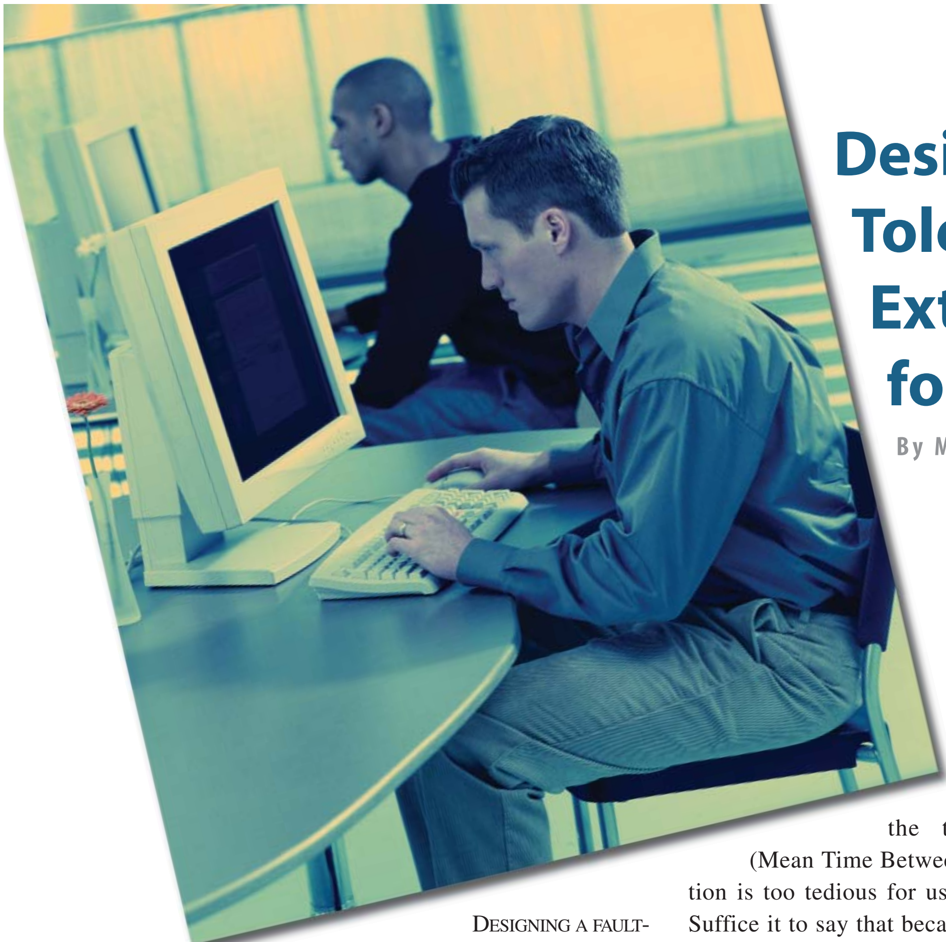

# Designing a Fault-Tolerant Channel Extension Network for Internal Recovery

By Mike Smith

DESIGNING A FAULT-tolerant channel extension for the mainframe environment incorporates all the best design principles and builds on them when configuring local and remote equipment as well as the network components. In simplest terms, this means ensuring that there are no single points of failure, there is adequate redundancy and capacity to support the environment when component failures do occur. The environment must also be manageable—at least to the extent that it can be documented and understood, diagnosed when necessary and supported on an ongoing basis.

Although very basic, a quick review of some of these configuration best practices can provide the foundation necessary to illustrate how to build on them going forward and to identify and eliminate single points of failure.

Figure 1 demonstrates the simplest of all possible configurations. In this configuration there is one of everything necessary to support the connectivity need, but no redundant channels have been implemented. Thus, any failure along the channel path will cause the system to go down.

The resilience provided by today's processors and storage systems is absolutely astounding when compared with the machines of just a few years ago. Today, most machines have dual power connections, separate power boundaries within the machine, and redundant processors for critical tasks making them extremely fault-tolerant. However, the way in which this configuration has been implemented has introduced several single points of failure.

For the sake of this discussion, let us say that three individual components have been identified, each of which creates a failure point: The channel adapter in the storage system, the channel adapter in the processor and the fiber optic cable itself.

Truthfully, discussing the true mathematical MTBF (Mean Time Between Failure) of this configuration is too tedious for us to deal with in this article. Suffice it to say that because of the implemented configuration, an environment has been created where the entire system can be "down" even though both the storage and processor are "up." Further, because three separate components have been identified, we can expect failures much sooner than the specified MTBF of any of the individual components.

