Evaluating remote data-replication solutions

Contents

Executive summary

This whitepaper describes the benefits of adopting the IBM SAN b-type Fibre Channel extension platform for remote data replication. It enumerates the values of b-type extension switches and applicable architectures, including both open system and mainframe. The IBM® System Storage® SAN42B-R extension switch is a purpose-built, enterprise-class product that is characterized by an essential feature set: excellent application performance, wide swath of scale, robust security, high reliability, network integration, comprehensive monitoring, application visibility and diagnostic tools.

An alternative to array native IP ports are IBM SAN b-type extension products connected to array native FC ports. These products provide replication performance, flexibility and reliability. IBM SAN b-type extension is optimized for a wide range of scale from small to large, which are applicable to nearly every replication environment.

IBM SAN b-type FC technology integrates perfectly into any IP network and provides an efficient data transport capable of full bandwidth use across great distances. The defining features that bring value to b-type extension switches are described in this document, including extension trunking, extension hot code load (HCL), adaptive rate limiting (ARL), quality of service (QoS), Fabric Vision technology from Brocade Communications Systems, Inc., and IBM Network Advisor. Our b-type technology also provides a full spectrum of security features and connectivity validation tools. Overall, IBM SAN b-type extension products make use of 20 years of distance connectivity innovation and thought leadership, as demonstrated by the fact that they are the market's preferred extension switch solution.

The situation

Consider a situation in which there are two or more data centers replicating data for disaster recovery (DR). This may be open system logical unit number (LUN) replication or mainframe volume replication. This paper applies equally to either environment. You might ask if remote data-replication (RDR) performance over distance is meeting your requirements, and if the recovery point objective (RPO) objective is optimal. Will more bandwidth fix the problem? And if so, how will costs increase? Perhaps data replication is taking too much time and occasionally falls behind the RPO objective. Is the replication across multiple arrays unbalanced? And does this negatively affect multisession consistency groups? Compression rate and ratio may be insufficient to provide adequate throughput. In addition, when data leaves the protection of your data center, there may be a compliance requirement to encrypt data in flight. Other questions might include:

- Does the combination of encryption and compression bring throughput to low levels?
- Is there a better way to configure, manage, monitor and troubleshoot your RDR network?

Architectures

The System Storage SAN b-type extension solution can scale from small to large. The smallest-scale deployment is two System Storage SAN06B-R extension switches with four 8 Gbps FC ports and two 1 Gb per extension FCIP ports connected directly to storage array replication ports. This offers two Gbps of FCIP WAN bandwidth. Assuming 2:1 data compression, four Gbps of replication bandwidth will be seen by the array. At this scale, cost is sensitive, and only a single SAN b-type extension network is connected to both controllers. The environment may grow and expand to two parallel RDR networks. Further scaling can be achieved by adding a port on demand (PoD) license, which upgrades the SAN06B-R from four 8 Gbps FC ports and two 1 Gb per extension (GbE) FCIP ports to 16 8 Gbps FC ports and six 1 Gb per extension FCIP ports. This more than doubles the original capacity. This deployment is small-scale, cost-effective, upgradable and extraordinarily powerful.

On the other hand, a very-large-scale deployment may use four SAN42B-Rs deployed in parallel pairs. Two parallel SAN42B-Rs can accommodate 160 Gbps of mainframe extended remote copy (XRC), mainframe tape, disk replication and open systems tape—all coming from multiple sources. The overall capacity doubles if you add a second parallel redundant network: 160 Gbps equals two SAN42B-Rs, and each SAN42B-R has a two-times data processor at 40 Gbps each. This is the bandwidth seen on the FC/FICON side. On the FCIP WAN side, this equates to 80 Gbps (two SAN42B-Rs times two data processors at 20 Gbps each), assuming 2:1 compression using the fast deflate algorithm. There is plenty of failover bandwidth for ARL or extension HCL. The SAN42B-R can directly connect to applications through either 24 16 Gbps FC ports or a production fabric. This deployment offers an extremely large-scale, redundant and cost-effective architecture.

The SAN42B-R provides the performance and tools to best transport storage data of all kinds to anywhere in the world, offering high reliability, great efficiency, outstanding performance and considerably easier management. IBM Network Advisor is a comprehensive tool that simplifies management and helps users proactively diagnose and troubleshoot issues to maximize uptime and increase operational efficiency. IBM Network Advisor pulls data from the Brocade Fabric Vision technology—including the Monitoring and Alerting Policy Suite (MAPS) and Flow Vision—into customizable dashboard views with deep drill-down capabilities that provide comprehensive visibility into network health and performance of storage replication.

