An Overview of QoS Capabilities in InfiniBand, Advanced Switching Interconnect, and Ethernet

Sven-Arne Reinemo, Tor Skeie, Thomas Sødring, and Olav Lysne, Simula Research Laboratory
Ola Tørudbakken, Sun Microsystems

ABSTRACT

A recent trend in interconnection network technologies is the inclusion of various mechanisms to support a variety of quality of service (QoS) concepts. This has been necessitated by an increasing number of application areas that require some level of performance guarantees from the network for parts of its traffic. In this article we describe and compare the capabilities and support for the QoS of three of the most important interconnection network technology standards of today. Equalities between the technologies are explained and differences are clarified.

INTRODUCTION

Research into the quality of service (QoS) in interconnection networks has gone through several phases. Two decades ago interconnection networks with point-to-point links were mainly used in massively parallel processors undertaking scientific calculations. At that time the interconnection network was considered a bottleneck in the computation and therefore most of the research and development effort focused on improving overall performance, while differentiation between different classes of traffic did not receive noteworthy attention. This resulted in rapid increases in interconnection performance and when new application areas related to multimedia and other real-time applications came along, the bandwidth of the interconnection network was widely believed to be abundant. Taking special measures to control the QoS was therefore considered unnecessary. It was not until the late 1990s that this view changed in a profound way, and at present there is an expectation that QoS-controlling mechanisms are part of any general-purpose interconnection standard.

When discussing QoS in interconnection networks, there are three properties of significant importance: *bandwidth*, *latency*, and *packet loss*. The granularity of the object on which these metrics are applied are single data streams, classes of traffic, or all-network traffic. In most interconnection technologies, there is a strict guarantee of no packet loss that is valid for all data traffic. Ethernet can be viewed as an exception to this, which we discuss below. With regard to latency and bandwidth, a combination of the mechanisms are often defined, ranging from strict guarantees for single streams [1] via relative guarantees for classes of traffic [2] to no guarantees/overprovisioning.

The capabilities that influence the technologies' ability to leverage QoS guarantees and differentiated treatment of traffic fall into three categories: flow-control, congestion management, and traffic differentiation.

Flow-control aims to reduce or eliminate packet loss that is a result of contention and overflowing receive-buffers in switches; however, its use can result in congestion in one part of the network spreading to others parts as link transfers slow down or are temporarily halted. This phenomenon is known as back-pressure. Congestion management aims to prevent or react to the onset of congestion, reducing or controlling its effect on overall throughput in the network. Traffic differentiation applies differential treatment to traffic in order to provide certain guarantees to particular streams. A complex application (e.g., a video server) deals with a multitude of traffic types, each with different requirements for timely delivery. Some of the traffic may be sensitive to delay but without strict bandwidth requirements, for example, network control and management traffic. Other types of traffic (e.g., video streams) may have strict bandwidth requirements whereas the latency requirements are relaxed.

In this article we present an overview and comparison of flow-control, congestion management, and traffic differentiation mechanisms in the InfiniBand Architecture (IBA) [3], Advanced Switching Interconnect (ASI) [4], and what has so far been specified for Backplane Ethernet (by the IEEE 802.3ap Backplane Ethernet Task Force). We also clarify the situations where the technologies give different names and descriptions for what is essentially the same mechanism.

This work is financed in part by the EU under the 6th framework program for IST as part of the SIVSS project. Sven-Arne Reinemo, Tor Skeie, and Thomas Sødring are all primary authors.

0163-6804/06/$20.00 © 2006 IEEE

Feature	IBA	ASI	Ethernet
Routing	Destination lookup	Source routing	Destination lookup
Link width (no. of lanes)	1x, 4x, 12x	1x, 2x, 8x, 16x, 32x	1x, 4x (for 3.125 Gb/s)
Bandwidth per lane	2.5, 5, 10 Gb/s	2.5, 5 Gb/s	1, 3.125, 10 Gb/s
Bandwidth	2.5–120 Gb/s	2.5–128 Gb/s	1–10 Gb/s
Maximum packet size	4096 bytes	2176 bytes	1522 bytes (9000 bytes is supported by many vendors)
Minimum packet size	24 bytes (20 bytes raw)	64 bytes	64 bytes
Transmission encoding	8B/10B	8B/10B	8B/10B, 64/66B for 10 Gb/s
Maximum cable length	Unspecified	Unspecified	5000 m
Maximum number of hosts	49,152	Unspecified	Unspecified
Maximum ports per switch	255	256	Unspecified

Table 1. *IBA, ASI, and Ethernet features overview.*

The rest of the article is organized as follows. First, we present a basic description of each of the three technologies under review. We then detail the technologies' ability to support flow-control, congestion management, and service differentiation. Thereafter, we discuss the mechanisms provided and how they support the higher-level properties of QoS that have motivated the development of these mechanisms before we conclude in the final section.

