WHITE PAPER

Extension Trunking

Extension Trunking is an advanced feature of the Brocade Extension platforms for both FC and IP extension, enabling bandwidth aggregation and lossless failover for increased resiliency over IP WANs. This paper details the operation and advantages of this technology in extended storage applications.

Table of Contents

Introduction

Over the last decade, extension networks for storage have become commonplace and continue to grow in size and importance. Growth is not limited to new deployments but also involves the expansion of existing deployments. Requirements for data protection will never ease, as the economies of many countries depend on the successful and continued business operations of their enterprises and thus have passed laws mandating data protection. Modern-day dependence on Remote Data Replication (RDR) means there is little tolerance for lapses that leave data vulnerable to loss.

In mainframe, open systems, and IP storage environments, reliable and resilient networks—to the point of no frame loss and in-order frame delivery—is necessary for error-free operation, high performance, and operational ease. This improves availability, reduces operating expenses and, most of all, reduces the risk of data loss. Brocade has been developing advanced extension solutions for more than 15 years and continues to be the thought leader in extension solutions for all storage applications. Brocade addresses today's important trends and requirements with the Brocade® 7840 and Brocade 7800 Extension Switches, as well as the Brocade FX8-24 Extension Blade for the Brocade DCX® 8510 Backbone and DCX Backbone family.

Extension Trunking Overview

Extension Trunking is one of the advanced extension features of the Brocade Extension platforms, providing the following benefits:

- Single logical tunnel comprised of one or more individual circuits
- Efficient use of Virtual E_Ports, known as VE_Ports.
- Aggregation of circuit bandwidth
- Failover/failback
- Failover groups and metrics
- Use of disparate characteristic Wide-Area Network (WAN) paths
- Lossless Link Loss (LLL)
- In-Order Delivery (IOD)
- Nondisruptive link loss
- Deterministic path for protocol acceleration

Extension Trunking, in essence, provides a single logical tunnel comprised of multiple circuits. Terminology can be confusing here. A single circuit is referred to as a tunnel. A tunnel with multiple circuits is referred to as a trunk, simply because multiple circuits are being trunked together. A tunnel or trunk is a single Inter-Switch Link (ISL) and should be treated as such in architectural designs. These extension ISLs can carry Fibre Channel (FC), Fiber Connectivity (FICON), and Internet Protocol (IP) storage traffic using IP extension.

Circuits are individual connections within the trunk, each with its own unique source and destination IP address. On older Brocade Extension platforms that had multiple connections, such as the Brocade 7500 Switch and the Brocade FR4-18i Blade, each connection was its own tunnel, which is no longer the case. Now, with the Brocade 7800 Extension Switch, Brocade FX8-24 Extension Blade, and Brocade 7840 Extension Switch, a group of circuits associated with a single VE_Port forms a single ISL, known as a trunk.

Because a tunnel/trunk is an ISL, each ISL endpoint requires its own VE_Port. Or, if it is an FC router demarcation point for a "remote edge" fabric, then it is called a VEX_Port. Remote edge fabrics are edge fabrics connected through a WAN connection to a VEX_Port. Extension Trunking operates the same, whether the virtual port is a VE_Port or a VEX_Port. If each circuit is its own tunnel, a VE_Port is required for each circuit. However, with Extension Trunking, each circuit is not its own tunnel, hence the term "circuit." Since a trunk is logically a single tunnel, only a single VE or VEX_Port is used, regardless of the fact that more than one circuit may be contained within the tunnel.

Figure 1 shows an example of two tunnels that are trunking four circuits in Tunnel 1 and two circuits in Tunnel 2. Each circuit is assigned a unique IP address by way of virtual IP interfaces (ipif) within the Brocade 7840, 7800, and FX8-24. Those IP interfaces are, in turn, assigned to Ethernet interfaces. In this case, each IP interface is assigned to a different Ethernet interface. This is not required, however. Ethernet interface assignment is flexible, depending on the environment's needs, and assignments can be made as desired. For instance, multiple IP interfaces can be assigned to a single Ethernet interface. There are no subnet restrictions. The circuit flows from local IP interface to remote IP interface through the assigned Ethernet interfaces.

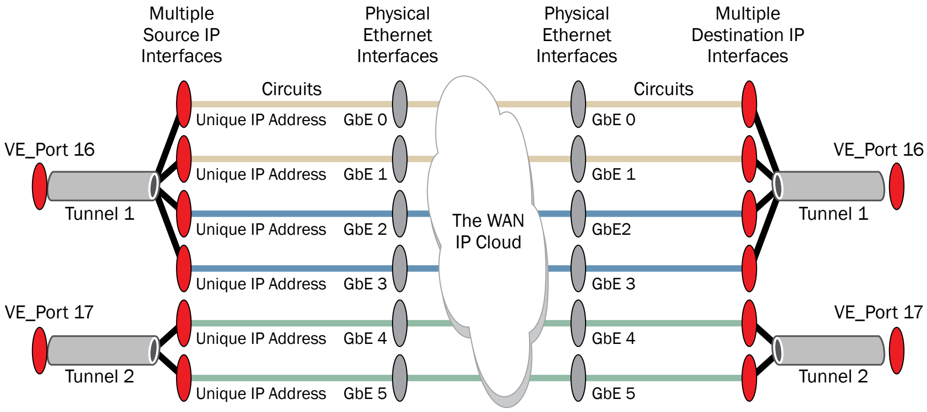

Figure 1: Trunks and their IP/Ethernet interfaces and circuits.

Within the architecture of the Brocade 7840, 7800, and FX8-24, there are Brocade FC Application-Specific Integrated Circuits (ASICs), which know only FC protocol. The VE/VEX_Ports are not part of the ASICs and are logical representations of actual FC ports. In actuality, multiple FC ports feed a VE/VEX_Port, permitting high data rates well above 8 gigabits per second (Gbps) and 16 Gbps, which is necessary for feeding the compression engine at high data rates and for high-speed trunks. On the WAN side, the Brocade 7840 features 10 Gigabit Ethernet (GbE) and 40 GbE interfaces, the Brocade FX8-24 features 10 GbE and 1 GbE interfaces, and the Brocade 7800 has 1 GbE interfaces. Think of VE and VEX_Ports as the transition point from the FC world to the TCP/IP world inside the extension devices.

IP extension switching and processing is done within the Ethernet switch, the Field-Programmable Gate Arrays (FPGAs), and the Data Processors (DPs). IP extension uses VE_Ports as well, even if no FC or FICON traffic is being transported. VE_Ports are logical entities that are more accurately thought of as a tunnel endpoint than as a type of FC port. IP extension uses Software Virtual Interfaces (SVIs) on each DP as a communication port with the IP storage end devices. The SVI becomes the gateway IP interface for data flows headed toward the remote data center. At the SVI, end-device TCP sessions are terminated on the local side and reformed on the remote side. This is referred to as TCP Proxying, and it has the advantage of local acknowledgments and great amounts of acceleration. WAN-Optimized TCP (WO-TCP) is used as the transport between data centers. Which side is "local" or "remote" depends on where the TCP session initiates. Up to eight of the 10 GbE interfaces can be used for IP extension

Local-Area Network (LAN) side (end-device or data center LAN switch) connectivity. Eight 10 GbE interfaces are reserved for WAN side (tunnel) connectivity.

Since a single VE/VEX_Port represents the ISL endpoint of multiple trunked circuits, this affords the Brocade 7840, 7800, and FX8-24 some benefits. First, fewer VE/VEX_Ports are needed, making the remaining virtual ports available for other trunks to different locations. Typically, only one VE/VEX_Port is needed to any one remote location. Second, bandwidth aggregation is achieved by merely adding circuits to the trunk. Each circuit does not have to be configured identically to be added to a trunk; however, there are limitations to the maximum differences between circuits. For example, the maximum bandwidth delta has to fall within certain limits. Circuits can vary in terms of configured committed rate, specifically Adaptive Rate Limiting (ARL) floor and ceiling bandwidth levels. Circuits do not have to take the same or similar paths. The RTT (Round Trip Time) does not have to be the same; the delta between the longest and shortest RTT is not limitless and must fall within supported guidelines.

Best practices accommodate nearly all practical deployments. The resulting RTT for all circuits in the trunk is that of the longest RTT circuit in the group. Bandwidth scheduling is weighted per each circuit, and clean links operate at their full available bandwidth. A common example is a tunnel deployed across a ring architecture, in which one span is a relatively short distance and the opposite span is longer. It is still practical to trunk two circuits, one across each span of the ring. Asynchronous RDR (Remote Data Replication) applications are not affected if both paths take on the latency characteristic of the longest path. Please refer to the Brocade Fabric OS Extension Administrator's Guide for more details pertaining to a specific Brocade Fabric OS® (Brocade FOS) version.

Link failover is fully automated with Extension Trunking. By using metrics and failover groups, active and passive circuits can coexist within the same trunk. Each circuit uses keepalives to reset the keepalive timer. The time interval between keepalives is a configurable setting on the Brocade 7840, 7800, and FX8-24. If the keepalive timer expires, the circuit is deemed to be down and is removed from the trunk. In addition, if the Ethernet interface assigned to one or more circuits loses light, those circuits immediately are considered down. When a circuit goes offline, the egress queue for that circuit is removed from the load-balancing algorithm, and traffic continues across the remaining circuits, although at the reduced bandwidth due to the removal of the offline link.

Keepalives and Circuit Timeouts

1. **How does the keepalive mechanism work on the trunk/tunnel/circuit?**
   A Keepalive Timeout Value (KATOV) is configured for each circuit in a trunk. The Extension Trunking algorithm uses that value to determine how frequently keepalive frames need to be sent. (The math used to calculate that value is not completely straightforward and is beyond the scope of this document.) The algorithm ensures that multiple keepalive frames are sent within the timeout period. If the receiving side does not receive any keepalives within the timeout period, the circuit is brought down.

2. **How are keepalives queued and handled through the network layers?**
   Keepalives are treated like normal data WO-TCP. They are not sent on any special path through the network stack; therefore, they are intermixed with normal data that is passing through a TCP for transmission. This means that the transmission of keepalives is guaranteed through the network by WO-TCP, so it cannot be lost unless a circuit is down. The keepalive process determines true delay in getting a packet through an IP network. This mechanism takes into account latency, reorder, retransmits, and any other network conditions.

   If packets take longer than the allotted amount of time to get through the IP network and exceed the KATOV, the circuit is torn down, and that data is requeued to the remaining circuits. Configuring the KATOV based on the timeout value of the storage application passing through the trunk is important and recommended. Having said this, the default value for FICON is 1 second, which should not be changed. The default value for RDR and tape is 10 seconds and can often be shortened. This change must be evaluated on a case-by-case basis and depends greatly on the characteristics of the IP network. The KATOV should be less than the application level timeout, when a trunk contains multiple circuits. This facilitates Lossless Link Loss (LLL) without the application timing out, ensuring that data traversing the IP network does not time out at the application level. TCP is the preferred transport for the keepalive mechanism, because it allows tight control of the time that a segment is allowed on the WAN.

