Compaq SANworks™

Data Replication Manager over an ATM Link 3/2000

Application Note

About This Document

This Application Note provides an overview of the ATM (Asynchronous Transfer Mode) link and its impact on performance when used with Data Replication Manager (DRM)—a storage-based data replication and workload migration solution for copying data online and in real time to remote locations via an extended Storage Area Network (SAN). The ATM link has not been qualified for configurations other than Data Replication Manager.

For complete details on Data Replication Manager, refer to the *Compaq SANworks Data Replication Manager HSG80 ACS Version 8.5P Operations Guide.*

This Application Note contains the following sections:

- Introduction to DRM with ATM Transport
  - Fibre Channel to ATM Gateway
    - Open Systems Gateway Interface
    - ATM Link Topologies
- Case Study
  - Customer Requirements
  - DRM Requirements
  - Obtaining Best Performance
- Installation and Configuration
  - WTI
  - Using the Fibre Channel-to-ATM Configuration Wizard
  - Site Readiness Checklist
  - Modifying VP and VC Values
  - Modifying ActiveVciBits Values
- Additional Resources
  - Website
- Glossary

Notice

Printed in the U.S.A.

Application Notes - Data Replication Manager over an ATM Link

First Edition (February 2000)

Second Edition (March 2000)

Part Number EK-DRMAL-AA.B01

Introduction to Data Replication Manager with ATM Transport

During normal data processing, data is simultaneously written to initiator (local) and target (remote) sites. While copies of data reside at both sites, host data access occurs through the initiator site, unless there is a failure or catastrophe that disables processing at that site. In the event of an initiator failure, another site can continue processing data in the interim.

Data Replication Manager provides rapid data access recovery and continued data processing after the loss of one or more components. Data Replication Manager uses the peer-to-peer remote copy function of the HSG80 controller to achieve data replication. HSG80 controller pairs at the initiator site are connected to their partner HSG80 controller pairs at the target site.

Data Replication Manager can replicate data up to 70 kilometers (approximately 44 miles) via an extended storage area network over direct Fibre Channel links or go greater distances with Fibre Channel-to-ATM Gateways.

Fibre Channel to ATM Gateway

Fibre Channel-to-ATM gateways, referred to as Open Systems Gateways (OSG), provide ATM connectivity into telecommunication networks for the *Compaq SANworks* extended SAN/DRM solution. ATM connectivity is ideal where contingency sites are required to be very long distances away to protect against wide area disasters, or where dark fiber connectivity is unavailable.

