White Paper

Backing Up and Restoring Multi-Terabyte Data Sets

SGI, LEGATO, and Others Break the 10TB-per-Hour Barrier

Executive Summary

Backup and restore performance have become increasingly important as data set size in high-performance computing (HPC) environments has grown. Until now, published benchmarks of backup and restore speeds have focused almost exclusively on tests concerning a relational database, ignoring the large and growing pool of file data. For this reason, SGI, LEGATO, StorageTek, Brocade, and LSI Logic Storage Systems undertook an effort to benchmark backup and restore performance using a storage area network (SAN) under conditions typical for HPC. A variety of scenarios using both image and file-by-file backup were tested to help customers better understand the strengths and limitations of each approach.

Using LEGATO NetWorker®, this benchmark achieved results that far exceeded previously published numbers in every category tested:

- Sustained throughput in excess of 10TB per hour was demonstrated for file-by-file backup
- Over 10TB of data was backed up within 60 minutes of initiating backup
- A 1TB data set was backed up in only seven minutes from start to finish
- Backup of a single 10TB SGI® XFS® filesystem demonstrated sustained throughput in excess of 6TB per hour
- Restore performance for all file-by-file tests ranged from 4 to 4.5TB per hour, up to two times the performance of the previous best result
- Image backups—using snapshots to ensure consistency—achieved 7.2TB per hour sustained throughput for backup and a record 7.9TB per hour sustained throughput for restore

These results were achieved using a 32-processor SGI® Origin® 3000 server, SGI® TP9500 RAID storage arrays (developed by LSI Logic Storage Systems, Inc., and marketed as SGI TP9500), 48 StorageTek® T9940B 2Gb Fibre Channel tape drives installed in a StorageTek® PowderHorn® 9310 tape library, Brocade® Fibre Channel switches, and LEGATO NetWorker® 7. All components are readily available from SGI or the individual vendors. SGI® Professional Services can assist in designing backup solutions that can meet or exceed the results reported in this paper and that are tailored to unique requirements. Combining these components with the SGI® CXFS™ filesystem, SGI can create a high-performance data-sharing environment in which all systems share access to data at SAN speeds while achieving exceptional backup and restore performance.

1.0 Taking Backup and Restore Performance to a New Level

Organizations that depend on high-performance computing for critical research and development are struggling to cope with data sets that are growing at astronomical rates. The data management problem is not just one of providing adequate storage, but also providing adequate protection of critical data through backup and restore. In many cases, the scope of important simulations must be limited because of an inability to effectively manage the huge volumes of data that can result from detailed models.

SGI has long recognized the unique problems associated with the backup and restore of multi-terabyte data sets. In 1997, SGI—in working with LEGATO, IBM, and Computer Associates—was the first to demonstrate the ability to back up a 1TB database in less than one hour. This record stood for over four and a half years until VERITAS demonstrated a 2TB-per-hour result in May 2002. Later in 2002, a group led by Computer Associates demonstrated a database backup and restore solution with sustained throughput of 2.6TB per hour for backup and 2.2TB per hour for restore. Early in 2003, Hewlett-Packard further upped the ante with a database benchmark that achieved throughput of 3.62TB per hour for backup and a more modest 1.29TB per hour for restore. While impressive by most standards, these historical results have been targeted for database backup and have ignored the large and growing pool of file data. Even performance at these levels may be inadequate for the needs of today's large technical data stores.

For this reason, SGI and its partners—LEGATO, Brocade, LSI Logic Storage Systems, and StorageTek—set out to demonstrate a data-protection solution for today's multi-terabyte environments. The objective was to improve upon the current industry-best backup and restore benchmarks by a factor of three. As part of this benchmark effort, a variety of test scenarios were carried out to help customers choose the best backup strategy for their particular needs. These results break all previously established performance records for both file-by-file and image backup and restore. Results of the various tests performed are summarized in figure 1, along with the results from the previously mentioned tests for comparison. Each test is described in more detail in the following section.