3. **Can having too much data going over a connection bring down a trunk/tunnel/circuit?** The answer is “yes” and “no.”
   If there are no issues in the IP network, and the tunnel is just not big enough to handle the workload that is driven from the application, then the answer is no. This means there is no congestion; ARL is configured properly, but the bandwidth is simply not adequate for the RDR or tape application. When all the available bandwidth is used, and Buffer-to-Buffer Credit (BBC) flow control is applying back pressure to incoming flows, preventing overflow, the tunnel nevertheless does not go down, due to a keepalive timeout. The time it takes to get a keepalive through is not affected by having large amounts of data queued in the egress scheduler. However, the time is impacted by the WO-TCP data queue, which is limited by the amount of data that can be queued but not yet sent to the network. This amount of data is relatively small (in the range of only a few milliseconds as of Brocade FOS v7.4) and does not use enough time to cause a keepalive timeout due to circuit saturation.

> The answer can also be “yes”: If the circuit is congested because ARL is misconfigured, allowing more data to enter the IP network than there is available bandwidth, and the network cannot handle the data rate, then network errors result. If these errors become severe enough, leading to buffer overruns and causing packet loss and retransmits, then one or more keepalives may be lost, and a timeout may result.

When a connection is lost, data inflight is almost always lost as well. Any frames in the process of being transmitted at the time of the link outage are lost, due to partial transmission. This causes an Out-Of-Order-Frame (OOOF) problem, because some frames have already arrived, then one or more are lost, and frames continue to arrive over the remaining link(s). Now the frames are not in sequence because of the missing frame(s). This is problematic for some devices, particularly mainframes, but there are others in the open systems world, resulting in an Interface Control Check (IFCC) or Small Computer Systems Interface (SCSI) error. For example, frames 1 and 2 are sent and received. Frame 3 is lost. Frame 4 is sent and received. The receiving side detects 1-2-4, which is an OOOF condition, because it was expecting frame 3 but received frame 4.

Extension Trunking resolves this problem with LLL. When a link is lost, inevitably frames inflight are lost as well, and the lost frames are retransmitted by LLL. Refer to Figure 2. Normally, when a TCP segment is lost due to a dirty link, bit error, or congestion, TCP retransmits that segment. In the case of a broken connection (circuit down), there is no way for TCP to retransmit the lost segment, because TCP is no longer operational across the link.

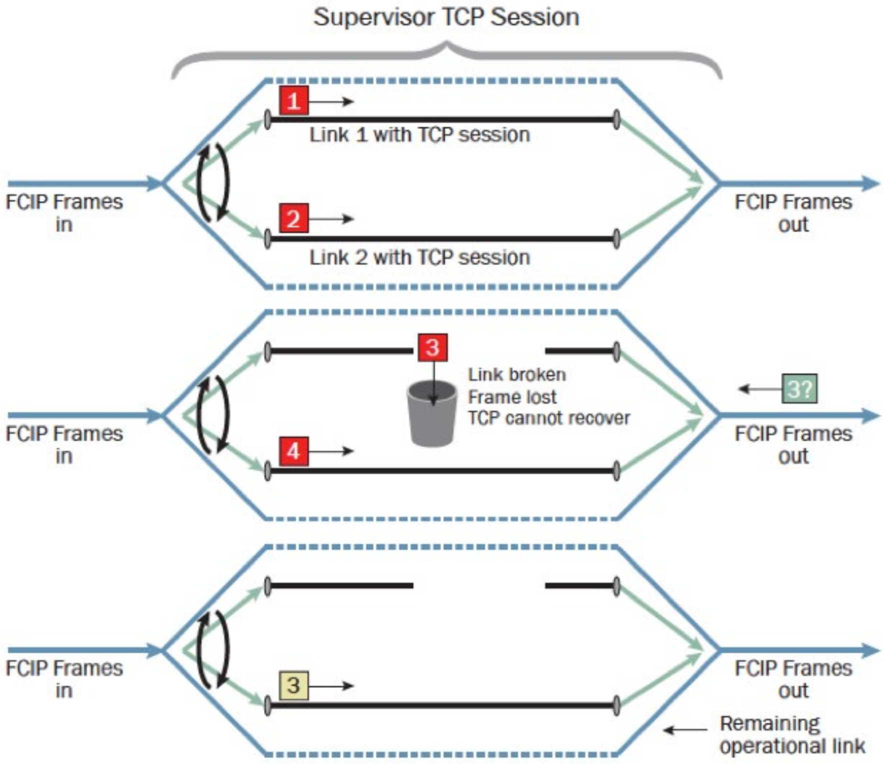

Figure 2. Lossless Link Loss (LLL) process.

Using proven Brocade technology that is often utilized with mainframes, an elegantly simple solution is to encapsulate all the TCP sessions, each session associated with an individual circuit within the trunk. This outer TCP session is referred to as the "Supervisor" TCP session and it feeds each circuit's TCP session through a load balancer and a sophisticated egress queuing/scheduling mechanism that compensates for different link bandwidths, latencies, and congestion events.

The Supervisor TCP is a Brocade proprietary technology that is purpose-designed, and a special function algorithm. It operates at the presentation level (Level 6) of the Open Systems Interconnection (OSI) model and does not affect any LAN or WAN network device that might monitor or provide security at the TCP (Level 4) or lower levels, such as firewalls, ACL, or sFlow (RFC 3176). It works at a level above FC and IP such that it can support both Fibre Channel over IP (FCIP) and IP extension across the same extension trunk.

If a connection goes offline and data continues to be sent over remaining connections, missing frames indicated by noncontiguous sequence numbers in the header trigger an acknowledgment by the Supervisor TCP session back to the Supervisor source to retransmit the missing segment, even though that segment was originally sent by a different link TCP session that is no longer operational. This means that segments that are held in memory for transmission by TCP have to be managed by the Supervisor, in the event that the segment has to be retransmitted over a surviving link TCP session. Holding segments in memory on the Brocade Extension platforms is not an issue, because of the large quantity of memory on these platforms. For instance, the Brocade 7840 Extension switch has 128 GB of memory available to accommodate the large BDPs (Bandwidth Delay Products) associated with multiple 10 Gbps connections over long distances.

Circuits can be active or passive within a trunk and can be configured with metrics. Within a failover group, all online circuits with the lowest metric value are active, for example, a value of 0. Circuits with a value of 1 are not active, and they only become active after all circuits with a value of 0 have gone offline. Refer to Figure 3. This shows an example of one-to-one pairing of metric 1 circuits with metric 0 circuits, so that for every metric 0 circuit that goes offline, it can be losslessly replaced by a metric 1 circuit. The metric 1 circuits have to take different IP network paths, considering that most likely an IP network outage caused the metric 0 circuit to go offline. Metrics and failover groups permit configuration of circuits over paths that should not be used unless the normal production path has gone down, for example, a backup path.

In a three-site triangular architecture, in which normal production traffic takes the primary path that is shortest in distance and latency with one hop, a metric of 0 is set. Sending traffic through the secondary path, which has two hops and typically is longer in distance and latency, is prevented unless the primary path is interrupted. This is done by setting a metric of 1 to those circuits. Nonetheless, the dormant circuits are still members of the trunk. Convergence over to the secondary path is lossless, and there are no out-of-order frames. No mainframe IFCC or SCSI errors are generated. Extension Trunking is required to assign metrics to circuits.

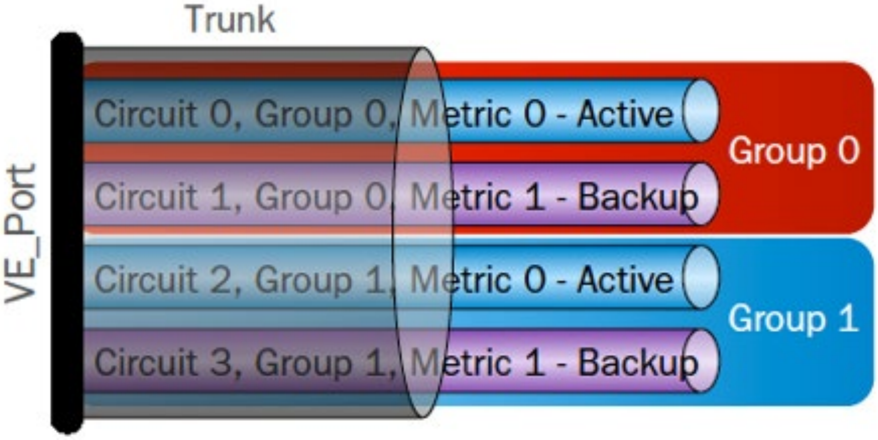
Trunk
VE_Port
Circuit 0, Group 0, Metric 0 - Active
Circuit 1, Group 0, Metric 1 - Backup
Circuit 2, Group 1, Metric 0 - Active
Circuit 3, Group 1, Metric 1 - Backup
Group 0
Group 1

Figure 3. Extension Trunking: Circuit metrics and failover groups.

Extension Trunking provides a single point of termination for multiple circuits. This is important for protocol optimization techniques like FastWrite, Open Systems Tape Pipelining (OSTP), and FICON Acceleration. These protocol optimization techniques are not relevant to IP extension. They are specific to certain FC and FICON applications only. Prior to Extension Trunking, it was not possible to perform protocol acceleration over multiple IP connections, because each connection was an individual tunnel with its own VE/VEX_Port. It was possible for traffic to take different paths outbound versus inbound, creating ambiguity in the network, which prohibited protocol acceleration. Protocol acceleration requires a deterministic path both outbound and inbound. With or without being attached to an edge fabric, Dynamic Path Selection (DPS) across VE or VEX_Ports on the Brocade 7840, 7800, or FX8-24 prevents a bidirectional deterministic path, even when using Port-Based Routing (PBR) or APTpolicy Device-Based Routing (DBR). DBR = PBR + Dynamic Load Sharing (DLS).

Using protocol acceleration, it is necessary to confine traffic to a specific deterministic path bidirectionally. This can be accomplished in a few ways:

- Use only a single physical path, including a single VE_Port pair.
- Use Virtual Fabrics (VFs) with Logical Switches (LSs) that contain a single VE_Port pair.
- Configure Traffic Isolation Zones (TIZs).

Note: All of these methods prevent the use of any type of load balancing and failover between VE_Ports.

The reason that optimized traffic must be bidirectionally isolated to a specific path is because protocol acceleration uses a state machine to perform the optimization. Protocol acceleration needs to know, in the correct order, what happened during a particular exchange. This facilitates proper processing of the various sequences that make up the exchange until the exchange is finished, after which the state machine is discarded. These state machines are created and removed with every exchange that passes over the tunnel/trunk. The Brocade 7840, 7800, and FX8-24 have the capacity for tens of thousands of simultaneous state machines or the equivalent number of flows. The ingress FC frames are verified to be from a data flow that can indeed be optimized. If a data flow cannot be optimized, it is merely passed across the trunk without optimization.

These state machines reside within each VE_Port endpoint. A state machine cannot exist across more than one VE/VEX_Port, and there is no communication between different state machines. As shown in Figure 4, if an exchange starts out traversing tunnel 1 and returns over tunnel 2, with each tunnel using a different VE_Port, the exchange passes through two different state machines, each one knowing nothing about the other. This situation causes FC protocol and SCSI I/O to break, producing error conditions.

