwidth, production time, and the costs associated with high-bandwidth communications ports. In short, to limit problems and costs associated with moving large data files around in a HPC environment, never move them at all. Simply allow shared file access by authorized users and leave the data on whatever storage device is most appropriate for the file types being stored. This means that true concurrent design can be practiced by engineering departments, petrogeologists can get immediate access to large oil field maps, and financial traders or analytic applications can simultaneously access time-critical data feeds.
- *Data striping.* Sun StorEdge QFS creates “stripe groups” for each file so data can be read or written in parallel across multiple channels and disk drives. The maximum tested configuration is a single file system with a 24-wide stripe group. Using FC-AL drives, this configuration yielded throughput in excess of 2 gigabytes per second — achievable either per single file system or as an aggregate across the file system. Sun StorEdge QFS is excellent at using however many parallel storage resources that an application requires. The scaling characteristics of the Sun HPC SAN are virtually linear as users add FC ports.
- *Metadata separation.* File system metadata typically resides on the same disks as user data. This condition often causes disk heads to constantly reposition themselves. During reads or writes, the disk heads are often forced to physically move back and forth between the places where metadata is



stored and where user data is stored. That's a problem. This physical motion is the slowest process in the entire computational chain, and thus creates the biggest opportunity for bottleneck. The result of this thrashing back and forth is increased latency and decreased performance. Sun StorEdge QFS instead separates file system metadata from user data so that both can be accessed and updated simultaneously on physically separate devices. That is, it brings parallelism into I/O management.

- *Variable block size.* When using file systems that mandate a single, fixed data block size, undue latency — not to mention wasted disk capacity — can be introduced when the file system block size is not tightly aligned with the array subsystem or application requirements. Varying the block size allows system administrators to tune the file system to the specific storage environment and data access patterns for maximum performance and capacity utilization.
- *Automatic Direct I/O.* When large file transfers are required, forcing data through buffer cache in the processor's memory can actually add latency as buffers are constantly flushed and refreshed. Sun StorEdge QFS has several options that allow system administrators to connect large data transfers directly to disk, bypassing processor-resident disk buffers.
- *Pre-allocation.* Using the Sun StorEdge QFS shared file service Application Programming Interface (API), an application can request that the Sun StorEdge QFS file system reserve a contiguous set of free disk blocks prior to a write operation. It can then write data, block after block, without causing the disk head to move. Later, when data is subsequently read, drive heads again require no movement. This type of careful optimization is often just what is needed for HPC applications.

Tivoli SANergy

Tivoli SANergy is third party software which enables heterogeneous data sharing in a SAN environment working with Sun StorEdge QFS. SANergy also enables a multiple writer capability to Sun StorEdge QFS. Supported platforms include Solaris, as well as Apple MacOS, IBM AIX, Red Hat Linux, SGI IRIX, Tru64 UNIX, and Microsoft Windows (NT and 2000).

SANergy consists of add-in software that resides on all participating application hosts, coupled to a metadata-processing component (or metadata controller) that lives outside of the application environment.



Sun StorEdge Utilization Suite Recovery and Archive Service (Sun StorEdge SAM-FS)

The recovery and archive service in the Sun StorEdge Utilization Suite software (Sun StorEdge SAM-FS) presents users with a virtual infinite disk, thereby dramatically easing the job of the administrator. Sun StorEdge SAM-FS automatically handles the otherwise onerous task of migrating files from one medium to another — whether from fast disk to slow, or from disk to tape — as needed to optimize storage performance and capacity. This allows administrators to manage both the human costs associated with monitoring disk space, and the costs associated with rapid and potentially uncontrolled usage of expensive storage media.

Sun StorEdge SAM-FS automatically migrates less-frequently used files from disk to higher-density, less-costly storage media such as tape. It optimizes tape library performance by storing groups of files logically in an open TAR format well known to UNIX users and almost universally processed by UNIX utilities. Sun StorEdge SAM-FS inherently works to place these files onto the same media and at the same time. Grouping and writing files in this way reduces the number of media mounts required to both write data to or read data from tape, and generally reduces positioning time required to get from file to file. Upon retrieval, individual files — not large groups of files or the entire file system — are read back to disk or disk cache.

