BROCADE

WHITE PAPER

# Brocade Adaptive Rate Limiting

Brocade ARL is an advanced extension capability that delivers dynamic rate limiting. Available on the industry-leading extension platforms—the Brocade 7840, Brocade 7800, and Brocade FX8-24—ARL provides higher performance and offers the advanced capabilities needed to safely share and efficiently utilize WAN bandwidth all the time. With Adaptive Rate Limiting, organizations can optimize bandwidth utilization and maintain full WAN performance of the link during periods when a path is offline due to a platform, IP network device, or array controller outage. Adaptive Rate Limiting uses dynamic bandwidth sharing between minimum and maximum rate limits to achieve maximum available performance during failure situations.

This paper provides an overview of Adaptive Rate Limiting capabilities and advantages that are included within Brocade Extension solutions.

## Table of Contents

## Introduction

ARL is a high-availability function used to maintain full bandwidth utilization, upholding availability of replication and backup operations during both normal and degraded conditions.

Common to all Remote Data Replication (RDR) systems is some type of WAN connection. WAN characteristics include bandwidth, propagation delay (often referred to as Round Trip Time [RTT]), and the connection type. Modern and ubiquitous connection types include MPLS (Multiprotocol Label Switching), DWDM (Dense Wavelength-Division Multiplexing), and Carrier Ethernet. There is yet another characteristic of considerable importance to RDR network design, the attribute that causes the most operational difficulties. Is the WAN dedicated to storage extension or shared? A dedicated or shared WAN connection is a very important consideration both for cost and operations. A WAN connection may be shared among multiple storage applications (such as Fibre Channel over IP [FCIP] disk and tape, native IP replication, or host-based replication) or shared with other nonstorage traffic (for instance, e-mail, web traffic, and so on).

More and more, enterprises are opting to use a shared WAN of considerable bandwidth for extension rather than a dedicated connection. There are many economic reasons for this. This paper discusses how Brocade ARL is used to facilitate the proper operation of extension across a shared WAN connection.

Implementing synchronous RDR (RDR/S) over any type of shared bandwidth infrastructure is not considered best practice, is not recommended by Brocade, and is not discussed at length in this paper. Tape and asynchronous RDR (RDR/A) can be implemented over a shared bandwidth infrastructure, and this implementation is the focus of this paper.

## Shared Bandwidth

Shared bandwidth between data centers is common in IP networks and WANs used by RDR and Backup, Recovery, and Archive; however, sharing bandwidth with these applications often presents difficult problems. Here are examples of shared bandwidth:

- Bandwidth shared between multiple extension flows originating from various interfaces on the same or different Brocade extension devices
- Bandwidth shared between multiple extension flows and other storage flows from host-based replication, native IP on arrays, Network Attached Storage (NAS) replication, Virtual Machine (VM) cluster data, and so on

- Bandwidth shared with non-storage-related applications, such as user traffic (for instance, web pages, e-mail, and enterprise-specific applications) and Voice over IP (VoIP)

Sharing bandwidth poses a problem for storage traffic. Extension traffic is, in essence, Inter-Switch Link (ISL) traffic. ISL traffic characteristically makes stringent demands on a network. For instance, storage traffic typically consumes a considerable amount of bandwidth. This depends on the number of volumes being replicated and the amount of data being written to the volumes within a given period of time.

In general, shared bandwidth is neither best practice nor recommended for storage; however, considerations of practicality and economics make it necessary for organizations to share their WAN links with other applications. Experience has demonstrated that placing storage traffic over shared bandwidth without proper planning and without implementing facilitating mechanisms for storage transmission either fails in practice or is plagued with problems. Mechanisms to facilitate storage transmission include ARL, massive overprovisioning of bandwidth, QoS, and Committed Access Rate (CAR). Each of these can be used alone or in combination. This paper focuses on Brocade ARL. ARL is the easiest mechanism operationally and the most reliable of the mechanisms that can be deployed.

## Brocade ARL Defined

Characteristics of Brocade ARL are described below and include the following:

ARL is a rate limiting technique exclusive to the Brocade 7840 and 7800 Extension Switches and the Brocade FX8-24 Extension Blade for the Brocade DCX 8510 and DCX Backbones.