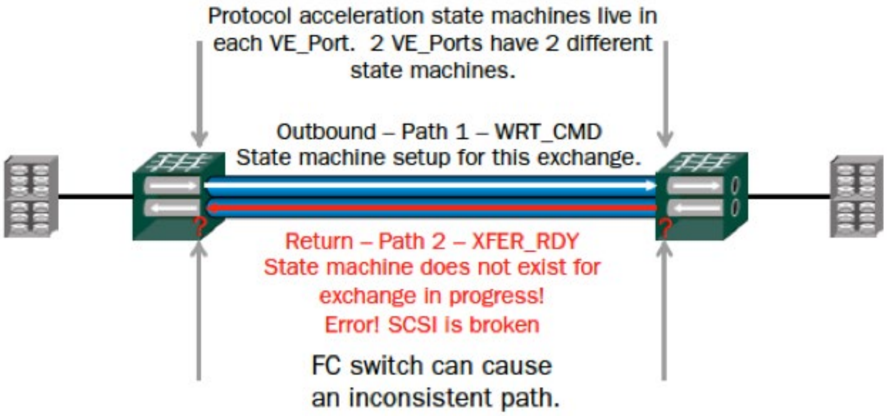
Protocol acceleration state machines live in each VE_Port. 2 VE_Ports have 2 different state machines.
Outbound – Path 1 – WRT_CMD
State machine setup for this exchange.
?
?
Return – Path 2 – XFER_RDY
State machine does not exist for exchange in progress!
Error! SCSI is broken
FC switch can cause an inconsistent path.

Figure 4. SCSI I/O breaks without deterministic path.

An advantage to Extension Trunking is that logically there is only a single ISL tunnel and a single endpoint at the VE or VEX_Port on each side of the trunk, regardless of how many circuits exist within the trunk. The state machines exist at the endpoints of the Supervisor TCP and remain consistent, even if an exchange uses circuit 0 outbound and circuit 1 inbound. Refer to Figure 5. Extension Trunking permits protocol optimization to function across multiple links that are load balanced and can fail over with error-free operation and without any disruption to the optimization process.

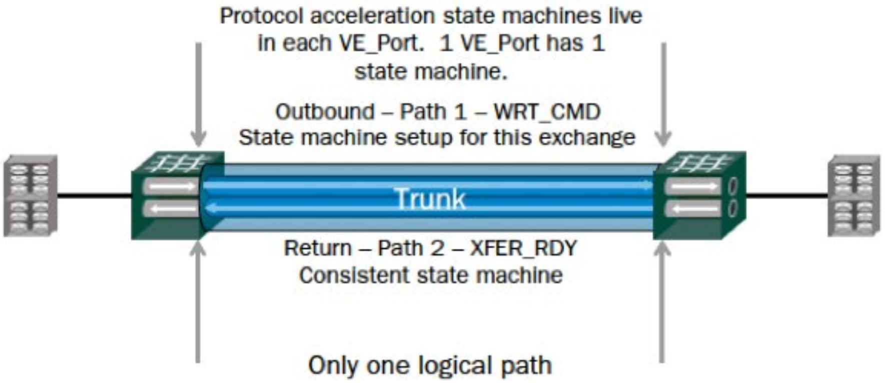
Protocol acceleration state machines live in each VE_Port. 1 VE_Port has 1 state machine.
Outbound – Path 1 – WRT_CMD
State machine setup for this exchange
Trunk
Return – Path 2 – XFER_RDY
Consistent state machine
Only one logical path

Figure 5: Protocol optimization with Extension Trunking as a single logical tunnel.

It is important to remember that if more than one VE/VEX_Port is used between two data centers, the Brocade 7840, 7800, and FX8-24 can route traffic indeterminately to either of the virtual ports, which breaks protocol optimization. In this case, one of the isolation techniques described above is required to prevent failure by confining traffic flows to the same trunk or tunnel. There is one tunnel or trunk per VE_Port or VEX_Port.

A key advantage of Brocade Extension Trunking is that it offers a superior technology implementation for load balancing, failover, in-order delivery, and protocol optimization. FastWrite disk I/O protocol optimization, OSTP, and FICON Acceleration are all supported on Brocade Extension platforms. Brocade protocol acceleration requires no additional hardware or hardware licenses. The same DP that forms tunnels/trunks also performs protocol acceleration for a higher level of integration, greater efficiency, better utilization of resources, lower costs, and a reduced number of assets.

Based on the Brocade FOS APTpolicy setting, one of three routing methods is used to route FC data to VE_Ports when there is more than one port:

- EBR (Exchange-Based Routing, the default): Originator Exchange ID/Source ID/Destination ID (OXID/SID/DID)
- DBR (Device-Based Routing = PBR + DLS): SID/DID
- PBR (Port-Based Routing): Source Port

Either on the same blade or across blades within the same chassis, VE_Ports/VEX_Ports can load share data flows. These can be used in situations in which protocol acceleration is not enabled and the storage application supports it. EBR is not supported in a mainframe (IBM System z) environment; DBR and PBR are supported.

FCIP Batching

The Brocade 7840, 7800, and FX8-24 use batching to improve overall efficiency, maintain full utilization of links, and reduce protocol overhead. Simply put, a batch of frames is formed, after which the batch is compressed and processed as a single unit. This single unit is a “compressed byte stream,” and in this byte stream it is no longer relevant where frames begin and end. Frame boundaries become arbitrary for the purpose of transport across the tunnel. A batch is comprised of up to 16 FC frames

(FCIP) or four FICON frames. All the frames have to be from the same FCP exchange’s DATA_OUT sequence. SCSI commands, transfer readies, and responses are not batched; they are expedited by immediate transmission or protocol acceleration. If the last frame of the sequence arrives, the end-of-sequence bit is set, and the batch is known to be complete and processed immediately, even if fewer than 16 frames have been received. Refer to Figure 6. In this example, two open systems batches are created. One is full with 16 frames, and the other has two FC frames, because of the set end-of-sequence bit.

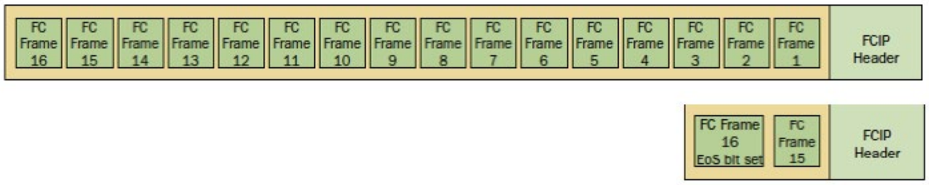

Figure 6: FCIP batch formation: creating two batches.

IP Extension Batching

IP extension batching is a bit different than FCIP batching. IP extension batches are created on a stream basis. 1024 streams are supported on the Brocade 7840, 512 streams per DP. A batch fills until it gets to 32 Kilobytes (KB), and then no more IP datagrams are added. Refer to Figure 7. In this example, the 16th IP datagram either meets or exceeds the 32-KB quantity; therefore, it is the last IP datagram to be added to the batch. If data is arriving at 10 Gbps, it takes about 25 microseconds (µs) to fill a batch. In the event that no more IP datagrams arrive for that stream, there is an up to 2 ms backstop timer that triggers the processing of that batch. Also, if TCP frame is received with a PUSH Flag set, it triggers the processing of that batch. After an IP extension batch is formed, the processing of that batch is identical to that of FCIP batches.

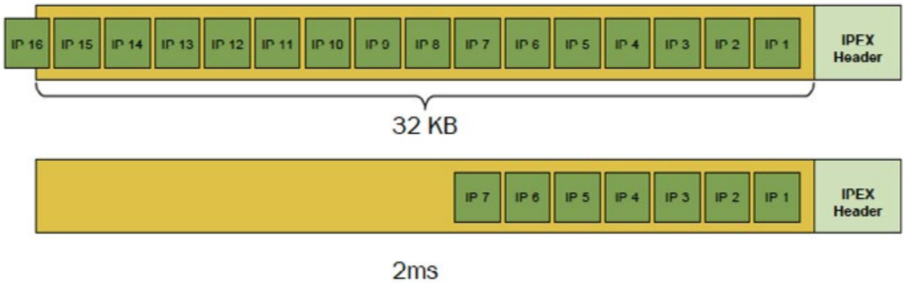

Figure 7: IP extension batch formation.

Batches are then parsed into TCP segments according to the Maximum Segment Size (MSS), which specifies only the TCP payload size without the header. MSS is not configurable and is based on the IP Maximum Transmission Unit (MTU) minus the IP and TCP headers. The Brocade 7800 and FX8-24 support a maximum MTU of 1500 bytes. The Brocade 7840 supports Jumbo Frames with a maximum MTU of 9216 bytes. Depending on the size of the batch, one batch might span multiple TCP segments, or it might fit within a single TCP segment. Each TCP segment is filled to its maximum size. Refer to Figure 8 for the Brocade 7800 and FX8-24 1500-byte MTU, and refer to Figure 9 for the Brocade 7840 9216-byte MTU using Jumbo Frames. This method generates the lowest amount of overhead, which results in superior performance and the highest efficiency achievable.

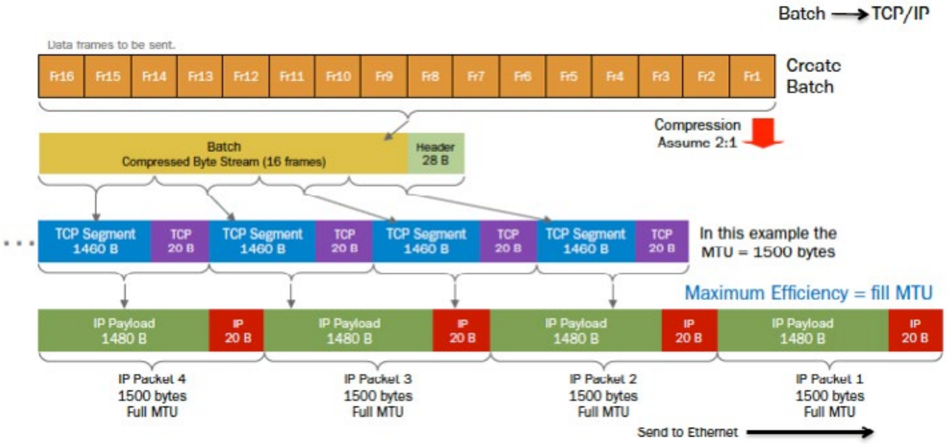

Figure 8: A batch into TCP/IP method (1500-byte MTU).