Architectural Overview

InfiniBand Architecture

The InfiniBand Architecture (IBA) was first standardized in October 2000 [3], as a merging of two older technologies called Future I/O and Next Generation I/O. As with most other recent interconnects, IBA is a serial point-to-point full-duplex interconnect. It was originally designed as a unified I/O fabric to replace everything from the PCI bus inside commodity servers and FibreChannel in storage area networks, to Ethernet in system area networks. While it has not been successful in achieving this, it has found its niche as an *intersystem* interconnect in storage and high-performance computing, where high-bandwidth and low-latency networks are key requirements. The feature set of each technology is presented in Table 1.

Advanced Switching Interconnect

Since its inception in the mid-1990s the PCI interconnect has evolved from a parallel, bus-based technology to a serial, switched one. In its latest incarnation, PCI Express IO vl.0 is a low latency, high-bandwidth *intrasystem* interconnect [5]. Advanced Switching Interconnect (ASI) is an intersystem interconnect that is built upon the same technology as PCI Express, reusing the same physical and link layers but with its own transaction layer to provide peer-to-peer communication within an ASI fabric [4].

Ethernet

Since its invention at Xerox PARC in 1973, local area networking (LAN) has been, by far, the most important application domain for Ethernet technology [6, 7]. Today, Ethernet is the dominant LAN technology for both wired and wireless networking but is now making inroads into new application areas ranging from clustering and storage networks to wide area networking. Furthermore, Ethernet is currently undergoing a standardization process for the backplane in the IEEE 802.3ap Backplane Ethernet Task Force.

Elements of Quality of Service

In the following we assess the flow-control, congestion-management, and differentiated-services mechanisms for IBA, ASI, and Ethernet, before we resume a discussion regarding the QoS capabilities of each of the technologies. The different terms used by each technology are summarized in Table 2.

Flow Control

IBA — In order to avoid packet loss due to buffer overflows, IBA uses a point-to-point credit-based flow-control scheme. In such a scheme, the downstream side of a link keeps track of the available buffer resources (credits) by decreasing a credit counter whenever buffer space is allocated and increasing the credit counter whenever buffer space is deallocated. Similarly, the upstream node keeps track of the available credits (i.e., the number of bytes it is allowed to send) and decreases this amount whenever it sends a packet. Whenever credits arrive from the downstream node, it increases the amount of available credits. A packet is never sent downstream unless there is room for it. At regular intervals the downstream node details credit availability to the upstream node. The update interval depends on the load — high loads increase the frequency of updates, while low loads reduce the frequency.

The concept of virtual lanes allows for grouping of virtual lanes as layers, thus making it possible to build virtual networks on top of a physical topology. These virtual networks, or layers, can be used for various purposes such as efficient routing, deadlock avoidance, fault tolerance, and service differentiation.

Term	IBA	ASI	Ethernet
Virtual link	Virtual lane	Virtual channel	Priority
Service class	Service level	Traffic class	Priority
Link layer packet	Packet	Packet	Frame
Network interface	Channel adapter	Endpoint	Network interface card

Table 2. *Terminology used by IBA, ASI, and Ethernet.*

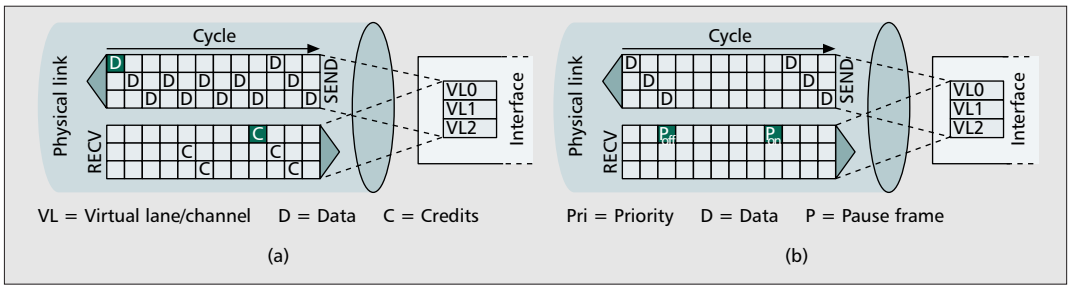


Figure 1. *Flow control in a) IBA and ASI; b) Ethernet.*

The use of flow-control ensures that packet loss is only a result of link-transmission errors and hence the available link bandwidth is used effectively as retransmissions are not necessary.