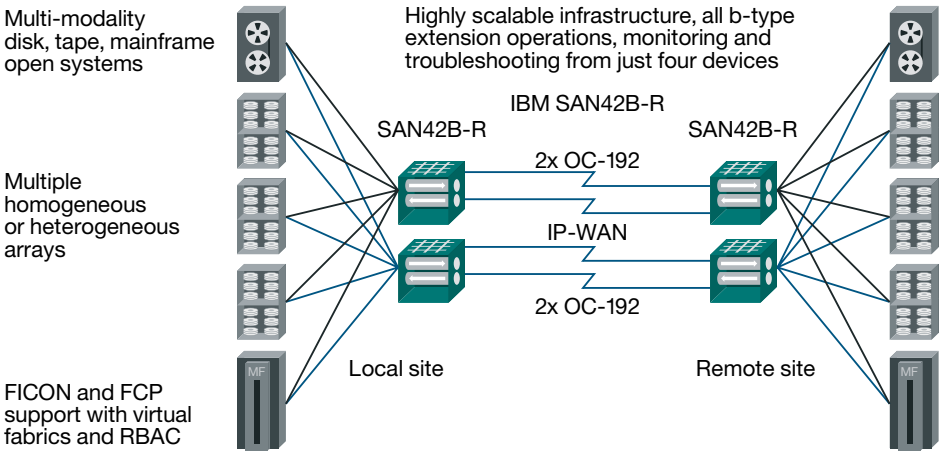

Figure 1. Large-scale SAN b-type extension deployment

IBM SAN b-type extension uniquely fits into large-scale storage deployments, as shown in Figure 1. Large-scale deployments often require the following: disk, tape, open systems, mainframe and so forth, heterogeneous arrays, high bandwidths, high throughput, nonstop operations, tools for operations and robust diagnostics. The RDR network referred to here can be integrated into production fabrics or kept completely separate, depending on what makes most sense. Separation can be achieved logically, using Virtual Fabrics (VF), or physically, using completely different switches. Either way, the RDR network and all fabrics can be managed from a single pane of glass using IBM Network Advisor.

Enterprise-class

The SAN42B-R is the storage industry's highest-throughput extension device. Compared to competing extension products and methods, IBM SAN b-type extension provides ample capacity for multiple FC ports, including those coming from multiple sources. The SAN42B-R contains the world's fastest FC switching ASIC, 64 network processors and 128 GB of high-speed RAM.

Performance

The advanced performance and network optimization features of the SAN42B-R enable replication and backup applications to send more data over metro and WAN links in less time and to optimize available WAN bandwidth. The SAN42B-R leverages the core technology of b-type Gen 5 Fibre Channel platforms, consistently delivering 99.9999 percent uptime in the world's most demanding data centers. It combines enterprise-class availability with innovative features and the industry's only WAN-side non-disruptive firmware upgrades to achieve always-on business operations and maximize application uptime. These capabilities enable a high-performance and highly reliable network infrastructure for disaster recovery and data protection. Let's review some other performance characteristics of this platform.

Purpose-built hardware

Developing purpose-built hardware and firmware offers the benefit of extreme performance and ultra-low latency. The architecture of the SAN42B-R is elegantly simple and enables ultra-low latency for synchronous applications. The data path within the SAN42B-R is concise and is implemented in a few fast, but highly effective, components.

Synchronous replication

IBM SAN b-type Fibre Channel platforms offer extension products with ultra-low added latency. The added propagation time gained through b-type extension products is within the tolerance for synchronous applications.

For synchronous applications, low added latency is just the beginning. Purpose-built hardware for compression and IPsec can be used with synchronous replication. This is made possible with native FC on b-type Gen5 switches, as well; however, combine this capability with extension trunking, and the synchronous solution becomes even more compelling. Extension trunking, which is explained in more detail later in this paper, is a technology exclusive to IBM SAN b-type platforms for establishing multiple circuits between two VE_Ports. Each circuit can take its own IP network path, usually dense wavelength-division multiplexing (DWDM). Multiple paths provide resiliency, fast error recovery from lost links, no data lost in flight from lost links (Lossless Link Loss [LLL]), data integrity, in-order delivery and true bandwidth aggregation of each circuit. These are all beneficial to fast response time and network reliability, which are demanded by synchronous storage replication. An example benefit of extension trunking in a DWDM environment is the prevention of interface control check (IFCC) in the event of an optic, cable or optical multiplexor failure.

Today, clients use b-type extension for synchronous applications with positive results. Keep in mind that the IP network itself must do the same. Adding synchronous applications to robust extension over a poor-performing IP network will result in a poor-performing synchronous application. While IP networks do not perform poorly by nature, they can perform poorly if they are not constructed well. IP networks can be properly built and configured to meet the requirements of synchronous storage. LLL and Extension Trunking are both exclusive to SAN b-type FC, and are described in the high-availability (HA) section of this whitepaper.

Encapsulation using FCIP

The IBM SAN b-type FC technology uses a unique method of forming streams of bytes from storage I/O. There is no concept of individual FC frame discrete encapsulation, as this would be too inefficient. A stream of bytes is formed, which is transported by WAN-optimized Transmission Control Protocol (TCP),or WO-TCP. Sixteen data frames form a stream called a “batch.” Each batch has a single FCIP header, which reduces headers by 16:1. The batch is then compressed.

By compressing the entire batch, it is possible to gain higher compression ratios. The SAN42B-R supports various deflate-based compression algorithms, namely Fast Deflate, Deflate and Aggressive Deflate. Each algorithm has a different trade-off of speed versus compression ratio. The stream fills TCP segments to their maximum segment size. The maximum segment size is the IP maximum transmission unit (MTU) minus the IP and TCP headers (IP plus TCP headers is about 40 bytes). The result is full-size IP datagrams and minimal overhead, regardless of the compression. The b-type encapsulation method excels in efficiency.

WAN-optimized TCP

Transmission Control Protocol is centric to the high-speed transport of large data sets that are common in storage extension. Through years of experience, our SAN b-type FC OEM vendor has developed an aggressive TCP stack called WAN-optimized TCP (WO-TCP). WO-TCP is a transport that cannot be outperformed by competing WAN optimization products. In other words, you receive negligible benefit from WAN optimization when using the SAN42B-R extension switch. Overall, SAN b-type FC technology is comparable from the perspective of the data transport bottom line. The total bytes transferred within the same period of time, over the same bandwidth, will be virtually the same compared to competing WAN optimization products. In addition to these benefits is the satisfaction that the cost of purchasing SAN b-type FC extension is considerably less than WAN optimization products.