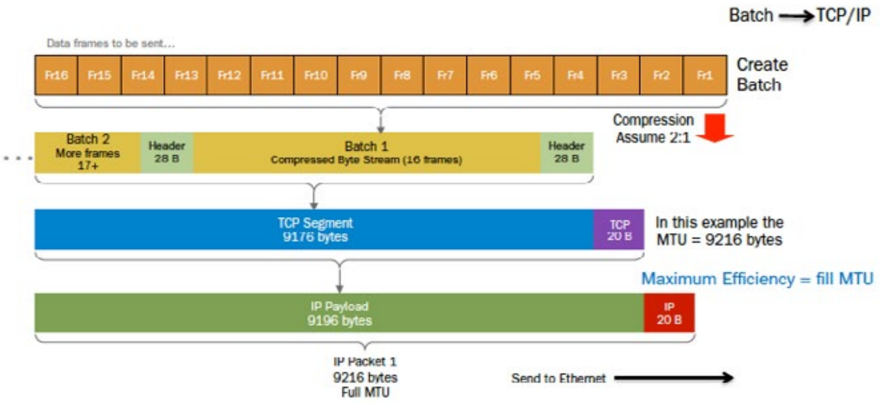

Figure 9. Batches into TCP/IP method (9216 byte MTU).

IP based UDP, UDT, and BUM (Broadcast, Unknown, and Multicast) traffic is assigned to non-TCP streams within WO-TCP. IP extension traffic must be defined in the Traffic Control List (TCL) to pass across the tunnel/trunk, including any UDP, UDT, or BUM traffic. Best practice is to limit communications across the tunnel to strictly the source and destination subnet(s) at each end of the tunnel. This prevents unnecessary and unwanted network utilization. Of course, it is simple to define the TCL as ANY to ANY, but this may have serious negative consequences to IP storage flows. However, a proper TCL configuration optimizes IP storage flows across the WAN.

Because batches are specific to a flow between an initiator and target, and because flows can be prioritized using Quality of Service (QoS), there are separate designations for IP extension (high, medium, low) and FCIP (high, medium, low). Additionally, a percentage is established to apportion the amount of bandwidth assigned to IP extension vs. FCIP during periods of contention. If there is no bandwidth contention, then any flow can use whatever amount of bandwidth it needs within the confines of ARL. Batches are specific to a stream and are fed into the correlating priority's WO-TCP session, based on what has been designated for that traffic in the TCL. This is referred to as Per-Priority TCP QoS (PTQ). PTQ is the only way that QoS can function within an IP network when using TCP. It is not possible for QoS to operate within an IP network if all the priorities are fed into a single TCP session; in fact, this could cause severe performance degradation. In Hybrid mode (IP extension + FCIP), each circuit has at least seven WO-TCP sessions, one for each priority plus one for class-F communications.

On the Brocade 7840 Extension Switch, a new TCP stack named WAN Optimized TCP (WO-TCP) was introduced. One of the primary developments of WO-TCP is IP extension streams. Streams is a technique that prevents collateral effects to other flows by Slow Drain Devices (SDD), which cause Head of Line Blocking (HoLB). TCP uses send (swnd) and receive (rwnd) windows as its flow control mechanism. If an IP storage device on the receiving side needs to reduce an incoming flow, it closes the rwnd. It is deleterious to all flows using TCP if there is only one window for all data being transported. Having only one rwnd means that all flows are slowed or halted, if just one end device asserts flow control. Clearly, this can be a major problem. The remedy is independent flow control for every stream; however, it is not practical to create a separate TCP stack for each flow. There are not enough compute or memory resources on the Brocade 7840 for that to be practical. It is practical to use a single TCP stack that implements virtual TCP windows for each stream. The Brocade 7840 accommodates 512 streams per DP, or 1024 streams per 7840. WO-TCP streams allow one flow to slow or stop while other flows are unaffected.

A "batch" is the unit of load balancing across circuits in a trunk. Batches are placed in the egress queues by the Supervisor TCP session, using a Deficit Weighted Round Robin (DWRR) algorithm; this is referred to as scheduling. The scheduler does take into account egress bandwidth and queuing levels for each of the circuits. When a queue becomes full, usually because of disparate circuit characteristics, the scheduler skips that queue to permit it to drain, while continuing to service other circuits. This maintains full link utilization for different bandwidth circuits and circuits with dissimilar latency. The Supervisor TCP session on the receiving side ensures in-order delivery of any batches that arrive out of order due to circuit disparity and queuing. As mentioned previously, circuits do not have to have the same bandwidth, ARL settings, or latency (RTT). This permits circuits to use a variety of infrastructures such as Dense Wavelength-Division Multiplexing (DWDM), Virtual Private LAN Services (VPLS), Multiprotocol Label Switching (MPLS), and carrier Ethernet.

If the queues for all the circuits become full and can no longer be serviced by the scheduler, the buffers fill, and eventually BBC R_RDYs are withheld from the source to stop data from overflowing the buffers within the Brocade 7840, 7800, and FX8-24. Buffer overflow results in lost FC frames, and Brocade Fibre Channel products rarely drop FC frames. There is a tremendous amount of buffer memory within the Brocade 7840, 7800, and FX8-24 platforms; however, there are advantages to limiting buffering to little more than the bandwidth-delay product (the amount of data that can be inflight) of the circuit. Robust buffers on the Brocade 7840, 7800, and FX8-24 can accommodate multiple long, fat pipes and a very large quantity of simultaneous flows. Additionally, limiting a flow's buffering permits the source device to better know what has or has not been sent, by having as little data as possible outstanding in the network. After queues drain to a particular point, FC frames resume flow from the source, and new batches are produced. All the while, there is no idle transmission of any circuit in the trunk, as none of the queues are ever completely emptied.

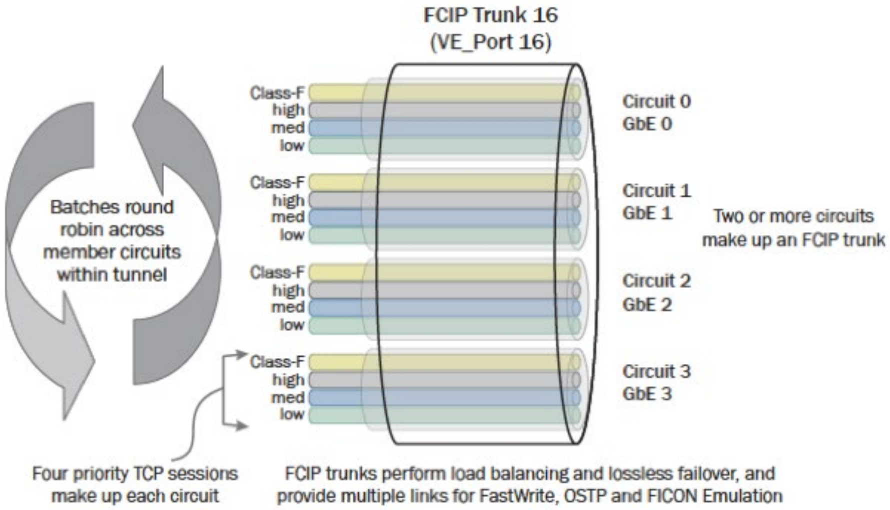
FCIP Trunk 16
(VE_Port 16)
Batches round robin across member circuits within tunnel
Class-F
high
med
low
Class-F
high
med
low
Class-F
high
med
low
Class-F
high
med
low
Circuit 0
GbE 0
Circuit 1
GbE 1
Two or more circuits make up an FCIP trunk
Circuit 2
GbE 2
Circuit 3
GbE 3
Four priority TCP sessions make up each circuit
FCIP trunks perform load balancing and lossless failover, and provide multiple links for FastWrite, OSTP and FICON Emulation

Figure 10: Load balancing batches across trunk circuits.

A trunk has multiple circuits, which raises the question, "How are Fabric Shortest Path First (FSPF) link costs computed with multiple circuits that have variable rate limits?" ARL has a minimum and maximum bandwidth level that is set for each circuit, so how does that affect FSPF costs?

In IP extension, both the control and data planes are separate from the FC/FICON and FCIP control and data planes. All IP extension control is accomplished via the TCL The TCL configuration defines IP extension behavior. Fabric services you normally associate with Fibre Channel traffic are not involved with IP extension. FSPF is just one fabric service that is specific to FCIP and is not relevant to IP extension. IP extension never passes through an FC switching ASIC or fabric.

An FSPF cost is calculated for the whole trunk, which is the same as the bandwidth of the VE_Port or VEX_Port. Circuits are not entities that are recognized by FSPF; therefore, circuits have no FSPF cost associated with them individually. FSPF cost is determined from the sum of the maximum bandwidth levels configured, and only from online circuits that have the lowest metric within the trunk. For example, all online metric 0 circuits have their ceiling ARL value added, and that aggregate value is used in the computation.

FSPF link cost is computed using the following function (refer to Figure 11):

- If the aggregate bandwidth is greater than or equal to 2 Gbps, the cost is 500.
- If the aggregate bandwidth is less than 2 Gbps and greater than 1 Gbps, the cost is 1,000,000 divided by the aggregate bandwidth amount in megabits per second (Mbps), from 500 to 1000.
- If the aggregate bandwidth is less than 1 Gbps, the cost is 2000 minus the aggregate bandwidth amount in Mbps, from 1000 up to 2000.

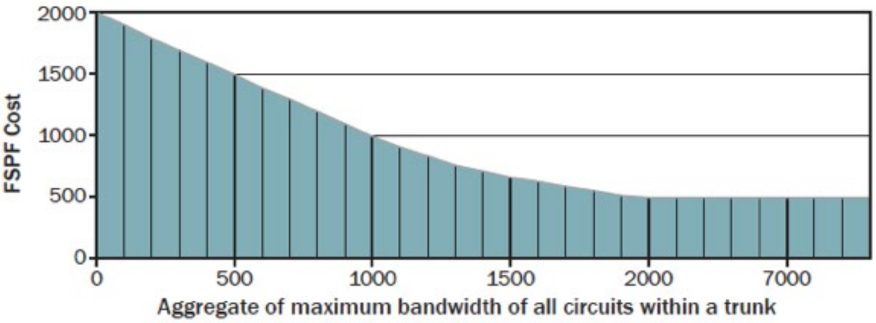

Figure 11: FSPF costs for trunks based on maximum aggregate bandwidth.

Each Brocade Extension platform has a maximum supported bandwidth for the VE_Ports. The bandwidth of a VE_Port is the aggregate of the ARL floor (-b) values from all the member circuits with the same metric. Do not confuse this with the maximum extension bandwidth that the platform can support. The maximum VE_Port bandwidths are as follows:

- Brocade 7800: The maximum VE_Port bandwidth is 6 Gbps.
- Brocade 7840: The maximum VE_Port bandwidth is 20 Gbps.
- Brocade FX8-24: The maximum VE_Port bandwidth is 10 Gbps.

If multiple circuits are assigned to the same Ethernet interface, the aggregate of the minimum bandwidth rates cannot exceed the interface speed. This prevents oversubscribing the interface when guaranteed minimum values are configured. It is possible, however, to configure the aggregate of the maximum bandwidth values beyond the capacity of the physical Ethernet interface. This permits a circuit to use bandwidth that another circuit is not using at that moment. For example, two circuits have their maximum bandwidth level set to full line rate. This is twice the level that the physical interface can handle. If both are demanding bandwidth, then they equalize at 500 Mbps each. If one uses bandwidth only during the day, and one only at night, then each gets the available bandwidth of that interface at that time.