This simple configuration also ignores the possibility that a single channel may not have sufficient capacity to support the data transfer rate at all times. However, since only a single channel path has been implemented, there are no separate concerns regarding sufficient capacity during a failure situation. In this case, there simply is not any capacity and the complete environment is down until repairs can be completed.

Figure 3 shows a slightly more robust configuration where two channel paths have been configured between the storage subsystem and the processor. Assuming the second channel path is connected to different channel adapters (serviced by different power boundaries in each machine) then the system is protected from failure of either of the channel paths. Instead, a complete outage is only possible if multiple concurrent failures occur along both channel paths.

By doubling the number of channel paths and eliminating all single points of failure in this configuration, we must now ask ourselves two capacity questions: First, do two channel paths provide sufficient capacity for normal operations at all times, and second, will the capacity of a single channel suffice if the other one fails?

Although this second configuration has eliminated all the single points of failure, it is still a bare bones configuration that is rarely seen.

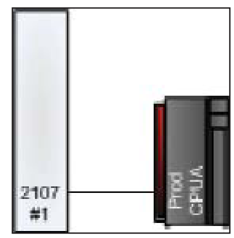


**Figure 1: Simplest possible configuration**

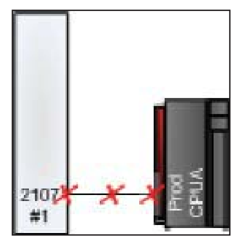


**Figure 2: Single points of failure in simple configuration**

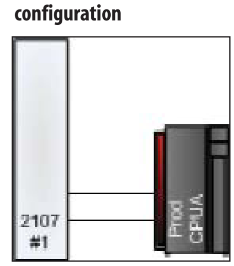


**Figure 3: Simplest configuration with no single points of failure**

A configuration that is more commonly encountered is illustrated in Figure 4. In this configuration, storage directors have been included to simplify the installation of a second processor. In order to make the configuration as resilient as possible, the channel connections have been spread across the two storage directors. Each storage subsystem is using four channels, with two channels connected to each of the storage directors. Likewise, two channel paths have been configured between each storage director and processor.

The capacity assumptions made here are that during normal operations the aggregate data transfer workload from both processors will not exceed the capacity of four channels (since there are only four channels running between the storage and the directors), and that in the unlikely event that an entire storage director fails, the remaining two active channels will be sufficient to support the workload until repairs can be completed.

This is a much more robust configuration, as the redundancy helps to provide the necessary levels of performance and reliability. Unfortunately, this environment is much more complex than the configuration described in Figure 1, but the additional complexity is necessary in order to achieve the desired level of resiliency. (This is true even though we have many more individual components that can fail, but fewer components can cause a catastrophic outage).

For the remainder of this article, we will consider that the configuration pictured in Figure 5 is the current architecture of the single-site production environment. This diagram is identical to Figure 4 except that a total of four disk subsystems are shown and the processors have been moved to the left of the directors. It is building upon this resilient framework that channel extension for Business Continuity and Disaster Recovery will be added.

Let us assume that management has selected a recovery site that is approximately 1500 miles away from the production data center and that performance of the production applications can not tolerate any additional processing delays. Because of the distance involved, any synchronous disk mirroring solution would introduce an additional 30ms of delay for each write I/O, therefore only asynchronous disk mirroring methodologies can be considered. (The rule-of-thumb is to add 1ms for each 100 circuit miles. Thus, the round-trip transit time would be 30ms: 2*1500=3000/100=30ms.)

After review of various vendor solutions and the necessary due diligence, management has selected Global Copy for zSeries (formerly called XRC or eXtended Remote Copy) to be implemented.

The next decision to be reached is the selection of the specific channel extension equipment to be used. Again, necessary due diligence is performed and McData's USD-X (UltraNet Storage Director-eXtended) equipment is selected as best matching the performance and availability requirements of the new environment.

Now that these strategic decisions have been made, the most important remaining task prior to designing the channel extension network is to determine the amount of bandwidth that will be required. This information will guide us towards certain specific configuration options as we build the environment.

