White Paper

Deploying Quality of Service and Congestion Control in InfiniBand-based Data Center Networks

Diego Crupnicoff, Sujal Das, Eitan Zahavi

The InfiniBand architecture defined by IBTA includes novel Quality of Service and Congestion Control features that are tailored perfectly to the needs of Data Center Networks. While Ethernet QoS and Congestion Control mechanisms are well deployed in Enterprise LAN and WAN environments, they need significant improvements to meet the needs of the Data Center Network. Through use of features available today in InfiniBand Architecture based hardware devices, InfiniBand can address QoS requirements in Data Center applications far better than any other interconnect technology available today.

Executive Summary

Quality of Service (QoS) is required to fix the problems of best-effort service, lower bandwidth and high latency seen in Ethernet networks. InfiniBand, because of its inherent bandwidth and latency benefits, has been well deployed in high-performance computing applications where high bandwidth and low latency are de-facto requirements. QoS is an important requirement in Enterprise networks where Ethernet based best-effort service is deployed. However, with the advent of grid computing and increasing performance needs in the Enterprise data center, InfiniBand-based solutions are gaining momentum in data center fabric applications where the distinctive nature of short range networks bring in unique requirements for bandwidth and latency and how congestion is handled. The InfiniBand architecture defined by IBTA (InfiniBand Trade Association) includes novel QoS features in line with the scalable and well-accepted Differentiated Services architecture defined by the IETF. Through use of such features available in InfiniBand hardware devices, and related flow control and congestion management features, InfiniBand can address QoS requirements in Data Center Fabric applications far better than any other interconnect technology available today.

Introduction

QoS refers to the capability of a network to provide better service to selected network traffic or applications over various networking or interconnect technologies. Such networking and interconnect technologies include Asynchronous Transfer Mode (ATM), Ethernet, SONET, InfiniBand and others. The primary goal of QoS is to provide priority to selected traffic including dedicated bandwidth and controlled latency. However, depending on the technology and application, QoS offers other benefits as well. For example, for networking technologies such as Ethernet, QoS provides better handling of packet loss or packet drop characteristics and improves upon best-effort service delivery. For Wide Area Networking (WAN) and Internet topologies, QoS related protocols like Virtual Circuits (VCs) for ATM networks and Multi

Protocol Label Switching (MPLS) for Ethernet networks provide tunneling and security services as well.

InfiniBand has been traditionally used in high-performance computing and clustering applications. This application focus and the inherent high bandwidth and low latency characteristics available in InfiniBand networks have resulted in either sparse or no use of QoS features available in InfiniBand devices. InfiniBand is quickly becoming the fabric of choice for grid, utility and virtualized computing because of numerous compelling benefits such as price/performance and convergence of storage, compute and networking infrastructures. As a result, InfiniBand is being increasingly deployed in Enterprise Data Center (EDC) environments where the cost and flexibility related benefits of grid, utility and virtualized computing are becoming important decision drivers for IT managers. Because of the wide breadth of applications supported in EDC environments and the requirements to guarantee service levels to mission critical applications, deployment of QoS in InfiniBand-based networks will become increasingly important. This White Paper addresses InfiniBand QoS features and how they can be used to meet the IT Manager's needs in the EDC environment.

QoS Concepts and Requirements

At the heart of any QoS implementation is the concept of traffic classes or flows. A combination of source and destination addresses, source and destination socket numbers, or a session identifier may be used to define a flow or traffic class. Or more broadly, any packet from a certain application, from an incoming interface, or from a certain user or user group can be defined as a flow or traffic class. In this paper, references to flow or traffic class could be any one of these definitions.

Fundamentally, QoS enables the IT Manager to provide better service to certain flows or traffic classes. This is done by either raising the priority of a flow or limiting the priority of another flow. Congestion management schemes raise the priority of a flow by queuing and servicing queues in different ways. Policing and shaping provide priority to a flow by limiting the throughput of other flows. Finally, the rate of packet injection of lower priority flows can be controlled at the source nodes themselves to prevent congestion from happening in the intermediary devices of the network.

It is important to note that QoS tools can help alleviate congestion problems, not eliminate them. The full-proof way to eliminate congestion problems is with infinite bandwidth and zero latency networks. As such, technologies like InfiniBand, that support the highest bandwidth and lowest latencies are least prone to congestion. When there is too much traffic for the available bandwidth, QoS is merely a bandage.

QoS Implementation in InfiniBand Architecture

Unlike in the more familiar Ethernet domain where QoS and associated queuing and congestion handling mechanisms are used to rectify and enhance the best-effort nature of delivery service,

InfiniBand starts with a slightly different paradigm. First of all, the InfiniBand architecture includes QoS mechanisms inherently. What this means is that QoS mechanisms are not extensions as in the case of IEEE 802.3-based Ethernet that only defines a best-effort delivery service. Because of inherent inclusion of QoS in the base layer InfiniBand specification, there are two levels to QoS implementation in InfiniBand hardware devices:

1. QoS mechanisms inherently built into the basic service delivery mechanism supported by the hardware, and
2. Queuing services and management for prioritizing flows and guaranteeing service levels or bandwidths

InfiniBand Service Delivery Characteristics

A discussion of bandwidth and latency characteristics is important because inefficiencies in those parameters are the very reason why congestion occurs in the first place and leads to creation of mechanisms for congestion control and quality of service. If all interfaces met the necessary bandwidth and latency requirements for applications to meet required service levels, the discussions in this paper would be redundant.

Bandwidth: The latest available InfiniBand HCA and switch solutions can deliver up to 20Gbps and 60Gbps bandwidth respectively using high-performance processors and PCI Express-based motherboards on end nodes. In the context of congestion management, switch bandwidth is more relevant and it is worth noting that this is at least 6 times more than what is available with the latest Ethernet switches. In addition, with FAT tree configurations and full CBB (Constant Bisectional Bandwidth) support, available bandwidth can be increased indefinitely while adding fault tolerance to data paths.