The key to achieving this level of performance is parallelism in the hardware and software components that make up the solution. Each component of the solution can sustain many parallel data streams without interruption. This paper examines the various tests that were carried out and the performance achieved for each and examines the hardware and software components that were used to achieve these results. The information included covers the highlights of all backup and restore results. A technical white paper is currently in preparation which will describe each of the tests and their respective configurations in greater detail.

2.0 Measured Backup and Restore Performance

The backup-and-restore approach best suited to a particular environment depends on a variety of factors, such as the total amount of data being stored and the frequency with which stored data changes. For this reason SGI and its partners tested a variety of scenarios that were considered to be of interest for various HPC environments. Each scenario measures backup and restore performance for a 10TB data set, not a relational database housing 10TB of data. Online database backup performance depends on the database management system and the ongoing workload. Offline database backup is equivalent to an ordinary image or file backup of the devices or files

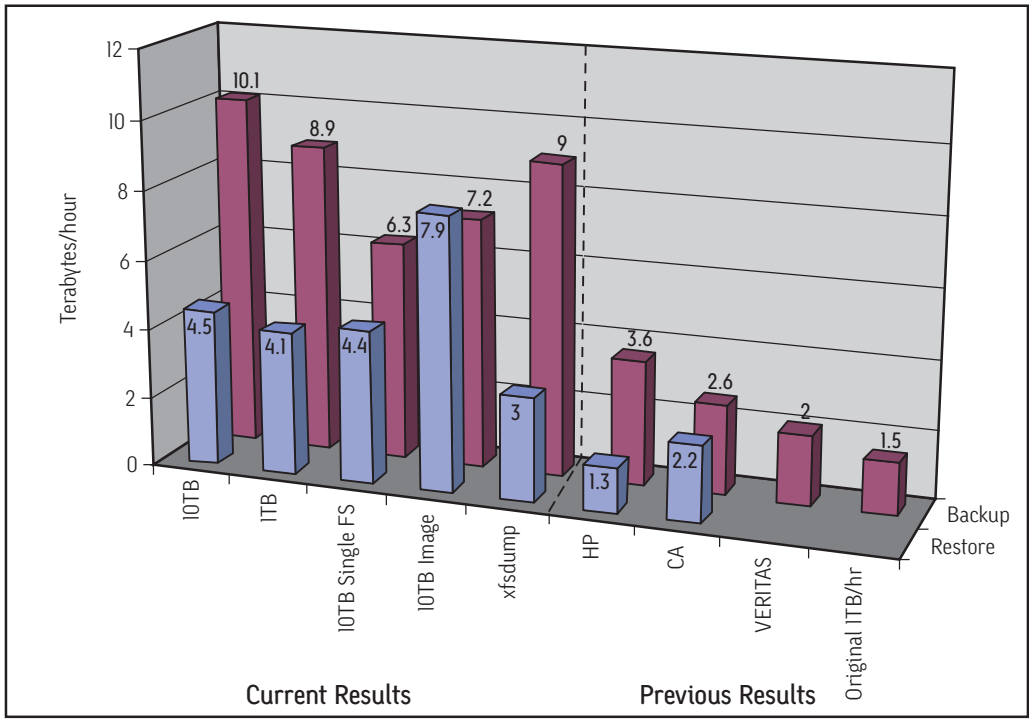

Fig. 1. Sustained throughput achieved by various tests described in this paper, plus SGI's previously demonstrated 1TB-per-hour database backup (1997)

where the database resides. These results simulate what a customer might experience with a data protection solution designed for an environment with a large amount of file data to back up. The specifics of each test and the rationale for choosing that particular scenario are described in the following sections. Two categories of backup and restore were tested:

1. **File-by-file backup and restore** using NetWorker® to protect and catalog each file in each test filesystem: The advantages of this approach are that it supports both full and incremental backups, and individual files can be easily restored. A possible disadvantage is that restore speeds are typically slower than backups.
2. **Image backup and restore** using NetWorker: The raw data from the storage volume is streamed directly to tape, resulting in an exact image of the volume including filesystem data structures. Image backups are restored by copying the entire image from tape back to the original storage volume or a similar-sized volume. Therefore, they generally offer the same or better restore performance as backup performance. However, individual files are difficult or impossible to restore unless there is a mechanism to locate the blocks that correspond to a particular file in the tape image.