The first step in determining the amount of bandwidth required is to perform a bandwidth analysis study. For the sake of simplicity, assume the results of the I/O study determined that the workload is very well balanced across the four storage subsystems, and the peak I/O rates require approximately 1.2Gb/second of channel capacity for each storage subsystem. Further, the write I/O workload accounts for 25% of the total I/Os. (This is significant as with any of the advanced recovery methodologies, only the write I/O's need to be mirrored to the recovery site.)

The bandwidth requirement can then be determined by taking the total I/O rates from each subsystem from the peak period(s) and multiplying by the write I/O percentage—in this case 25%. Therefore, the total amount of bandwidth required to support this channel-extended write I/O workload is something approaching 1.2 Gb/second of uncompressed data.

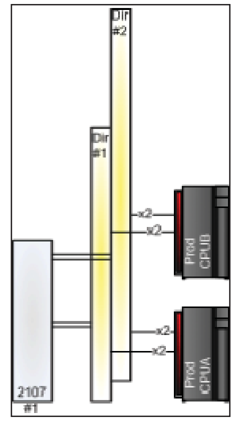

Figure 4: More common configuration with storage directors

It is important to understand that the I/O bandwidth analysis was based on SMF (System Management Facility) data and that many of the usage peaks were "smoothed-out" due to the SMF recording interval. The actual "instantaneous" peaks could have been much higher than reported by the study. Even so, Global Mirror for zSeries will manage these peaks and maintain data consistency so we can expect 1.2 Gb/second of network capacity to be adequate.

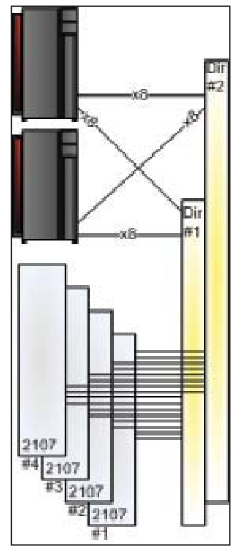

Figure 5: Larger configuration with multiple DS8000's and processors

With this understanding, we can satisfy the network bandwidth requirements with two OC12 circuits. Each OC12 circuit provides 622 Mb/second of capacity. Since the availability of these circuits is essential to disk mirroring and Business Continuity, they should be configured with diverse routes and protected via APS 1+1. (APS 1+1: Automated Protection Switching is a communications method where each circuit is backed-up with a completely redundant circuit. In the event of a fiber-cut or other circuit problem, the communications gear automatically switches to the alternate or protection circuit with no data loss). In this manner, the circuit provider should be able to commit to "near-zero" unplanned outages.

In this high availability configuration, four pair of channel extenders have been configured. Each USD-X is configured with four FICON adapters and one active Gigabit Ethernet adapter. A second Gigabit Ethernet adapter has been added to provide an alternate path for added resiliency should the primary path fail.

In order to provide as much clarity as possible, the channel connections have been color coded as follows: The local channels (from the earlier configurations) are shown in Black. The new channel connections are shown in Red, Blue, Green and Brown. One color is assigned to each pair of channel extenders. The first two pair of USD-Xs (Nodes 10+80 and 20+90) will service the Red and Blue channels and will travel OC12#1. The second two pair of USD-Xs (Nodes 30+A0 and 40+B0) service the Green and Brown channels using OC12#2.

Four additional FICON channels have been implemented between each storage subsystem and the local directors for the channel extension environment. Although the first four channels have sufficient available capacity to support the additional XRC workload, it is important to evaluate the performance as seen by the remote host.

The read I/O's are being issued by the System Data Mover (SDM) at the remote site and must traverse the network. Network latency is based on the circuit distance and is added to the I/O response time. So each successful read takes a minimum of 30ms plus the local response time for a total of 31 or 32ms.

If the fiber connections between the storage and directors were shared, a channel busy condition could be returned in response to the SDM read operation. Remember that when the operating system detects a busy condition on the channel path that the I/O is re-driven on an alternate path. This is not a significant issue for the local processing with 1ms or 2ms response time, but the distance penalty will come into play dramatically when this occurs to the remote host supporting the SDMs. In this example, the normal 31-32ms reads become 64 or 96ms (or even more) depending on the number of times the I/O must be re-driven.

The best design principle here is to eliminate any possible contention that will cause these I/Os to be re-driven. From that viewpoint, all the extended channels should be treated as if they were simple point-to-point channels, even though they go through directors, switches, and perhaps thousands of miles across the network.

For the remainder of this discussion, we will largely ignore the local "Black" channels from the configuration and concentrate on the extended channels.

