WHITE PAPER

SCALABLE FABRIC DESIGN—OVERSUBSCRIPTION AND DENSITY BEST PRACTICES

This paper investigates the justification of oversubscription within the server access layer of a Fibre Channel fabric. It publishes storage benchmark numbers as well as real-world application I/O performance numbers from the Cisco Systems® manufacturing module’s enterprise resource planning (ERP) system. The laboratory storage benchmark test results are based on standard TPC-C tests. As for the real-world application I/O performance numbers, Cisco’s financial accounting practice is completely automated and is a sought-after model by many Fortune 100 companies. The data collected from Cisco’s ERP system includes the I/O performance numbers during the year-end processing and closing of Cisco’s books for the company’s fiscal year 2002. All results validate the practice of oversubscribing server fabric ports to accommodate the fan-out ratio recommendations of storage subsystem vendors and to provide the customer with an optimized hardware infrastructure with balanced price-performance metrics.

INTRODUCTION

One of the most critical needs in today’s enterprise data center is the ability to optimize the use of existing computing infrastructures. This optimization leads to better price-performance metrics and may help to offset possible costly infrastructure expansion projects.

A major infrastructure component that represents a significant opportunity for cost savings is the storage infrastructure, which plays an important role in determining the optimal usage of the computing infrastructure. The storage infrastructure includes storage subsystems, the storage area network (SAN) infrastructure, and host bus adapters (HBAs). It is a common practice for storage subsystem vendors to share one port on the storage subsystem among multiple HBAs on multiple servers. After all, the initial purpose for a SAN was to increase connectivity to storage subsystems that were growing in storage capacity at a faster rate than their external connectivity capabilities. This SAN fan-out ratio of storage ports typically ranges from 6:1 to 12:1 server-to-storage subsystem ports. This ratio balances different server platforms and applications across these subsystem ports to fully utilize available bandwidth while enabling the maximum throughput of each HBA to achieve near-line-rate throughput at a given time. The ratio also implies that the Fibre Channel switch ports that the server HBAs are connected to are being underused most of the time. To achieve the best price-performance metrics, it is common practice to oversubscribe server HBA-connected switch ports, as long as storage subsystem ports have access to full line-rate performance and the Fibre Channel switching fabric is non-blocking.

Today’s enterprise IT organizations require automation of most of the day-to-day activities of the enterprise. This has prompted application environment vendors—namely the hardware, software, storage, and network vendors—to create faster and more scalable products. With these new products, systems architects must balance the performance of the application environment with what is adequate to run the business at the optimum price-performance metric.

Making the right selection of components for the application infrastructure requires careful estimation of performance and scalability requirements and the evaluation of potential hardware, software, and network components that will satisfy the needs of the enterprise. During this elaborate capacity planning exercise, the systems architect estimates the individual and aggregate feeds and speeds of the application servers and storage devices that make up the end-to-end application infrastructure. Some of these components (the storage devices) are estimated to function at near-rated speeds. However, most application servers are oversubscribed in terms of fabric bandwidth requirements, due to the bursty and often underused fabric utilization. Oversubscription helps lower the cost of the overall application infrastructure while helping to ensure that application servers receive the appropriate I/O channel resources to meet their application needs. A computing environment comprises a CPU, memory, and the I/O subsystem resources. The systems architect must evaluate these resources, which heavily
influence fabric bandwidth requirements, and design a system that adequately suits the requirements of the application for which the system is being designed.

This paper highlights the design considerations that are important from an I/O subsystem point of view when building an enterprise SAN infrastructure. With the aid of data collected from applications in a large enterprise data center, this paper also illustrates the common practice of under utilizing the I/O channel to maintain application reliability, availability, and performance.

COMPUTING SYSTEM COMPONENTS

A computing system, made up of processing, memory, and I/O subsystem components and resources, may be on a single system or may be part of a distributed architecture.

- **Processing (CPU)**—The processing resource is defined as the number of CPUs in a computing system. The number of CPUs necessary for a business function is chosen based on the amount of processing power needed to perform a given amount of work in a given amount time.

- **Memory**—The amount of memory is chosen based on the memory requirements of applications being executed on the system. Generally speaking, due to the various caches in a typical computing system, the higher the amount of memory, the faster the I/O response time of an application.