Three key metrics were measured for both backup and restore performance for each test scenario:

1. **Average throughput**—the total amount of data transferred divided by the total amount of time needed for the operation to complete
2. **Sustained peak throughput**—the throughput observed when the backup is running at a sustained rate for a minimum period of 60 minutes
3. **Data transferred in first 60 minutes**—the amount of data transferred in the first 60 minutes after the test is initiated

To understand the rationale for these metrics, consider the performance profile for the file-level 10TB-per-hour backup result illustrated in figure 2.

As figure 2 illustrates, it takes several minutes for the backup to ramp up to full throughput while tapes are being loaded and the various parallel operations are initiated. Likewise, an additional ramp-down period exists at the end of the process as individual jobs complete at

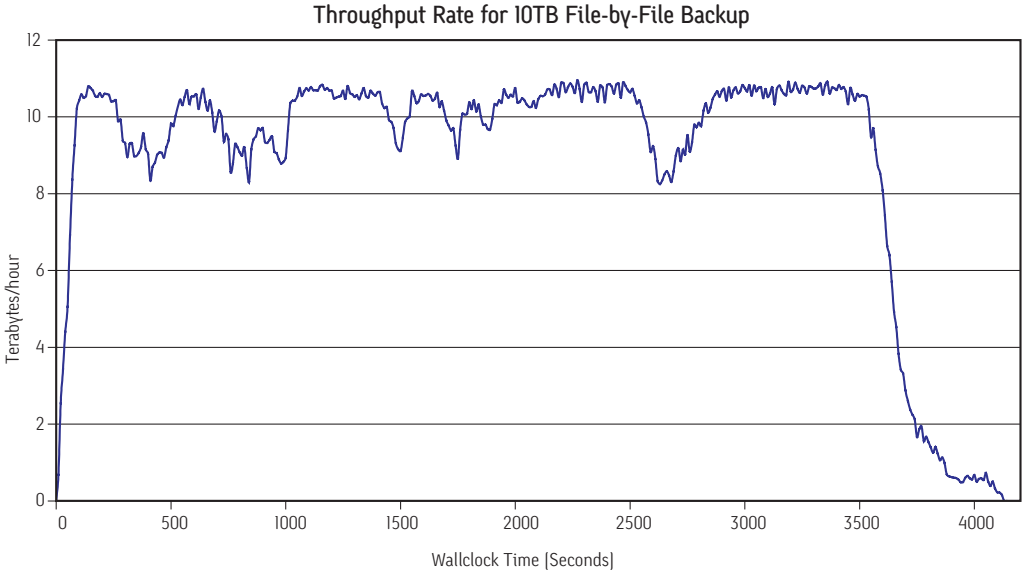

Fig. 2. Performance profile for file-by-file 10TB-per-hour backup test

different times. As a result, the average throughput may significantly less than the sustained peak throughput. The impact of these ramp-up and ramp-down periods on the total backup time obviously depends on the amount of data to be backed up and time period for completion. For data sets larger than the 10TB data set used in these tests, ramp up and ramp down would be correspondingly less significant, while they become more significant for smaller backups. The three test metrics should therefore provide a more complete picture to help assess performance on data sets of various sizes.

2.1 A Real-World Data Set

The data set used for this benchmark was provided by a large HPC customer and contains a variety of application data files and personal productivity files. The customer currently manages hundreds of terabytes of data, which is typical of HPC customers across many industries, including government, energy, sciences, manufacturing, and media and entertainment.

Files in the data set ranged in size from 2GB to over 42GB. Multiple filesystems were used in most of the tests. The results of all tests are summarized in table 1, and a description of each test scenario and results are given in the following subsections.