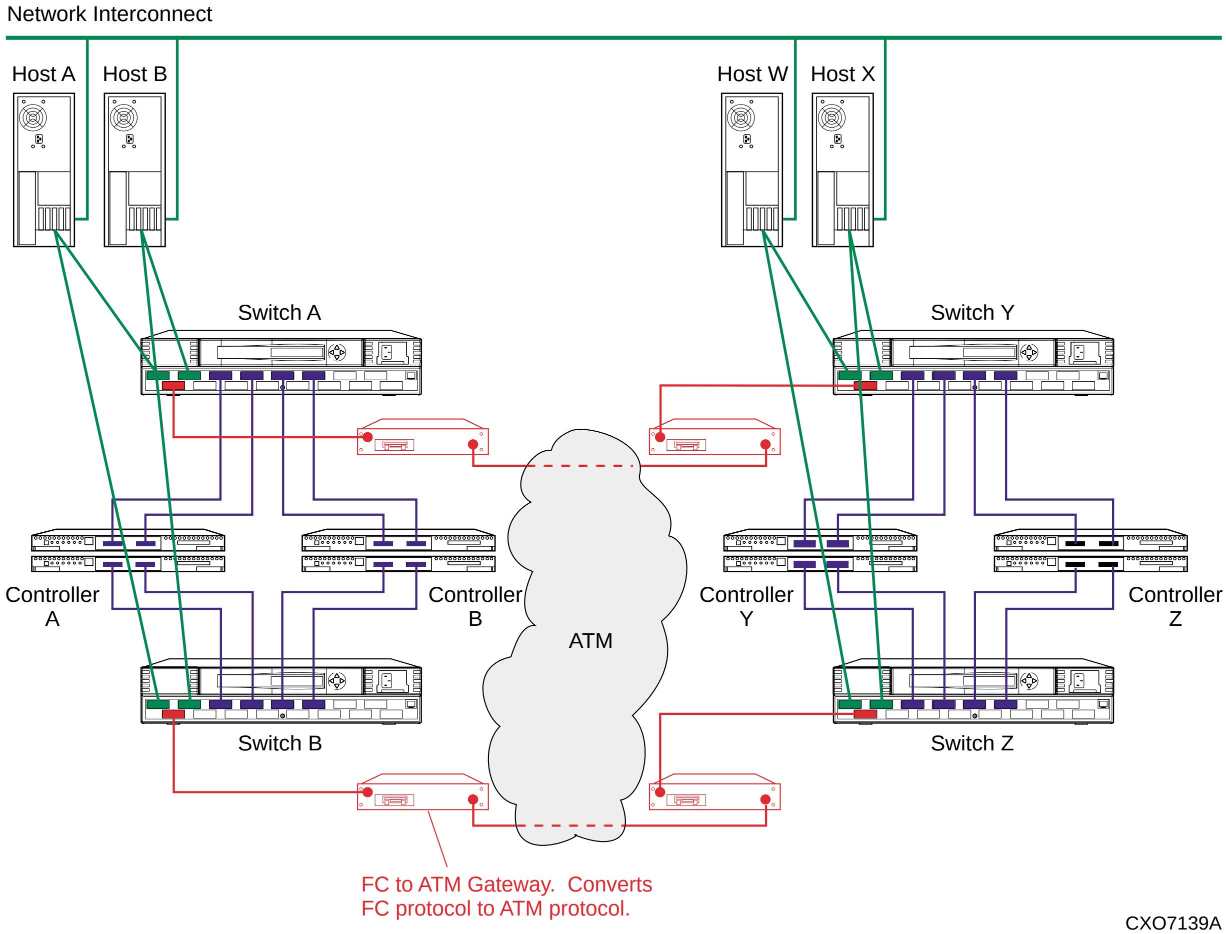
ATM is a communications networking technology that carries information (voice, video, and data) in 53-byte cells. The fixed-length cell allows a network to carry any type of information within the cell and also to provide stringent service qualities that can differ by application.

Each virtual connection in an ATM network has a service category. The performance of the connection is measured by the established *QoS* (Quality of Service) parameters. *Constant Bit Rate* (CBR) is the service category used by the current DRM solution over ATM. The only requirement for CBR is Peak Cell Rate (PCR). This value will be determined by performance requirements.

The DRM solution uses virtual connections in an ATM network that are predefined and permanently left in place. This is referred to as a *Permanent Virtual Circuit* (PVC). The 53-byte cell contains a 5-byte header that contains all the information to get the cell from the source to the destination. Every cell header contains an address identifier used by the ATM switches to track how cells move from one physical link to the next. Before any communication exchanges take place, the connectivity information stored in the switches must be manually loaded. The network administrator must select the end points that require connectivity to set up PVCs. These address identifiers are known as VPIs and VCIs (Virtual Path Identifiers and Virtual Channel Identifiers).

DRM requires two ATM links at a rate between 1.544 megabits per second (193 kilobytes per second) and 126 megabits per second (15.75 megabytes per second).

FIGURE 1. Data Replication Manager configured over an ATM link



Open Systems Gateway Interface

The OSG interface supports:

- 50-micron multi-mode fiber-up to 500 meters (Fibre Channel and ATM sides)
- 62.5-micron multi-mode fiber-up to 200 meters (Fibre Channel and ATM sides)
- Performance target data rate from 1.554 megabits per second to 155 megabits per second.
- Point-to-point connectivity only (no alternate routing supported at present)
- Fibre Channel Class 3 traffic
- Frame data field size of 2048 bytes

The OSG interface does not support:

- Class F frames

ATM Link Topologies

Dual Links

The most disaster tolerant of the ATM Link configurations, dual links provide protection in the event that one link becomes unavailable. No ATM switch is required.