ARL is primarily used when a WAN link shares bandwidth across multiple extension interfaces or with other user traffic, in other words, when a single interface does not have its own dedicated WAN link. It is very rare for an extension interface to have its own WAN link, although in small environments with no network redundancy this does occur.

ARL is not part of WAN Optimized Transmission Control Protocol (WO-TCP) or Storage Optimized TCP (SO-TCP), nor does ARL work by adjusting TCP windows or altering TCP functionality. However, ARL does use TCP to obtain current traffic conditions occurring across the IP network path.

Brocade ARL limits traffic to a circuit's maximum configured bandwidth, referred to as the ceiling. The maximum bandwidth of a circuit depends on the platform, 10 gigabits per second (Gbps) on the Brocade 7840 and FX8-24, and 1 Gbps on the Brocade 7800. The aggregate of the maximum bandwidth values across multiple circuits assigned to an Ethernet interface can exceed the bandwidth of an interface and is limited to two times the maximum speed of a VE_Port. A VE_Port on the Brocade 7840 has a maximum

speed of 20 Gbps. It has a maximum speed of 10 Gbps on the Brocade FX8-24 blade and 6 Gbps on the Brocade 7800.

Multiple circuits can be assigned to the same Ethernet interface, or a circuit can be assigned to a dedicated Ethernet interface. The design of the extension network and the number of available Ethernet interfaces and their speed dictate how circuits are assigned to Ethernet interfaces. Brocade ARL permits circuits that are active during one portion of the day to use the available bandwidth, while on the same interface during a different time of day a different circuit uses the available bandwidth. No one circuit exceeds the interface rate, but the aggregate of multiple circuits can exceed the individual physical interfaces—up to two times the maximum speed of the VE_Port.

Brocade ARL reserves a configured minimum amount of circuit bandwidth, referred to as the floor. The minimum for the Brocade 7800 and the Brocade FX8-24 is as low as 10 megabits per second (Mbps), and the minimum for the Brocade 7840 is as low as 20 Mbps. The aggregate of the minimum (floor) bandwidth values for multiple circuits assigned to an Ethernet interface cannot exceed the bandwidth of the interface. The minimum values cannot oversubscribe the interface, which always permits at least that much bandwidth to exist for that circuit. Proper extension network operation is always assured when operating at floor values.

Tape and replication traffic are sensitive to latency. Even asynchronous traffic can be negatively affected, primarily in terms of reduced throughput. There are Small Computer Systems Interface (SCSI) and Fibre Connection (FICON) protocol optimization methods such as FastWrite and FICON emulation that mitigate the effects of latency; however, optical propagation delay is not the only source of latency. Buffering delays also cause latency in the IP network during periods of congestion. Retransmission due to packet loss also causes added latency. Rate limiting in general must be used to counter these problems; otherwise, network performance degrades. In a properly designed and deployed IP network that has ample bandwidth and is not being oversubscribed, performance and behavior are the same as a native FC network across the same distance. This should be the goal for high-performance extension.

Proper rate limiting of extension traffic into a WAN network prevents packet loss due to congestion and buffer overflow. Feeding unlimited or excessive data into an IP network that is not capable of the capacity causes packet loss. This results in retransmissions, which adds to WAN utilization and exacerbates the problem. To make the situation worse, when congestion events occur, TCP limits transmission as a flow control mechanism in an effort to dissipate congestion and prevent it from immediately happening again. Using TCP as a flow control mechanism is extremely inefficient and results in poor performance. Simply said, you cannot feed more data into a network than it is capable of transmitting; otherwise, data is lost in transmission and performance suffers. Therefore, it is essential to provide the network with a comfortable data rate that avoids such problems. Since the inception of extension, rate limiting has been accomplished through a simple static method. This approach is inadequate and is rarely used in modern times.

Three common scenarios exist in which Brocade ARL is effective:

1. Storage plus nonstorage traffic over a single WAN link
2. More than one extension interface feeding a single WAN link that has been dedicated solely to storage
3. A combination of these two scenarios

In all three situations, the link may be oversubscribed, and rate limiting is required to prevent congestion and packet loss.