2.2 File-by-File Backup and Restore (10TB)

In this test, a 10TB data set was divided between 48 filesystems. NetWorker, running on the SGI® Origin® server, was used to initiate a file-by-file backup of each filesystem. This test achieved sustained peak throughput in excess of 10TB per hour for backup and also exceeded the overall goal of backing up more than 10TB in the first 60 minutes, including the ramp-up time needed to initialize all tape drives. These numbers are more than three times the level of performance achieved for any previously reported backup benchmark of

Table 1. Summary of performance results for all tests

	NetWorker File Backup 10TB	NetWorker File Backup 1TB	NetWorker File Backup 10TB Single Filesystem	NetWorker Image Backup w/Snapshot	xfsdump
# of Filesystems	48	48	1	48	48
Sustained Peak Throughput (Backup)	10.09TB/hr	NA	6.26TB/hr	7.24TB/hr	8.95TB/hr
Sustained Peak Throughput (Restore)	4.52TB/hr	NA	4.43TB/hr	7.90TB/hr	3.96TB/hr
Average Throughput (Backup)	9.00TB/hr	8.86TB/hr	6.10TB/hr	6.14TB/hr	7.83TB/hr
Average Throughput (Restore)	3.81TB/hr	4.07TB/hr	3.99TB/hr	6.67TB/hr	3.79TB/hr
Data Transferred in First 60 Minutes (Backup)	10.05TB	NA	6.20TB	7.16TB	8.92TB
Data Transferred in First 60 Minutes (Restore)	4.50TB	NA	4.36TB	7.88TB	3.92TB

any kind. As is often the case for this type of backup, restore performance was less than backup performance with a sustained rate of 4.52TB per second. This number is still well in excess of the industry's previous best for restore performance.

2.3 File-by-File Backup and Restore (1TB)

This test is identical to the previous test, except that the data set size is reduced to 1TB. This test was intended to demonstrate how quickly a smaller data set could be backed up to tape, including initialization time. The entire 1TB backup was accomplished in just over seven minutes and restore was accomplished in less than 16 minutes. These numbers are impressive, considering that just a few years ago the best reported 1TB backup took over an hour. Only average throughput was measured due to the short duration of the tests.

2.4 Backup and Restore of a Single Large Filesystem

Many customers find that managing one or a few large filesystems is preferable to managing a large number of smaller ones, although the number of filesystems needed may also be dictated by organizational or other constraints. Because of the high performance and scalability of the SGI XFS filesystem, customers often find they can achieve the necessary performance and simplify operations with a single large filesystem. Both XFS and the SGI CXFS shared filesystem scale to address up to 18 exabytes as a single filesystem and support a single file size of nine exabytes. For this test, the entire 10TB data set was stored in a single XFS filesystem and backed up using LEGATO NetWorker. The filesystem was divided into 48 directories, where each of the directories included 16 files for an aggregate size of 214.5GB, and was assigned to an individual tape drive for backup. The single filesystem contained 10.3TB of data with a capacity of 12.3TB.

Even when protecting a large single filesystem, the sustained peak throughput for backup performance was an impressive 6.26TB per hour. While this is less than the performance achieved with multiple, separate filesystems, it's still more than double the best previously reported backup speed. Restore performance is equivalent to that seen with the other file-by-file tests included in this benchmark that were performed using multiple filesystems.

2.5 Image Backup with Snapshots

For this scenario, the 10TB data set was again divided between 48 filesystems. SGI® XVM Snapshot (see Section 4) was used to create a point-in-time snapshot of the volume containing each filesystem. NetWorker was used to perform an image backup of the raw data from each snapshot to tape. The time necessary to create the snapshots was not reflected in the image backup benchmark, though snapshots typically take less than one second to generate. This approach generally delivers faster restore times in comparison with file-level backup methods, making it of particular interest to customers that may require faster restores of a large filesystem or data set as part of a disaster recovery or data replication scheme.