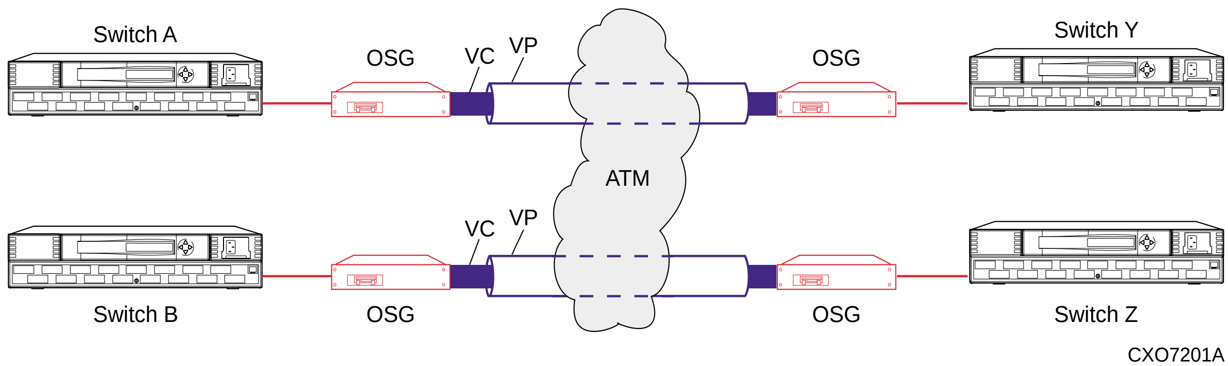


FIGURE 2. Dual link configuration

Shared Virtual Channel Links

A shared link allows for one telecommunication link between sites with the line bandwidth divided between the two Fibre Channel fabrics. The data rate will be shared according to the Peak Cell Rate allocation. Because there is only one link, this configuration is the least disaster tolerant.

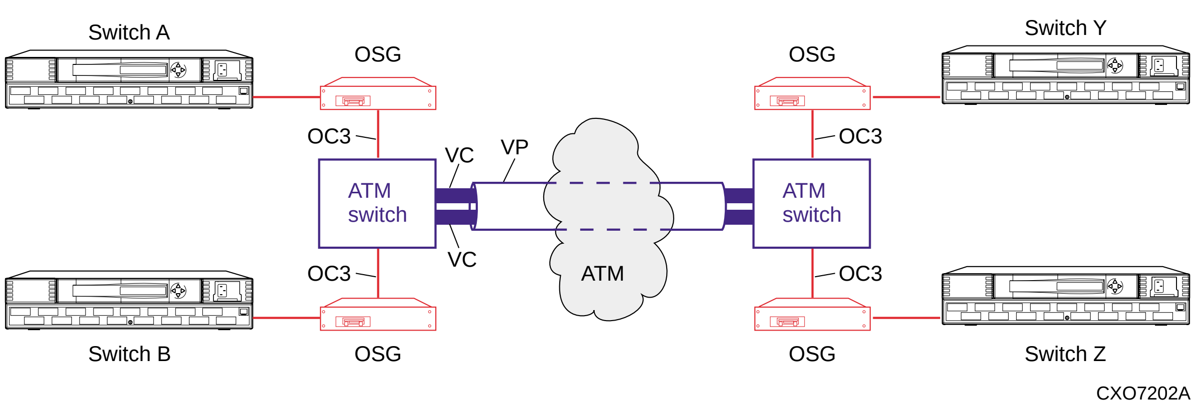


FIGURE 3. Shared virtual channel link configuration

Fractional Virtual Channel Links

Fractional lines allow for leasing the exact bandwith or access speeds needed to support a user's requirements.

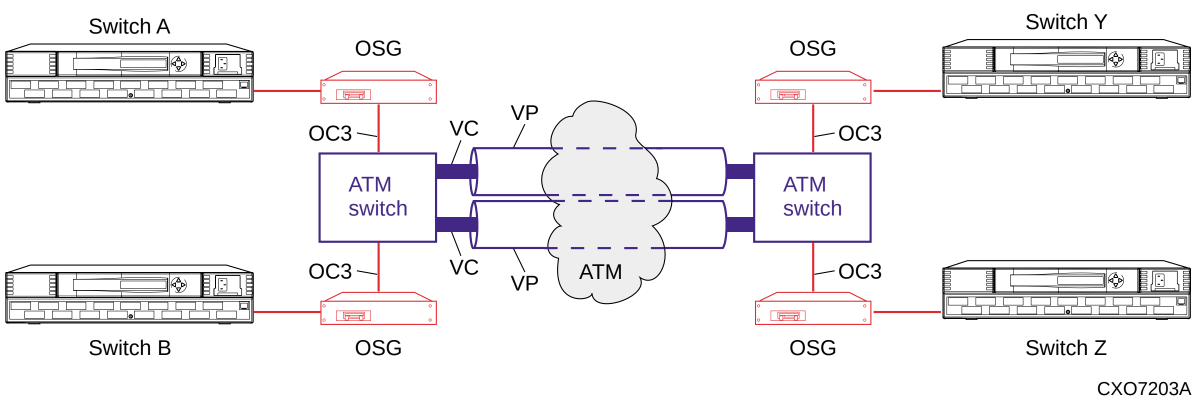


FIGURE 4. Fractional virtual channel link configuration

ATM Inverse Multiplexing

Multiple ATM T1/E1 rates provide aggregate bandwidth when higher bandwidth is not available or there is no need for faster access rates.