Sun StorEdge™ 3900 Series Storage Systems

Sun StorEdge™ 3900 series storage systems are easy to order, deploy, and manage. Integrated call-home and remote diagnostic capability provides high levels of application availability. The Sun StorEdge 3900 series run high-bandwidth applications in the Solaris Operating Environment on single or clustered servers. Unlike other systems, Sun StorEdge 3900 series storage systems cost-effectively scale bandwidth performance in a positive, consistent, linear manner.

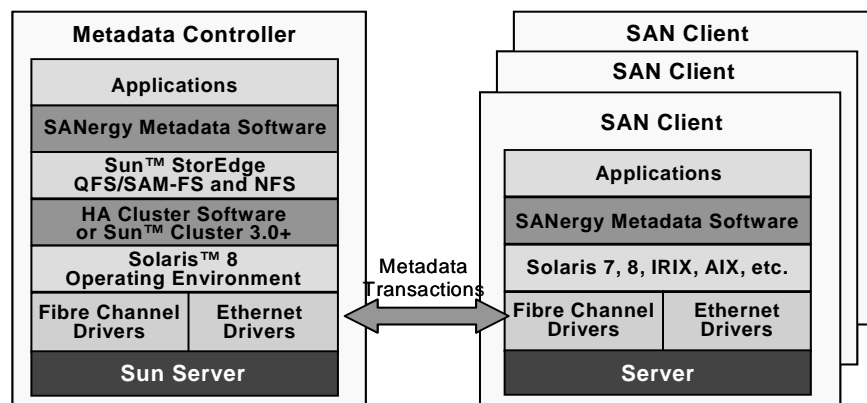
The Sun StorEdge 3900 series storage systems are designed for running single or clustered high-bandwidth applications such as HPC. As capacity is added, the full-fibre, non-blocking, positive scaling system architecture enables performance to increase consistently and predictively. Each modular storage unit provides two 1 GB RAID controllers and eighteen disk drives, which adds a 200 MB/sec increment of bandwidth.



The Sun StorEdge 3900 enterprise systems offer great configuration flexibility, allowing administrators to optimize their disk systems for the maximum available performance. Each controller module can be configured to stream large block-sequential I/O to yield throughput rates closely approaching the maximum sustainable channel rate for both reads and writes. Throughput scales linearly as channels and arrays to RAID stripe groups are added. Caching options include mirrored read/write cache or read-only cache.

The Sun StorEdge 3900 series storage systems exhibits the greatest — and most noticeable — performance improvement when I/O requests are greater than or equal to 64 KB in HPC applications. There are exactly eight drives served by each hardware Fibre Channel RAID controller. Each controller contains 1 GB of read/write dynamic adaptive cache, which automatically adjusts to match the workload by changing the cache segmentation on the fly, resulting in very high sustainable throughput rates. Each RAID controller employs three separate 64-bit PCI buses so administrative traffic does not interfere with application traffic, reducing effective overhead by as much as 30 percent. The controller's XOR engine is a separate processor with its own cache which provides an extremely high degree of parallelism through the controller's pipelined XOR, allowing up to 128 I/Os to be stored and concurrently piped (i.e., up to 128 XOR calculations can be performed simultaneously) without interruption to data cache access.

Hardware and Software Stack



***The Result: HPC-Class
Heterogeneous, shared-file system performance!***



Advantages of the Sun HPC SAN

Compatibility with existing resources. The Sun HPC SAN supports multiple, disparate application hosts and multiple, disparate storage subsystems — as it has from day one. The wide range of system types supported by the Tivoli SANergy software allows the maximum use of legacy systems, thereby extracting value from already paid-for assets. The range of support also increases the ability to deploy heterogeneous environments for testing and other reasons.

Reduced dependence on existing network for data movement. When using SAN technology, the data stays fixed, but the channel connections that access the data move. Changes in channel access occur at the speed of a Fibre Channel switch. In essence, a multi-gigabyte file can be transferred from one application to another in a matter of microseconds.

