

Cisco MDS 9000 Series Multilayer Switches

Sprint Storage Transport Case Study: Joint Technology Research and Development and Cisco Partnership

Sprint Storage Transport Services Case Study

Executive Summary

Data distribution, data protection, business continuance, and disaster-recovery strategies are critical components of today's information-centric businesses. The ability to efficiently replicate important data on a global scale not only helps ensure a higher level of protection for valuable corporate information, but also promotes increased usage of backup resources, reduces the impact of a catastrophic failure at a single site, and lowers the total cost of storage ownership.

In this case study, Sprint's Computing Systems Technology Evaluation (CSTE) lab, Cisco Systems®, and Hitachi Data Systems have partnered to demonstrate the feasibility of using Fibre Channel over Internet Protocol (FCIP) to accurately and efficiently replicate large amounts of data across transcontinental distances, and by implication, transoceanic distances. (See Appendix A for more details of FCIP.)

This study will show how successful replication of a large Oracle database over a distance of 3600 miles (5760 kilometers) was achieved using Cisco® MDS 9509 multilayer director switches with the IP Services module, Hitachi's TrueCopy replication technology, and Sprint's nationwide SONET WAN.

Using the IP network in conjunction with FCIP employs a far less expensive method of reliably and accurately transferring large amounts of data over significant distances. This is of utmost importance because all businesses, not just the very largest enterprises, need to protect their data in the face of natural disasters or acts of terrorism. A business-continuance and disaster-recovery plan has now become a part of a businesses operating budget regardless of the business's size.

Until more recently, business-continuance and disaster-recovery plans that transfer critical data to diverse locations have been feasible only for the largest enterprises. The costly WAN connections and equipment-transferring storage area network (SAN) traffic across significant distances have traditionally been accomplished with technology that puts the SAN traffic directly on a SONET system, directly onto dense wavelength-division multiplexing (DWDM) systems, or even onto dark fibers. FCIP technology avoids all these expensive alternatives and makes use of the low-cost and ubiquitous IP network to transfer



SAN data. This one technology brings true business continuance and disaster recovery within reach of the small-business and midsized-business (SMB) customer. It also provides a much lower-cost alternative for the large enterprise.

Typical Customer Scenario

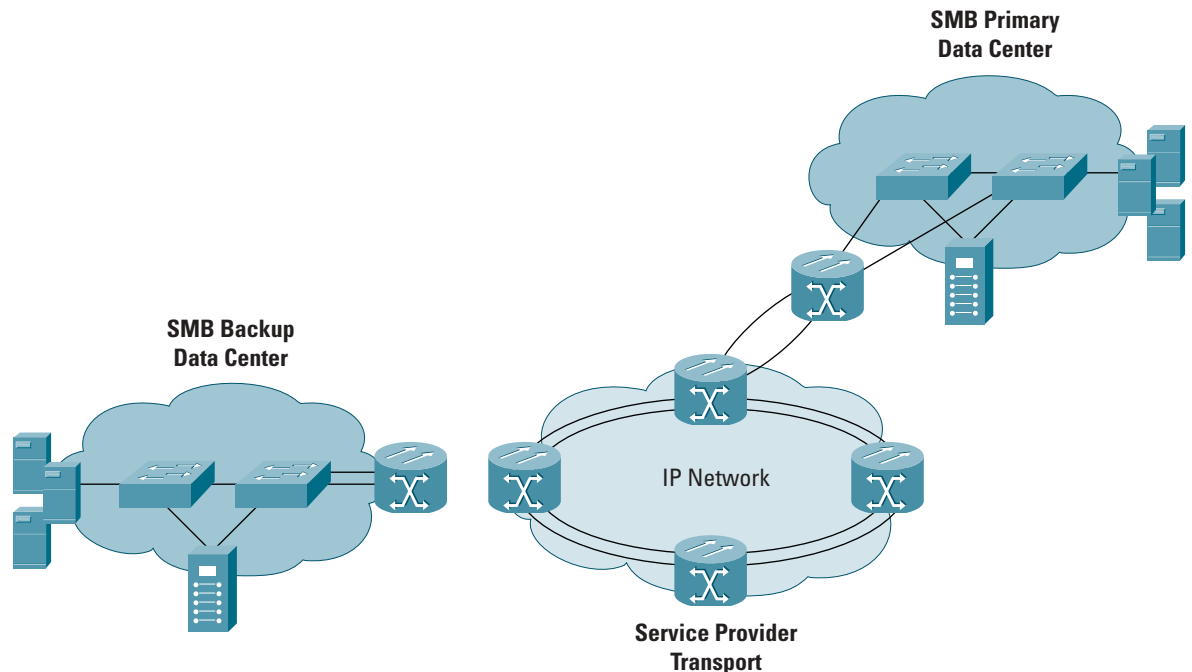
SMBs represent approximately 95 percent of all companies in the United States today. These organizations usually have fewer than a hundred employees and are highly focused on producing a specific good or service.