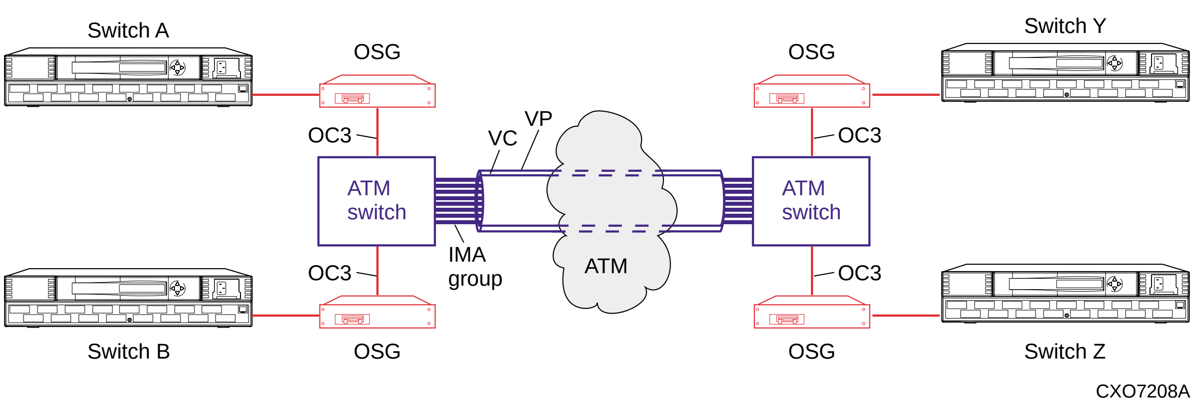


FIGURE 5. ATM inverse multiplexing configuration

Case Study

Example Criteria:

- 150 miles (241.35 kilometers) between sites
- Write percentage of 30%
- Size of I/O equals 8 kilobytes

DRM Requirements:

DRM requires two separate PVCs. Each Open Systems Gateway pair and the pair's corresponding virtual path can sustain a maximum data rate of 15.75 megabytes per second.

The telecommunications provider must supply VPIs and VCIs through their network for each PVC.

Obtaining Best Performance

When setting up a DRM configuration, changing the operational mode between synchronous and asynchronous can have a significant impact on performance as seen by the user. Unfortunately, there are no hard and fast rules for which setting is best. Instead, the best setting depends on a combination of the I/O workload, the type of inter-site communication, and the distance of the inter-site link.

This section provides the background information necessary to obtain the best performance over a DRM connection with an ATM link.

Workload profile

In normal operation, read requests from the host will be satisfied from the local system. Read performance will therefore be very similar to performance on a non-DRM system. Write operations, however, must be replicated to the remote site. It is this replication using differing communication transports over varying distances that will introduce performance variations.

As the percentage of writes increases, the amount of data sent to the remote site will also increase, potentially reducing performance. The first step in assessing the impact of DRM is, therefore, to analyze the existing I/O workload. In particular, the percentage of writes, the size of these writes, the write request rate, and the existing write response time should be determined. With this information it will be possible to predict the resulting performance using DRM under different configurations with a reasonable degree of accuracy.

Distance considerations

When data is transferred from the local to remote sites, the SCSI protocol requires a minimum of four trips over the path. These are conceptually as follows:

1. Ask the remote site if it is ready to receive data.

2. Wait for the response from the remote site.

3. Send the data to the remote site.

4. Wait for acknowledgment from the remote site.

When sending data over fiber, the transmission time is approximately 5 microseconds per kilometer (KM). Since a minimum of four trips is required for each SCSI data transfer, this translates to 20 microseconds per kilometer, or about 32.2 microseconds per mile. As an example, if a remote site is located 150 miles away from the local site, the total time will be 4,830 microseconds for every data transfer. Since a “typical” I/O operation on a non-DRM configuration with write-back cache takes approximately 500 microseconds, it is easy to see that long distances can have a significant effect on performance.