As IBA is a layered networking technology, flow-control is performed per *virtual lane* (channel). The concept of virtual lanes allows a physical link to be split into several virtual links, each with its own buffering, flow-control, and congestion management resources. Figure 1a shows an example of per virtual lane (VL) credit-based flow-control where VL0 runs out of credits after cycle 1 (depicted by a bold D) and is unable to transmit until credit arrives in cycle 9 (depicted by a bold C). As the other lanes have sufficient credit they are unaffected and are able to use the slot that VL0 would otherwise use. Transmission resumes for VL0 when credit arrives.

The concept of virtual lanes also allows for grouping of virtual lanes as layers, thus making it possible to build virtual networks on top of a physical topology. These virtual networks, or layers, can be used for various purposes such as efficient routing, deadlock avoidance, fault tolerance, and service differentiation. Virtual lanes and service differentiation are discussed further below.

ASI — ASI also uses a link-by-link credit-based flow-control mechanism to prevent packet drops as the status of a downstream ingress queue is reflected in its upstream neighbor. Credit is updated periodically and, if needed, additional replenishment is sent, as frequent transmitters will consume their allocated credit quicker than intermittent ones.

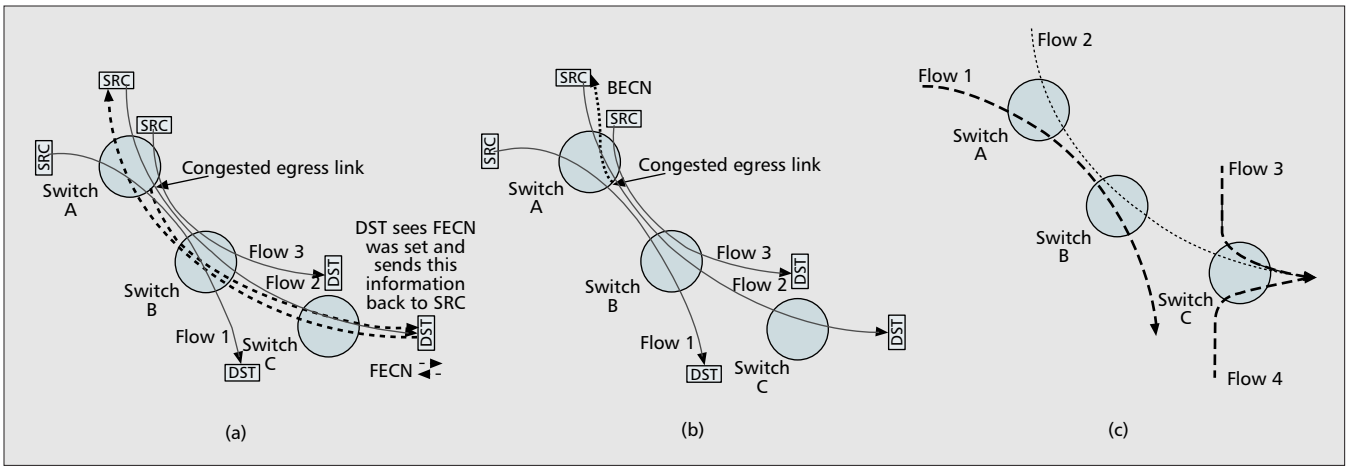
One of the limitations of a lossless network is that it can lead to a deadlock situation when a number of resources are in a circular dependency. A typical deadlock scenario is when two nodes that initiate a request/response transaction with each other are unable to complete the transaction because the response is blocked by the other request. ASI has an end-to-end application deadlock avoidance scheme built into its queuing structures in such a way as to break this dependency, ensuring that one of these operations can always bypass the other. Typically, certain types of PCI operations necessitates the use of such a mechanism. This deadlock avoidance is achieved by using a *bypassable* queue that maintains flow-control credit for two types of packet, *bypassable-ordered* and *bypassable-bypass*, where a bypassable-bypass is not allowed impede the progression of a bypassable-ordered packet.