Reduced I/O processing overhead. Both the SAN approach and the Sun StorEdge Performance Suite Shared File System Service reduce the processing overhead required to transfer data to and from the application host, allowing the host to devote more processing cycles to what counts — the HPC application.

Enhanced change management. Because data is accessed globally, rather than from host-resident storage, applications can either remain on their existing platforms or with a minimum of disruption can be migrated to higher performance processors. Capacity and performance can be added incrementally without a system outage.

Fluid reconfigurations. HPC environments typically require more fluid system configurations than do more traditional transaction processing environments. With Sun HPC, storage configuration changes can also be made without disturbing applications. Changes in configuration can take the form of adding additional system capacity, trying out new equipment from different vendors, and redeploying gear to match the needs of a given problem. The Sun HPC SAN allows for highly dynamic configuration of the processing elements of the HPC solution without disrupting the data or requiring high bandwidth data movement — an increasingly difficult proposition as dataset sizes reach into the hundreds of terabytes.



Multi-dimensional scalability. The Sun HPC SAN easily handles many-terabyte configurations, and it can be scaled upwards to the petabyte range without disrupting operations. Performance can also be scaled linearly up to gigabytes per second for each file system and application host, again without disruption.

Automated exploitation of technology advances. Higher-density, lower-cost storage media can easily be integrated. The Sun HPC SAN supports a variety of storage media types, from very high-performance disk arrays to very high-density tape libraries. The Sun HPC SAN automatically chooses the most advantageous repository, allowing administrators to architect very high capacity online data stores without committing all data to disk arrays. The use of Sun StorEdge SAM-FS recovery and archive service in the Sun StorEdge Utilization Suite automates the migration of data when it is no longer in day-to-day use.

Efficient operations. Efficiency is the watchword throughout the Sun HPC design. The separation of file system metadata from user data minimizes head movement on storage devices. Efficient file system operation minimizes system time required for I/O. Efficient and flexible volume management accommodates a wide range of storage configuration requirements while consuming minimal system resources. And the Sun StorEdge Performance Suite software's Sun-QFS shared file system service ensures the continuous flow of data as volumes grow via file system striping. Efficient operations mean more resources are left over for application processing.

Effective centralized utilization of storage assets. Large quantities of storage may need to be allocated to store intermediate results. In a typical environment with multiple problem sets, a single storage pool can address storage needs much more efficiently because maximum storage requirements for parts of the computing environment will typically occur at different times. Efficient, centralized management of storage allocation is an important element of this mix.

Proven and refined. Sun understands the HPC environment holistically, including HPC's critical data access and management requirements. And the Sun solution — including both the HPC SAN product set and Sun implementation and integration services — has been repeatedly proven and successful in large-scale, real-world enterprises.



Conclusion

Back in its early days, high performance computing was all about compute, compute, and more compute. The problems and the computers that solved them were smaller, and so were datasets that recorded them. No longer. Today HPC users face vast problems, whether in molecular modeling, computational fluid dynamics, weather forecasting, petroleum exploration, mechanical simulation, or the hundreds of other high-return HPC applications. The emerging life sciences fields of genomics and proteomics bring even higher-end requirements.

Vast problems to solve mean vast datasets and extreme data handling requirements. Whatever the HPC application, establishing reliable, high-performance, low-overhead access to many-terabyte datasets is now the most pressing requirement for progress.

The Sun HPC SAN helps to solve this key problem. It provides substantial value through:

- Unmatched performance and scalability through the selection and pre-integration of “best-in-class” components for HPC
- Increased efficiency. There is no need to alter the storage system architecture as additional computational servers are added to the data center even when those are different types of servers
- Reduced costs, which result from the elimination of hardware redundancy and through the consolidation of previously fragmented storage systems
- Reduced time to market for applications, which accrues from the ability to add servers and storage capacity without disruption. The ability to share files across a SAN also speeds time to market by enabling faster access to critical data

References

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



Sun Microsystems Incorporated
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 0 1252 420000
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