- WAN latency exceeds 100 milliseconds (ms).
- WAN quality is poor.
- WAN is prone to errors.
- WAN has excessive jitter.

WAN optimization equipment is expensive compared to b-type extension. IBM SAN b-type extension makes WAN optimization totally unnecessary. Adding WAN optimization introduces complexity, another point of failure and another asset to configure, manage, monitor and troubleshoot. If WAN optimization already exists, b-type extension will unnecessarily consume that resource, which other applications can use, instead.

WO-TCP integrates with ARL, and the synergy of these two technologies creates an industry-dominating transport for storage. No similar transport exists on any storage array-based native IP replication. WO-TCP demonstrates the enterprise-class SAN42B-R.

Refer to Figure 2. Enterprise-class b-type extension exclusively offers the following:

- Outstanding performance
- Ultra-low latency b-type extension devices with IPsec, supporting synchronous applications
- LLL, which prevents IFCC when a circuit is disrupted
- Extension HCL for nonstop operations during firmware updates
- WO-TCP, the industry's highest-performing TCP stack

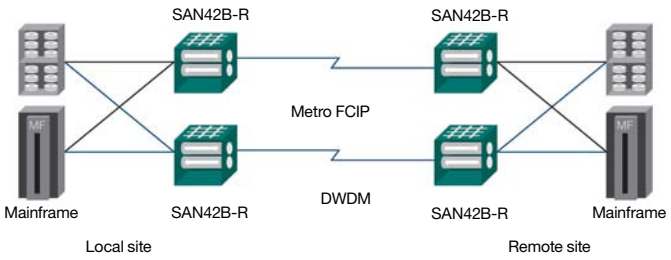

Figure 2. Enterprise-class SAN b-type extension

Bandwidth delay product (BDP)

In TCP, the amount of outstanding unacknowledged data that is needed to fully use a WAN connection depends on the bandwidth delay product (BDP). BDP is easily calculated by multiplying the bandwidth by the IP network's round-trip time (RTT). For example, a 1 Gbps WAN link with 160 ms RTT has a BDP of 20 MB. This means that you must have at least 20 MB of data in flight to fully use the bandwidth of this link. Any less than 20 MB results in "droop." Droop is the inability to fully saturate the WAN connection.

How does BDP apply differently to the SAN42B-R than to array-based native IP replication? IBM SAN b-type extension has superior BDP capacity built into WO-TCP, and it can maintain line-rate for a 10 Gbps WAN connection across 160 ms RTT, without droop. Consider that this connection could be an OC-192 between the US and Hong Kong, which has an RTT of about 160 ms. This works out to 1250 MBps times 0.16 seconds = 200 MB. The SAN42B-R actually has more than 200 MB of BDP capacity, if you take into account added memory for retransmits and other TCP window elasticity.

Because it has such a large BDP capacity, the SAN42B-R has the ability to maintain high throughput in such situations. WO-TCP is aggressive, and its aggressiveness is applied directly to the storage administrator's mission to expedite the protection of data.

Maximum transmission unit (MTU)

MTU is the largest-size IP datagram that an IP network can support end-to-end. If you are unsure what your path MTU (PMTU) is, the SAN42B-R can automatically determine the path MTU by using the PMTU feature.

Perhaps your IP network has a ring or multiple links with an active-passive architecture. For example, your storage applications might be using one of two links. The other link either remains passive, and your service provider cuts you a low-cost deal, or nonstorage applications use this path. Storage is permitted to take the alternate link only when the primary link goes down. IBM SAN b-type extension circuits have metrics and failover groups to automate failover for such architectures. When an extension circuit goes down, another circuit within the same failover group comes online. No data in flight is lost, and all data remains in order, even if frames are dropped. All circuits, including backup circuits, are independent and can be uniquely configured for each environment—in this example, primary and alternate.

Protocol optimization

When using extension external to storage arrays, combining other storage applications such as tape and mainframe over the same tunnel (one VE_Port) or different tunnels (different VE_Ports) is cost effective. IBM SAN b-type extension has protocol optimization for open systems tape pipelining (OSTP) read-write and FICON acceleration (XRC, tape read-write and Teradata); furthermore, optimization of these protocols simultaneously is supported. IBM SAN b-type extension can discern these different applications and apply protocol optimization accordingly. These applications can be extended great distances, mitigating the effects of latency while maintaining full bandwidth utilization.

Scale and operations

There are many aspects to scale that you might need to address. How small is your environment? How large? How much are the customer's replication needs growing over time? Can it be managed at scale? Will it be cost effective at scale? As outlined in this paper, all aspects of scale are addressed by SAN b-type extension, which offers a wide range of solutions that are cost effective. In addition, IBM has customers and experience that span the entire spectrum of these scales.

Throughput

IBM SAN b-type extension products scale on the FCIP WAN side from 100 Mbps up to 40 Gbps. On the FC/FICON side, they can scale from 2 Gbps to 80 Gbps, depending on compressibility of the data.

Compression

The SAN42B-R switch has been developed with specialized compression algorithms. These algorithms vary in processing rate and compression ratio, and are the most-aggressive compression algorithms available in the industry. On the SAN42B-R, there are three compression algorithms:

- Fast deflate
  - Rate: FC/FICON maximum ingress rate is 40 Gbps precompressed per DP
  - Ratio: Typical is approximately 2:1
- Deflate
  - Rate: FC/FICON maximum ingress rate is 16 Gbps precompressed per DP
  - Ratio: Typical is approximately 3:1
- Aggressive deflate
  - Rate: FC/FICON maximum ingress rate is 10 Gbps precompressed per DP
  - Ratio: Typical is approximately 4:1

Multimodality

An investment in high-performance extension technology usually means it must be employed across the enterprise to include the various modalities, such as:

- Mainframe volume replication
- Various open systems disk replication
- Mainframe tape
- Open systems tape

All of these can easily be accommodated by SAN b-type extension and managed by different administrator groups within an enterprise by using Virtual Fabrics and role-based access control (RBAC).