During operation, when credit is abundant for both bypassable-bypass and bypassable-ordered packets, the queuing structure acts like a FIFO, but when credit reserves for the bypassable-bypass are depleted, packets in the bypassable-ordered queue with sufficient credit are allowed to pass the stalled packets until such time when bypassable-bypass credit is reallocated. These stalled packets are then given strict priority, as they have already been subjected to a delay.

It is expected that fabric management and some PCI-based operations will use the bypassable queue, while another credit-controlled unicast FIFO, the *ordered* queue, is available for other types of traffic.

Ethernet — The Ethernet link-level flow-control mechanism was specified for use within lossless networks. Flow-control is time-based, in the sense that the receiving node sends an explicit *off-message*, directing the sender to stop all transmissions for a certain amount of time so as to avoid swamping the receiver with frames that it cannot buffer. If the transmission of frames can be resumed before the time specified in the off-message has run out, an *on-message* is sent. This approach is different from the credit-based flow-control principle applied in ASI and IBA technologies, where the control-messages sent upstream reflect available buffer space in the downstream node.

In Ethernet, xon/xoff messages come in the form of a *pause frame* that includes a *pause-time*.



■ **Figure 2.** *Congestion management mechanisms: a) FECN; b) BECN; c) SBFC.*

The pause-time is the time that the upstream node must wait before sending the next frame. For the exchange of pause frames to work correctly, the messages must take into account the fact that there is a delay between the transmission of pause frames and their activation. This delay is a function of the propagation and processing time. It should be noted that, in order to sustain the link bandwidth, xon/xoff requires $2 \times N$ buffer space, where credit-based flow-control requires N buffer space.

The current Ethernet flow-control is limited to flow governance on a per-port basis, which undermines effective QoS provisioning when flow-control is turned on. Figure 1b shows Ethernet priorities in combination with flow-control. Since the flow-control mechanism is port-wide, the lack of buffer space for one priority affects the whole link, and any P_{off} message shuts down the link, while any P_{on} message restarts the link. We will further discuss the consequences of this below.

Congestion Management

IBA — As flow-control is a mechanism to avoid packet loss due to buffer overflows, congestion management is a mechanism to aid switches and links in the network from becoming overloaded and depleting their credit supplies. When congestion develops, a depletion of credit supplies starts and the queues begin to fill up. This process spreads upstream through the network and results in the creation and growth of congestion trees which eventually terminate at the end-nodes. Obviously, this is a bad situation for a network, as the growth of congestion trees can quickly preclude transmission of other flows (in the same virtual lane) that are not even destined towards the congested area.

IBA supports three mechanisms to control congestion. These are forward explicit congestion notification (FECN), static rate control, and a head-of-queue drain mechanism.

Forward explicit congestion notification is used to inform a packet's destination that it was subjected to congestion while traversing the network. This is achieved by setting an FECN flag in the packet's header. This flag is observed by the destination, which can signal the source about congestion either by sending a congestion notification packet or, when acknowledgments are used, by setting the congestion flag in the next acknowledgment for the packet in question (see Fig. 2a for an illustration of FECN). In this figure, the link from switch A to switch B is oversubscribed. The FECN flag is set at this point and when the destination sees a packet that has this flag set, it sends the FECN status to the source. For the sake of clarity, we only show FECN for one flow.

A switch also has the ability to inform a source end-node of various levels of congestion through the use of up to 16 different congestion thresholds. A switch can identify itself as either the root (cause) of congestion or as a victim of congestion. If a virtual lane has exceeded a given buffer occupancy threshold and has available credits, the switch considers itself the root of congestion. Otherwise, if the virtual lane has exceeded a given threshold and is out of credits, it considers itself a victim of congestion.

Static rate control is used at the end nodes to reduce the injection rate after receiving either a congestion notification or an acknowledgment with the congestion flag set. The arrival of successive congestion notifications leads to further reductions, while a subsequent rate increase is based on a time-out that is relative to the latest congestion notification. The available static rates are selected from a congestion-control table, where each entry defines an acceptable injection rate.

