

Providing Quality of Service over InfiniBand™ Architecture Fabrics

Joe Pelissier
Intel Corporation
5200 N. E. Elam Young Parkway
Hillsboro, OR 97124-6497

Abstract -- The provision of Quality of Service (QoS) in data communication networks is currently the center of much discussion and research in the industry. InfiniBand™ Architecture (IBA) enables QoS support through a rich set of mechanisms to segregate traffic flows into traffic classes and to provide hop level forwarding control over these individual classes. This paper describes the QoS problem and describes a generalized approach to its solution. The IBA mechanisms applicable to QoS are described in relation to the generalized model. Finally, the Differentiated Services architecture, as described in RFC 2475, is introduced and an overview of a possible implementation of DiffServ over IBA is described.

I. THE QOS PROBLEM

Providing quality of service in modern data communication networks is a concept that has grown have many connotations. However, to a large extent, the problem can be divided into four broad categories of traffic types and two significant network parameters.

Overall, the types of traffic that transverse data communication networks can be divided into four categories: dedicated bandwidth time sensitive (DBTS), dedicated bandwidth (DB), best effort (BE), and challenged (CH).

DBTS traffic is that which requires a given minimum bandwidth (and is often fairly constant in its actual bandwidth usage) and which must be delivered within a given latency in order for the data to be useful. Examples of such data streams include video conferencing and interactive audio such as Voice over IP (VoIP). The key differentiation of this class of traffic is the timeliness in which the traffic must be delivered (on the order of 10's of milliseconds) is typically known by the application and must be communicated to the network. In either example, if the latency becomes excessive, the value of the communication diminishes.

Dedicated bandwidth traffic is that which requires a given minimum bandwidth but is not particularly sensitive to latency. This includes traffic that must have a bounded latency but where the actual value of the bound is not critical. Playback of a video clip from a server is an example of such a class of traffic. Note that in this example, having a bounded latency is important for the playback application (so that sufficient local buffering may be allocated for smooth playback). However, the quality of the playback is insensitive to the latency so long as the latency bound is known (the quality of the playback is identical regardless of whether the latency is 10ms or 10 seconds). One can argue that this does imply a latency sensitivity with respect to the buffering capability of the playback device; however, given the relative cost of memory and the performance of modern data communication networks, this argument is valid for a very small application set.

Note that there is a subtle but very important difference in the latency requirements of DBTS and DB traffic classes. DBTS requires a given latency that is generally known by the application and must be communicated to the network. The latency for DB traffic may be communicated by the network to the application since all the application requires is any reasonable bound. This distinction for DB traffic is important for two reasons. First, many DB applications actually have no rational basis on which to request a given latency. And second, enabling the network to provide the latency value permits greater flexibility to the bandwidth/latency reservation mechanism to better utilize existing network resources.

The third general class of service is best effort. This accounts for the majority of traffic handled by data communication networks today: file and printing services, web browsing, etc. This traffic tends to be bursty in nature and largely insensitive to both bandwidth and latency.

The final class of service is challenged. This includes traffic sources that, for administrative reasons, are intentionally degraded so as not to interfere with best effort traffic. An example of such traffic could include disk backup activities. Using such a class of

service, it is possible to perform disk backup during normal business hours without impacting production applications.

Reviewing the above four classes, two significant network parameters emerge: bandwidth and latency. Jitter and packet loss are also frequently included as significant network parameters. However, for most DBTS traffic, if the latency is successfully bounded, then, by definition, the jitter is also sufficiently bounded (since jitter must be less than the latency). There are two major sources of packet loss: that from random bit errors and that from congestion. In modern data communication networks, the random bit error packet loss is generally insignificant for DBTS and DB traffic, and for other traffic types is eliminated using reliable transports (such as TCP). Packet loss from congestion may be eliminated for DBTS and DB traffic simply by preventing congestion using a bandwidth reservation mechanism.

Therefore, the major requirements for supporting QoS may be summarized as follows:

- DBTS traffic: ensure that the network will deliver the bandwidth and latency requested by the application.
- DB traffic: ensure that the network will deliver the bandwidth requested by the application within the latency provided by the network (i.e. the bandwidth/latency mechanism)
- BE traffic: ensure that network bandwidth not utilized by DBTS and DB traffic is efficiently made available to BE traffic.
- CH traffic: ensure that any remaining network traffic not utilized by the other classes is efficiently made available to CH traffic.