Special features can be applied to these modalities to ensure proper operation. In the case of FICON, you can apply FICON Accelerator, FICON CUP and FICON Management Server. In the case of open systems disk and tape, you can apply FCIP-based FastWrite and OSTP. You can use Virtual Fabrics to separate ports into their own logical switch (LS). You can configure LSs for FICON traffic and LSs for FC traffic, with different required settings. Member circuits of VE_Ports located in various LSs can all share the same Ethernet interfaces. This is vital, considering that an Ethernet interface may be 10 GbE or 40 GbE and is meant to supply connectivity for many trunks across the different Virtual Fabric LSs. In addition, to parse out the various circuits coming through that Ethernet link at the next hop DC LAN switch, VLAN tagging is used. All the functionality needed to support multimodality environments is available on SAN b-type extension.

Configuration simplicity

Configuration of IBM SAN b-type extension is considerably simple, compared to alternative solutions. You can configure SAN b-type extension in two ways: one method uses the Command-Line Interface (CLI), and the other method uses IBM Network Advisor. IBM Network Advisor is a GUI method of configuration for users that prefer GUI methods. The example configuration shown below includes the following:

- The Ethernet interfaces are set to 10 GbE (10 Gbps SFP+ are required).
- An IPsec policy is created.
- There is one trunk, defined by VE_Port 24. The trunk has two circuits.
- Two logical IP interfaces (ipif) are created—one for each circuit (192.168.0.2 and 192.168.0.3) —and the 10 GbE interfaces 0 and 1 are used (GE0 and GE1).
- Two IP routes are created to point to the local router gateway (192.168.0.1), one for each circuit. The remote side is 192.168.1.0/24. The MTU of this IP network supports jumbo frames at 9,216 bytes.
- There are two data processors (DPs) per SAN42B-R (DP0 and DP1). This trunk uses DP0.
- A tunnel is created with a minimum bandwidth (-b) of 5 Gbps and a maximum (-B) of 10 Gbps, specified in kilobits per second (Kbps).

- The trunk uses Fast Deflate compression, and IPsec is enabled. The first circuit (numbered "0" it is not entered in the CLI command) is added automatically when the tunnel is created. A second circuit (numbered "1" it is specified in the CLI command) is added next, which turns the tunnel into a trunk.

Of course, many features and functions can be deployed, which adds to any configuration. In most cases, adding functionality (VLANs, QoS) is as simple as adding additional arguments to the commands, shown below.

SAN42B-R configuration example of a two-circuit trunk

```
portcfgge ge0 --set -speed 10G
portcfgge ge1 --set -speed 10G

portcfg ipsec-policy pol1 create –k “think up some pre shared key for both sides”

portcfg ipif ge0.dp0 create 192.168.0.2/24 netmask 255.255.255.0 mtu 9216 portcfg
iproute ge0.dp0 create 192.168.1.2 netmask 255.255.255.255 192.168.0.1

portcfg ipif ge1.dp0 create 192.168.0.3/24 netmask 255.255.255.0 mtu 9216 portcfg
iproute ge1.dp0 create 192.168.1.3 netmask 255.255.255.255 192.168.0.1

portcfg fciptunnel 24 create --local-ip 192.168.0.2 --remote-ip 192.168.1.2 -b 5000000

-B 10000000 -c –fast-def -i enable pol1

portcfg fcipcircuit 24 create 1 --local-ip 192.168.0.3 -–remote-ip 192.168.1.3 -b 5000000 -B 10000000
```

Doing the mirror of this on the remote side creates a trunk. Overall, the configuration of SAN b-type extension is fairly simple, even with some of the advanced features discussed in this document (ARL, compression, IPsec, Extension Trunking, jumbo frames).

As for the IP network itself, the SLA with your IP networking department is not more stringent for SAN b-type extension relative to array native-IP requirements. In fact, the IP network SLA for SAN b-type extension is less, if not the same, due to the robust ability to drive across less capable IP networks. Considerable planning might be involved to obtain the right IP network deployment for RDR. IBM Network Advisor is a comprehensive management tool enabling storage administrators to manage their infrastructure end-to-end.

Security

Any data leaving the safe confines of the data center should be protected by using encryption. Encryption does not apply to the public Internet, only. Private WAN connections are not secure outside of your data center. Unsecured data leaving your data center potentially could cause data breaches and even unwanted publicity for an enterprise. IBM SAN b-type extension products have developed hardware-based IPsec for secure data in flight across extension Inter-Switch Links (ISLs).

IPsec

IPsec operates at line rate and introduces only a couple of microseconds (µs) of added latency, making it useful for synchronous applications. IPsec uses AES-GCM-256, Diffie-Hellman 2048-bit Modular Exponential (MODP),

Internet Key Exchange version 2 (IKEv2), Hashed Message Authentication Mode Secure Hash Algorithm 512 (HMAC-SHA2-512) and Transport Mode, and it is rekeyed every few hours without disruption. A pre-shared key (PSK) is configured per tunnel and trunk on each side.