Communication type

The specific type of communication may also have a significant impact on performance. When using ATM over an OC-3 link, the data rate is limited to a maximum of 155 megabits per second. Straight Fibre Channel, on the other hand, has a native speed slightly over 1 billion bits (Gb) per second. If the communications link is ATM, the hardware that translates between Fibre Channel and ATM protocol introduces an additional latency of about 2.41 milliseconds for all four trips. The effects of these different transport mechanisms must also be taken into account when considering performance.

Taking the previous example of 150 miles (4.83 milliseconds), the addition of ATM will increase the latency by 2.41 milliseconds to a total of 7.24 milliseconds. Additionally, the maximum data rate will be lowered from the Fibre Channel maximum of 100 megabytes per second to approximately 13 megabytes per second (taking protocol overhead into account).

Synchronous operation

In the synchronous mode of operation, command completion is not returned to the host until the data has been stored on both the local and remote sites. If the host computer issues multiple write requests at the same time, all of these requests will be pipelined (multiple requests on the connection at the same time). Even though multiple commands may be outstanding, completion for each command will only be returned when all data for that command has been received and acknowledged by the remote site. Since the commands will be pipelined, there will be very little queuing, so the response time of each request will be determined by the total time for the four trips. This means that the length and type of the communications path is the primary factor for the response time of each command.

The maximum number of commands that may be outstanding simultaneously is controlled by the OUTSTANDING_IO parameter on the controller. If the number of commands issued by the host exceeds this value, those commands will be queued internally by the local controller for later transmission.

Asynchronous operation

In order to reduce response times, asynchronous operation will return I/O completion to the host as soon as the data has been sent to the local controller, but before it has been received by the remote system. The result is that the host computer will see only the response time associated with the local system. The time delays associated with data replication will be masked, since the replication is accomplished in the background.

Since completion of the transfer has been returned to the host, the host may send additional requests, secure in the knowledge that the previous data was safely written. Because of this, the controller must ensure that all data that was previously acknowledged as being complete, but that is currently “in transit” to the remote system, is delivered in the correct order. This is done by allowing only one data transfer at a time (for a particular association set) to be sent to the remote site. Any additional data transfer commands are queued internally within the controller and sent one at a time as the previous transfers complete.

Because the host sees a very fast response time, it may issue I/O requests at a rapid rate. The replication, on the other hand, proceeds at the rate dictated by the inter-site link. As a result, a large queue of requests that are waiting for replication may build within the controller.

This number will limited by the OUTSTANDING_IO parameter. The number of requests that have been acknowledged to the host by the controller as being completed, but have not yet been sent (or acknowledged) by the remote site is compared to the value of OUTSTANDING_IO. If the number of requests is greater, then subsequent requests sent by the host will not be acknowledged as completed until they are sent to and acknowledged by the remote site.

A simple example may help to clarify this:

Assume that OUTSTANDING_IO is set to a value of five, and that a host issues a request, waits for completion from the controller, then immediately issues another request. Furthermore, assume that the DRM configuration is on ATM over a very long distance, requiring a significant amount of time to perform replication. In asynchronous mode, each request issued by the host will be acknowledged as completed by the controller very quickly. As a result, the host will issue five requests before the remote site has completed the first request. If the host then issues another (sixth) request, this will exceed the value of OUTSTANDING_IO. The local controller will not return completion of this request to the host, and will queue it behind the previous requests that are awaiting replication. Completion of the sixth request will not be returned until such time as all five preceding requests have been sent to and acknowledged by the remote site, and the sixth request itself has been sent and acknowledged. The result of this is that the host will see a response time for the first five requests equal to the time taken by the local controller, while the sixth request will have a response time equal to the sum of the transit times from local to remote controller for all preceding requests.

Assume that the time for the local controller is 500 microseconds, while the total time for replication to the remote controller is 5 milliseconds. In this case, the host will see a response time of 500 microseconds for each of the first five requests, since completion will be returned immediately. The sixth request will have response time slightly less than 30 milliseconds, since it must wait for all previous requests to be completed by the remote site. In other words, the sixth request will have a response time sixty times greater than the first five.

Performance analysis

With the preceding taken into account, the performance picture becomes quite complex, since there are so many factors that interact. In order to analyze the performance of a DRM configuration and select the optimum parameters, all of the contributing factors must be fully quantified. This is a multiple step procedure, requiring detailed knowledge of the I/O workload and proposed DRM configuration.