Latency: Through the use of proven and mature RDMA, zero copy, kernel bypass and transport offload solutions, HCAs in end nodes can sustain very low latencies. The combination of low-latency HCA features with the cut-through forwarding mechanism available in InfiniBand switches results in very low end-to-end latency for applications – less than 3 microseconds. Switch only latency is in the order of 100-200 nanoseconds. Typical end-to-end Ethernet latency is at least 5 times higher than InfiniBand.

Service Delivery: InfiniBand is a loss-less fabric, that is, it does not drop packets during regular operation. Packets are dropped only in instances of component failure. As such, disastrous effects of retries and timeouts on data center applications are non-existent. It also supports a connection-oriented reliable transport mechanism that is implemented in hardware. This, along with transport window sizes optimized for wire-speed performance, enables very fast reaction times to handle bursty traffic and movement of congestion points. This level of service ensures that bandwidth and latency performance are maintained at the highest levels. Contrary to the best-effort paradigm, InfiniBand enables multi paths and topology designs with dedicated links that eliminate congestion possibilities for critical applications. We discuss these service delivery benefits further in the following discussion.

Cost: Because of the simple cut-through forwarding architecture, availability of efficient link level flow control mechanisms (described later) and the loss-less nature of the InfiniBand fabric, InfiniBand switches require significantly lower buffer memory and as such provide the highest price/performance benefit compared to any other technology that offers QoS.

InfiniBand Architecture Basics

InfiniBand Architecture's (IBA) basic unit of communication is a message. A message may contain between 0 and 2GB of data. Messages are segmented into packets. The payload of each packet must contain the maximum number of bytes negotiated for the path MTU. Segmentation and reassembly of packets is done by IBA hardware and hence MTU size restrictions are not detrimental to performance in anyway. The most common path MTUs are likely to be 256 bytes and 2048 bytes.

A fabric comprising of a set of HCAs interconnected with Switches is called a subnet. Subnets are interconnected at the higher layer with routers. All IBA packets contain a local route header (LRH) that includes the information necessary to forward a packet through switches. Additionally, a global route header (GRH) is provided that contains the information necessary to forward a packet through IBA routers. With few exceptions, the GRH is only present on packets that are to be routed between subnets.

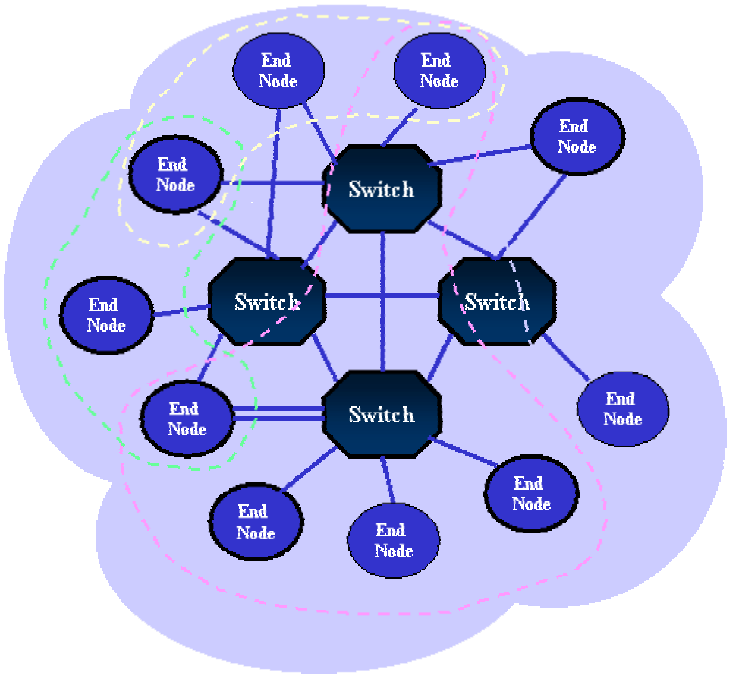

Figure 1: InfiniBand Fabric

IBA utilizes a memory-based user-level communication abstraction. The communication interface in IBA is the *Queue Pair* (QP), which is the logical endpoint of a communication link. The QP is a memory-based abstraction where communication is achieved through direct memory-to-memory transfers between applications and devices. A QP is implemented on the host side of an InfiniBand channel adapter (such as an HCA). The port side of the channel adapter implements what are called Virtual Lanes. Figure 2 is a pictorial depiction of Virtual Lanes. Virtual Lanes enable multiple independent data flows share same link and separate buffering and flow control for each flow. A VL Arbiter is used to control the link usage by the appropriate data flow.

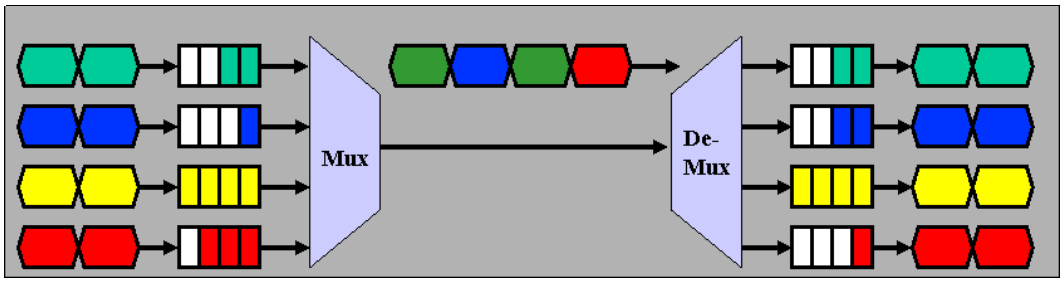

Figure 2: Virtual Lanes

The following Figure 3 is a pictorial depiction of the channel adapter and use of QP (Queue Pairs) and Virtual Lanes (VLs). There is further explanation of VLs and their application later in this paper.