Ethernet Interfaces

Each VE/VEX_Port can have one or more circuits. There are no VEX_Ports on the Brocade 7840. Each Ethernet interface on the Brocade 7840, 7800, and FX8-24 can be used by circuits from multiple VE/VEX_Ports (also known as an extension tunnel/trunk). Each VE/VEX_Port can use multiple Ethernet ports. Each VE/VEX_Port can have multiple IP interfaces (also known as circuits). Each IP interface or circuit in a VE/VEX_Port is associated with an IP address. In turn, those IP addresses are assigned to specific Ethernet interfaces. The types of Ethernet interfaces and the number of interfaces of each type supported on the Brocade 7840, 7800, and FX8-24 are listed in the following table. Table 1 lists the number of VE/VEX_Ports and the maximum bandwidth supported per DP or VE/VEX_Port on different Brocade Extension platforms.

40 GbE interfaces on the Brocade 7840 cannot be used for IP extension LAN connections to data center switches from which local IP storage devices connect.

Table 1. Ethernet Interfaces on the Brocade 7840, Brocade 7800, and Brocade FX8-24

	Brocade 7800	Brocade FX8-24		Brocade 7840	
Data Processor	Data Processor 0	Data Processor 1	Data Processor 0	Data Processor 0	Data Processor 1
GbE Interfaces	6	10	0	16 1/10 GbE interfaces shared by	
10 GbE	N/A	Two 10 GbE interfaces: Circuits can be configured to use either DP.			
40 GbE	N/A	N/A		Two 40 GbE interfaces shared by both DPs	
Maximum Bandwidth per VE/VEX_Port	6 Gbps	10 Gbps	10 Gbps	20 Gbps	20 Gbps
Maximum WAN Bandwidth per DP	2 Gbps for 4/2 6 Gbps for 16/6	10 Gbps	10 Gbps	20 Gbps	20 Gbps
Maximum IP Extension Bandwidth per DP	N/A	N/A	N/A	20/20 Gbps LAN/WAN	20/20 Gbps LAN/WAN
Maximum Number of VE/VEX_Ports	2 for Brocade 7800 4/2 8 for Brocade 7800 16/6	10	10	No VEX_Ports Default 5 per DP at 20 Gbps VE_Port maximum bandwidth 10 per DP at 10 Gbps VE_Port maximum bandwidth	No VEX_Ports Default 5 per DP at 20 Gbps VE_Port maximum bandwidth 10 per DP at 10 Gbps VE_Port maximum bandwidth

Note: 10 GbE interfaces on the Brocade FX8-24 and 40 GbE ports on the Brocade 7840 require an optional license.

For Brocade Extension Trunking, the DP that owns the VE_Port controls all its member circuits. There is no distributed processing, load sharing, or LLL across DPs. Failover between DPs is done at the FC level by the Brocade FC switching ASIC, provided that the configuration permits it. The only shared components may be the Ethernet interfaces.

On the Brocade 7840 and FX8-24, multigigabit circuits can be configured from 1 to 10 Gbps in 1 Gbps increments. The circuit aggregate bandwidth from a DP cannot be more than 20 Gbps on the Brocade 7840 and 10 Gbps on the Brocade FX8-24. If less than an entire gigabit increment is used by ARL, an entire gigabit increment is still consumed internally. For example, if your WAN is an OC-48 (2.5 Gbps) then a 3 Gbps is used internally, however, ARL can be set at any increment of 1 Mbps. For an OC-48, the maximum ceiling (-B) value would be set to 2,500 Mbps. Rate limiting is very granular for optimal WAN utilization.

On an Ethernet interface, not all circuits have to be members of the same trunk. There can be multiple trunks on an Ethernet interface. Consider the extreme example where ten separate 1 Gbps circuits are associated with a 10 GbE interface, each with their own VE/VEX_Port, which creates ten tunnels. These ten tunnels can be used to connect ten different locations. Best practice is to use 1 VE_Port to connect a fabric between two sites. Scaling, redundancy, and failover are facilitated by the member circuits. For example, on a Brocade 7840, two 10 Gbps circuits are created with a metric of 0, and both stem from a VE_Port 24. Circuit 0 is assigned to 10 GbE interface 2, and Circuit 1 is assigned to 10 GbE interface 3. In the IP infrastructure, the circuits take different OC-192 paths across different carrier networks. The two circuits join up again at the remote side. Logically, this is a single 20 Gbps ISL between the two data centers.

The IP network routes the circuits over separate network paths and WAN links based on the destination IP addresses and possibly other L2/L3 header attributes. Ethernet interfaces on the Brocade 7840, 7800, and FX8-24 provide a single convenient connection to the data center LAN for one to many circuits.

There are no specific requirements for data link connectivity to LAN and DWDM devices, other than that those devices should view the Brocade 7840, 7800, and FX8-24 Ethernet ports as if a server Network Interface Card (NIC) is connected and not another Ethernet switch or router. No Ethernet bridging, Spanning Tree, or IP routing is occurring on the Brocade 7840, 7800, and FX8-24 platforms and the Ethernet interfaces are the origination and termination point of TCP flows—the same as most servers.

Other Features and Extension Trunking

Features such as compression, IPsec (IP Security), VLAN tagging (802.1Q), FCR (Fibre Channel Routing), Virtual Fabrics (VF), and QoS all function with Brocade Extension Trunking and without any caveats.

Compression is configured on a per-tunnel/trunk basis. A different compression mode can be selected for each tunnel/trunk based on its bandwidth profile.

IPsec is provided by Brocade at no additional cost. It is included from the entry level Brocade 7800 4/2 all the way up to the Brocade 7840 and Brocade FX8-24 platforms. IPsec encrypts/decrypts in hardware, adds a negligible amount of latency (approximately 5 µs) and runs at full line rate; therefore, there is no performance degradation when using IPsec, even for synchronous applications. Brocade Extension Trunking is fully compatible with IPsec. IPsec costs only the time and effort to configure the feature, making it prudent to enable IPsec in most cases.

VLAN tagging (IEEE 802.1Q) inserts a tag into the Ethernet header and is fully supported by Brocade Extension Trunking. There is no requirement for circuits to participate in VLAN tagging. If tagging is needed, circuits can be individually configured into the same or different VLANs.

The Brocade 7840 is the newest-generation extension product. The Brocade 7840, along with the previous-generation Brocade 7800 and FX8-24 have a focused design for high-performance extension. These products are not specifically FC Routers (FCR). FCR functionality has been designed into the Brocade 8 Gbps and Gen5 (16 Gbps) FC switching ASICs found in most Brocade products that operate at these port speeds, which include the Brocade 7800, 7840, and FX8-24. The requirement for a specialized FCR platform no longer exists, and FCR is easily enabled by applying the Integrated Routing (IR) license.

By enabling FCR on the Brocade 7840, 7800, or Brocade DCX 8510/DCX Backbone chassis in which Brocade FX8-24 blades have been placed, it is easy to build a best-practice Edge-Backbone-Edge architecture. IR enables the VEX_Ports to be configured for Backbone-Remote Edge or Edge-Backbone-Remote Edge architectures. The Brocade 7840 does not support VEX_Ports.

Brocade Virtual Fabrics are useful when implementing Edge-Backbone-Edge with the Brocade FX8-24 blade, as shown in Figure 12. So that EX_Port connections can be made to an edge Logical Switch from the backbone Base Switch, it is not necessary to purchase a separate Brocade switch or Brocade DCX 8510 to install the Brocade FX8-24 blades into. Instead, it is more cost-effective to put the Brocade FX8-24 blade directly into a Brocade DCX 8510 or DCX that is part of the edge fabric and create a backbone from the Base Switch. The VE_Ports on the Brocade FX8-24 Extension blade and EX_Ports are members of the Base Switch. All other device connection ports and E_Ports that participate in the edge fabric belong to the Fab-A Logical Switch.

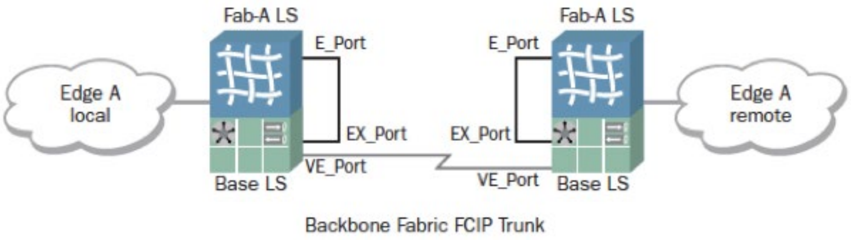

Figure 12: Using Virtual Fabric Logical Switches to implement Edge-Backbone-Edge.

The Brocade 7840, 7800, and FX8-24 have various QoS functions, including Differentiated Services Code Point (DSCP), 802.1P (L2 Class of Service [CoS]), and PTQ. DSCP and L2 CoS perform marking of IP packets and VLAN tags, respectively, and are fully compatible with Brocade Extension Trunking.

PTQ is a special QoS function that is exclusive to Brocade Extension and that was developed especially for high-performance tunnels/trunks. For QoS to function properly across an IP network, it is essential that each priority have its own flow. This is not possible using a single TCP session, because there would be only a single merged flow and a single flow-control. PTQ provides a TCP session for each priority flow. There are seven flows within each circuit: FC and FICON (class F, high, medium, and low) and IP extension (high, medium, and low). The IP network can manage the flows based on their QoS settings. Brocade Extension Trunking is fully compatible with PTQ.

Architectures

High Availability Design

A popular design is the four Brocade 7840 HA (High Availability) architecture, as shown in Figure 13. This design can easily accommodate two 10 Gbps WAN connections; however, in practice many enterprises use only a single WAN connection. The number of WAN connections is a requirement that is determined on a case-by-case basis and that is mostly predicated on cost versus risk.

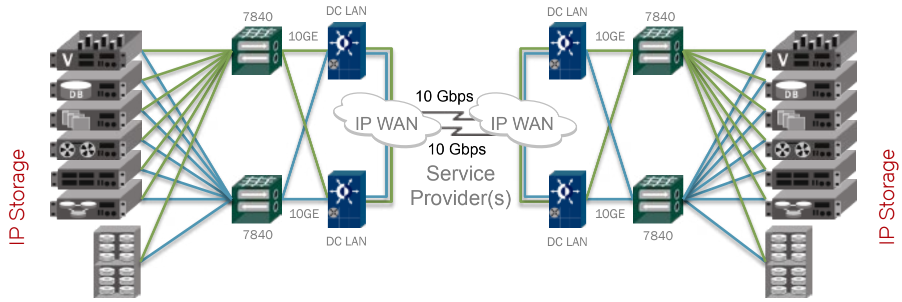

Figure 13: Four Brocade 7840 High Availability Architecture with FCIP & IP extension.

The Brocade 7840 has bandwidth licensing tiers with 5 Gbps as the base product tier. If future growth dictates, or applications such as tape are added, the Brocade 7840 can be upgraded by way of a software license to accommodate larger scales up to 10 Gbps and 40 Gbps tiers.

The four Brocade 7840 HA architecture uses a single trunk that takes advantage of a two-path (dual core) IP network. This provides redundancy for each storage controller on the IP network side. The extension trunks are equivalent to a single FC connection between array controller pairs at each site. As shown in the diagram, there are two sites, and each site has two Brocade 7840s, one attached to controller “A” and one attached to controller “B”. There are two extension trunks, one on each Brocade 7840, and each trunk has two circuits. Each circuit has its own dedicated Ethernet interface. The A controllers use the green circuit trunk, and the B controllers use the blue circuit trunk, as shown in Figure 13. The circuits are forwarded by the data center LAN towards specific WAN connections.