A typical SMB may have one or a few locations in a metropolitan area, with one or more locations in other metropolitan areas. One location is usually the “home office” where the majority of the business activity occurs, and other locations can be considered “satellite” offices where business activity is funneled to the home-office location.

An SMB typically has an IP link connecting each of its satellite locations to the home-office location. This IP connectivity supports the SMB’s data network and possibly the telephone system. An SMB usually does not have similar connectivity for its storage, due to cost constraints. It stores most of its data at the home-office location, which may be consolidated onto a SAN, and small SAN islands at the satellite locations.

Storage needs for an SMB then fall into two categories: local SAN requirements and WAN connectivity. Figure 1 depicts a typical SMB scenario.

Figure 1
SMB Infrastructure Environment





Challenges

Disparate SAN Islands

As companies implemented SANs within their organizations, disparate SAN islands began to surface within the data-center environment. However, because there was no cost-effective way to interconnect these SAN islands, resources such as tape libraries and disk arrays could not be shared across the entire SAN infrastructure.

SAN Extension Beyond Metropolitan Distances (Greater than 60 Kilometers)

Currently, most of the typical SAN-extension infrastructures are remote point-to-point disk-arrays replications. In most cases, these configurations are confined within a metropolitan area that requires expensive additional fibre channel extension such as DWDM.

Keeping SAN Extension Cost-Effective

The cost of SAN extension has been very high, upward of hundreds of thousands of U.S. dollars per connection when considering the cost of the gateway hardware and the transport service. Many organizations have had few options to use their IP network for fibre channel traffic, and had to implement additional fibre channel infrastructures.

Complexity in Management

Most existing SAN-extension infrastructures are composed of disparate components (fibre channel switches, other switches, SAN extension box, and routers) and central management is either awkward or nonexistent. This complexity in management can potentially add to the deployment's cost and limit the scalability of the solution.