The sustained peak throughput for this test was 7.24TB per hour for backup and 7.90TB per hour for restore. The increase in restore performance in comparison to other tests illustrates the fact that restores of single, large images are faster because they eliminate the overhead necessary with traditional file-level backup.

2.6 Backup Using xfsdump (10TB)

The SGI XFS filesystem ships with a simple backup utility called xfsdump. The companion utility, xfsrestore, provides restore capabilities. While xfsdump lacks the advanced file cataloging features of a full-featured backup application such as LEGATO NetWorker, it does have the ability to recognize and back up extended attributes native to both XFS and CXFS filesystems. Many backup applications do not look for extended attributes, and therefore do not recognize or copy these attributes in the backup operation. Both xfsdump and xfsrestore are useful in conjunction with applications that utilize extended file attrib-

utes, such as the SGI® Data Migration Facility (DMF) data management solution, a multitiered data life-cycle management technology that administers migration and provisioning policies.

Because these utilities are in widespread use in DMF environments and elsewhere, comparison tests were executed to determine the performance of both xfsdump and xfsrestore. These tests compare favorably with file-by-file backups using LEGATO NetWorker under the same conditions. Sustained peak throughput for both backup and restore are somewhat lower than with NetWorker (8.95TB per hour for backup and 3.96TB per hour for restore), but this level of performance is still well beyond that reported in any previous benchmarks.

3.0 A Hardware Architecture Optimized for Parallel I/O

The hardware solution that was used to achieve the backup and restore results described in this paper is illustrated in figure 3. A storage area network (SAN) fabric composed of Brocade 2Gb Fibre Channel switches was used as an interconnect between the SGI Origin 3000 server, the SGI TP9500 storage array, and the 48 T9940B tape drives in the StorageTek PowderHorn 9310 tape library. All components were chosen for their proven ability to support a large number of I/O streams in parallel and configured to ensure that no bottlenecks occurred under test conditions.

SGI Origin 3000 server: An Origin 3000 server with 32 processors, 8GB of system memory, and 48 2Gb Fibre Channel host bus adapters (HBAs) acted as the backup server for all tests. The NUMAflex™ architecture of the SGI Origin 3000 servers is well known for its tremendous I/O capabilities. A system can easily be tailored to meet almost any I/O requirement. Origin systems capable of achieving sustained I/O bandwidth in excess of 7GB per second (24.6TB per hour) to and from a single file-system have been demonstrated, ensuring that the Origin server has the I/O bandwidth to exceed the requirements for these tests. During testing, this configuration consistently