- **I/O subsystem**—An I/O subsystem is made up of the storage subsystem, the transport layer (including SAN switches), and HBAs in servers. The transport layer consists of the physical connections between HBAs and storage subsystems, and can consist of direct attached storage (DAS) devices using copper or fiber cables. It can also consist of servers connected through SAN switches to form a SAN. With SANs, server HBAs are connected to SAN switches and are zoned with target storage devices to access required logical unit numbers (LUNs).

The majority of application infrastructure oversubscription occurs at the SAN fabric layer. Fabric oversubscription occurs when HBAs with high-bandwidth capacity are connected to a shared SAN fabric that cannot provide the aggregate amount of bandwidth to support all connected HBAs transmitting simultaneously at wire rate to either the attached ISLs or disk devices. The likelihood of all connected HBAs simultaneously transmitting at near wire rate is low. One reason is that HBAs cannot generate high throughput, due to limitations imposed by server platform architectures and application I/O profiles. The acceptable oversubscription at the server access layer of a fabric is dependent on the fan-out ratio of server-to-storage ports, which is often defined by storage subsystem vendors.

The following section provides definitions of different aspects and contexts of oversubscription and blocking. In many cases, various switch products in the market exhibit a combination of these items.

**Port Oversubscription**

Fabric switch port oversubscription occurs when the amount of internal switching fabric bandwidth allocated to a given switch port is less than the device connection speed at that port. For example, if a port on a Fibre Channel switch has a connection speed of 2 Gbps but is unable to achieve wire-rate 2 Gbps of performance, then the port is said to be oversubscribed. In many cases, extra Fibre Channel IDLE ordered sets are inserted to pace Fibre Channel frame transmission to support the sub-wire-rate performance. All ports on any Cisco MDS 9000 Series switch are able to sustain wire-rate performance for any frame size, and therefore do not suffer port oversubscription.

**Switch Oversubscription**

Fabric switch oversubscription occurs when the overall switching bandwidth of the switch is less than the aggregate bandwidth of all ingress switch ports. This means that a subset of the total number of ports can run at full wire-rate simultaneously, but not all ports can. The important switch characteristic is to be able to balance the amount of bandwidth fairly among all the switch ports. Many Fibre Channel switches in the market today are oversubscribed at the switch level either intentionally or unintentionally, due to architectural inabilities. The Cisco MDS 9000 Series uses oversubscription on the DS-X9032 32-port switching module and on the Cisco MDS 9100 Series fabric switches to increase port density and reduce complexity and overall solution costs.
Network Oversubscription

Network oversubscription occurs at several points of consolidation within a fabric network, including across a single switch and at Inter-Switch Link (ISL) uplink points. Network oversubscription refers to a point of bandwidth consolidation where the ingress bandwidth is greater than the egress bandwidth. For example, at an ISL uplink from an edge layer switch to a core, the oversubscription of the ISL is typically on the order of 7:1 or greater. In a single director fabric, the fan-out ratio of server to storage subsystem ports is directly related to the network oversubscription and is typically on the order of 10:1 or higher. Network oversubscription is normal and unavoidable—it is a direct by product of the primary purpose for deploying a SAN. An important characteristic of the network related to oversubscription is its ability to fairly allocate its bandwidth resources among all clients of the SAN.

Blocking Switch

A switch is said to be blocking when inefficient hardware design causes ingress traffic to be blocked or stalled, due to preceding traffic that is destined to slower or congested receivers. Blocking represents a condition where ingress and egress bandwidth capacity exist, but the switch is unable to forward at the desired rate due to hardware or queuing inefficiencies. Through the use of a technology called Virtual Output Queuing (VoQ), none of the Cisco MDS 9000 Family of switches suffers from this blocking effect.

Fan-Out Ratio

The fabric fan-out ratio typically refers to the number of server HBAs connected either directly to a storage subsystem or to a SAN, relative to the number of storage subsystem connections to the SAN. This ratio can vary from 1:1 for DAS to well over 10:1. Storage vendors typically determine this ratio, based on the server platform and performance requirement.

APPLICATION I/O SUBSYSTEM—DESIGN CONSIDERATIONS

There are several important considerations for planning an application and storage infrastructure rollout. A balance must exist between computing resources, I/O channel capacity, storage capacity, and storage subsystem performance, all of which directly affect the overall cost of the solution. It is difficult to estimate such resource requirements, especially when considering peak and sustained usage patterns. Following are several considerations for each type of computing infrastructure resource.

• **Processing (CPU) resources**—The principal criteria for planning CPU resources is to estimate the amount of work required from the application being executed on the computing infrastructure. Multiprocessor systems can distribute application load or functions among multiple processors to help scale the application.