Solution

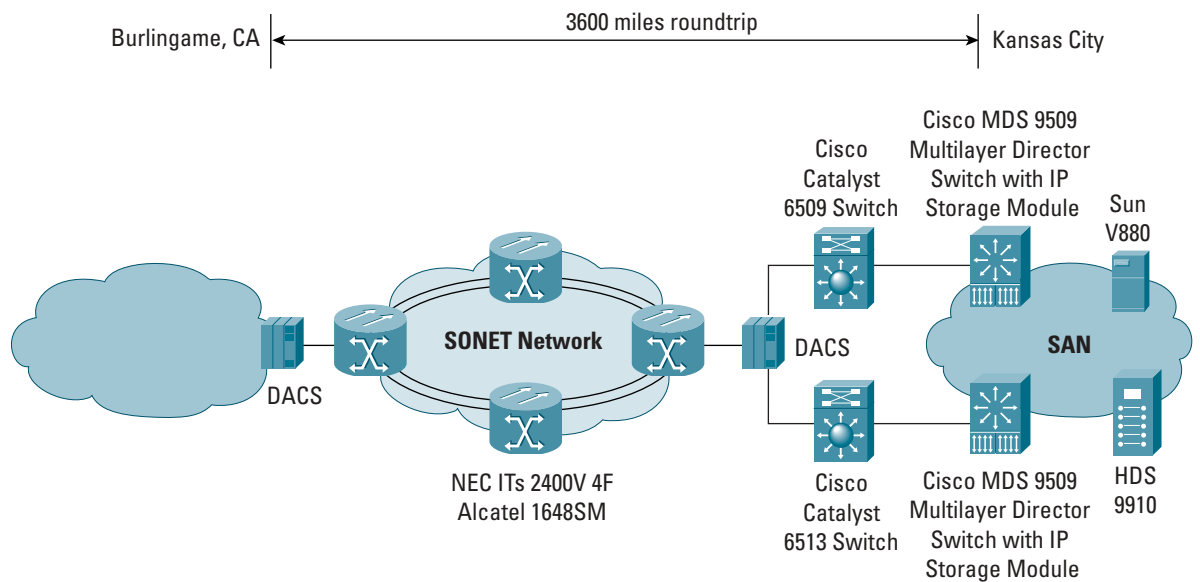
To cost-effectively provide SAN-extension services over transcontinental distances, Sprint and Cisco have created a solution that uses Sprint's existing IP and SONET infrastructure along with the FCIP capabilities of Cisco MDS 9000 Series multilayer switches to transfer fibre channel data using Hitachi's TrueCopy replication technology.

Storage and network resources across the various SAN islands enable better resource usage of tape libraries and disk arrays. With the IP-over-SONET infrastructure as the primary transport, the SAN can extend far beyond the typical metropolitan area to reach transcontinental distances. The Cisco MDS 9000 Series IP Storage Services module services 24 separate FCIP tunnels simultaneously, helping achieve greater scalability and lower cost per FCIP tunnel. Because the module is integrated into the Cisco MDS 9000 Series, management is simplified and centralized.

Fibre channel frames from the end devices of the fibre channel fabric are encapsulated within TCP/IP packets and transported over the SONET infrastructure to the remote switch (Figure 2).



Figure 2
Transcontinental FCIP Replication



Storage WAN Environment

The SAN test equipment used for the test described in the section consisted of two Cisco MDS 9509 multilayer director switches with IP services modules, a Hitachi Lightning 9900 Series, Model 9910 storage array, and a Sunfire V880 host. This equipment was a subset of the equipment in the SAN, which connected various Windows XP and Windows 2000 hosts, multiple UNIX hosts (Linux, AIX, Solaris), as well as several different storage arrays from Hitachi Data Systems, EMC, and IBM. The SAN was connected to the Cisco MDS 9509 switches, which had an IP Services blade installed in each switch.

Hitachi's 9910 storage array was connected via a 2-Gbps fibre channel interface on one of its ports to one of the Cisco MDS 9509 switches. From that switch there was a 1-Gigabit Ethernet IP Services module connection to a Cisco Catalyst[®] 6509 Switch, which then connected via OC-3 packet over SONET (POS) to a Tellabs DACS. The DACS put the OC-3 POS inside an OC-12 as an STS3c and connected to an NEC SONET node at the Sprint campus in Kansas City. That NEC SONET ring then connected to two more Alcatel SONET rings between Kansas City and Burlingame, California, where the STS3c signal was received by a DACS. The DACS in Burlingame, CA then cross-connected the STS3c signal to another STS3c inside of the same OC-12 and sent it back to Kansas City. The Kansas City DACS broke it out into an OC-3 and sent it to a Cisco Catalyst 6513 Switch, which then sent it via Gigabit Ethernet to the second Cisco MDS 9509, which in turn sent it to a separate port on the Hitachi Model 9910 via another 2-Gbps fibre channel interface.

An Oracle database of about 128 GB was constructed using the Sunfire V880 for processing power and the Hitachi array for storage. The Hitachi array was running asynchronous Hitachi TrueCopy.

The remainder of this section lists all of the equipment in the test environment. The equipment falls into three main categories: equipment that comprises the IP network, including the WAN; SAN networking equipment, hosts, and storage array; and applications used in the test environment.