Sharing a link with both storage traffic and nonstorage traffic is the most problematic. User data tends to be unpredictable and very bursty, causing bouts of congestion. QoS can be used to give storage a higher priority. CAR can be used to logically partition the bandwidth over the same WAN connection. Brocade has developed WO-TCP, which is an aggressive TCP stack that facilitates and continues storage transmission—even after experiencing congestion events—by maintaining throughput in the most adverse conditions. Other TCP flows on a shared link will retreat much more relative to WO-TCP. This frees more bandwidth for storage during contentious times. WO-TCP cannot overcome User Datagram Protocol (UDP)-based traffic, because UDP has no flow

control mechanisms. Therefore, it is best to prevent collisions with UDP by using Brocade ARL as the mediator.

Brocade ARL is easily configured by setting a ceiling (-B or –max-comm-rate) and a floor (-b or –min-comm-rate) bandwidth level per circuit. Brocade ARL always maintains at least the floor level, and it never tries to exceed the ceiling level. Everything in between is adaptive. Brocade ARL tries to increase the rate limit up to the ceiling until it detects that no more bandwidth is available. If it detects that no more bandwidth is available, and ARL is not at the ceiling, it continues to periodically test for more bandwidth. If ARL determines that more bandwidth is available, it continues to climb towards the ceiling. On the other hand, if congestion events are encountered, Brocade ARL drops to the floor based on the selected back-off algorithm. Subsequently, ARL attempts to climb again towards the ceiling.

On the Brocade 7840, ARL back-off mechanisms have been optimized to increase overall throughput. The Brocade 7800 and Brocade FX8-24 have only one back-off method, which is reset to floor. On the Brocade 7840, ARL has been enhanced with new intelligence, which permits rate limiting to retrench in precise steps and preserves as much throughput as possible. Experience shows that a complete reset back to the floor value on a shared link is most often not required. Preserving bandwidth and reevaluating whether additional back-off steps are required is prudent.

ARL now maintains RTT stateful information to better predict network conditions, allowing for more intelligent and granular decisions. When ARL encounters a WAN network error (congestion event), it looks back at pertinent and more meaningful stateful information, which will be different relative to the current state. Rate limit decisions are fine-tuned using the new algorithms on the Brocade 7840. The Brocade ARL back-off algorithms are as follows:

- **Static Reset (Brocade 7800, 7840, FX8-24)**

  Static Reset is most appropriate for latency-sensitive applications such as FICON. Static Reset quickly addresses congestion issues due to excessive data rates. Static Reset is the fastest way to resolve the negative effects of congestion. FICON has strict timing bounds and is very sensitive to latency. FICON cannot wait for multiple iterations of back-off to finally get data through. It is most prudent to just reset the rate limit, avoid congestion, and resend the data. The bandwidth is the trade-off here. Time is the benefit.

- **Modified Multiplicative Decrease (MMD) (Brocade 7840 only)**

  For MMD, with each RTT across the IP network the rate limit is decreased by

  50 percent of the difference between the current rate limit and the floor value. An example: The floor is 6 Gbps, and the current rate limit is at 10 Gbps. A congestion event occurs. The first back-off goes down to 8 Gbps. Another RTT occurs, and another congestion event occurs also. The second back-off goes down to 7 Gbps (and so forth). A quick RTT is critical for quick and effective back-off. The MMD method of back-off is best with RDR (such as EMC SRDF, HDS HUR, IBM GM, and HP CA) on links less than 200 ms RTT. When the link is less than 200 ms, the contention on the WAN connection is resolved in a quick and efficient manner while, if possible, attempting to maintain as much bandwidth as possible.

- **Time-Based Decrease (Brocade 7840 only)**

  For ultra-long distance links that are greater than 200 ms RTT, MMD is not effective in backing off quickly enough to prevent excessive congestion. To solve this problem, the

back-off algorithm is time-based instead of RTT based. The Time-Based Decrease (TBD) algorithm calculates a number of steps (Steps = 1 second ÷ RTT). The back-off of the rate limit is computed as: Step-down = (current rate – floor value) ÷ steps. An example: On a 300 ms RTT link, the current rate is 10 Gbps and the floor rate is 6 Gbps. The first back-off is 1.2 Gbps down to 8.8 Gbps. The second back-off is 0.84 Gbps down to about 8 Gbps, and so forth. The rate at which WAN contention is resolved is directly related to the RTT. TBD is optimal for most intercontinental applications.