The steps required to configure a DRM solution for optimum performance consist of both pre-installation and post-installation analyses. The pre-installation phase consists of measuring the current I/O workload on the units to be replicated, calculating the minimum transport bandwidth required to support the existing I/O load, and predicting the response times for the DRM system for both synchronous and asynchronous replication. Once the system is installed, additional steps may be taken to tune the configuration for maximum performance.

The pre-installation phase consists of the following steps:

1. The first step is to determine the I/O workload parameters that are relevant to DRM. These parameters should be measured only on those units that are to be replicated, since non-DRM units are not of interest to this analysis.

The parameters to be measured for each unit include:

- ❑ Transfer size of writes.
- ❑ Request rate of writes.
- ❑ Either the average response time or the average queue depth of writes.
- ❑ Either the average response time or the average queue depth of reads (for a later, more comprehensive analysis).
- ❑ Transfer size and request rate of reads (for a later, more comprehensive analysis).

The data may be measured with existing operating system utilities (such as the Windows NT Performance Monitor) or third party utilities (such as ViewPoint from Datametrics). It is important to gather this data during “typical” and “critical” periods. Since the data gathered will be used for later analysis, the results of the analysis are dependent on the quality of the collected information. If there are several distinct I/O phases (such as interactive and batch), data should be collected separately for each of these periods, and the resulting data analyzed independently.

2. The next step is to determine the minimum transport bandwidth required. The bandwidth required is the product of the write request rates and the write transfer sizes for all replicated units. As an example, if the busiest period shows a total write request rate of 150 requests per second with an average transfer size of 8 KB, the required bandwidth would be 1200 KB/Sec, or 1.2 megabytes per second, so neither ATM nor straight Fibre Channel would impose significant bottlenecks. Note that due to other considerations, this rate may not be maintainable. This step is merely to ensure the minimum bandwidth required to meet the existing data rate.

3. The latency for synchronous replication may be calculated once the distance between the local and remote site is known. For straight Fibre Channel, the latency in milliseconds is the distance in miles multiplied by 0.0322 (distance in kilometers must be multiplied by 0.02). If ATM is used, 2.41 milliseconds must be added to this figure to account for the ATM protocol conversions. For a 150 mile distance, as an example, the fiber latency would be 4.83 milliseconds, and 7.24 milliseconds for ATM as earlier illustrated.

   Although there may be large latencies associated with synchronous replication, the ability to send multiple requests from the host simultaneously allows a very high request rate to be supported.

4. For asynchronous replication, the host will see very little latency, since completion will be returned once the local controller has received the data. Only one data transfer may be outstanding between the local and remote site in asynchronous replication mode however, so the maximum request rate is dependent on the amount of time taken for that replication. The maximum rate can be calculated by taking the reciprocal of the latency calculated in step 3. Using 4.83 milliseconds (0.00483 seconds) as an example, 1 divided by 0.00483 equals about 207 requests per second. If the connection is made via ATM, the latency from step 3 of 7.24 milliseconds results in a maximum request rate of about 138 requests per second.

   The rate that is calculated is the maximum rate that can be supported while retaining full asynchronous response times. If the rate rises above the calculated value, then more requests will be arriving at the controller every second than can be sequentially sent over the communications link. This will result in queues of requests building at the local site, and when the number of requests exceeds the value of OUTSTANDING_IO, the request will be delayed for a significant period of time (as detailed earlier), resulting in a large increase in response time.

It is important to note that most I/O is quite “bursty.” Although the average arrival rate may be below the thresholds calculated in this step, bursts of I/O requests may cause an instantaneous peak rate that will cause lengthy response times.

A general rule of thumb is to aim for an average rate of no more than 80% of the rates calculated in this step. From the examples given, an average write rate of more than 165 requests per second (fiber) or 110 write requests per second (ATM) may result in long response times. Once the maximum write request rates have been calculated, the minimum transport bandwidth should (once again) be determined. As explained in step 2, this is done by multiplying the average write transfer size by the write request rate. For the examples given, a maximum rate of 165 write requests per second yields a data rate of 1.32 megabytes per second. For a rate of 110 on ATM, the minimum required bandwidth will be 880 kilobytes per second.

5. If the data gathered in step 1 indicates that the write request rate for an association set will exceed the maximum request rate for asynchronous requests calculated in step 4, then better performance will be realized by setting the operation mode to synchronous. On the other hand, if the request rate from step 1 is less than the maximum rate for asynchronous, then setting the operation mode to asynchronous will provide the best performance.

6. The optimum value of OUTSTANDING_IO depends on both the type of unit and the spatial locality of requests sent to individual units.