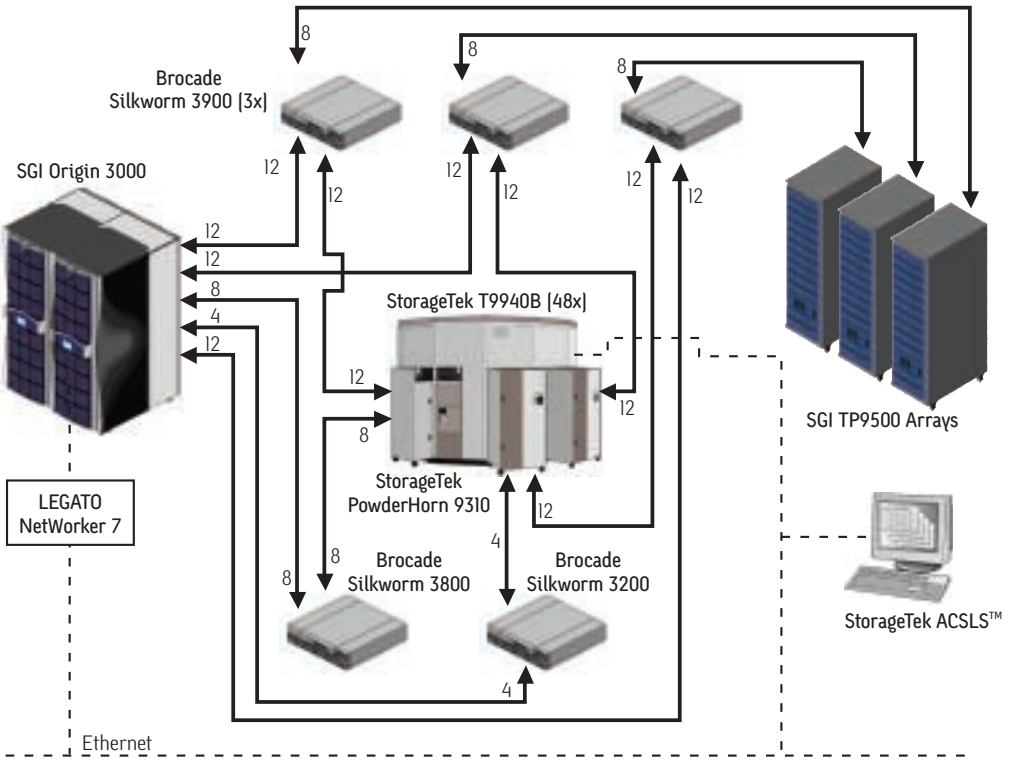

Fig. 3. Solution architecture (numbers indicate the number of individual channels between each pair of targets)

read data off the disk subsystem at a rate of over 10.4TB per hour into memory, then piped it out at over 10.4TB per hour through the fabric to the tape drives. The concurrent read and write operations performed in duplex demonstrates an aggregate sustained throughput rate of over 20.8TB per hour (5.8GB per second) through the Origin server's I/O subsystem.

SGI® Total Performance 9500 (TP9500) RAID storage array: The high-performance disk storage subsystem used in this benchmark was developed by LSI Logic Storage Systems, Inc. The storage system is configured according to SGI's specifications and sold by SGI and its partners as the SGI TP9500 storage array. SGI TP9500 is a full end-to-end 2Gb Fibre Channel solution, which uses parallel high-performance controllers and advanced cache management to achieve the throughput needed for these tests. The configuration included 120 Fibre Channel disk drives with a raw capacity of 17.5TB, capable of sustaining a read rate of over 10.4TB per hour through the storage area network. A total of six dual-controller RAID enclosures funneled data through 24 Fibre Channel connections to the Fibre Channel mesh fabric. In addition to tremendous I/O throughput, SGI TP9500 also offers advanced features for high availability and data protection.

StorageTek T9940B tape drives: All backup and restore operations were performed using StorageTek T9940B tape drives with 2Gb Fibre Channel interface installed in a StorageTek PowderHorn 9310 tape library. A single silo can be configured with up to 1200TB of native capacity and 960 individual tape drives with 100TB-per-hour native throughput. For these tests, a single library was configured with 48 of the T9940B tape drives, where each drive used a single 200GB cartridge for each backup and restore test. At the time of its release in September 2002, the T9940B Fibre Channel tape drive offered the fastest data transfer rate (30MB per second native and 70MB per second with compression) and the highest-capacity cartridges (200GB native) of any tape drive on the market.

Brocade® SilkWorm® Fibre Channel fabric switches: A SAN fabric composed of three SilkWorm® 3900 switches, one SilkWorm® 3800 switch, and one SilkWorm® 3200 switch was used to interconnect the Origin server with the disk and tape storage systems for optimal performance. All three switch models offer full 2Gb performance in a nonblocking architecture. Hardware-enforced zoning was used to ensure that each tape drive achieved optimal performance.

4.0 Advanced Software for Optimal Backup and Restore

The software used in this testing was chosen for its ability to support a high degree of parallelism. Two main software products were used.