Auto Mode should be used and is considered best practice in most instances. Auto Mode chooses the correct back-off algorithm based on the detected and configured characteristics of the circuit. Auto Mode functions as follows:

- If FICON acceleration is enabled, or the Keep Alive Time Out Value (KATOV) is set to 1 second or less, Static Reset is used automatically. If FICON is used without FICON acceleration, the KATOV should be set to 1 second, which automatically implements Static Reset.
- If FICON is not used, and the KATOV is greater than 1 second, MMD is set.
- Within WO-TCP's first few RTTs upon bringing the circuit up, if the measured RTT is greater than 200 ms, the algorithm automatically changes to TBD.

After 10 milliseconds (ms) of idle time, ARL begins to return to the floor rate, using a step-down algorithm. The Brocade 7800 and FX8-24 take 10 seconds to return to the floor, and the Brocade 7840 takes 1 second. When data transmission starts, ARL climbs towards the ceiling. To climb from floor to ceiling, the Brocade 7840 takes 1 second, and the climb is smooth and linear. The climb on the Brocade 7800 and FX8-24 takes 10 seconds and is also smooth and linear. Climbing at faster rates is prone to creating detrimental congestion events and therefore is not done.

ARL is not intended to be an extremely reactionary mechanism for the purpose of precisely following network conditions. ARL's purpose is to maximize overall average bandwidth on shared links, to not interfere with other WAN applications, and to recover bandwidth from offline circuits. The process of detection, feedback, and readjustment is not practical for instant-to-instant adjustments. ARL is reactive enough to provide a positive experience on shared WAN connections and for reestablishing application bandwidth from circuits that go offline.

Brocade ARL plays an important role in extension across shared links with other traffic. If congestion events are detected, Brocade ARL backs off the rate limit and permits other traffic to pass. Typically, within Enterprise IT there is an agreement on the maximum bandwidth that storage can consume across a shared link. However, if other traffic is not using their portion of the bandwidth, storage traffic may use free bandwidth. When storage gets to use extra bandwidth from time to time, asynchronous applications (RDR/A) have a chance to catch up. RDR/A applications may fall behind during interims of unusually high load. Brocade ARL facilitates these types of SLAs (Service Level Agreements) without the cost of overprovisioned WAN connections.

In the case of multiple extension interfaces feeding a single link, the sources can be from a variety of storage devices, such as tape and arrays. Rarely does a host communicate with storage across a WAN due to poor response times and reliability. Alternatively, the source may be a single array with "A" and "B" controllers that are connected to respective "A" and "B" Brocade 7840, 7800, or FX8-24s in a High-Availability (HA) architecture, but that utilize only a single WAN connection. In any case, the link is oversubscribed and congestion occurs.

## The Need for Adaptive Rate Limiting

Static rate limiting or no rate limiting does not provide optimal performance in most situations. Consider these three situations:

- Oversubscribed during normal operation
- Undersubscribed during maintenance or failure
- Changing application demands

"Oversubscribed during normal operation" means that when all devices are online, and there are no failures in the network, the available bandwidth is oversubscribed by some amount. Figure 1 shows multiple 10 GbE interfaces on Brocade 7840s running at line rate (a similar situation can occur on the Brocade 7800 and FX8-24). This example uses two OC-192s (10 Gbps each) dedicated to storage applications. If multiple 10 GbE circuits feed each OC-192 for redundancy and resiliency, the link is over-subscribed, and the system suffers poor performance. This worst-case scenario is a result of misconfiguration and bad practices. Generally, TCP does a poor job of maintaining performance while performing flow control; therefore, the goal is to prevent TCP from having to perform chronic flow control. The motivation for such a configuration is to maintain full utilization, if any of the following devices go offline: an extension switch, an IP network device, or an array controller. Now, consider the next scenario.