IP/SONET Infrastructure

- NEC ITS-2400 Series 4F [NOTE: please use official NEC product name]and an Alcatel 1648 SONET Multiplexer (SM)
- Cisco Catalyst 6500 Series switches
 - Cisco Catalyst 6509 Switch
 - Cisco Catalyst OS Software Version 6.2.2
 - Cisco IOS® Software Release 12.1(11b)E11
 - Cisco Catalyst 6500 Series Supervisor Engine 2
 - Cisco Catalyst 6000 Multilayer Switch Feature Card MSFC2
 - WS-X6516 16-port Gigabit Ethernet module
 - FlexWAN module
 - PA-POS-OC3
 - Cisco Catalyst 6513 Switch
 - Cisco Catalyst OS Software Version 6.3.5
 - Cisco IOS Software Release 12.1(11b)E11
 - Cisco Catalyst 6500 Series Supervisor Engine 2
 - Cisco Catalyst 6000 Multilayer Switch Feature Card MSFC2
 - WS-X6516 16-port Gigabit Ethernet module
 - FlexWAN module
 - A-POS-OC3

DACS

- Tellabs Titan 5500

SAN Infrastructure

- Cisco MDS 9509 Multilayer Director Switch (code 1.1.0-180)
- Cisco MDS 9000 16-port 2/1-Gbps FC Module, SFP/LC (DS-X9016)
- Cisco MDS 9000 32-port 2/1-Gbps FC Module, SFP/LC (DS-X9032)
- Cisco MDS 9000 IP Storage Services Module (DS-X9308-SMIP)

Hitachi Storage Infrastructure

- Hardware
 - 9910, number 60270: microcode level- 01-18-48-00/00
 - CHPid – 8 ports 2gig FC (tachyon 2GB- SE-Fiber T 4 ch each)
 - ACP- 1 pair (DKRD2D – J072FC)
 - 1024 MB shared memory
 - 4-GB cache
 - 72-GB disk type
 - 6 array groups, OPEN E's, RAID 5 15 Ldev each



- 16 logical control units
- Dispersed mode
- Software

Hitachi TrueCopy

- Resource manager 9000 V2 RM2
- Hi Command, Device Manager Version .2.1
- ShadowImage (SI)
- SAnTinel (HSAN)

Server and Database Infrastructure

- SUN V880 with Solaris 8, release 2/02
- CPU: 4x 900 Mhz UltraSPARC III+
- Memory: 8 GB
- Fibre channel host base adapter : QLogic 2340
- Oracle version: 8.1.0
- Data Buffer Size: 4096000 bytes
- Total Database Size : 152,168 MB

Tests Performed

- Replication (Hitachi TrueCopy)
The Hitachi TrueCopy replication times for the complete replication of the 128-GB Oracle database
- Database exercising (Oracle)
Database testing included row selections, deletes, inserts, and updates

FCIP Parameters

- Path Maximum Transmission Unit (PMTU) was set to 3000 to ensure that 2148-byte fibre channel frames were not segmented unnecessarily.
- Maximum Window Size that is used to compensate for the latency within the IP network was adjusted on the FCIP profile of the tunnel to reflect the maximum bandwidth of the link (155 Mbps) and a roundtrip delay time of 75 milliseconds (ms).

Performance Observed

- Theoretical maximums—This is the maximum throughput that could be achieved theoretically according to written calculations. For 1-Gigabit Ethernet connections, the maximum throughput in a single direction is calculated to be 105 MBps. For OC-3 connections, the maximum throughput in a single direction is calculated to be 15.5 Mbps.



- Replication (Hitachi TrueCopy)—With delays of 75 ms, TCP Maximum Window Size of 1334 Kb, zero percent drop/retransmit transfer began at 16:05:27 for a total of 2 hours and 44 minutes. The array has 14.1-GB disks physically installed 9 concatenated RAID 5 ldevs the storage array software reported a total of 128 GB of data transferred.

$$\frac{128 \text{ GB} * 1024 \text{ MB/GB}}{2 \text{ hr} * 3600\text{sec/hr} + 44\text{min} * 60\text{sec/min}} = 13.3 \text{ MBps}$$

- Database exercising—read and write times for various sizes of data across increasing distances. See Table 1 for results.