Starting at the storage subsystems on the left side of Figure 6, each subsystem has four extended channels cabled through the FICON directors and from there to the four channel extension devices. On the network-facing side of the channel extenders, there are two Gigabit Ethernet connections running between the USD-X and the network switches. The protocol running across this interface is Gigabit Ethernet; however, the actual speed of the connection is limited by the bandwidth assigned to it.

Because of these capacity limitations, it is important for the channel extension equipment not to overdrive the available network resources. In a Gigabit Ethernet environment this would lead to packet loss and force the channel extenders to retransmit packets. Various tuning parameters exist within the channel extenders to control the amount of data that can be transmitted across the network-facing connections. In this example, each USD-X would be allowed to transmit about 280 Mb/second.

The formula for computing this is:

1. Each OC12 = 622Mb/second.
2. We allocate half of that capacity (622 / 2 = 311 Mb/second) to each USD-X pair.
3. Approximately 90% of that capacity is available for data payload (311 * 0.9 = 279.9. Round up to 280 Mb/sec).

So the total effective network bandwidth capacity is 1120 Mb/sec. This is a little bit less than the previously stated 1.2 Gb/sec. One might think that there could be insufficient capacity during periods of peak activity and become alarmed at how the environment would operate in various failure scenarios. However, the USD-X compresses the data payload prior to shipping the data across the network. This introduces additional useable capacity into the environment and provides additional resiliency.

The USD-X generally achieves excellent data compression. However, the level of compression that is possible is dependant on the customer data and can vary from moment to moment. Assuming that

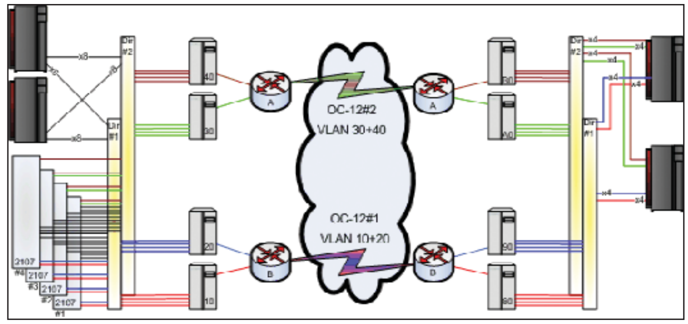


**Figure 6: Multi-site configuration for DR with channel extension and remote XRC recovery**

the mix of customer data is such that the USD-X can achieve a compression ratio somewhere in the range of 2:1 to 3:1 then the extended environment should not only have sufficient capacity for normal processing, but also ample headroom to allow for growth and provide additional resiliency when experiencing failures within the network or of other component.

The Global Mirror for zSeries (XRC) application also provides various facilities that will help to minimize the impact of any component failures and the resulting loss of network bandwidth.

Referring back to Figure 6 and looking to the right of the local channel extenders is a pair of network switches. In addition to providing a connection point for each of the Gigabit Ethernet cables, these devices separate each OC12 circuit into two logical circuits or VLANs (A VLAN is a group of network resources that behave as if they were connected to a single, network segment—even though they may share physical network resources with other components).

In the middle of the diagram is the network "cloud." This may consist of many hundreds of pieces of telecommunications equipment, but luckily, that's for the circuit providers to manage.

The next three sets of components; switches, channel extenders and FICON directors are paired with those same items on the device-side and should not require additional discussion.

The CPU(s) on the far right are located at your recovery center. This is where the host component of Global Mirror for zSeries runs. It is the function of this application to read the changed data from the cache of the disk subsystems at the primary site and continuously perform updates to the secondary disk subsystems (which are not shown in order to reduce the complexity of the diagram) while keeping the mirrored data time consistent across the four mirrored subsystems.

The best practices developed by MVS systems programmers over the years continue to be important today. Just as it is critical to not engineer a single point of failure into an otherwise fault-tolerant local environment, these same design goals are critical when configuring a complex channel extension environment for Disaster Recovery. Although the configuration can become rather complex, it can be manageable with sufficient documentation and understanding.

**Questions or comments? Please e-mail editor@NaSPA.com.**

**NaSPA member Mike Smith has over 35 years of experience in IT with the last 12 years being devoted to Storage and Business Continuity. He has designed and implemented various electronic recovery solutions including bi-directional GDPS/XRC mirroring at a major financial institution on the west coast. For more information please visit recoveryspecialties.com.**