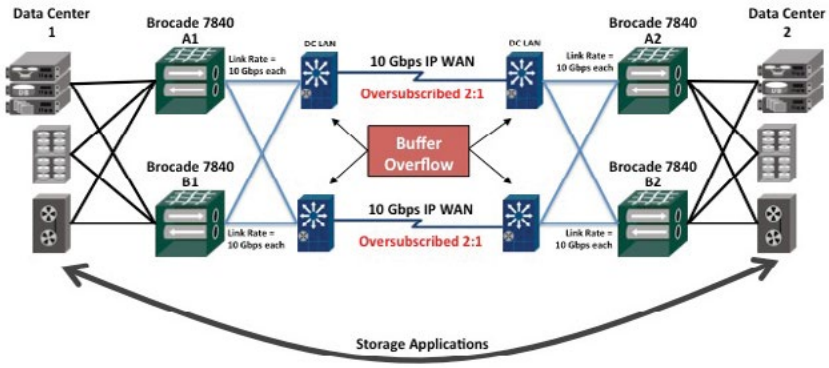


**Figure 1:** Oversubscribed WAN link scenario.

In the event that a Brocade Extension Switch (a Brocade 7840, 7800 or FX8-24), LAN switch/router, or array controller is taken offline for maintenance, technology refresh, or due to failure, then the result is that a path goes down. In the previous scenario, during an outage the oversubscription goes away, and there is no TCP flow control problem. The issue is that normal operation does not work properly, due to the oversubscription. Clearly, it is impractical to have a network that works properly only during device outages.

In a different case, "undersubscribed during maintenance or failure," normal operation has no oversubscription permitting proper operation. The relevant issue is that during times of failure or maintenance, undersubscription of the WAN occurs, which presents a problem of a different kind. This undersubscription may be acceptable for some organizations, but it is unacceptable for many organizations. It is imperative that during periods of

maintenance and failures, no increase in the risk of data loss occurs. This is exactly the situation that is faced when half the available bandwidth disappears, as shown in Figure 2 and Figure 3.