Adaptive Rate Limiting (ARL) is necessary in this architecture, since a total of four 10 Gbps Ethernet interfaces from two Brocade 7840s vie for the same bandwidth across the WAN. Please refer to the Brocade Adaptive Rate Limiting technical brief for a detailed explanation of ARL uses and operation.

This design has a capacity, from the point of view of the storage array, of twice the bandwidth of the WAN. This assumes 2:1 compression is achievable using the Brocade 7840 Fast Deflate algorithm. Other compression algorithms get higher compression ratios; however, they would not accommodate the 20 Gbps WAN bandwidth in this example. The applicable compression mode depends on the available WAN bandwidth. The compression ratio obtainable is specific to the actual data that is being compressed. Brocade makes no promises, guarantees, or claims as to the achievable compression ratio for customer-specific data.

If data security is needed, IPsec is provided. Best practice is to use Brocade IPsec between extension platforms. Typically, devices such as firewalls are not capable of maintaining throughput at the demanding rates of storage replication and become a bottleneck. Firewalls are not permitted to alter the TCP stream in any way, because that is not supported.

WAN optimization products are not supported or needed with Brocade Extension. Generally, WAN optimization products provide little benefit to Brocade Extension, which is already highly optimized. WO-TCP is not improved upon by using WAN optimization devices. And, the nature of most data traversing storage extension does not lend itself to effective deduplication, for example, transactional data. The Aggressive Deflate algorithm found on the Brocade 7840 is capable of nearly the same data reduction across the WAN as the WAN optimization products on the market today. If WAN optimization currently exists in the IP infrastructure, best practice is to either configure those devices to bypass extension traffic or to shunt extension traffic past the WAN optimization devices.

Site-to-Multisite Architecture

Figure 14 illustrates a site-to-multisite architecture. This example shows only the A side of the SAN (Storage Area Network); there is a B side mirror. The bandwidths across the WAN must be doubled to account for both A and B fabrics, as shown in Table 2. This architecture uses Brocade Extension Trunking that originates from the main data center and terminates in one of three remote sites (A, B, and C). Both disk RDR and tape flows are traversing the network, and a significant amount of bandwidth is required for this purpose. Extension Trunking is used to aggregate bandwidth and provides failover between multiple physical connections to the LAN.

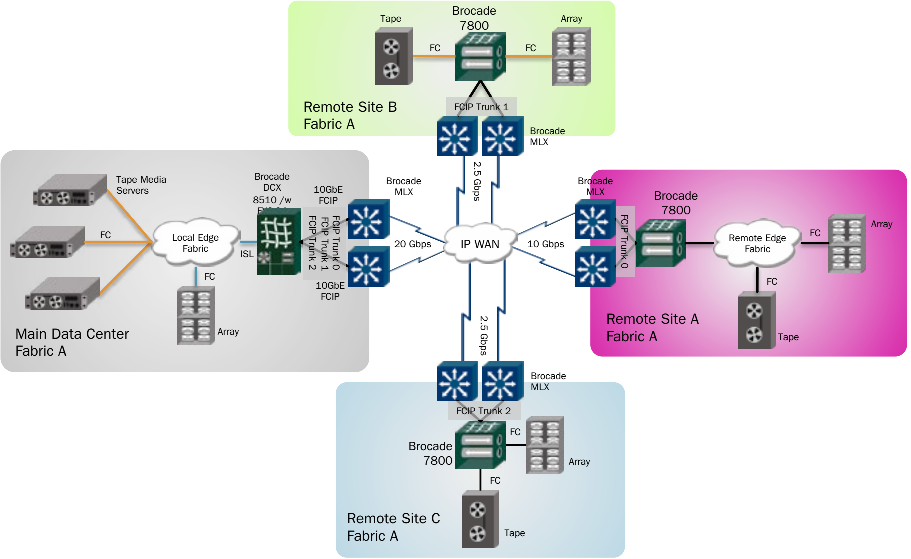

Figure 14: Site-to-multisite design using Extension Trunking, fabric A only.

Table 2. Failure Impact on Bandwidth Availability: Site-to-Multisite Dual Fabric Design Using Extension Trunking and ARL

	Normal Operations					Failure Analysis			
	Trunk Bandwidth			ARL (min) Mbps	ARL (max) Mbps	Site A Optic Failure	Site B/C Optic Failure	Site A Optic Failure	Site B/C Optic Failure
	Site A	Site B	Site C						
Fabric A									
DP 0	5	DP 0 is dedicated to Site A							
XGE0 circuit	3x.833			833	1000	0		0	
XGE1 circuit	3x.833			833	1000	3		0	
DP 1		1.25	1.25						
XGE0 circuit		1 x .625	1 x .625	625	1000		0		0
XGE1 circuit		1 x .625	1 x .625	625	1000		0.83		0
Fabric B									
DP 0	5	DP 0 is dedicated to Site A							
XGE0 circuit	3x.833			833	1000	3		3	
XGE1 circuit	3x.833			833	1000	3		3	
DP 1		1.25	1.25						
XGE0 circuit		1 x .625	1 x .625	625	1000		0.83		1
XGE1 circuit		1 x .625	1 x .625	625	1000		0.83		1
Normal Operation and Failure Analysis									
Total (Gbps)	10	2.5	2.5			9	.25	6	2
WAN Link BW	10	2.5	2.5			10	2.5	10	2.5
BW Delta (Failure)						-1 Gbps	0	-4 Gbps	-5 Gbps
% BW Available						90%	100%	60%	80%

The main data center's WAN is dedicated to storage applications and accommodates more than 15 Gbps of bandwidth (BW). Each remote data center has ample bandwidth as well at 10 Gbps for A and 2.488 Gbps for B and C. There is only a single IP connection designated for storage at the remote sites, due to a cost versus risk determination. This is a common practice at most enterprises.

The 10 GbE interfaces (referred to as XGE interfaces) at the main data center contain 5 circuits each. Three 1 Gbps circuits go to Site A. Brocade FX8-24 10 GbE interfaces can be configured with multigigabit circuits, however, that is not supported when the receiving side is a Brocade 7800, which has only 1 Gbps interfaces. Multiple 1 Gbps circuits are used to connect Brocade 7800s using Brocade Extension Trunking, logically combining the 1 Gbps circuits into a single connection. The other two circuits are 1 Gbps: One goes to Site B and one goes to Site C.

Site A uses Trunk 0 on DP0 and has six 1 Gbps circuits. Remember that this refers to only fabric "A," and there is a mirror of this on fabric "B," both of which feed a single WAN connection. There is no mirror of the WAN, only the SAN. To Site A, 10 Gbps is the maximum that can be passed over the WAN, because it is an OC-192. On the Brocade FX8-24, three of the trunk's circuits are assigned to interface XGE0 and three to interface XGE1 for redundancy. Since optics are the most common failure in this type of system, it is prudent to have interface-level redundancy.

ARL is implemented in this design. If there is a failure or maintenance of a 10 GbE interface, an optic, a cable, a remote Brocade 7800, or an entire Brocade FX8-24 blade, then the other circuits passing over the WAN increase their output to make up for as much lost bandwidth as possible. If there is no failure on fabric "B" and only one of the two XGE interfaces on fabric A are offline, ARL adjusts the bandwidth up to 9 Gbps out of the available 10 Gbps. The failure condition is near-perfect, providing 90 percent of the bandwidth for sustained operations. If an entire Brocade FX8-24 blade fails or the Site A Brocade 7800 fails, 60 percent of the bandwidth is maintained. During normal operations with no failures, ARL keeps each circuit at an aggregate 5 Gbps on each fabric (A and B), consuming all the bandwidth. This is a rate limit of about 833 Mbps, which is automatically set by ARL during such conditions.

Site B uses Trunk 1 on DP1 and has two 1 Gbps circuits using 625 Mbps from each, for a total of 1.25 Gbps (½ OC-48 for fabric A to site B). One circuit is assigned to interface XGE0 and one is assigned to interface XGE1. If there is no requirement for OSTP, because there are only 2 circuits, the remote site can deploy a Brocade 7800 4/2. In this example, there is tape, and OSTP is required.

ARL is implemented in this design. If there is a failure or maintenance of an XGE interface, an optic, a cable, a Brocade FX8-24 blade, or a remote Brocade 7800, then rate limiting automatically adjusts to provide as much bandwidth as possible. If one XGE interface or optic goes offline, 3 Gbps of interface bandwidth remains: 2 Gbps on one fabric and 1 Gbps on the other. This provides enough bandwidth for full utilization of an OC-48. If a Brocade FX8-24 blade or Brocade 7800 goes offline on either fabric A or B, 2 Gbps remains, reducing the operating environment down by only 0.5 Gbps and maintaining 80 percent of operational bandwidth.

Site C uses extension Trunk 2 on DP1 and is the same scenario as Trunk 1 above.

Note that it is not absolutely necessary to use the XGE interfaces in this example. The ten 1 GbE interfaces on the Brocade FX8-24 blade could be used instead. Also, it is not required to use both DPs, as a single complex can process enough data alone to accommodate the A side of this architecture, which is 7.5 Gbps. However, with failure scenarios, the use of two DPs provides a more robust solution. The XGE interfaces do provide for consolidated connectivity within the data center. If the remote sites use Brocade FX8-24 blades instead of the Brocade 7800 switch, the use of XGE interfaces with multigigabit circuits becomes very compelling and best practice.

The way the circuits are routed in the IP network is key to the redundancy and resiliency of this architecture. There are two core switches or routers, each connected to an XGE interface on the Brocade FX8-24 blade. Remember that there is a mirror B side that connects to these same core Ethernet switches or routers.

The IP network routes the circuits to their destination using traditional methods. Nearly all types of WAN architectures are supported. The only architecture not supported is one that uses Per-Packet Load Balancing (PPLB) within the IP network. PPLB tends to work against TCP in most cases, by causing excessive and chronic Out-Of-Order-Segments (OOOS), which leads to retransmits, delay of data to ULPs (Upper-Layer Protocols), higher response times, excessive CPU utilization, increased overall bandwidth utilization, and lower throughput. Flow-based load balancing in the IP network is fully supported and does not cause these issues.

Although not applicable to this specific example, Brocade Fabric OS normally performs EBR (Exchange-Based Routing) internally between the two VE_Ports on each DP within the Brocade FX8-24 Extension blade. These DPs are engines that process FC into FCIP and perform all the transport functions. EBR operates automatically between E_Ports, EX_Ports, VE_Ports, and VEX_Ports and is the default APTpolicy. Considering extension trunks, when there are equal-cost FSPF paths, data load shares across those VE_Ports. EBR is an exchange-based FC routing technique (do not confuse it with FCR), and it directs traffic between DPs based on a SID/DID/OXID (Source ID, Destination ID, and Exchange ID) hash. If one of the XGE interfaces, optics, cable, or the IP path were to fail; EBR would failover traffic to another equal-cost path. EBR is not used in FICON applications, and the APTpolicy is set to either DBR or PBR. Plus, if protocol acceleration is used, some method of traffic isolation is required.