Use IPsec for extension. IPsec is part of circuit formation and protects data from virtually every type of attack, including sniffers, data modification, identity spoofing, man-in-the-middle and denial-of-service attacks. IPsec requires no additional licenses or costs and is simple to configure. IPsec plus extension trunking gives you the ability to granularly load-balance encrypted storage flows across all the trunk's member circuits. Up to 20 Gbps is supported for a single trunk, and two such trunks are supported per SAN42B-R. This is a large amount of encrypted load-balanced data bandwidth (40 Gbps) for a single box. IPsec provides prudent security for most organizations and costs nothing extra with SAN b-type extension.

FOS security features

IBM SAN b-type Fabric Operating System (FOS) offers a large number of security features, such as RBAC. These features are beyond the scope of this document, but you should know that they exist.

High availability

There are many aspects to building a highly available RDR and tape network. Availability can be enhanced by network redundancy, resiliency of components, failover and failback functionality, continuous operations during firmware updates and preservation of bandwidth.

Extension HCL

Extension Hot Code Load (Extension HCL) was introduced to the storage industry with the SAN42B-R. Firmware upgrades can be done without tunnel disruption. A firmware update can take too much time to have a large extension connection go down. Years ago, WAN links had much less bandwidth, and it was not paramount to maintain connectivity during firmware updates. The interim backlog of data was acceptably small. However, by today's standards the amount of backlog data during a firmware upgrade can be significant, around half a terabyte or more when using one 10 Gbps connection. At many enterprises, to comply with RPO policy and to maintain a comfort level for storage administrators, nonstop operations are required. The SAN42B-R is the only product on the market that maintains extension connectivity during a firmware upgrade.

Extension HCL in SAN b-type FC products is lossless and always keeps data in order. No data is lost during the firmware update process, and all data sent to Upper Layer Protocol (ULP) is consistent and in order. This means that Extension HCL can be used in mainframe environments, without causing IFCC.

Extension trunking

With SAN b-type extension trunking, each storage I/O accesses all the WAN bandwidth that is seen by all the circuits belonging to a tunnel. An extension tunnel is defined by its VE_Port endpoint. The tunnel has a maximum bandwidth of 20 Gbps on the SAN42B-R extension switch, 10 Gbps on the SAN b-type extension blade, and 6 Gbps on the SAN06B-R extension switch.

Having multiple circuits per tunnel enables high availability. Extension trunking spreads data across all circuits, and those circuits can be dispersed across various paths and service providers; there is no requirement for equal bandwidth or latency among the circuits. Load balancing uses deficit weighted round robin (DWRR) on a per-batch basis. This is a granular load balancing method with the ability to failover or failback without data loss or out-of-order data. This capability is essential for mainframe environments and makes for more durable open system RDR environments as well.

If an IP path goes down at any level (service provider, local or remote, switches or routers, optics, cables, and so on)—and circuits are dispersed across different service providers, routers,

switches and paths—then no outage will occur, provided at least one path remains up. ARL will optimally readjust the bandwidth usage based on the remaining path or paths. Extension trunking is lossless: No data will be lost, and all data will be received by the upper layer protocols (ULP) in-order. The storage applications will not time out and will not perform error recovery.

Adaptive rate limiting (ARL)

Where rate limiting occurs in the network is important, and that point is after storage flows have been aggregated and before the IP network. IBM SAN b-type extension should be connected as closely to the WAN as possible. This way, the aggregate of all data flows is managed holistically with security and QoS effectively applied.

ARL automatically adjusts the rate limiting on all associated circuits replicating across the IP network, regardless of the ingress FC device and the WAN path or paths. ARL automatically adjusts rate limiting when other SAN b-type extension circuits go online or offline, or the available IP bandwidth that is being experienced changes. ARL works across all SAN b-type extension products using the same WAN infrastructure.

Shared WAN connections with nonstorage applications are common. ARL is designed to work on WAN connections that are shared with other IP storage and nonstorage applications. The SAN42B-R can be configured so that during an outage, high-priority applications maintain their bandwidth while lower-priority devices sacrifice theirs. ARL dynamically adjusts rate limits independent of each circuit, permitting efficient use of WO-TCP across a variety of ever-changing WAN environments. In this example, during the WAN service outage the overall bandwidth is halved and the SAN b-type ARL, integrated with WO-TCP, best uses the available bandwidth while maintaining nonstop operations.

Metrics and DF Bit

Not all WAN connections are provisioned equally. This may be due to the capabilities of the service provider or intermediate devices, or due to the cost associated with various connections. This means that backup circuits may need to have different configurations than primary circuits. Consider an example of two WAN connections. The primary connection supports jumbo frames of 9,216 bytes (MTU = 9216). The other is a less expensive secondary connection that does not support jumbo frames (MTU = 1500).

There are two ways to deal with this. The first example described here is not the SAN b-type way. Do not set the DF bit (don't fragment bit) in FCIP datagrams. The double negative (don't set the don't fragment bit) means that it is permitted to fragment these packets. If IP datagrams exceed the network's supported MTU, routers will fragment datagrams to conform to the supported MTU. Fragmentation is a resource-expensive operation and is not done in router hardware. It is done in software, upon arriving at the destination device. The destination is forced to reassemble these fragments, which takes time and processor resources. Generally speaking, IP fragmentation is a highly inefficient process that is not intended for high-speed, high-rate data transfers. Fortunately, there is a better way to handle this situation.