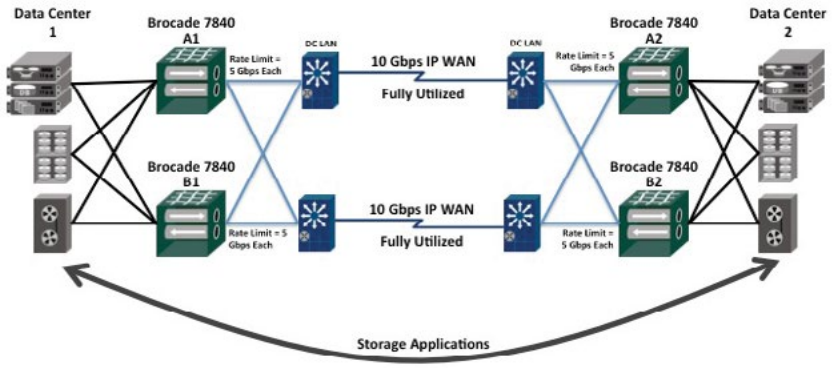


**Figure 2:** Oversubscribed WAN link scenario.

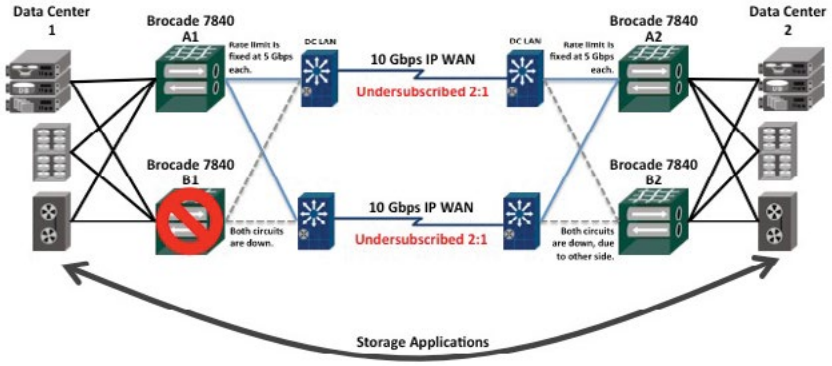


**Figure 3:** Outage situation: Undersubscribed with fixed Rate Limit (no ARL).

If, suddenly, only half the bandwidth is available, and assuming that more than half the bandwidth is used on a regular basis, this creates a situation in which an RDR/A application has to start journaling data (that is, Data Set Extension on SRDF) because of incomplete transmissions. This pending data is stored on the limited space of disk journals or cache side files. Depending on the load during the outage, these journals can quickly fill, as they are not meant to provide storage for longer than just minutes to possibly a couple of hours. The journal has to be large enough to hold all the writes that are replicated during the outage. Once the journal fills to maximum capacity, the RDR/A application has no choice but to sever the local and remote volume pair relationship. When connectivity is restored, a resynchronization of all dirty tracks on every volume pair is required. During the period of time that extends from the time the journals start to fill, through the regaining of connectivity, to completion of resynchronization of the pairs, recent data is not protected and RPO becomes long. Reestablishing volume pair

status can take a lengthy period of time, especially if the volumes are very large and many tracks have been marked as changed (dirty). During all this time, data is not protected, which exposes the enterprise to a significant liability. Many companies are interested in maintaining available bandwidth at all times and wish to avoid the risk associated with not doing so.

The problem described in the previous paragraph can be solved by simply purchasing a link with twice the bandwidth than is actually needed or will be used, and by setting static rate limiting to half that bandwidth. During times of failure, upgrades, or maintenance—when one or more paths are offline—only half the bandwidth can be used, due to static rate limiting. This does not pose a problem with an overprovisioned link, because twice the needed bandwidth is available, thus half the bandwidth works well. Obviously, this is not an optimal or cost-effective solution, because during normal operation half the bandwidth goes to waste.

Brocade ARL permits more efficient use of that bandwidth, reducing Operational Expenses (OpEx) and delivering a higher Return on Investment (ROI). With Brocade ARL, a properly sized WAN can be provisioned with the ceiling set to the WAN bandwidth and (assuming two extension interfaces per WAN link) the floor set to half of the WAN bandwidth. Refer to Figure 4 and Figure 5. If four extension interfaces are present, the floor is set to one-fourth of the WAN bandwidth.

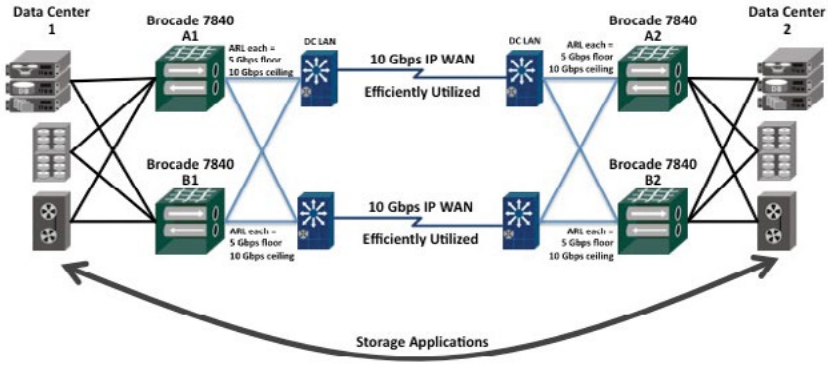


**Figure 4:** Normal operations: Extension network using Brocade ARL.

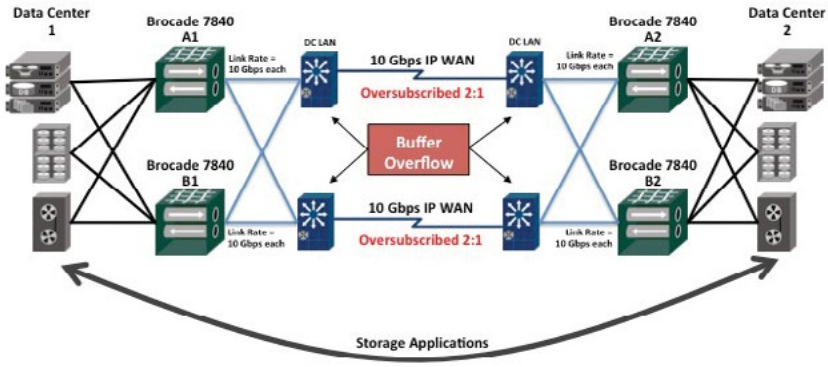


Figure 5: Offline device operation: ARL adjusted for an offline Brocade 7840.