In this use case, EBR would not be used, because there are not multiple equal-cost paths within fabric A at site A, or within fabric A to site B, and so on. There is just one path: one VE_Port making one logical ISL. Redundancy is built into the overall infrastructure of A plus B fabrics. Storage applications have the ability to load balance data across multiple paths. This means that the storage applications divide the workload evenly across both A and B fabrics and provide failover and failback.

The Brocade 7800 switch and FX8-24 blade do not support IP extension, and there are no IP storage flows in this example.

Ring Architectures

Rings are another common architecture used for RDR and tape, due to the proliferation of DWDM infrastructure in many countries. There are many different kinds of offerings from service providers involving rings, for example, customers can buy service across only one span of the ring, permitting the service provider to sell the other span to a different customer. Or both spans can be commissioned at a higher cost for the purpose of aggregating bandwidth or driving higher availability. At a reduced cost, only one span may be active, leaving the other span passive until it is needed for failover. Optionally, at additional cost, both spans may carry the same data. If one span fails, the other span continues with only a slight disruption. This is referred to as “protected.” The current example discusses a ring in which both spans are active/active.

As shown in Figure 15, there is a single extension trunk with two circuits. Each circuit has been assigned to its own dedicated 10 Gbps Ethernet interface, and the trunk is capable of 20 Gbps of aggregated bandwidth. The Brocade 7840 with Extension Trunking and ARL is an appropriate solution for this. Traffic is load balanced across the two circuits on a per-batch basis. Of course, using ARL, the bandwidth can be rate limited to a smaller amount: as low as 20 Mbps per circuit. The amount of bandwidth on each circuit does not have to be the same. The ARL configuration on both ends of a circuit does have to be the same.

The Ethernet interfaces connect directly into the DWDM equipment provided by the service provider. This piece of equipment is often referred to as Customer Premise Equipment (CPE). The service provider must provision a 10 Gbps circuit across the blue span and a 10 Gbps circuit across the red span. Sometimes a full 10 Gbps circuit may not be practical, and less bandwidth must be considered. The operation is the same but with less bandwidth. In the event of a fiber cut, which brings down one of the ring’s spans, Extension Trunking fails all traffic over to the remaining circuit, recovers any frames lost in transit. This means that from the perspective of the storage array or mainframe, all frames arrive in-order. When the ring is repaired and the span comes online again, Brocade Extension Trunking automatically adds the circuit back into the trunk and resumes transmitting and load balancing data. The keepalive mechanism is persistent and detects when the span comes online again.

Using metrics, it is possible to configure this exact architecture such that only one span of the ring is active at a time. If the active span fails, the alternate span starts carrying the data. When that span comes back online, the original span solely carries the data again.

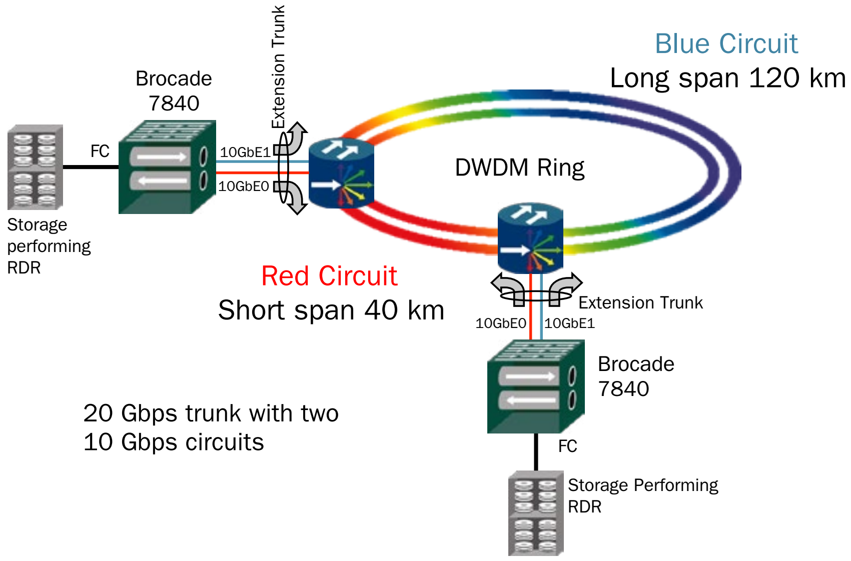

Figure 15: Ring architecture utilizing Brocade Extension Trunking over dissimilar circuits.

Three-Site Triangle with Failover Metrics

Many large enterprises, especially financial institutions involved with frequent and sizable transactions, have requirements for both synchronous RDR (RDR/S) and asynchronous RDR (RDR/A). Why use both? RDR/S replicates every transaction in real time but is limited in distance. RDR/A cannot replicate every I/O, but it can go much farther distances. There are many issues to consider when using this type of architecture.

RDR/S is required to confidently capture all writes to disk, except for the write that is in progress as the disaster disables the server/cluster performing the current write. Other than the last write in progress, every acknowledged write to disk is safely replicated to a remote location. A relatively local (metro area) remote location for backup operations is often referred to as a “bunker site.” There are many issues to consider with this type of architecture also.

Due to the limitations of the speed of light, RDR/S has a limited distance before it causes poor response times for user applications. 100 km (62 mi.) is about the average maximum distance; however, the distance actually depends on a number of factors not discussed in this paper. RDR/S provides a significantly better RPO (Recovery Point Objective), which may represent a significant amount of transaction revenue.

Considering the possibilities and types of disasters that can occur in the world today, it is prudent for large enterprises to replicate data to a third data center that is assured of being outside the perimeter of the event. To extend data beyond the bunker site, it is necessary to use RDR/A. Most of the major storage suppliers offer software that manages both RDR/S and RDR/A, so that if any one connection or site fails, replication continues without the need to reinitialize any of the volume pair relationships (which can take a considerable amount of time).

As shown in Figure 16, this example uses the Brocade 7840 for the RDR/S, which is done across the metro DWDM network connected via Brocade switches to provide the needed BBCs. These mission-critical architectures also deploy A and B RDR fabrics for redundancy by way of completely physically isolated networks.

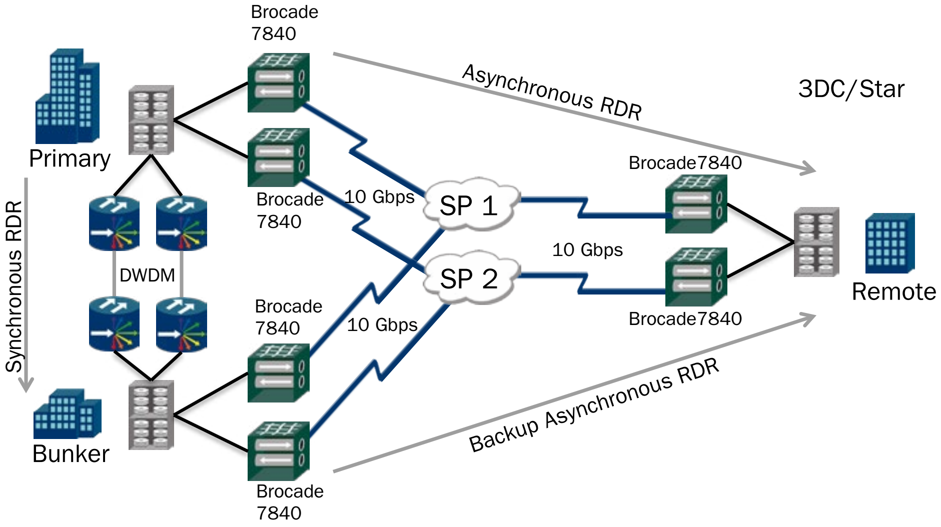

Figure 16: Three data center RDR architecture.

The infrastructure used to transport RDR/S and RDR/A also differs. RDR/S connections are short, and many service providers can offer DWDM for such distances at a reasonable price. However, it is often more expensive for native FC or FICON compared to IP. RDR/A can operate with much less bandwidth relative to RDR/S. RDR/A requires an average demand to be provisioned, and RDR/S requires peak demand to be provisioned. RDR/A needs only the high average over a finite period of time. The period of time cannot be so long that the average is artificially low—due to respites during low workload periods. On the other hand, it cannot be so short that the average approximates the actual peak, like in RDR/S.

Another consideration when determining the adequate bandwidth for RDR/A is the capability of the storage array. Some storage arrays can journal data that is waiting to be sent across the link; however, this should be the exception, not the rule. More importantly, elongation of cycle times should be considered. The amount of data that has to be sent during each cycle can be considered a gauge of the average. If the available WAN bandwidth is below this average, cycle times tend to elongate. If the cycle times tend to always finish on time, and the link utilization is low, the average may have been overestimated, leaving ample room for growth. These issues must be evaluated on a case-by-case basis.

Design questions about RDR/A traffic might include this one: Should RDR/A traffic between primary and remote sites be rerouted across the DWDM link to the bunker site and then forwarded to the remote site to maintain primary-to-remote site replication? Or should rerouting be prevented, allowing failover and letting the bunker site manage replication between the remote site? This depends on a few factors, as follows.

First, is enough bandwidth available on the DWDM network to fail over the RDR/A traffic onto it and maintain peak RDR/S demand? Second, what is best practice for the RDR application software? Best practice is often to prevent the RDR/A flows from being rerouted to the bunker site and then to the remote site, unless allocated bandwidth is available that is specific to this purpose. This is because all too often encroachment onto the RDR/S bandwidth is not workable.

RDR/A and RDR/S traffic are located on different networks, therefore, nothing has to be done to prevent RDR/A traffic from being rerouted through DWDM. If RDR/A needs to be rerouted across the DWDM network to maintain replication with the remote side, then connectivity to an IP router is needed to facilitate that failover.

Ping-ponging happens when a WAN connection to a site goes down, and traffic ping-pongs to other remote sites before finding its way to the destination. In Figure 17, one of the links (A→C) goes down; however, the FC routing tables in the Brocade 7840, 7800, and FX8-24 can still reach the final destination by ping-ponging the traffic back and forth across the WAN (A→D→B→C). The FSPF cost is greater, but it is still viable, as it is the only path. Depending on the WAN links, this may introduce considerable latency, since a single RTT now becomes 3xRTT. Additionally, it causes undesired WAN utilization effects and breaks protocol acceleration by performing optimization on exchanges that are already being optimized. Repeated protocol acceleration is not supported. Ultimately, this is not a best-practice architecture. The best-practice architecture uses Virtual Fabric Logical Switches (VF LSs) to create isolated paths that do not permit rerouting when a WAN connection fails. Alternatively, this can be fixed using Traffic Isolation Zones (TIZs) with failover disabled, but this method has the drawback of involving more complex operations.