Note: Table 1 shows results for additional distances to give context to the long-reach (3600-mile) distance results.

Table 1 Oracle Test Results Tabl

Test Case and Storage Configuration	1 Switch	25-Km Spool	50-Km Spool	3600-Mile OC3 Link
Select single row, query on indexed column	7	8.25	8.5	8.5
Select multiple rows, query on nonindexed column	2777.75	2820.5	2876.25	2763.5
Select multiple rows, query on indexed column	29.5	28	30.25	33
Select long data (text, binary/blob, clob)	6941.5	7044.75	7294	7039.25
Select distinct	17414	17418.5	17949.75	17362.25
Select group by	23823.75	24187.25	24196.25	24293
Select count	6700.75	6851.75	6875.75	6905
Select order by	44.25	44.25	45.25	48
Select maximum on key	6.25	6.25	6.5	6.75
Delete with key	44	24	27	177
Insert with data select from other table	34	16	14	18
Update multiple records	5425	6152	7851	4565
Update one record with key in where clause	31	15	14	13

Note: Execution time is in milliseconds

Results

With the Hitachi TrueCopy disk replication, average throughput for the 128-GB database was 13.3 MBps. It should be noted that the real throughput observed was near wire rate throughout the majority of the data transfer (96 percent), with a drop coming in the last 4 percent of the data transferred. Subsequent transfers of the database showed similar results. On careful observation, the test team noted that the drop in throughput coincided with the disk activity drop off. The team believes the decline in disk activity was simply due to how the data is physically



stored on the disks, with some disks being depleted before others until there are just a few disks sending data. Another factor may have been that the same storage array was both ends of the connection at once. In other words, the reading and writing was ultimately performed on the same set of physical disks.

Additional testing using separate Hitachi arrays separated by 1800 miles (2880 kilometers) brought results that were wire rate throughout the transfer and did not show any signs of slowing toward the end of the transfer. It should be noted that in these additional tests the physical arrangement of the data on the arrays was the same as in the 3600-mile test. One item of interest that was not intuitively obvious was the fact that initial tests were under-running the bandwidth capacity of the OC-3. Some tuning of the storage array was necessary to achieve optimal performance across the WAN. A change was made on the storage array to increase the number of outstanding I/Os from 4 (the default setting) to 16. This allowed the storage array to send more unacknowledged data before requiring an acknowledgement that data had been received. In other words, it allowed the pipe to always be filled up rather than sending data and waiting for the acknowledgement.

The Oracle testing showed that the database response time of the local disk array and that of the remote disk array across the FCIP link is very comparable, suggesting that running the database remotely is a feasible task.

What Sprint Offers

In addition to server, network, storage, and SAN technical expertise, as well as application-management expertise, Sprint also provides the storage network management expertise to operate and maintain a highly available, scalable, and dynamic data center infrastructure for the end customer.

What Cisco Offers

The Cisco MDS 9000 Series IP Storage Services Module uses the open-standard FCIP protocol to break the distance barrier of current fibre channel solutions and enable interconnection of SAN islands over extended distances. With the scalability to reach up to 24 FCIP tunnels with a single IP services module, interconnection between large numbers of SAN islands becomes feasible at a competitive cost. Equally compelling is the consolidation of the FCIP services onto a single platform in which one management interface for the various intelligent services can be centrally administered and provisioned.

What Hitachi Data Systems Offers

Hitachi Data Systems offers robust, proven disk-replication software—Hitachi TrueCopy—to take advantage of the long-haul SAN extension capability of Sprint's IP over SONET and Cisco MDS 9000 Series intelligent SAN infrastructure.