• **Memory resources**—Typically, the size of the physical memory is directly proportional to the amount of users using the system. Extra memory is required for the applications themselves, including the ability to run multiple instances of an application (such as a database).

• **I/O subsystem resources**—The capacity planning of the I/O subsystem is one of the most important planning steps—the I/O subsystem response time, throughput, and IOPS (I/Os per second) are critical to the overall work done by the application. Typically, the I/O subsystem is the slowest component of the computing environment. It needs to address multiple I/O profiles relating to numbers of I/O operations, I/O sizes, I/O latency, and total I/O throughput. These I/O characteristics are closely intertwined. Often, systems and storage architects tune the I/O subsystem by spreading the I/O load over multiple HBAs in a server and multiple physical disk drives or storage ports on a storage subsystem. In addition to the performance requirements, the data resiliency and availability requirements are paramount during the design of a SAN.

• **Availability**—It is imperative to provision multiple paths to the same data to help ensure I/O channel resiliency. There should be at least two paths available from server to disk to access the same set of targets and LUNs. In such a configuration, each data path is under utilized by design. The HBAs are also under utilized or undersubscribed in order to achieve optimum application performance and high availability. It is a common practice to architect server-to-storage multipathed access such that each HBA will not have a throughput of more than 50
percent of its maximum capacity during the normal operations. The design of multiple undersubscribed paths helps to ensure an optimal performance level during normal and degraded modes, such as a path failure. During the SAN design phase, special attention must be given to the fan-out ratio prescribed by storage subsystem vendors. The fan-out ratio can range from 4:1 to upward of 12:1, but most commonly is between 4:1 and 8:1. These oversubscription ratios are partly driven by the common practice of under-subscribing multiple HBA I/O paths for the reasons mentioned above.

- Performance—Storage subsystem vendors publish rated IOPS performance numbers that directly relate to the maximum throughput and fan-out ratios recommended for storage subsystem ports. These parameters form the basis of required information to complete the I/O subsystem design. Once the IOPS and throughput estimates of the applications are determined, the I/O must be spread across the number of HBAs to achieve optimal high availability and performance metrics. A higher IOPS number does not necessarily relate directly to higher I/O bandwidth requirements. In many cases, higher IOPS numbers are required for smaller I/O sizes and the effective bandwidth requirements are less, due to the effects of the storage array I/O service times.

Based on these design considerations, HBAs and their corresponding switch access ports are generally undersubscribed by design to optimize both the application availability and performance capabilities. For this reason, the increase in effective port density and reduction in solution cost achieved by the use of the Cisco MDS 9500 Series of directors along with the server-optimized DS-X9032 32-port switching module outweighs any considerations over the inherent small 3.2:1 oversubscription in the switching module. The fan-out recommendations made by storage subsystem vendors, combined with the common practice of purposefully under-subscribing server HBAs and their related switch ports, makes the possibility of sustained wire-rate performance on all fabric access ports not only not required but most likely not achievable.

APPLICATION I/O SUBSYSTEM—DESIGN VALIDATION

Cisco has used industry-standard benchmarks with real-world application I/O data to validate the previously mentioned design considerations and aid in sizing the I/O requirements of an application environment; the results are presented in this paper. One of the industry-standard benchmarks used to model a large-scale online transaction processing (OLTP) relational database is TPC-C, an OLTP system benchmark defined by the Transaction Process Council (TPC) that simulates a complete environment in which a population of terminal operators executes transactions against a database. The benchmark is centered on the principal activities (transactions) of an order-entry environment. These transactions model the processes of entering and delivering orders, recording payments, checking the status of orders, and monitoring the level of stock at warehouses.

The laboratory tests conducted by Cisco were designed to validate design considerations, as well a common fan-out ratio used for storage subsystems. An Oracle 8i TPC-C database of scale factor 6000 (simulating 6000 warehouses) was used to perform the tests. The database was approximately 600 GB in size and built on a Sun Enterprise 6500 server with 20 400-MHZ CPUs with a total of 20 GB of memory. The database block size used was 4 KB. The storage subsystem used in all tests was an HDS 9970V array with four 2-Gbps Fibre Channel ports. Each Fibre Channel port was configured with 28 Open-E LUNs presented to the server. The server was equipped with four 1-Gbps JNI FCE-1063 HBAs, which were zoned to the four storage subsystem ports (Figure 1). A Cisco MDS 9509 Multilayer Director Switch equipped with two DS-X9016 16-port modules was used for the test.