IBA also supports a head-of-queue drain mechanism that ensures a switch queue is drained after a timeout. This is useful in the case where a destination has stopped consuming packets.

ASI — Congestion as a result of credit depletion is typically a transient condition, so ASI attempts to handle it using a localized congestion-control mechanism known as status-based flow-control (SBFC). The use of SBFCs allows a downstream switch to signal the congestion status of one of its egress ports to its upstream neighbor, thus allowing it to change its scheduling mechanism so that packets destined for the congested (downstream) port are given a lower priority. As a result, congestion will only have an impact on packets contributing to the congestion while packets heading for an uncongested port will remain unaffected. The use of SBFCs is illustrated in Fig. 2c, where the onset of flows 3 and 4

Even though end-to-end congestion notification is not explicitly supported within ASI, a regular ASI packet has an FECN field that can be used to inform a destination end-node that, while it was en route, it was subjected to congestion.

traversing switch C results in congestion in switch B affecting the throughput of flow 1 (that is not even destined towards the congestion). The use of SBFCs provides switch A with a one-hop look-ahead with regards to congestion, making it possible to lower scheduling priorities for packets destined for switch C.

The SBFC procedure should be invoked in advance of credit depletion so as to prevent or limit the growth of congestion trees. SBFCs are issued to order a total cessation of traffic or for limited time periods.

End-nodes also play a role in managing congestion as, ultimately, they contribute to the problem. Traffic injection into the fabric is limited by the use of connection queues with an associated injection rate. These connection queues can isolate traffic by class, destination, and chosen route. Each queue has an associated token bucket that controls the average rate of injection while also supporting a limited number of bursts.

Even though end-to-end congestion notification is not explicitly supported within ASI, a regular ASI packet has an FECN field that can be used to inform a destination end-node that, while it was en route, it was subjected to congestion. This information could be used by the destination node to signal the source node, requesting it to limit its injection rate for the given flow.

Ethernet — Currently the Ethernet standard does not offer any congestion-management mechanisms but, with the development of Backplane Ethernet, congestion-management techniques are under investigation by the IEEE 802.3ar Congestion Management Task Force. This work has not been completed as yet, but one of the goals is that congestion-notification messages are targeted directly to the end-nodes causing the congestion.

One solution under consideration is the use of a backward explicit congestion notification (BECN), which has a shorter control loop than the IBA/ASI FECN approach since there is no need to go all the way to the destination before returning a congestion notification to the source. In Fig. 2b we illustrate the use of BECN where a message is sent to the source from the point where the congestion is observed, in this case, switch A. The inclusion of a mechanism to reduce traffic injection that is based on the degree of congestion is also under consideration.

Congestion management supported at the link level ensures that Ethernet also remains agnostic with respect to higher-layer protocols, which is an important issue for interconnect protocols.

Service Differentiation

IBA — We have already seen IBA's support for virtual lanes, a key feature for supporting service differentiation. The independent resources dedicated to each virtual lane in combination with IBA's mechanisms for differentiation between lanes forms the core of IBA's QoS capabilities. Certain traffic may be assigned a low-priority best effort, for example, while others may require strict latency and jitter guarantees.

The three key mechanisms to achieve service differentiation in IBA are service level to virtual lane mapping, the weighting of virtual lanes, and their classification as either high or low priority. A service level denotes the type of service a packet receives as it travels toward its destination, which is similar to the packet marking approach used in Differential Services as specified by IETF. Since there is not necessarily a one-to-one relation between the number of virtual lanes and service levels, a table is kept which contains a mapping from one to the other. IBA supports a maximum of 16 virtual lanes and service levels, where lane 15 is a control-lane dedicated to management traffic.

Apart from the control-lane, which has strict priority over everything else, other virtual lanes can be classified as either having high or low priority. Thus, by assigning a packet to a certain service level and setting the service level to map to a particular virtual lane, packets can be classified with either a high or low priority. High-priority traffic will preempt low-priority traffic, but in order to ensure forward progression of low-priority packets, a parameter called limit of high priority (LHP) is used. The LHP is the maximum number of packets that can be scheduled on high-priority lanes before a packet must be selected from a low-priority lane.

Arbitration between individual virtual lanes of the same priority is carried out using a weighted fair arbitration scheme. Each virtual lane is scheduled in table order and assigned a weight indicating the number of bytes it is allowed to transmit during its turn.