LEGATO NetWorker: Protecting more than 30,000 enterprise sites worldwide, NetWorker is the premier solution for information protection in multi-terabyte, heterogeneous environments. NetWorker supports a broad range of platforms, filesystems, tape libraries, and disk-backup devices, and it can be configured to carry out a large number of backup and/or restore operations in parallel, to ensure that any set of operations can be carried out in the shortest possible time. A single NetWorker server is highly scalable and can automate backup, restore, and archival activities of hundreds to thousands of network clients in DAS, NAS, and SAN environments. Data can be sent to the server itself or to independent storage nodes to optimize network bandwidth. storage nodes also enable protection of large systems to locally or SAN-attached tape or disk storage devices.

The Origin 3000 server was configured as the NetWorker server and directed to perform backups of local filesystems. The image backup and restore tests were done using the commercially available NetWorker version 7.0 Power Edition. For the traditional backup and restore tests, NetWorker 7.0 with a direct I/O feature was used. This feature is available via patch for version 7.0 and will be native to version 7.1, scheduled for release in fall 2003.

SGI XVM Snapshot: XVM Snapshot is an optional extension for the SGI XVM volume manager that provides snapshot capabilities at the volume level. Volume-level snapshots are more convenient and efficient than snapshot products that work at the LUN level, since a single volume can span many LUNs. XVM Snapshot can take a snapshot of an entire XFS or CXFS filesystem, minimizing the number of snapshots needed, especially for environments that use a single large filesystem. Among other purposes, snapshots are frequently used to provide a consistent image for backup of an active filesystem. When a snapshot is in place, the original versions of changed blocks are automatically copied to a snapshot partition (copy on write). When the snapshot is accessed for backup, restore, or other purposes, the original blocks are read in lieu of the changed blocks, preserving a consistent image of the volume from a single point in time.

For this testing, XVM Snapshot was used in combination with LEGATO NetWorker for image backups. XVM Snapshot can also be used in combination with LEGATO NetWorker or xfsdump during file-by-file backups. NetWorker has built-in capabilities to recognize when individual files are changing to help ensure consistent backups of live filesystems in situations where snapshots are not used.

5.0 Beyond Benchmarks

All hardware and software used in this benchmark is readily available. Therefore, unlike many laboratory benchmarks, customers will be able to duplicate this configuration with the same or similar components and can expect to achieve similar levels of performance.

The key to achieving the levels of performance demonstrated by these benchmarks is providing the parallelism to achieve the desired level of throughput. Because of its unique NUMAflex architecture, a single SGI Origin server can scale to these levels and beyond. The other components chosen for these benchmarks offer similar levels of parallelism for optimum scalability and data consolidation, creating a backup environment that is not only highly scalable, but also simple—especially given the level of performance achieved—and easy to manage.

The Origin backup server included both XFS and CXFS filesystems, but all tests were run using XFS. The SGI CXFS shared filesystem has demonstrated performance levels similar to the XFS filesystem when implemented in a SAN. While it was not explicitly tested during this benchmark, SGI believes that results similar to the actual achievements in this benchmark would be achieved by using CXFS instead of XFS. Other distributed filesystems don't provide the performance expected from a native filesystem. CXFS provides data sharing capabilities across a SAN with the performance of a native filesystem. CXFS creates the opportunity to design a heterogeneous shared data environment in which all computer systems have high-speed access to shared data (multiple gigabytes per second), plus the backup and restore performance required for extremely large data sets. An Origin server in such a configuration acts as a backup server, accessing all data at native-filesystem speed via CXFS. The server can also be used for other purposes when high-speed backup and restore are not in progress.

Achieving optimal backup and restore performance tailored to real requirements in the real world is not a trivial task. Most sites already have a substantial investment in servers and disk and tape storage systems that must be preserved. SGI Professional Services can provide the storage expertise to help customers choose the appropriate hardware and software components for their needs and to integrate those components in their existing environment to maximize performance.

Corporate Office
1600 Amphitheatre Pkwy.
Mountain View, CA 94043
(650) 960-1980
www.sgi.com

North America 1(800) 800-7441
Latin America (52) 5267-1387
Europe (44) 118.925.75.00
Japan (81) 3.5488.1811
Asia Pacific (65) 6771.0290

3523 07/23/2003]

J14309