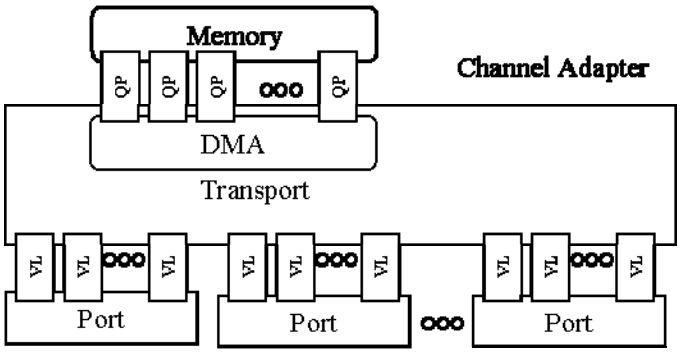

Figure 3: Use of Queue Pairs and Virtual Lanes in InfiniBand Channel Adapters

Virtual Lanes and Service Levels

IBA provides two fields for marking packets with a class of service: the service level (SL) field in the LRH and the traffic class field (TClass) in the GRH. The SL field is a four-bit field that may be arbitrarily used to indicate a class of service. IBA does not define a specific relationship between SL value and forwarding behavior; this is left as deployment policy to enable a wide

variety of usage models. There is, however, a defined mechanism in the specification to administratively specify a mapping between the SL values and the available forwarding behaviors in switches. The TClass field is an eight-bit field that serves the same purpose for routers as the SL field does for switches.

At the subnet layer (i.e. switches), IBA defines forwarding mechanisms to support a rich set of behaviors including various options to implement QoS and congestion control. These mechanisms can be divided into three major components:

- Virtual lanes (VL),
- Virtual lane arbitration, and
- Link level flow control.

This section discusses Virtual Lanes and Virtual Lane arbitration, which is depicted in Figure 4. Link level flow control is discussed in a later section.

IBA switches may implement between one and 15 VLs (Mellanox devices currently implement 8 VLs). A VL is an independent set of receive and transmit resources (i.e. packet buffers) associated with a port.

In addition to SL, the LRH contains the VL field that indicates the virtual lane number from which the packet was transmitted. Upon reception, the packet is placed in the port's receive buffer corresponding to the virtual lane indicated by the VL field. As a packet transits the switch from input port to output port, the packet may transfer from one virtual lane to another. Each switch in the fabric contains a table (referred to as the SL to VL mapping table) that selects the output port virtual lane based on the packets SL, the port on which the packet was received, and the port to which the packet is destined. This mapping function permits interoperability on fabrics consisting of switches supporting various numbers of virtual lanes. Note that an implication of this is, while the VL indication in a packet may change from hop-to-hop, the SL indication remains constant within a subnet. Note that packets within one virtual lane may pass packets in another virtual lane as they transit a switch.

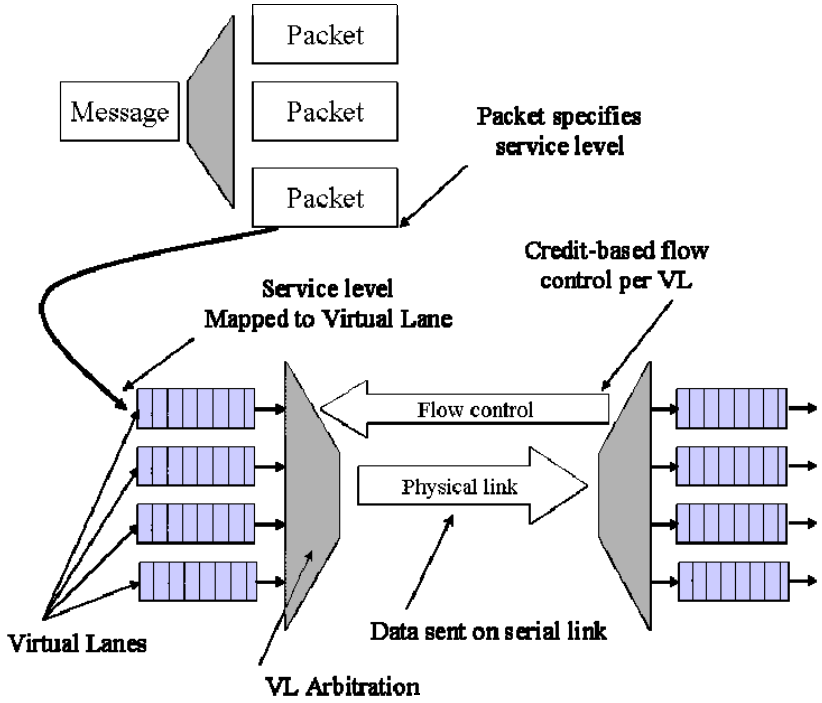

Figure 4: Service Levels, Virtual Lanes and Arbitration

Virtual lane arbitration is the mechanism an output port utilizes to select from which virtual lane to transmit. IBA specifies a dual priority weighted round robin (WRR) scheme. In this scheme, each virtual lane is assigned a priority (high or low) and a weight. Packets from the high priority virtual lanes are always transmitted ahead of those from low priority virtual lanes. Within a given priority, data is transmitted from virtual lanes in approximate proportion to their assigned weights (excluding, of course, virtual lanes that have no data to be transmitted).

Identifying Traffic Class or Flows

Traffic classes or flows need to be identified before bandwidth guarantee mechanisms using Service Levels and Virtual Lanes can be applied. Scalable and high performance implementations of QoS are based on the concept of IETF defined Differentiated Services as defined in RFC 2475. A discussion on Differentiated Services is beyond the scope of this white paper. However, one of the key concepts of Differentiated Services is labeling of packets. The mechanism to label the packets for class of required forwarding behavior may be as simple as including a few bits in the packet for this purpose. Examples of this for Ethernet include the IEEE 802.1p priority field in 802.3, the TOS field in IPv4, the TClass field in IPv6 [4] and the DSCP (Diff Serv Code Point) byte within the TOS or TClass fields. In InfiniBand, the Service Level or SL fields and the TClass fields serve the same purpose.