ASI — ASI supports differentiated services using traffic classes that isolate flows into various classes. Eight traffic classes are mapped to virtual channels, enabling flow prioritization throughout the fabric. Up to 20 virtual channels are supported: eight ordered, eight bypass, and four multicast. Fabric management messages must use the bypassable queuing structure and assigned the highest priority (TC7). These TC7 packets are given strict priority during scheduling.

ASI supports two different egress scheduling mechanisms: table-based and minimum-bandwidth. The table scheduler is a legacy PCI Express scheduler; while a minimum bandwidth scheduler allows for differentiated services for the various virtual channels while also providing a minimum share to each one to avoid starvation, usually by using a variant of weighted fair queuing. The ASI specification recommends the use of the minimum bandwidth scheduler.

During link negotiation, both processors negotiate the number of virtual channels to be employed where the lowest common number of virtual channels will be chosen and processors may adapt their queuing structures to ensure compliance.

Ethernet — In 1998 Ethernet was extended with a priority mechanism in order to support differentiated services. It is based on priority tagging of packets and an implementation of multiple queues within the switches in order to discriminate packets based on eight different levels of priority. The standard specifies Strict Priority Queuing (SPQ), but WFQ is also supported by most vendors.

There is an obvious conflict between the Ethernet priority tagging concept and the use of a port-based flow-control mechanism. Recall that the flow-control (the pause frame) is non-discriminatory. As it is port-based it will "shut down" the

link without any regard to the priority of the traffic (illustrated in Fig. 1b). This causes a situation where congestion spreads, which may result in service degradation as both low and high priority traffic will be subjected to an increase in latency and jitter. The relevant standardization bodies are responding to this problem with a suggestion that the Ethernet flow-control concept is extended to support a granularity similar to the priority mechanism (i.e., eight levels) realizing class based flow-control. This is implementable using some reserved fields in the MAC control frame.

Another consequence of Ethernet's port-based flow-control is that it cannot offer virtual layer networking for the purpose of effective deadlock-free routing and performance enhancements, as opposed to IBA and ASI [8]. The concept of virtual layer networking should not be confused with Ethernet's VLAN (Virtual Local Area Networking) feature, which is primarily a mechanism for assigning end-nodes in VLAN groups (identified by the VLAN tag) limiting the broadcast domain [9]. VLAN can also be used to implement multiple spanning trees in an Ethernet network yielding performance gains compared to single spanning tree routing, as it shuts down fewer links [10]. However, since each VLAN does not have a separate queue the union of multiple spanning trees may deadlock when flow-control is turned on.

QoS Capabilities Assessment

The definition of QoS is very broad and interpreted differently dependent on the application domain, but in terms of interconnection networks, four properties are central; lossless operation, effective bandwidth distribution, minimum latency and minimum jitter.

A network should operate fairly and efficiently. Fairness is achieved by distributing the bandwidth based on service classes while efficiency equates to keeping latency and jitter low, by preventing the onset of congestion. An inescapable fact is that packets are subjected to delay while traversing a network, as packets are queued and scheduled according to various priorities.

A network becomes over-subscribed as a result of the generation of excess traffic at its edge. This traffic can result in the onset of congestion and mechanisms should be in place to limit performance degradation and prevent saturation. Typically, this over-subscription is transient but it can have a detrimental effect on performance.

We have seen some of the capabilities that the technologies have to deal with these issues, bandwidth distributed between varying classes allows for segregation of traffic as the onset of congestion in one does not affect the other. The three technologies have various mechanisms in place with regard to congestion. Common amongst these mechanisms is that they attempt to reduce delay and jitter by preventing the network from reaching a saturated state.

A loss-less network reduces the need for retransmissions and, when compared to lossy networks, we can expect increased network utilization but at a cost of congestion tree growth in the face of congestion.

Table 3 summarizes the QoS features available in the three technologies. They are all capable of supporting lossless networking, IBA and ASI deploy a credit-based flow-control concept and Ethernet an XOn/XOff technique. Although the XOn/XOff scheme produces less control traffic than credit-based flow-control, it requires twice the buffer space. Apart from this the differences are negligible, and the three technologies can be considered equal with regards to lossless operation.