The SAN b-type way is as follows. Set the DF bit in FCIP datagrams. Setting the DF bit is not a configuration option. FCIP datagrams always have the DF bit set for optimal operation. Simply put, if datagrams do not pass across the network, then the circuit is misconfigured. Oversized datagrams will not be fragmented and are dropped. In this example, two circuits are configured in the same failover group and with different metrics. The primary circuit is configured with metric 0 and an MTU of 9,216 bytes. The backup circuit is configured with metric 1 and an MTU of 1,500 bytes. A WAN path change occurs, resulting in jumbo frame to standard frame support (9,216 to 1,500 bytes). This prevents the passage of FCIP datagrams on the primary circuit. When these datagrams stop,

the circuit (metric 0) goes down as soon as the keep-alive time-out value (KATOV) expires (which is set for 1 second). At this point, the IP network has already converged to the secondary path; now the backup circuit (metric 1) must be brought back online. The backup circuit is brought online and data resumes without any data loss and before the RDR application times out.

The SAN42B-R automatically, immediately and repeatedly tries to bring circuits back online after going down. When the primary path returns to an online state, the primary circuit with metric 0 will retry, succeed and come back online. When a metric 0 circuit comes online and is in the same failover group as a metric 1 circuit, the metric 1 circuit will go offline. The transitions from metric 0 to 1 and 1 to 0 is a lossless process due to the LLL.

The SAN42B-R can reroute between different MTU paths without any disruption, frame loss, or out-of-order frames. In this example, during interims of degraded IP network operation, storage is forced to use a less optimal MTU path. Nevertheless, operations stay online, and no data is lost in transit. During degraded IP network operations there is no need for IP fragmentation, which is inefficient. During normal IP network operations there is no need for smaller-than-supported MTU packet sizes, which is inefficient. It is most efficient to use circuits configured specifically for the primary and backup environments.

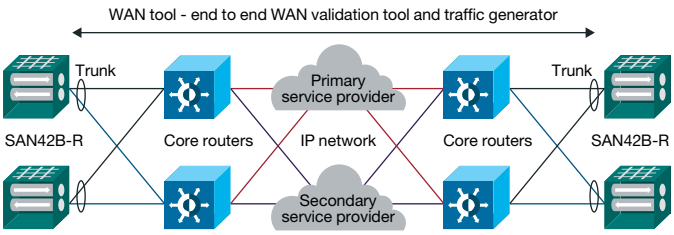

Figure 3. SAN b-type reroute between different MTU paths

Reliability, availability and serviceability (RAS)

The SAN42B-R is one component of an overall system that works together to guard against disruption. The SAN42B-R has certain features that can facilitate the quick resolution of support issues and the root-cause determination of faults or degradation.

Fabric Vision technology is supported on IBM SAN b-type extension products to help maximize uptime, simplify management and provide unprecedented insight and visibility across the storage network. With powerful built-in monitoring, management and diagnostic tools, organizations can proactively monitor, increase availability and dramatically reduce costs.

Monitoring and Alerting Policy Suite (MAPS)

Customers ask, "How can we resolve support issues more quickly and effectively?" OEM support organizations struggle to resolve cases before they become critical issues and before the RDR application is already down. This situation is further aggravated by the inability to pinpoint quickly whether the problem is a network or storage issue. Both customers and OEMs can benefit from the ability to monitor and effectively troubleshoot the local FC connections and network device health proactively —as well as the ability of the IP network to meet its SLA.

It is important to build intelligence into these networking systems. When a data connection starts to experience errors of any kind, the proper action may not be readily apparent until the situation becomes a major outage. Years of practical experience must be applied, because there is a large permutation of errors and effects. MAPS for FOS and IBM Network Advisor were introduced to provide a comprehensive suite of monitors, alerts, actions and reporting. MAPS assists operations in achieving higher availability, quicker troubleshooting and infrastructure planning. It provides a prebuilt, policy-based threshold monitoring and alerting tool that proactively monitors the health of the storage extension network, based on a comprehensive

set of metrics at tunnel, circuit and QoS layers. Administrators can configure multiple fabrics at one time, using predefined or customized rules and policies for specific ports or switch elements.

MAPS monitors use, packet loss, RTT, jitter and state changes for tunnels/trunks, circuits and PerPriority-TCP-QoS (PTQ). Each PTQ priority (class-F, low, medium, high) is monitored independently and includes throughput, duplicate Acknowledgments (ACKs), packet count, packet loss and slow-starts.

MAPS can be used in many situations. One example is the fencing of circuits that exhibit errors. MAPS is simple and easy to deploy with preset threshold levels and responses (conservative, moderate and aggressive). As needed—though not required—virtually every element is customizable in MAPS.

Flow vision

There are advantages to using SAN b-type extension, including scale and visualization of flows through tunnels. A SAN b-type extension tunnel enables administrators to visualize each application. To ensure SLAs are being met, storage administrators monitor network and flow behavior. This would be difficult to accomplish if managed from each originating device and port.

Troubleshooting network flows is often a difficult and daunting endeavor. Making matters worse, storage administrators are not familiar with IP networks, and the IP network administrators are not familiar with storage. These two groups have different cultures and operating guidelines. It is difficult for storage administrators to depend solely on network administrators to maintain their replication environment, which makes flow, TCP, circuit and tunnel monitoring and visualization considerably more important.

When troubleshooting storage flows, imagine that the flows fall into one of two categories: victims or perpetrators. If something goes wrong in the network, every flow becomes a victim. However, sometimes there is nothing wrong with the network, and flows fall victim to perpetrators. Perpetrator flows are flows that use excessive resources to the point that other flows fall victim. This frequently happens downstream from the storage handoff to IP networking. IBM SAN b-type extension provides features, functionality and tools to deal with storage SLAs. Flows within the protection of SAN b-type extension tunnels meet their SLAs when they come up against perpetrator flows.