Packets are subject to multi-field classification (MFC) to identify flows or traffic classes in the sending nodes. Once flows are identified, packets are labeled (using the SL field in InfiniBand

or the DSCP field in Ethernet, for example) in the sending nodes. Downstream nodes, through which the packets traverse, inspect the packet label (the SL field in InfiniBand or the DSCP field in Ethernet, for example) and give the packet the right forwarding behavior by assigning it to the right queue (in case of Ethernet) or VL (in case of InfiniBand). This mechanism used in the downstream nodes is called the Behavioral Aggregate (BA) model. InfiniBand switches and receiving nodes use this BA model when using SL to VL mapping for packets.

IBA defines the BA model through use of SL and VL facilities. However, it does not define any methods for traffic classification in sending nodes. This is left to the implementer. Common classification methods use TCP or UDP port numbers to identify critical application flows. InfiniBand supports protocols such as Sockets Direct Protocol (SDP) and IP-over-IB (Internet Protocol over InfiniBand) that expose TCP and UDP based sockets interfaces. MFC schemes can be implemented with SDP or IP-over-IB protocols to provide identification of flows and marking of Service Level Fields for IP based applications. Similarly, block storage and other application flows can be identified by working with corresponding protocols such as SRP (SCSI over RDMA Protocol) or iSER (iSCSI over RDMA protocol) which are well deployed over InfiniBand today.

Prioritizing Flows and Guaranteeing Bandwidth

Virtual Lanes and SL to VL Mapping

InfiniBand links are logically split into Virtual Lanes (VLs). Each VL has its own dedicated set of associated buffering resources. The IB fabric can be regarded as an overlay of multiple logical fabrics that only share the physical links. Arbitration among VLs is done with a dual priority WRR scheme as described below. Packets are assigned to a VL on each IB device they traverse based on their SL, input port and output port as shown in Figure 5 below.

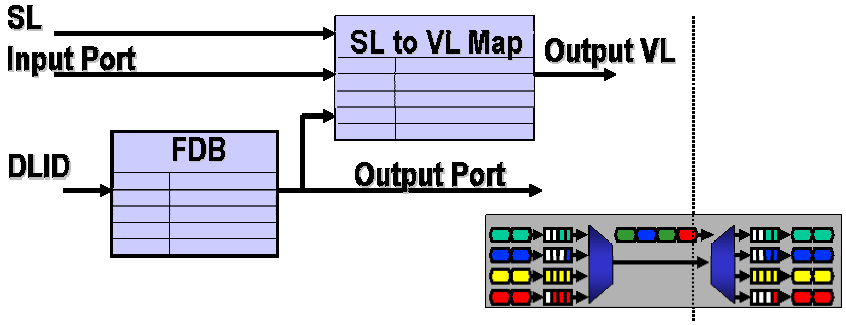

Figure 5: Service Level (SL) to Virtual Lane (VL) Mapping

This mapping allows (among other things) seamless interoperability of devices that support different numbers of VLs. The mapping is architected for very efficient HW implementation.

DLID and SL fields are among the first few bytes of incoming packet header. The Output Queue is determined by a simple lookup of SL2VL table.

Management and Data VLs have different attributes. They are listed in the table below:

Table1: Virtual Lane Attributes

	Management VL	**Data VLs**
Number Allowed	1	1,2,4,8 or 15
Buffering	Separate	Separate
Flow Control	None	Credit Based
Arbitration	Highest	Weighted Fair layered on 2 level priority
Buffer Size	Min: one packet	Min: Credit for one packet

Link Level Flow Control

InfiniBand Link Level Flow Control (LLFC) is implemented per VL. LLFC works as a first-degree mechanism to deal with congestion without dropping packets. Transient congestion is effectively dealt with by LLFC. Feedback-based mechanisms cannot deal with the time constants of transient congestion. Since it is on a per VL basis, LLFC maintains complete isolation of different traffic from each other. Transient congestion on one VL does not have any impact on traffic on another VL.

Subnet management packets have a dedicated VL, are not subject to LLFC, and are treated as the highest priority (thus guaranteeing management traffic progress regardless of fabric state).

VL Arbitration

Packets on different VLs share the same physical link. Arbitration is done through a dual priority WRR scheme. The scheme provides great flexibility and was designed with a HW implementation in mind.

VL Arbitration is controlled through a VL Arbitration table on each InfiniBand port. The table consists of three components: High-Priority, Low-Priority and Limit of High-Priority. The High-Priority and Low-Priority components are each a list of VL/Weight pairs. Each list entry contains a VL number (values from 0-14), and a weighting value (values 0-255), indicating the number of 64-byte units which may be transmitted from that VL when its turn in the arbitration occurs. The same VL may be listed multiple times in either the High or Low-Priority component list, and it can be listed in both lists (see VL Arbitration example below). The Limit of High-Priority component indicates the amount of high-priority packets that can be transmitted without an opportunity to send a low priority packet.

The High-Priority and Low-Priority components form a two-level priority scheme. Each of these components (or tables) may have a packet available for transmission. Upon packet transmission, the following logic is used to determine which table will be enabled to transmit the next packet:

- If the High-Priority table has an available packet for transmission and the HighPriCounter has not expired, then the High-Priority is said to be active and a packet is sent from the High-Priority table.
- If the High-Priority table does not have an available packet for transmission, or if the HighPriCounter has expired, then the HighPriCounter is reset, and the Low-Priority table is said to be active and a packet is sent from the Low-Priority table.