Figure 1
TPC-C Test Setup
The TPC-C test was conducted with one, two, and four HBAs to show the scaling effect within the server of adding more I/O channels to the application. The results of the TPC-C test in terms of transactions per second (TPS) and I/O throughput in terms of MBps and IOPS is shown in Figure 2.

Figure 2

![TPC-C Test Results Diagram]
Figure 2 illustrates two important points. The first point is that the overall aggregate I/O throughput generated by a server of this caliber with this amount of CPU and memory resources when running a heavily loaded database application does not exceed the line-rate performance of a single 1-Gbps HBA. Whether using one or four HBAs, the I/O bandwidth does not exceed 47 MBps, aggregating input and output. This is a typical result for most databases that use smaller block I/O sizes with higher IOPS loads. This server was equipped with much more CPU and memory resources than many typical enterprise servers, and the storage subsystem was dedicated to this test and not being consumed by any other application traffic.

The second point illustrated by the results is that moving from one to four HBAs yielded an improvement of approximately 40 percent in TPS numbers. This increase in TPS also resulted in an improvement of approximately 55 percent in aggregate I/O throughput, which was well within the performance capacity of a single HBA operating at 1 Gbps. It is clear that the IOPS capacity of a single HBA was the inhibiting factor, and eliminating it by adding HBAs to the system yielded a significant overall application performance improvement.

The I/O generated by the TPC-C database used a fixed I/O size of 4 KB. To profile I/O traffic patterns with different I/O sizes, a second batch of tests was performed using the same network configuration, server and storage components, and this time using the vxbench I/O test tool to generate I/O load. The vxbench tool was configured to exactly reproduce the same I/O profile generated by the TPC-C benchmark; however, the I/O sizes were varied. The purpose of this test was to show the varying I/O loads using I/O sizes other than the 4-KB size used in the TPC-C test. Vxbench is a multithreaded storage performance tool that enables the user to customize the workload, number of threads, and different I/O sizes. The TPC-C I/O traffic profile consists of approximately 66 percent read and 33 percent write I/Os with one to two simultaneously active I/Os per LUN on the storage subsystem.

Figure 3
Vxbench Test Results

The test was performed against all 112 LUNs configured on the HDS 9970V array (these were the same LUNs used for the TPC-C test) presented by array across four Fibre Channel ports. Each of the array’s Fibre Channel ports had 28 LUNs being presented to the server. Using the TPC-C I/O traffic profile and the vxbench tool, I/O was generated across all four ports to all 112 LUNs with a number of one to two outstanding active I/Os per LUN. Figure 3 shows the results of this test, depicting the throughput in MBps. The results show the aggregate bidirectional I/O throughput demands using various I/O sizes. The intention was to simulate different applications that may use I/O sizes other than the 4-KB size used in the first test. The results found at 4 KB are similar to the results measured with the TPC-C database traffic. Different operating systems use varying default I/O sizes for their file systems. For example, the UFS file system for Solaris uses an 8-KB I/O size whereas the Microsoft Windows NTFS file system uses a 64-KB I/O size.
The results in Figure 3 show that a set of HBAs within a server does share the I/O bandwidth, but it does not translate to an N \times Throughput capability. The I/O bandwidth of a server as powerful and resource-loaded as the one used for this test, running a staged I/O performance with varying I/O sizes with no background computing tasks, still did not generate more than 250 MBps of aggregate I/O performance. These tests were performed with no background tasks executing on the server or storage platforms. This configuration provides results that are higher than can be expected within a normal production environment, where servers and storage devices have to share their internal resources and performance capacity among multiple applications and I/O workloads, respectively.

The results presented in the TPC-C test and the vbench test show that the prescribed fan-out ratios recommended by storage subsystem vendors are valid and optimal for most commonly available commercial applications. In addition, the 3.2:1 oversubscription of the Cisco MDS 9000 Series DS-X9032 32-port switching module and Cisco MDS 9100 Series fabric switches provides ample bandwidth to satisfy most enterprise applications.

**APPLICATION I/O SUBSYSTEM—CASE STUDY**

The following real-world example shows applicability of oversubscription in the deployment of SAN.