Appendix A FCIP Protocol

Fibre Channel over IP encapsulates fibre channel frames and transports these frames within TCP packets (Figure 3). An FCIP tunnel effectively acts as an Inter-Switch Link (ISL) between two fabric switches and as such, end devices see each other as they would between two local switches interconnected with standard ISLs. Each end of the FCIP link is associated to virtual e-port and these ports communicate between themselves using SW_ILS frames: ELP, ESC, BF, Registration Confirmation (RCF), and Fabric Shortest Path First (FSPF), just as any other e-ports. Unlike the standard Fibre Channel Interface Protocol, Fibre Channel over IP (FCIP) uses the underlying TCP/IP protocol to provide congestion control and in-order delivery.

Compared to other IP-based storage protocols such as Small Computer System Interface over IP (iSCSI) or Internet Fibre Channel Protocol (iFCP), FCIP is considered the most transparent in that fibre channel switches, fibre channel hosts, and targets do not need to be reconfigured or changed in any way to take advantage of this SAN extension protocol.

As in all IP networks, performance can vary based on the types of switches and routers, the number of hops the packets must traverse, and the level of congestion in the network. Today, storage transport performance over IP networks—especially over public networks—is limited by the variable latency of service provider networks. As IP and Ethernet equipment continues to evolve, higher levels of quality of service (QoS), cost of service, provisioning, and circuit emulation should provide the latency guarantees required by synchronous storage applications. Regardless, FCIP is currently a cost-effective technology for asynchronous applications such as remote data backup and disk-to-disk replication.

For Storage Networking Industry Association (SNIA) definitions of FCIP, visit: http://www.snia.org/tech_activities/ip_storage/FCIP_whitepaper.pdf.

Figure 3

Conceptual Diagram



Corporate Headquarters
 Cisco Systems, Inc.
 170 West Tasman Drive
 San Jose, CA 95134-1706
 USA
www.cisco.com
 Tel: 408 526-4000
 800 553-NETS (6387)
 Fax: 408 526-4100

European Headquarters
 Cisco Systems International BV
 Haarlerbergpark
 Haarlerbergweg 13-19
 1101 CH Amsterdam
 The Netherlands
www-europe.cisco.com
 Tel: 31 0 20 357 1000
 Fax: 31 0 20 357 1100

Americas Headquarters
 Cisco Systems, Inc.
 170 West Tasman Drive
 San Jose, CA 95134-1706
 USA
www.cisco.com
 Tel: 408 526-7660
 Fax: 408 527-0883

Asia Pacific Headquarters
 Cisco Systems, Inc.
 Capital Tower
 168 Robinson Road
 #22-01 to #29-01
 Singapore 068912
www.cisco.com
 Tel: +65 6317 7777
 Fax: +65 6317 7799

Cisco Systems has more than 200 offices in the following countries and regions. Addresses, phone numbers, and fax numbers are listed on the **Cisco Web site at www.cisco.com/go/offices**

Argentina • Australia • Austria • Belgium • Brazil • Bulgaria • Canada • Chile • China PRC • Colombia • Costa Rica • Croatia
 Czech Republic • Denmark • Dubai, UAE • Finland • France • Germany • Greece • Hong Kong SAR • Hungary • India • Indonesia • Ireland
 Israel • Italy • Japan • Korea • Luxembourg • Malaysia • Mexico • The Netherlands • New Zealand • Norway • Peru • Philippines • Poland
 Portugal • Puerto Rico • Romania • Russia • Saudi Arabia • Scotland • Singapore • Slovakia • Slovenia • South Africa • Spain • Sweden
 Switzerland • Taiwan • Thailand • Turkey • Ukraine • United Kingdom • United States • Venezuela • Vietnam • Zimbabwe

All contents are Copyright © 1992–2003 Cisco Systems, Inc. All rights reserved. Cisco, Cisco Systems, the Cisco Systems logo, Catalyst, and Cisco IOS are registered trademarks of Cisco Systems, Inc. and/or its affiliates in the U.S. and certain other countries.

All other trademarks mentioned in this document or Web site are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company.
 (0304R) ETMG 203154—LSK 10/03