Weighted fair arbitration is used within each High or Low Priority table. The order of entries in each table specifies the order of VL scheduling, and the weighting value specifies the amount of bandwidth allocated to that entry. The table entries are processed in order.

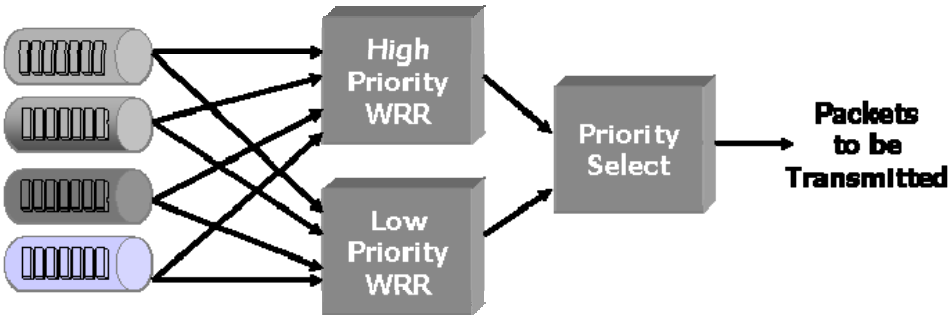

Figure 6: Use of WRR for Packets to be Transmitted

VL Arbitration Example

The following is an example of how VL arbitration can be used. See Figures 7 below.

VLs 0 and 1 are high priority. Minimum weight is used, so effectively as long as there are packets on VL0 and VL1 ready to be sent, they will alternate on the wire. When no packets from VL0 and VL1 are available to be sent then bandwidth is split between VLs 2-7. VL2 is been allocated 50%, VL3 16.66% and the rest is evenly divided between VLs 4-7. Note that VL2 bandwidth allocation is split in 3 entries to provide adequate bandwidth to VL2 and thereby improve the jitter and latency characteristics of data flowing through VL2.

High Priority			Low Priority	
VL	Weight		VL	Weight
0	1		2	8
1	1		3	8
			2	8
			4	4
			5	4

			2	8
			7	4
			6	4

Figure 7: VL Priority Assignment

Figure 8 below shows examples of required weight for 50MB/s available bandwidth when using different link speeds. To simplify calculations, the table depicts throughput of 0.25, 1.0 and 3.0 GB/s for link speeds of 2.5, 10 and 30 Gbps respectively. A 20% weight for a link with 0.25 GB/s throughput can meet bandwidth requirement of .2*0.25 GB/s = 50 MB/s.

Bandwidth Requirement: 50 MB/s		
Link Speed (Gb/s)	Throughput (GB/s)	Required Weight
2.5	0.25	20%
10	1.0	5%
30	3.0	1.7%

Figure 8: Link Speed versus Required Weight Example

Multi-pathing

InfiniBand supports multi-pathing natively. Multi-paths can be used for Fault tolerance (APM), load balancing and congestion avoidance through right provisioning of the network (Multiple parallel links can be used to increase CBB (Constant Bisectional Bandwidth) on potential hot spots – see section on use of fat tree topologies later in this paper.

Static Rate Control

InfiniBand supports static rate control, which enables a fabric manager to configure the network to avoid oversubscribed links in the fabric. For example, a path with a 4X link feeding a 1X link can be constrained to a 1X rate.

Link Level Flow Control

IBA is, in general, a lossless fabric, i.e., IB switches do not drop packets as a general method of flow control (there are exceptions to this to handle extreme cases of congestion such as what might occur in the presence of a component failure). To achieve this, IBA defines a credit based link-level flow control. Credits are issued on a per virtual lane basis; consequently, if the receive resources of a given virtual lane are full, this VL will cease transmission until credits are available again. However, data transmission on the other virtual lanes may continue. This permits traffic with latency or bandwidth guarantees using one set of virtual lanes to be unaffected by congestion of best effort traffic on other virtual lanes.

Congestion Control in InfiniBand Architecture

The InfiniBand Congestion Control Annex from IBTA defines a comprehensive and scalable method of congestion control that guarantees deterministic bandwidth and latency. This is in contrast to TCP-based congestion control mechanisms deployed in Ethernet networks which are based on dropping of packets – a mechanism acceptable in best-effort nature of delivery service scenarios only.

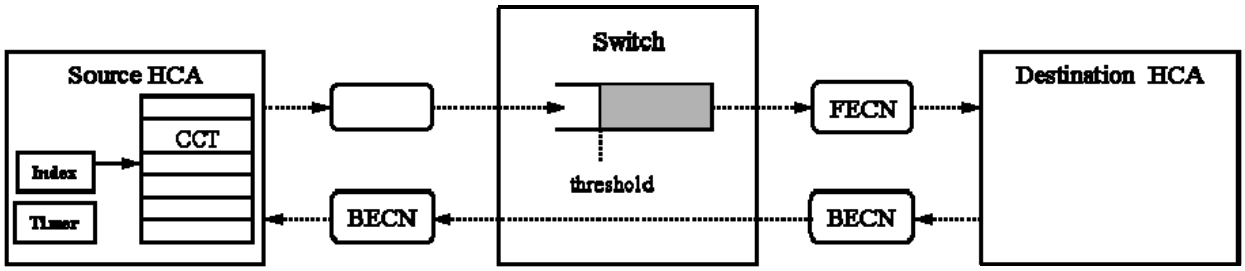

Figure 9: IBA Congestion Control Architecture

Figure 9 illustrates how the IBA Congestion Control Architecture (CCA) operates. This is a three-stage process: When congestion is detected in a switch, the switch turns on a bit (in packets) called Forward Explicit Congestion Notification (FECN). When the packet arrives at the destination HCA, it responds to the source with packets having a different bit set called Backward Explicit Congestion Notification (BECN). When the sending or source HCA receives a BECN, it responds by throttling back its injection of packets. This reduces congestion. Over time, the source gradually reduces it's throttling, which may again cause congestion to be detected. If all parameters – such as the rates of throttling and reduction in throttling over time – are appropriately set, the network should settle into a stable state that keeps the sources quenched. The setting of such parameters is under the control of a Congestion Control Manager (CCM), which establishes their values.