Feature	IBA	ASI	Ethernet
Flow control	Credit-based	Credit-based	XOn/XOff
Congestion control	FECN	FECN, SBFC	Under review
Rate control	Static	(Semi) Dynamic	Under review
Service classes	16	8	8
Virtual channels	16	20	8
Priority scheduling	WFQ	Table, MinBW	SPQ, WFQ (optional)

Table 3. *Quality of service features of IBA, ASI, and Ethernet.*

One side effect of flow-control is that, when a link runs out of buffers or is paused, a congestion tree will spread upstream from the point of congestion, reducing throughput and increasing latency for even the smallest flow. Both IBA and ASI support forward explicit congestion notification (FECN), which allows switches to notify end-nodes about congestion. Furthermore, IBA supports static rate control that enables end nodes to reduce injection rates based on predefined congestion thresholds and static rates.

ASI supports a parameterized token bucket algorithm for injection rate control as well as an SBFC mechanism to handle transient congestion. This gives a switch a one hop look-ahead when making its scheduling selection, allowing it to hold back packets which are likely to experience congestion downstream. In Ethernet congestion control is nonexistent, but ongoing work is needed to add this feature in the future. Currently, backward explicit congestion notification (BECN) with rate reduction, as well as end-node injection rate control, is under consideration. The BECN method improves upon FECN by introducing a shorter control loop, as the notification message goes directly from a switch to the source end node, without traveling via the destination. A dynamic injection mechanism is possible with the inclusion of information regarding the state of the observed congestion. However, until this is a reality, Ethernet has no support for congestion control.

Arguably, the most important feature needed is service differentiation, which allows two (or more) types of traffic to be treated differently. In IBA this is supported by a combination of service levels and virtual lanes, which can be categorized as high or low priority. IBA supports a total of 16 service levels that can be mapped to a maximum of 16 virtual lanes. The use of weighted fair queuing allows for a fair allocation of bandwidth, while high and low priorities make it possible to separate latency-sensitive traffic from other traffic. Furthermore, IBA supports strict

As with most technology, there is no "one glove fits all" answer to which is the best; rather, the applicable problem area should govern the choice of technology. Fundamentally, previous investment and familiarity with the technology will play a big part in this decision.

priority scheduling of high-priority traffic when it is deemed necessary, but this should be subject to admission control so as to avoid starvation.

ASI supports eight traffic classes, which can be mapped to 16 unicast virtual channels: eight ordered and eight bypassable. Each channel is supported by a weighted fair-queuing scheme in the form of the minimum bandwidth scheduler, but makes no distinction (except for fabric management messages) between high and low priority apart from the weights.

Ethernet supports up to eight priorities but its use is limited by the coarse flow-control. As described above, Ethernet's pause-based flow-control is port-based, which makes it impossible to exert flow-control without penalizing all eight priorities. This should be remedied in the future as per-priority flow-control is currently under consideration by the 802.3ar task force. But until this work is realized, Ethernet's service differentiation support is poor as compared to the offerings from IBA and ASI.

Meeting latency guarantees in lossless networks is difficult because of the effect of back-pressure. The congestion trees created by back-pressure increase latency for all packets that are part of the tree and, as the network size increases, the theoretical upper bound of latency grows exponentially [11]. A combination of service differentiation, congestion management, and efficient switch architecture contributes greatly to reducing this problem.

Conclusions

We have presented an overview of IBA, ASI, and Ethernet with a focus on their QoS features, followed by a short discussion of their strengths and weaknesses. The main features of each technology are summarized in Tables 1 and 3. It should come as no surprise that Ethernet has the weakest support for QoS, but as this is currently subject to attention, it could change in the near future. IBA and ASI have a rather similar feature set, with ASI supporting an advanced congestion management scheme, while IBA has a great deal of flexibility when it comes to configuration for service differentiation.

The success of these technologies as interconnects is dependent on many factors, and not all are related to their performance abilities. Currently, Ethernet has some QoS limitations that may impede its attempt to break into the intrasystem market, but these issues are under review. ASI has evolved from the PCI intrasystem to an intersystem interconnect, a fact which gives it a lot of strength but is no guarantee for further adoption. IBA, however, has found its niche in storage and high-performance computing, thus making it hard for other technologies without equal or superior features to make an impact. But, as with most technology, there is no "one glove fits all" answer to which is the best; rather, the applicable problem area should govern the choice of technology. Fundamentally, previous investment and familiarity with the technology will play a big part in this decision.