To further validate fan-out and fabric oversubscription at the storage and switch levels, respectively, the I/O statistics of one of Cisco’s main ERP servers were captured during the year-end book closing process for the company’s fiscal year 2002. Cisco has a completely automated financial system that ties together all major inventory, sales, manufacturing, and other systems. Cisco is able to close its books for an entire year through the touch of a button. The financial system developed and used by Cisco is designed to keep the company running efficiently. Cisco’s financial systems and practices have been a sought-after model by many Fortune 200 companies. Because Cisco conducts most of its business over the Internet, these systems are critical to Cisco and have a high level of application and I/O importance placed on them.

![Figure 4](image-url)
The I/O data was collected from the manufacturing module of an Oracle Applications 10 system. The front end of this ERP application consisted of an Oracle Applications 10 application running on a Sequent PTX server. The back end consisted of an Oracle 8i database running on an HP Superdome 2-node Oracle Parallel Server (OPS) active/standby cluster. Each HP Superdome 32000 was equipped with 32 875-MHz CPUs and 48 GB of memory. The cluster was also equipped with 22 HBAs, 18 of which were used for accessing storage for this application and are used to increase the resiliency of the I/O channel and to distribute the I/O load.

At the time of this data capture, the database was approximately one terabyte in size. The statistics were collected for one week leading to the end of the fiscal year from the manufacturing module of the Oracle Applications 10 system. The architecture of this ERP system is shown in Figure 4. The captured I/O statistics per HBA are shown in Figure 5, which shows that the I/O is evenly distributed across the 18 HBAs. The I/O load itself was never more than approximately 18 MBps per HBA, while the average transaction rate was around 270 TPS.

Figure 5

Cisco ERP Application Statistics

The I/O load statistics in Figure 5 show that an application as critical as the Cisco ERP system—during the most process-intensive time of the fiscal year—did not generate a large amount of I/O bandwidth. The Cisco application architects also followed the common practice of distributing I/O load across purposefully undersubscribed HBAs within the system. This data also points to the bursty nature of most database I/O activity. To design a network to meet peak I/O bandwidth demand all the time is wasteful in terms of switching resources and is overly costly to implement. As shown in Figure 5, even though the TPS numbers remain relatively constant, the I/O throughput was bursty with peaks being experienced as very short bursts. With the average I/O per HBA being around 4-5 MBps, the total aggregate I/O demand for this application during its most process-intensive time was approximately 90 MBps (18 HBAs x 5 MBps). This amount of I/O could easily be supported across a single 1-Gbps HBA if required.
This application validates the fan-out ratio prescribed by storage vendors and fits well within the embedded oversubscription of 3.2:1 in the DS-X9032 32-port line card and the Cisco MDS 9100 Series of fabric switches. This is not to say that all applications do not drive high amounts of I/O bandwidth. However, I/O bandwidth is bursty in nature and not directly tied to the computing power of the application server. Application types, their associated I/O profiles, and the computing platform architecture dictate the I/O bandwidth requirements.

CONCLUSION

This paper presented several examples of I/O subsystem architectures, capabilities, and traffic profiles based on benchmarks and real production data to illustrate oversubscription. The results gathered from industry-standard I/O benchmark tests and from the production environment of a large enterprise application show that HBAs in most enterprise application servers are undersubscribed either by design or by the limitations of the HBA or computing platform. Because HBAs are commonly deployed in at least a pair for resiliency, they are often designed to be loaded no more than 50 percent. This configuration helps to ensure that any failure of a primary path will not overload any backup I/O paths. Even in systems with more than two HBAs, most of these are also designed to sustain a maximum of 50 percent of full load per HBA for resiliency and IOPS loading reasons. Often, the requirement to distribute load across multiple HBAs is not based on actual bandwidth requirements, but on IOPS requirements. This common design practice fully validates the fan-out ratios for server-to-storage connectivity prescribed by storage vendors. Additionally, this design practice supports the inherent oversubscription in the DS-X9032 32-port host-optimized switching module for the Cisco MDS 9000 Series and the host-optimized ports within Cisco MDS 9100 Series fabric switches, both of which are at a small (3.2:1) oversubscription.

It is not only safe to design oversubscription into a SAN—it is also compulsory. Oversubscription helps offset infrastructure costs and is an integral part of every networking topology. The very nature of a SAN is to fan out the connectivity of fewer storage subsystems to numerous server connections. While performance is important, so are the capabilities of a switched fabric to provide services, including congestion avoidance, preferential services, and blocking avoidance. Cisco provides a full SAN switching product line with the Cisco MDS 9000 Series, a line that is optimized to build scalable SAN infrastructures and to provide industry-leading performance, resiliency, security, and manageability.