Congestion Marking Function: InfiniBand Switches detect congestion on a Virtual Lane (VL) for a given port when a relative threshold set by the CCM has been exceeded. The threshold is specified per port between 0 and 15; 0 indicates that the switch is not to mark any packets on this port, 15 specifies a very aggressive threshold. The marking rate is also configurable.

Congestion Signaling: Upon receipt of a packet with the FECN bit set, the destination HCA responds back to the source of the packet with Backwards Explicit Congestion Notification (BECN) that returns back to the specific queue pair that was the source of congestion. This may be piggybacked on normal ACK traffic for connected communication, or may be in a special congestion notification (CN) packet for unconnected service. The HCA processes Queue Pair information, and is the source of packets in normal operation, so this is natural to its function.

Injection Rate Reduction: When the source receives a packet with the BECN bit set, the injection rate of the flow is reduced. The reduction can be applied either to the specific queue

pair sourcing the flow, or to all QPs (queue pairs) on the port using a particular service level (which maps to a virtual lane). The amount of reduction is controlled by a table associated with the HCA, called the Congestion Control Table (CCT), whose content is loaded by the Congestion Control Manager. Each time a BECN is received, an index into the table is incremented. The amount of the increment is the CCTI_Increase value, also set by the CCM. Each entry in the CCT defines an Inter-Packet Delay value that is inserted by the HCA between packets. As BECNs are received, the index is repeatedly incremented, indexing further into the CCT where it will usually encounter larger and larger delay values set by the manager.

Injection Rate Recovery: Each HCA contains a timer. Each time it expires a duration set by the CCM, the index into the CCT is reduced by an amount also set by the CCM. This causes use of lower values in the CCT as Inter-Packet Delays. The CCM has set the CCT tables so that this reduces the added inter-packet delay. Eventually, if no more BECNs are received, the index reaches zero and delays are no longer inserted.

Use of Fat Tree Topologies in InfiniBand Based Clusters

High performance computing clusters typically utilize Clos networks, more commonly known as "Fat Tree" or Constant Bisectional Bandwidth (CBB) networks to construct large node count non-blocking switch configurations.

The major difference between fat-trees and traditional tree architecture is that fat-trees more resemble real trees. In a traditional tree (as used in Ethernet for example), the link bandwidth is fixed no matter how high the tree is. This will cause traffic congestion problems occurring at root switch. In a fat-tree, the link bandwidth increases when upward from leaves to root. And thus, the root congestion problems can be relieved. In a fat-tree-based interconnection network, leaf nodes represent processors, internal nodes are switches, and edges correspond to bidirectional links between parents and children. Figure 10 shows the difference between traditional trees and fat-trees. In Figure 10(a), a binary tree is shown. A binary fat-tree is shown in Figure 10(b). From Figure 10(b), we can see that the edge (link bandwidth) gets thicker (higher) when closer to the root.

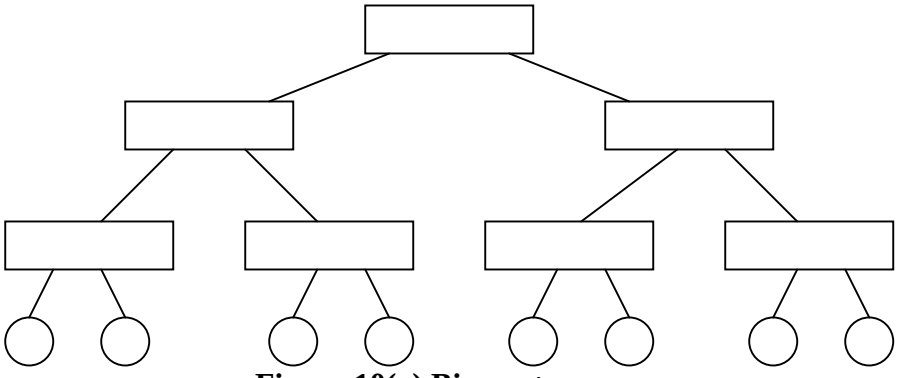

Figure 10(a) Binary tree

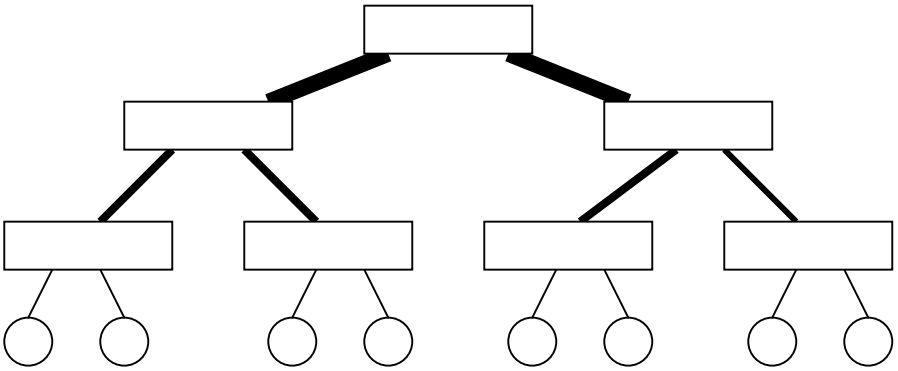

Figure 10(b) Binary fat-tree

Routing in a fat-tree is relatively easy since there is a unique shortest path between any two processing nodes i and j as shown in Figure 10(c) below. The routing consists of two phases, the ascending phase and the descending phase. In the ascending phase, a message from processing node i to processing node j goes upward through the internal switches of a fat-tree until the least common ancestor m of processing nodes i and j is found. In the descending phase, the message goes downward through the internal switches to processor j. Figure 10(c) shows a routing example in a fat-tree.