Flow Vision enables administrators to identify, monitor and analyze specific application flows to simplify troubleshooting, maximize performance, avoid congestion and optimize resources. The SAN42B-R has the capability to monitor specific LUN flows between F_Ports that are communicating end-to-end across the extension network. It is also possible to monitor flows coming in from an E_Port. At LUN-level granularity, I/O operations per second (IOPS) and data rate can be monitored. Flow Vision includes the following features:

- Flow Monitor provides comprehensive visibility into flows across storage extension networks, including the ability to learn flows automatically and monitor flow performance without disruption. Administrators can monitor all flows from a specific storage device that is writing to or reading from a destination storage device or LUNs, or across a storage extension network. Additionally, administrators can perform LUN-level monitoring of specific frame types to identify resource contention or congestion that is affecting application performance.
- Flow Generator is a built-in traffic generator for pretesting and validating storage extension infrastructure—including route verification, QoS zone setup, extension trunking configuration, WAN access, IPsec policy setting and integrity of optics, cables and ports—for robustness before deploying applications.

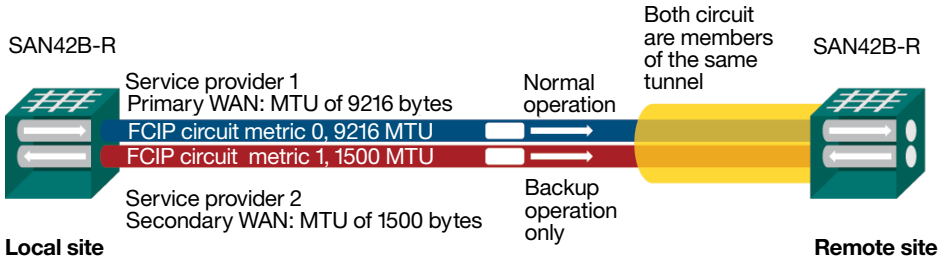

Figure 4. IBM FC SAN b-type extension integrate into the IP network

Qualification and validation tools

IBM SAN b-type extension offers a variety of tools to validate and troubleshoot IP networks, including those in the following sections. These tools can answer the following questions.

- What are the interface speeds? 40 GbE, 10 GbE or GbE?
- What is the path's MTU? If path MTU is used it can help determine the appropriate MTU size that is supported on the IP WAN.
- What about primary and backup paths? They are lossless when IP paths switch.
- Do you need QoS? PTQ assigns each priority to an autonomous TCP session.
- Do you need Ethernet-based QoS? Use 802.1P Layer 2 (L2) Class of Service (CoS).
- Do you need IP-based QoS? Use DSCP.
- Do you need to determine Full Duplex and Pause Frames? Enable/disable GbE Autonegotiation.
- Does one physical connection support multiple circuits over different VLANs? Use 802.1Q VLAN tagging.
- Is Network Address Translation (NAT) required? Assign inside versus outside devices.

Refer to Figure 4. IBM SAN b-type extension WAN tool example.

WAN tool

WAN tool was introduced with the SAN42B-R and accurately tests multiple IP network paths. WAN tool generates traffic at specified rates between a pair of IP addresses. It reports achieved throughputs, jitter, experienced latencies, congestion, packet losses and network reordering, and supports pertinent circuit characteristics, including PMTU, VLAN tagging, IPv4/IPv6, IPsec and jumbo frames. The main purpose of WAN tool is to validate the IP network before deploying a circuit. It is also useful as a diagnostics tool when you have a reliability issue with a circuit.

WAN tool simulates extension traffic exactly how the IP network would see it, such that the test results are truly relevant. WAN tool runs in the background and allows multiple simultaneous test sessions—up to eight sessions (four sessions per DP)—to coexist. Each test session equates to a single circuit. The total concurrent test capacity is eight circuits or two fully loaded tunnels or trunks. These connections are a UDP-like

simulation to facilitate detection of congestion, out-of-order delivery and packet loss; however, WAN tool runs the same TCP as the circuits do, so that IP network security mechanisms do not prevent testing. IP network security devices are tested, too.

Ping and trace route

IBM SAN b-type extension supports both ping and trace route, which are well-known IP networking tools. Ping is an Internet control message protocol (ICMP) echo that is used to determine if an IP datagram can successfully reach the destination and subsequently return. This is typically the first tool you use to validate end-to-end connectivity.

Trace route is similar to ping, except that the time to live (TTL) on the IP datagram is incremented by one from a starting value of 1 with each iteration. When a router receives a datagram with a TTL of 1, it drops the datagram and returns an ICMP message to the source, indicating the drop. That message has the IP address of the router responding to the drop, thereby informing trace route of the path along which the drops occurred.

Interfaces

The SAN42B-R has 16 1 GbE/10 GbE, two 40 GbE and 24 16 Gbps FC ports. You can run multiple 10 Gbps circuits across the two 40 GbE interfaces. Port and optic redundancy is now a reality with 20 Gbps VE_Ports that can span multiple Ethernet interfaces. Depending on the available interfaces that already exist in your data center, the most appropriate speed and number of interfaces are available on the SAN42B-R.