   For JBOD units, if there is a possibility that two or more requests might be sent rapidly to the same location on the unit, the value of OUTSTANDING_IO should be set fairly low. A value of 10 or less is recommended.

   For RAID-5 sets, if there is a possibility that two or more requests might be sent rapidly to the same strip, the value of OUTSTANDING_IO should also be set fairly low. To calculate the size of a strip, multiply the chunk size of the RAID-5 set by one less than the number of members in the set. For a six member RAID-5 set with a chunk size of 256, for example, the size of a strip would be 256 times 5, or 1,280.

   In any event, performance of the unit should be monitored with tools such as VTDPY and DSTAT, and the value of OUTSTANDING_IO adjusted as required to attain the maximum performance.

7. Once the preceding calculations are complete, it may be of interest to calculate the overall maximum request rate. Since all calculations were based on only the write request rate, the combined request rate is determined by dividing the write request rate by the fraction of writes.The result will be the maximum request rate that this configuration will support.

   As an example, if the write fraction is 0.3 (30%), then a write rate of 165 per second for fiber will result in a maximum request rate of 165 divided by 0.3, or 550 requests per second. For a write rate of 110 on ATM, the maximum request rate will be 367 requests per second.

Summary

When choosing between synchronous and asynchronous replication, the best performance depends on a variety of workload factors. Since the setting of this parameter is dynamic, it is relatively easy to switch between the two operational modes and view the resulting performance change. Guidelines for setting the mode are not intended to be hard and fast rules, but rather should be thought of as a starting point.

It should also be noted that as workload conditions change over time, changes might be needed in the operation mode to maintain optimum performance.

Installation and Configuration

WTI

The Western Telematics switch (WTI) is required to service and configure the Open Systems Gateway. The WTI is a 16-port, asynchronous switch that mounts in a standard 19-inch rack along with the Open Systems Gateway (OSG). To enable remote servicing of the OSG, a modem must be installed. You can obtain a local phone number from your carrier for the modem.

Using the Fibre Channel-to-ATM Configuration Wizard

Use the Fibre Channel-to-ATM Configuration Wizard to set up both gateways. The following information is required to run the Wizard:

- ❏ IP address
- ❏ Subnet address
- ❏ Default gateway
- ❏ Node number
- ❏ Internal Clocking (set to *no* when connected to ATM switch)

Site Readiness Checklist

- ❏ Copy of ATM QoS contract supplied to Compaq?
- ❏ VPI:___________VCI:___________PCR:___________
- ❏ Who is the ATM switch manufacturer?
- ❏ What is the model number of the switch?

Modifying VP and VC Values

Edit the *atmclink* file to modify the VPI, VCI, QoS, and Peak Cell Rate (PCR) values. Refer to the *Site Readiness Checklist* in this Application Note for VPI and VCI values.

1. Type the following command at the root level to access the *atmclink* file:

```
fs cd  /cnt/configs/active/slot01/cfg
ed atmclink
```

2. The *atmclink* file will contain the following:

AtmPvcVPI = *value*

AtmPvcVCI = *value*

AtmPvcQOS = *value (change to CBR)*

3. The *atmclink* file will also contain the following:

AtmPvcUcrPcr = *value*

This value is the PCR of the gateway. Its value is given in kilobits per second. Enter a value in the acceptable range of 1,544 to 126,000 kilobits per second. The value required for this field is the result of a performance analysis similar to the one done in the previous Case Study.

NOTE: Overhead values that should be considered when specifying Peak Cell Rate with the telecommunications provider are as follows:

Sonet overhead = 150.336 megabits per second/ 155.52 megabits per second = .967

ATM overhead = 48 cells/ 53 cells = .094

Fibre Channel 8b/10b encoding = 8 bits/10bits = .80

Fibre Channel overhead = 2048 bytes /(2048 + 48 bytes) = .977

These values as well as telecommunication-specific values must be used when calculating the bandwidth that the telecommunications provider must furnish to ensure the required DRM bandwidth.

As in our Case Study, if the required bandwidth over an ATM link of 150 miles is 880 kilobytes per second, then to maintain that rate a telecommunications provider must furnish that rate plus overhead or 1.87 megabits per second.

Modifying ActiveVciBits Values

Edit the *atmoc3* file to modify the *ActiveVciBits* values as indicated in Table 1 below. Refer to the *Site Readiness Checklist* in this Application Note for VPI and VCI values.