References

[1] F. J. Alfaro, J. L. Sanchez, and J. Duato, "Qos in Infiniband Subnetworks," *IEEE Trans. Parallel and Distrib. Sys.*, vol. 15, no. 9, Sept. 2004.

[2] S.-A. Reinemo *et al.*, "Admission Control for Diffserv-Based Quality of Service in Cut-Through Networks," T. M. Pinkston and V. K. Prasanna, Eds., *Proc. High Performance Computing — HiPC 2003: 10th Int'l. Conf.*, Hyderabad, India, Dec. 2003, LNCS, Springer, vol. 2913, pp. 118–29.

[3] InfiniBand Trade Assn., "Infiniband Architecture Specification," 1.2 ed., Oct. 2004.

[4] ASI Special Interest Group, "Advanced Switching Core Architecture Specification," rev. 1.1, Aug. 2004.

[5] D. Mayhew and V. Krishnan, "PCI Express and Advanced Switching: Evolutionary Path to Building Next Generation Interconnects," *11th Symp. High Perf. Interconnects*, 2003.

[6] IEEE 802.3-2002, "LAN/MAN CSMA/CD Access Method," 2002.

[7] ANSI/IEEE Std 802.1D, "Media Access Control (MAC) Bridges," 1998.

[8] O. Lysne *et al.*, "Layered Routing in Irregular Networks," *IEEE Trans. Parallel and Distrib. Sys.*, vol. 17, no. 1, 2006, pp. 51–65.

[9] IEEE 802.1Q-2003, "Virtual Bridge Local Area Networks," 2003.

[10] S. Sharma, "Viking: A Multispanning-Tree Ethernet Architecture for Metropolitan Area and Cluster Networks," *INFOCOM 2004*, vol. 4, Mar. 2004, pp. 2283–94.

[11] S.-A. Reinemo, T. Skeie, and O. Lysne, "Applying the Diffserv Model in Cut-Through Networks," *Proc. 2003 Int'l. Conf. Parallel and Distrib. Processing Techniques and Apps.*, June 2003, pp. 1089–95.

Biographies

Sven-Arne Reinemo (svenar@simula.no) recieved a Cand.Sci.. degree in computer science from the University of Oslo in 2000. He is currently a Ph.D. student at Simula Research Laboratory and the University of Oslo, where he is participating in the ICON project. His current research focus is on quality of service in interconnection networks.

Tor Skeie (tskeie@simula.no) is an associate professor at Simula Research Laboratory and the University of Oslo. He received an M.S. degree in computer science in 1993 and a Ph.D. degree in computer science in 1998, both from the University of Oslo. He has several years of experience as a researcher in the interconnect domain. His work mainly focuses on scalability, effective routing, fault tolerance, and quality of service in switched network topologies.

Thomas Sødring [M] (tsodring@simula.no) is a postdoctoral researcher at Simula Research Laboratory working on the ICON project. He received a B.S. degree in computer applications in 1998 and a Ph.D. degree in computer applications in 2002, both from Dublin City University. His primary research interests are now within the area of interconnects and the development of generic solutions for routing, fault tolerance, and quality of service. He is a member of the ACM.

Olav Lysne [M] (olavly@simula.no) is a research director at Simula Research Laboratory and a professor in computer science at the University of Oslo. He received an M.S. degree in 1988 and a Dr.Sci. degree in 1992, both from the University of Oslo. His early research was in the field of algebraic specification and term rewriting. In recent years he has mainly been working in the area of interconnects, focusing on effective routing, fault tolerance, and quality of service. In this field he has been a member of the program committees of several of the most renowned conferences, has participated in a series of European projects, and has published around 70 academic papers.

Ola Tørudbakken [M] (ola.torudbakken@sun.com) is a senior staff engineer for Sun Microsystems Inc., Oslo, Norway. He received an M.S. degree from the University of Oslo, Department of Informatics in 1994, and since then has been working with high-performance interconnects, switch fabrics, networking, and server systems. From 1994 to 1996 he worked as a research scientist at the Sintef Research Center, Oslo, Norway. He worked at Dolphin Interconnect Solutions from 1996 to 2000, where he held various R&D positions. He currently holds one U.S. patent, and has more than 16 U.S. patents pending. He has published several papers in the field of interconnection networks and has participated in several standardization activities.