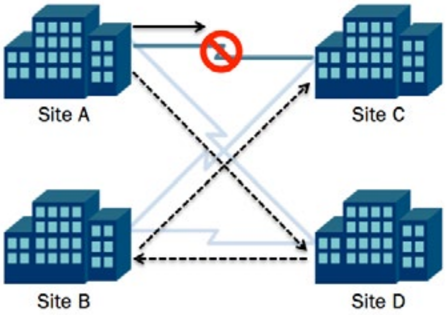

Figure 17: Ping-Ponging

FICON Dual 40 GbE Links

Mainframe environments are often high-bandwidth environments that can take advantage of the Brocade 40 GbE FCIP interfaces on the Brocade 7840 Extension Switch. In this design, the VF LSs at each site are connected together using one 20 Gbps trunk for each Logical Fabric across two redundant 40 GbE interfaces (see Figure 18). The “Red” Logical Fabric is dedicated to IBM XRC (zGM), and the “Green” Logical Fabric is dedicated to FICON tape or RDR/OSTP. An additional LS can be created to accommodate both FICON Tape and RDR/OSTP if needed. One Brocade 7840 has a total of 40 Gbps of WAN side capacity, 20 Gbps per DP. 20 Gbps of bandwidth is associated with the Green Logical Fabric and 20 Gbps is associated with the Red Logical Fabric.

Each 20 Gbps trunk is defined by a single VE_Port, allowing protocol acceleration to function properly. Each VE_Port lives within a Logical Switch. Data flows cannot cross LS boundaries, confining them to a deterministic path. Each trunk is comprised of two 10 Gbps circuits that take different WAN and IP network (switches and routers) paths. The circuits from the Logical Switches share the 40 GbE interfaces. The 40 GbE interfaces stay in the Default Logical Switch and have the ability to accommodate circuits from other Logical Switches. This is referred to as Ethernet Sharing.

It is possible for either the IP network to fail over the data flows upon a WAN link outage, or for the IP network to remain static and let Extension Trunking perform the failover/failback. Practical experience has demonstrated that a static IP network is more available than one that tries to reconverge. Letting Extension Trunking manage the failover and failback is a more reliable architecture. Brocade Extension Trunking is lossless and guarantees in-order delivery. Failover groups and circuits with metric 1 have been configured on the opposite 40 GbE interfaces and IP network paths to maintain usable application bandwidth in the event that one path goes offline and the other path remains online. The 40 GbE interfaces on the Brocade 7840 can accommodate this, because during normal operation those interfaces are operating at a maximum of 20 Gbps, and during failover they can operate at 40 Gbps. Each 10 Gbps circuit member of the trunk is assigned to each of the two 40 GbE interfaces on the Brocade 7840. There are two 10 Gbps circuits on each 40 GbE interface, one from each FCIP trunk (Red and Green). If one of the WAN paths or a data center switch/router goes offline, the circuits across the online path maintain connectivity to both FICON tape and XRC. In addition, the metric 1 circuits come online and maintain the same bandwidth as during normal operation.

Once the circuits enter the core L2/L3 switch, because each circuit has its own source and destination IP address pair, the individual circuits can be routed across the different WAN connections as necessary. There are three popular methods for statically confining data flows to a path, even in the event that the path fails:

- Use VLANs with no router interface: Data flows are forced towards a specific interface.
- Use two narrow static routes: The first route has default administrative distance for directing trunk subnets to the intended interface for the specific WAN path. The second route is an identical route with a larger administrative distance pointing to the NULL interface. If the first route is down, the second route ensures that the flows are not sent to a wider route or a default route.
- Use Policy-Based Routing.

If VLAN tagging is chosen as the method for sorting data flows to specific WAN connections, and various circuits use the same physical Ethernet interface, it is necessary to identify the circuits' frames using VLAN tagging (802.1Q). Brocade Extension products can tag the Ethernet frames from each circuit, so that upon entering the data center's Ethernet switch, the frame is directed to the correct VLAN for the WAN connection. QoS (802.1P and/or DSCP) can also be marked on these frames. The VLAN's path is determined and configured by the IP network administrators.

The FICON frames that are trunked across multiple WAN connections, as in this example, are always delivered to the ULP in the proper order in which they were sent. This is a function of the Supervisor TCP session that manages the data across all the circuits in a trunk. This is a requirement in mainframe environments and a unique Brocade technology.

FICON applications require lossless failover to prevent any frame loss. When one or more frames are lost, and traffic resumes immediately after the failover or rerouting of data, this causes what appears to be Out Of Order Frames (OOOFs). OOOF results in an IFCC on mainframes, which is a significant event. To prevent IFCCs, it is better to completely halt the traffic, rather than rerouting and keeping flows moving, which is not entirely intuitive. Because of this, TIZs are set to not fail over when a path is no longer available. VF LSs facilitate deterministic paths with no failover. In this example, both links have to go down before traffic completely stops.

Brocade FICON Acceleration supports both XRC and tape across the same circuits simultaneously. There is no requirement to separate the data flows; however, it is the practice of many mainframe administrators to keep the data flows separate. In fact, if desired, simultaneous FICON and Open Systems data flows across the same VE_Port are fully supported by Brocade and IBM. Special consideration must be given in such blended architectures, but that discussion is beyond the scope of this paper.

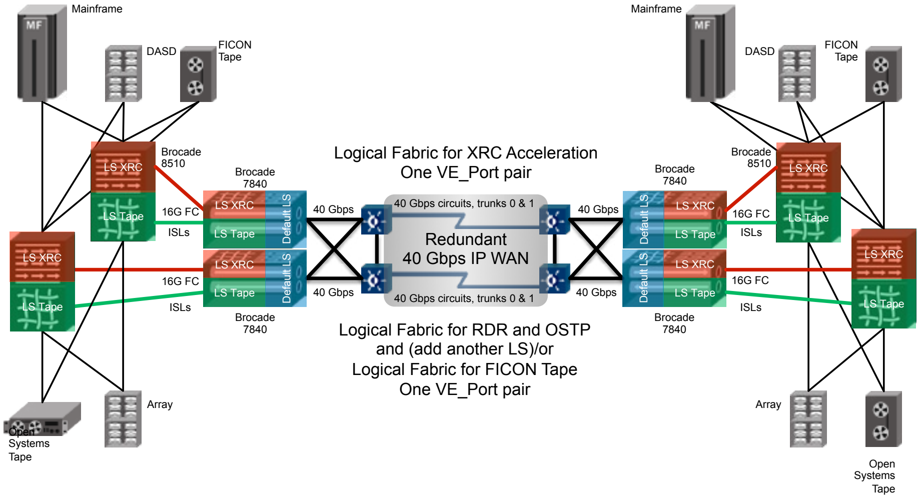

Figure 18: Protocol Intermix of FICON and Open Systems across 40 Gbps Brocade Extension.

IP Extension Architecture

In this IP extension architecture, the customer has a primary data center and a colocation site that is in or adjacent to a cloud service provider. Normally, the propagation delay and quality of WAN connections make remote Network-Attached Storage (NAS) replication impractical as a backup modality. The speeds at which NAS replication can perform locally in a data center versus performing across the WAN are typically very different. NAS replication across a WAN connection tends to be considerably slower. Using the Brocade 7840 with WO-TCP, which includes streams and TCP acceleration, results in "local performance" over distance. This is the fastest throughput possible across a WAN.

NAS is used as backend storage for various servers in the primary data center. There is an isolated Brocade VDX® 6740 Switch backend IP storage fabric used for NAS connectivity between VMs, the NAS heads, and the Brocade 7840s; other IP storage applications may be using this fabric as well. At the remote side is the same architecture of IP storage fabric and NAS heads. The VMs at the cloud provider (remote side) come online only when needed. The remote VMs can access the NAS heads and access the backup data, bringing the enterprise applications back online.

IP extension provides the following functionality to storage administrators:

- Increased NAS replication performance across distance
- Security of the NAS data inflight using IPsec
- Compression of the NAS data inflight
- Multipath High Availability using Brocade Extension Trunking and ARL
- Operational excellence via Brocade Network Advisor, MAPS, Flow Vision, Flow Generator, and Extension Dashboard
- Diagnostic and troubleshooting tools

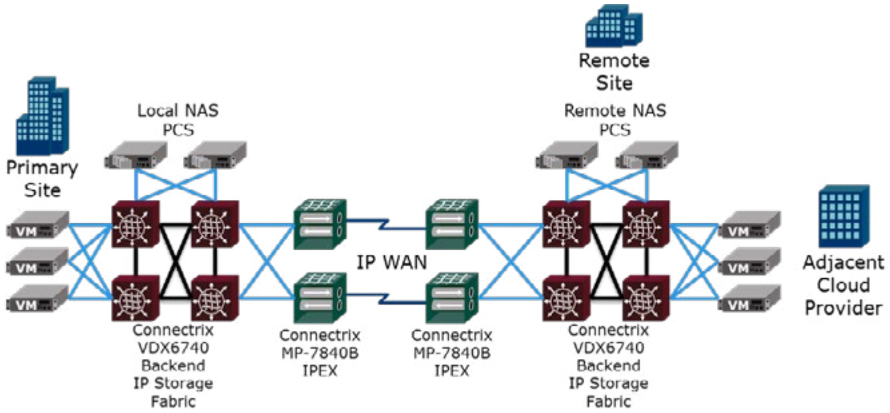

Figure 19: Tape Replacement using NAS over Brocade 7840 IP extension.

Conclusion

Brocade Extension solutions offer many advanced features that have been developed for a wide variety of critical environments found in small to medium businesses as well as the largest, most demanding enterprises. As a thought leader and innovator in extension technology, Brocade was the first to develop technologies such as Extension Hot Code Load (FCIP tunnels/trunks stay up during firmware upgrades), IP extension, Extension Trunking, FC Routing, FICON emulation, 10 and 40 GbE tunnels and trunks, ARL, FastWrite, OSTP, advanced compression techniques, extension IPsec, PTQ, and much more. These critical capabilities deliver unmatched performance, efficiency, and flexibility and drive down capital and operating expenses. Ultimately, Brocade Extension products enable storage and mainframe administrators to achieve the goals placed on them by their enterprises.

In addition to industry-leading extension technology, Brocade has deep technical expertise and offers assessment, design, and implementation services to help organizations optimize an existing SAN or architect a new SAN that meets the requirements of specific environments. When it comes to distance extension solutions, no one can match the Brocade experience and best-practice deployment knowledge of Brocade or can offer as thorough a service for all storage and tape platforms.

About Brocade

Brocade networking solutions help organizations transition smoothly to a world where applications and information reside anywhere. Innovative Ethernet and storage networking solutions for data center, campus, and service provider networks help reduce complexity and cost while enabling virtualization and cloud computing to increase business agility. Learn more at www.brocade.com.

Corporate Headquarters
San Jose, CA USA
T: +1-408-333-8000
info@brocade.com

European Headquarters
Geneva, Switzerland
T: +41-22-799-56-40
emea-info@brocade.com

Asia Pacific Headquarters
Singapore
T: +65-6538-4700
apac-info@brocade.com

Notice: This document is for informational purposes only and does not set forth any warranty, expressed or implied, concerning any equipment, equipment features, or service offered or to be offered by Brocade. Brocade reserves the right to make changes to this document at any time, without notice, and assumes no responsibility for its use. This information document describes features that may not be currently available. Contact a Brocade sales office for information on feature and product availability. Export of technical data contained in this document may require an export license from the United States government.