1. Type the following command at the root level to access the *atmoc3* file:

```
fs cd  /cnt/configs/active/slot01/cfg
ed atmoc3
```

2. The *atmoc3* file will contain the following:

AtmVciBits = *value*

Table 1 ActiveVciBits Values

ActiveVciBits	VPI Range	VCI Range
12	0	0- 4095
11	0-1	0- 2048
10	0-3	0-1023
4	0-255	0-15

NOTE: For information about ed commands, refer to the *UltraNet Open Systems Gateway Command Reference Guide.* For detailed installation information consult your *UltraNet Open Systems Gateway Hardware Installation Users Guide.* Information is also available at:

www.CNT.com

Websites

Compaq

Check the Compaq website for more information on the complete line of Fibre Channel storage products, product certification, technical information, updates, and documentation. This information can be accessed through our website at:

http://www.compaq.com/products/storageworks/

ATM Forum

Over 600 organizations belong to the ATM Forum, an international standards organization that promotes the use of ATM products and services. ATM Forum can be accessed at the following website:

http://www.atmforum.com/

Disclaimer

Compaq does not warrant Third Party Products, which Customers may purchase in the making of a DRM solution. These products are used on an “AS IS” basis unless otherwise specified in the Third Party’s Warranty Attributes. Third Party Products may be warranted by the third party as specified in the documentation provided with the Third Party Products.

Glossary

ATM (*Asynchronous Transfer Mode*) Communications networking technology for both LANs and WANs that carries information in fixed-size cells. ATM uses logical connections to provide quality of service guarantees that enable disparate traffic such as data, voice, and video to be carried over the same local or wide area network.

CBR (*Constant Bit Rate*) Category of ATM service. Supports a constant or guaranteed data rate. CBR is designed to support applications that need a highly predictable transmission rate.

Default Gateway Default path that a computer or router uses to forward and route data between two or more networks that have different protocols.

Fibre Channel Technology for very high-speed, switching-based serial transmissions.

IP address Numeric address that identifies a particular computer on the internet.

Latency The amount of time required for a transmission to reach its destination.

OC-3 Optical carrier that provides high-speed bandwidth at 155 megabits per second.

PCR (*Peak Cell Rate*) Peak Cell Rate is the maximum transmission speed of a virtual connection. PCR is a required parameter for the CBR service category.

PVC (*Permanent Virtual Circuit*) Logical connection manually defined by the network administrator. The PVC is created by specifying the VPI and VCI.

QoS (*Quality of Service*) Each virtual connection in an ATM network has a service category. The performance of the connection is measured by the established QoS parameters, which are outlined by the ATM Forum. Performance issues include data rate, cell loss rate, cell delay, and delay variation (jitter). Categories of ATM service are: Constant Bit Rate (CBR), Variable Bit Rate-Real Time (VBR-RT), Variable Bit Rate- Non-Real Time (VBR-NRT), Available Bit Rate (ABR), and Unspecified Bit Rate (UBR).

SCSI (*Small computer system interface*) A processor-independent standard protocol for system-level interfacing between a computer and intelligent devices including hard drives, floppy disks, CD-ROMs, printers, scanners, and others.

Speed of light through fiber Approximately 200,000 kilometers per second or 5 microseconds per kilometer.

Subnet mask (also known as *address mask*) A subnet is an IP network that can be reached through a single IP address. All the members of the subnet share the mask value. Members of the subnet can then be referenced more easily. A subnetwork is a network that is part of another network, connected through a gateway, bridge, or router.

UBR (*Unspecified Bit Rate*) Category of ATM service. Supports connections that have no performance requirements.

UltraNet Wizard *Fibre Channel-to-ATM Configuration Wizard* is an *UltraNet* application that lets you designate the default configuration settings for Fibre Channel-ATM on the Open Systems Gateway.

VCC (*Virtual Channel Connection*) Defines a logical networking path between two endpoints on the network. ATM cells travel over this connection.

VCI (*Virtual Channel Identifier*) Field of the cell header that stores the virtual channel address.

VPC (*Virtual Path Connection*) Series of virtual paths joined together to form the logical groups of circuits defined for each link of the network. Moves traffic from one link to another.

VPI (*Virtual Path Identifier*) Field of the cell header that stores the virtual path address.

WTI Switch The Western Telematics (WTI) switch must be installed to set up and service the ATM gateway. The WTI switch is a 16-port serial switch that can be used to configure or service the OSG unit locally or remotely.