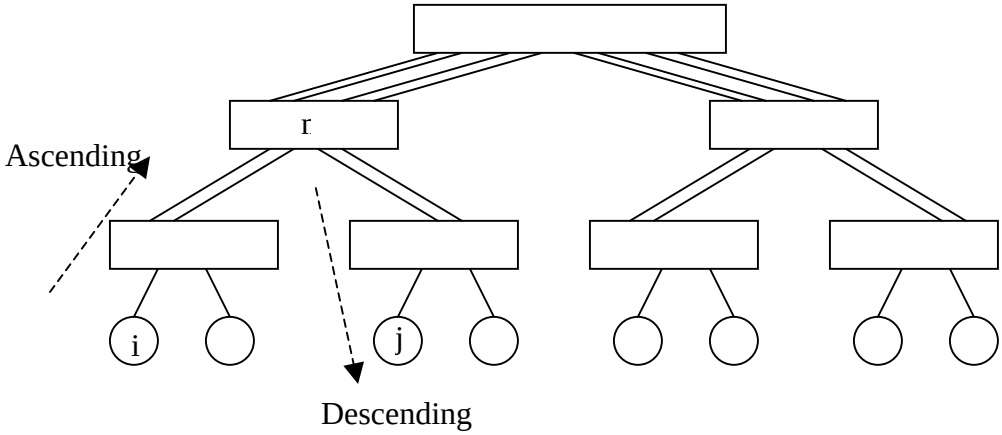

Figure 10(c) Routing example in a fat-tree

The use of Fat Tree topologies in InfiniBand clusters enables it to scale bandwidth as needed and provide packet forwarding using the shortest available path. Unlike Ethernet that uses traditional tree architectures, InfiniBand backbones are less prone to congestion. In addition, use of protocols such as spanning tree to prevent loops which result in packet forwarding behaviors that do not necessarily use the shortest paths is completely avoided.

Interconnect Requirements for the Data Center Network

As shown in Figure 11, a simplified view of a typical data center network is comprised of multiple blade servers or nodes using a data center switch or mesh network as the interconnect. Each blade server can be comprised of multiple processor and memory blades connected using a server switch fabric. Blade servers can be pure compute servers or may be storage servers with either local storage or network-attached storage. The data center switch or mesh network is also connected using a gateway to the corporate LAN or WAN network, and aggregators to aggregate traffic from clients and their applications that are serviced by the server blades.

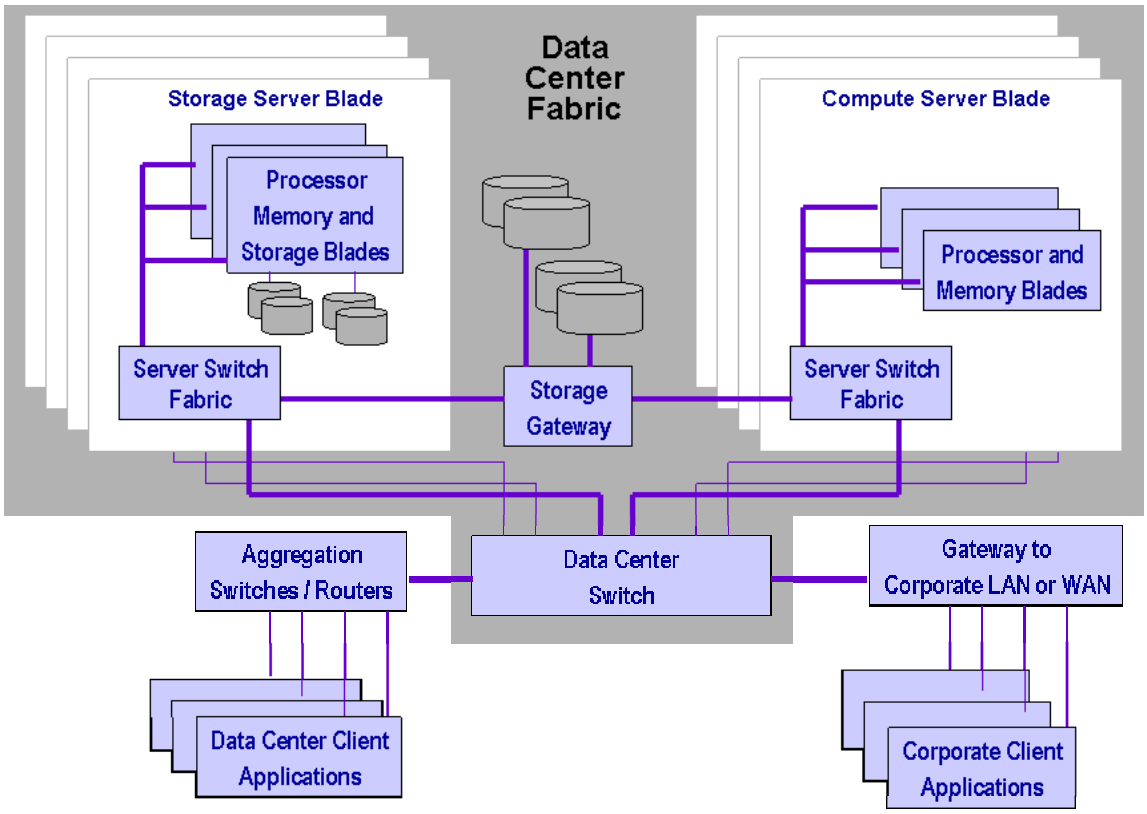

Figure 11: Data Center Network

The blade Servers, server switch fabric and the data center switch form the core of the data center network. Data center networks are short range networks which places them in a distinct position compared to Enterprise LAN and WAN technologies. The following table shows some typical characteristics.