Of course, manual intervention by resetting the rate limit settings can help here as well; however, the downside is that this requires more management, more chance of error, more change control, middle of the night and holiday problems, and increased operational costs.

Consider buying and using only the bandwidth that is actually needed. Some number of interfaces may go offline from time to time for a variety of reasons. If the remaining online interfaces automatically adjust their rate limit to compensate for other interfaces that are temporarily offline, then the bandwidth will be efficiently consumed all the time. Brocade ARL addresses this problem by automatically adjusting the rate limiting on extension interfaces. ARL maintains bandwidth during outages and maintenance windows, which accomplishes several goals:

- It reduces the likelihood that an RDR/A application has to journal data.
- It reduces the risk of data loss.
- It facilitates the provisioning of the proper amount of bandwidth versus excessive unused amounts.

Brocade ARL performs another important task by allowing multiple storage applications (such as tape and RDR) to access bandwidth fairly. As an application's demand diminishes, the other applications can utilize the unused bandwidth. If two storage applications assigned to different extension interfaces and utilizing the same WAN are both demanding bandwidth, they adaptively rate limit themselves to an equalization point. No application can preempt the bandwidth used by another application below its floor rate. However, if the demand of one application tapers off, the other application can utilize unused portions of the bandwidth. In Figure 6, you can see that as the blue flow tapers off, the red flow starts to increase its consumption of bandwidth. Conversely, if the blue application later needs more bandwidth, it preempts the red application's use of bandwidth back to the equilibrium point, but not beyond it. In this way, bandwidth is fully utilized without jeopardizing the minimum allotted bandwidth of either of the storage applications.

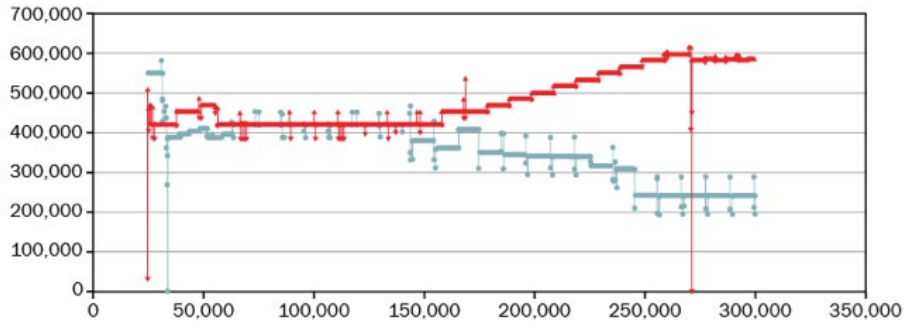


Figure 6: Application demand sharing a single WAN link.

## Summary

For growing enterprises, dedicating WAN bandwidth to data flows is inefficient and cost-prohibitive. It becomes crucial to avoid overprovisioning and efficiently share WAN bandwidth across multiple storage applications or across storage and nonstorage applications. Brocade ARL is an advanced extension capability that delivers dynamic rate limiting. Available on the industry-leading extension platforms—the Brocade 7840, Brocade 7800, and Brocade FX8-24—ARL provides higher performance and offers the advanced capabilities needed to safely share and efficiently utilize WAN bandwidth all the time. ARL is used in nearly all extension networks that connect to redundant IP network switches/routers to optimize bandwidth usage. ARL is used in mainframe and open systems environments for both disk replication and tape applications.

## About Brocade

Brocade networking solutions help organizations achieve their critical business initiatives as they transition to a world where applications and information reside anywhere. Today, Brocade is extending its proven data center expertise across the entire network with open, virtual, and efficient solutions built for consolidation, virtualization, and cloud computing. 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



BROCADE