PerPriority-TCP-QoS (PTQ)

Where QoS is enforced greatly matters. An optimal place to enforce QoS is at the point at which different storage applications converge just prior to being directed into the WAN. Frequently, the IP network does not have QoS configured, at least for the storage applications. Therefore, at a minimum it is important to deliver data to the IP network sequenced according to the storage administrator's priorities. IBM SAN b-type extension is located at the endpoints of the data transport, the TCP points of origin and termination. These endpoints are the most effective places to QoS-mark data and apply it to various applications. The association of proper QoS markings to specific circuits, either primary or backup, is easily done here. IBM SAN b-type products have PTQ, in which each priority receives its own autonomous WO-TCP session. Cooperating with IP network administrators, QoS values for 802.1P/DSCP can be vetted and deployed if QoS is being enforced in the IP network. Using b-type extension makes this an automated process.

PTQ assigns each QoS priority its own autonomous TCP session within each circuit.

Additionally, each priority within each circuit can be independently marked with 802.1P (L2 CoS), DSCP (Layer 3 [L3] DiffServ) or both, as needed. This robust QoS schema permits storage connections to travel different paths, because the connections have different QoS characteristics and requirements. It is not effective to enforce multiple QoS priorities within a single TCP session. Using multiple TCP sessions that are not under the same supervisor tunnel is also not effective. PTQ is jointly supervised across all WO-TCP sessions for all circuit members of a tunnel. PTQ is monitored by MAPS; refer to the MAPS section of this paper for more detail.

Pause frames (IEEE 802.3X)

IBM SAN b-type extension supports autonegotiation on GbE interfaces. GbE autonegotiation is not used to negotiate link speed; the speed is always GbE. GbE autonegotiation is used to determine link duplex (full or half) and pause-frames (802.3x) support. By default, b-type technology enables autonegotiation and supports pause frames either on or off, and only full duplex.

Autonegotiation is enabled by default on most data center Ethernet switches, and pause frames are disabled by default on those same switches. Some Ethernet switches do not support IEEE 802.3x pause frames. In practice, it is unlikely that network administrators will enable pause frames on their Ethernet switches. Pause frames can lead to head of line blocking (HoLB) on DC LAN switches, causing all flows to sporadically stop on an Ethernet inter-switch link (ISL), resulting in poor performance.

Unfortunately, in nearly all cases pause frames are disabled by default on the connecting DC LAN switches. There is a much better way to deal with storage flow control across an IP infrastructure, using ARL. Refer to the ARL section in this paper for more detail.

Ethernet sharing and VLAN tagging (IEEE 802.1Q)

IBM SAN b-type extension supports VLAN tagging (802.1Q). VLAN tagging is frequently used when a single physical connection carries data from different VE_Ports (VE_Ports define trunk endpoints, therefore, these are different trunks); most likely those different VE_Ports live in different Virtual Fabric LSs. These tunnels will share a common Ethernet interface, because the interface bandwidth is large—10 Gbps or 40 Gbps—and can easily accommodate multiple trunks. In this case, the Ethernet connection cannot be placed into one particular VLAN on the data center LAN switch port. By using VLAN tagging, each destination VLAN can be sorted upon receipt within the LAN switch. Each circuit from the SAN42B-R will specify its VLAN for a distinct path through the IP network. Multiple circuits from various tunnels can share the same large Ethernet interface, if desired.

Network address translation (NAT)

IBM SAN b-type extension supports NAT within the IP network. There are specific facilitating functions for proper integration in these environments. IBM SAN b-type extension devices communicate with each other based on a priority level based on the IP addresses on the SAN42B-R IP interfaces. This priority scheme can become ambiguous when NAT is performed; therefore, there are special commands to define the NAT endpoints.

Link aggregation

Link aggregation (IEEE 802.1ax LACP, link aggregation group [LAG]) is not supported on b-type extension products. LAG is not needed if you are using extension trunking, because the purpose of LAG is accomplished more effectively with extension trunking. Extension trunking performs not only the link aggregation, but a number of other important storage-specific functions, as well (for example, single logical link, LLL and In-Order Delivery [IOD]). Extension trunking is integrated into both the FC side and the LAN side, making it superior to LAG for storage applications. LAG solves only part of the problem that extension trunking solves, and LAG does it less effectively. LAG is flow-based. Extension trunking is batch-based, which is more granular. If a link disconnect occurs, LAG is not lossless for data in flight; extension trunking is lossless (LLL). All links in a LAG have to be in the same configuration; circuits in a trunk are not restricted and can be unique.

Summary

This paper has covered the intrinsic advantages of the b-type extension platforms. This applies equally to open system or mainframe environments that use array-to-array replication. There are numerous innovative technologies found in IBM SAN b-type FC purpose-built hardware and firmware. IBM SAN b-type extension solutions with this technology are of enterprise-class quality and demonstrate excellence in performance, reliability and availability, security, scale and operational management. The ability to integrate into any IP network is enabled through a variety of features and validation tools. There is a comprehensive management platform with Fabric Vision technology from Brocade Communications Systems that provides unprecedented insight and visibility across the storage network. The platform uses powerful built-in monitoring, management and diagnostic tools that enable organizations to simplify monitoring, increase availability and dramatically reduce costs.

For more information

To learn more about evaluating remote data-replication solutions, please contact your IBM representative or IBM Business Partner, or visit the following website: **ibm.com**/systems/storage/san/b-type/index.html

For more information on Fabric Vision technology, please visit: **ibm.com**/systems/storage/san/b-type/fabricvision

Additionally, IBM Global Financing can help you acquire the IT solutions that your business needs in the most cost-effective and strategic way possible. For credit-qualified clients we can customize an IT financing solution to suit your business requirements, enable effective cash management, and improve your total cost of ownership. IBM Global Financing is your smartest choice to fund critical IT investments and propel your business forward. For more information, visit: **ibm.com**/financing

Route 100
Somers, NY 10589

Produced in the United States of America
November 2015

TSW03321-USEN-00