Table 2: Network Characteristics Requirements

	Data Center	Enterprise LAN	WAN
End-to-end latency	Low	Medium	High
Sensitivity to packet drops	High	Medium	Low
Sustained data rate	High	Medium	Low

Because of the need for end-to-end low latency, key interconnect technology requirements are:

- Highest bandwidth
- Lowest latency using RDMA and transport offload technologies in end points
- Cut-through forwarding performance in intermediary devices like switches that forward packets
- Eliminate retries and timeouts in the transmission process

SAN applications deployed in the Data Center are characterized by large packet sizes and are sensitive to packet drops. As such, congestion management schemes that do not rely on dropping of packets are critical. Also, link-level flow control methods need to be able to differentiate traffic so that packet flows for critical applications can be maintained and only non-critical application packets are paused.

To enable high sustained data rate, the key interconnect technology requirements are:

- Highest bandwidth
- Loss-less fabric
- Deterministic performance

A key criterion for deterministic performance is availability of mechanisms for layer-2 level end-to-end congestion management, without any software or higher level overheads. Minimal use of buffers for packet storing and fast reaction times for transparent handling of congestion points in the network are critical because of bursty traffic.

A second criterion for deterministic performance is the use of fat tree topologies for bandwidth growth and the use of shortest-path forwarding mechanisms (in contrast to spanning tree forwarding mechanisms).

Finally, the short-range nature of data center networks and the need for scalability and bandwidth makes Differentiated Services (DiffServ) based QoS mechanisms more suitable, as compared to Integrated Services (IntServ) methods, where resource reservation and signaling methods are used for guaranteeing bandwidth for applications over long-range networks such as WANs.

Because of the above differences in how acceptable thresholds for critical performance parameters vary, congestion handling mechanisms in data center networks need to be specifically designed for that environment and requirements.

The following table provides a list of key performance, QoS and congestion management parameters relevant to data center networks, and depicts, based on the discussion above, how Ethernet and InfiniBand interconnects meet the requirements.

Table 3: Comparison of Ethernet and InfiniBand

Feature	Ethernet	InfiniBand
Raw Bandwidth or Throughput and Price/Performance	End Nodes: 1Gbps using available technology. RDMA and TOE technologies required for 10Gbps throughput are in their infancy and prohibitive from the cost standpoint. Switches: 10 Gbps available today but per port costs are above $2000.	End nodes: 20 Gbps available today at considerably lower price points than 10Gbps Ethernet. Mature deployments of RDMA and transport offload available today. Switches: 60 Gbps solutions available today with per port costs at about $200.
Raw end-to-end latency	Greater than 10 microseconds. Switches implement a mix of store and forward and cut-through forwarding. Switch latencies are in the order of hundreds of microseconds.	Can deliver < 3 microseconds today for end-to-end latency. Switches implement cut-through forwarding, enabling 100-200 nanosecond latency.
Eliminate retries and timeouts in the transmission process	No. Congestion control is based on dropping of TCP packets, resulting in retries and timeouts being common. No Layer 2 end-to-end congestion handling mechanisms	Yes. IBA is a loss-less fabric. Congestion handling is not based on dropping of packets. End-to-end congestion handling is based on Layer 2 mechanisms.
Deterministic performance	No. QoS and congestion management is based on dropping of packets. Use of spanning tree implies that L2 forwarding is not necessarily through the shortest path. Link control is not class or flow based.	Yes. Loss less fabric. End-to-end congestion control using L2 mechanisms. Use of fat tree topologies for meeting bandwidth requirements. Granular, class-based link level flow control through use of VLs.
Deterministic transport performance – fast reaction times	TCP transport is in software, uses large window sizes, and therefore response times for handling congestion can be slow.	Congestion is handled proactively and adaptively through use of FECN and BECN marking and signaling implemented in hardware. Hence reaction times

		can be very fast.
Handle overflow of arriving traffic	Yes. Through extensions to base Ethernet specifications.	Yes. Available as part of standard and base InfiniBand specification.
Manage traffic based on traffic class and priority. Assign strict priority to important traffic	Yes	Yes
Manage traffic using fair queue servicing methods	Yes	Yes
Ensure that queues do not starve for bandwidth and that traffic gets predictable service	Yes. High priority queues are serviced using weighted round robin (WRR) and other methods. Congestion is handled using WRED (Weighted Random Early Detection) which is based on dropping of packets. So, level of service may become unpredictable if congestion occurs.	Yes
Prevent multiple low priority flows from swamping out a single high priority flow	Yes	Yes
Differentiated Services based scalable and high performance QoS mechanisms	Yes. Not part of IEEE 802.3 Ethernet specifications. Achieved through extensions defined by IETF.	Yes. Behavioral aggregate based forwarding behavior is part of IBTA specification. Classification requirements in Data Center Fabric can be easily implemented.
Flow Control	Link level flow control only. Not capable of stopping or pausing specific type of traffic.	Per VL based flow control, implying that traffic with latency or bandwidth guarantees using one set of virtual lanes can be unaffected by congestion of best effort traffic on other virtual lanes.
Detect congestion on a per queue per application and/or per originating node basis	Yes.	Yes.
Control packet injection rate from sending nodes on a per queue, per application and/or per originating node basis	Per originating node and per application basis only. Performed by host software (TCP).	Yes. Granularity at the Queue Pair level. Performed by IBA defined hardware.

Conclusion

The InfiniBand architecture defined by IBTA includes several novel Quality of Service and Congestion Control features that are tailored perfectly to the needs of Data Center Networks. While Ethernet QoS and Congestion Control mechanisms are well deployed in Enterprise LAN and WAN environments, they have been designed as add-ons to base Ethernet specifications and need significant improvements to meet the needs of the Data Center Network. Through use of features available in InfiniBand hardware devices, InfiniBand can address QoS requirements in data center fabric applications far better than any other interconnect technology available today.