II. A GENERALIZED APPROACH TO ADDRESSING QOS

The requirements stated in the previous section may be addressed with three major components: a bandwidth/latency reservation mechanism (BLRM), a mechanism to label data (in general, packets) identifying a class of required forwarding behavior, and a mechanism to provide the identified forwarding behavior.

The BLRM ensures that the fabric maintains sufficient resources to meet its current bandwidth and

latency commitments, and to either accept or reject requests for new commitments. An example of such a mechanism is Resource reSerVation Protocol [1].

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 include the priority field in 802.3 [2], the TOS field in IPv4 [3], the TClass field in IPv6 [4], and the DS byte in DiffServ [5][6]. More sophisticated devices may implement “implicit labeling”, i.e. determining the class of service indirectly based on other fields in the packet.

The mechanism for providing the specified forwarding behavior may be as simple as independent sets of receive and transmit queues on each port of network forwarding elements. Packets are placed in queues based on the label mechanism described above. Packets are transmitted from the higher priority queues ahead of lower priority queues. More sophisticated queue arbitration may be implemented, e.g. weighted round robin; however, a simple priority scheme is sufficient for basic QoS.

In the general case, the BLRM maintains state for each link in the fabric indicating the total committed bandwidth (potentially including certain traffic characteristics such as burst size and rate, etc.) and the available bandwidth on each link for DBTS and DB traffic.

Refer to Fig. 1 one for an example of how the BLRM might operate For this example, assume that each of the two switches supports four independent queues (0 through 3) on each port. Further assume that each queue operates with absolute priority with respect to the others, with queue 0 being the highest priority. Further assume that each link is capable of a total bit rate B , that a given quantity of that bit rate, B_d , is reserved for DBTS and DB traffic where $B_d < B$. At the start of time of this example, there is no traffic operating at DBTS or DB.

Now, lets assume an application on node 1 requests a DBTS communication transmitting to node 4 with a bandwidth b_1 and a latency of l_1 . This request is sent to the BLRM, which may either approve or deny the request. The BLRM reserves priority 0 for this DBTS.

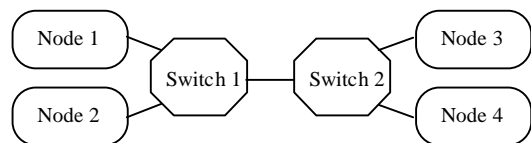


Figure 1: Example Network

Since this is the highest priority available, the BLRM can assume that so long as the sum of all the reserved bandwidths for DBTS, Σb_i , is less than B_d for each link, the bandwidth request may be honored.

Likewise, the maximum latency along the path may be computed. This is simply the sum of the latency through each switch (and potentially the flight delay across each link; for simplicity, this is assumed to be insignificant). To do this, the "burst time" of each flow must be known. The burst time is the length of data (in time) that a given flow will transmit in a continuous stream (it is assumed that the burst size, in bytes, and frequency of bursts are provided as the specification of bandwidth requested; from this, burst time is easily computed). The total latency that a given DBTS flow through a switch may encounter is the total of the burst lengths of all other DBTS flows through the destination port plus the time required to transmit a maximum sized packet. (This covers the case that a lower priority packet has just begun transmission at the same time a higher priority packet is received). If the total path latency is less than l_1 , then the latency request may be honored.

Assuming that the above two conditions are met, in addition to any other administrative policies that may be implemented, the BLRM may approve the request. If either of the conditions cannot be met, then the BLRM must deny the request. Once approved, the application may commence its communication marking each packet appropriately to use queue 0.

For subsequent DBTS requests, the same process takes place. In addition, the BLRM must verify that accepting the new request will not invalidate any previously made commitments.

A similar process is used for DB requests. In this case, the BLRM assumes that all GB traffic queue number 1 (i.e., the second highest priority). For calculating available bandwidth, both the DBTS and DB totals are included. Unlike DBTS, the BLRM provides the latency parameter, which may be simply administratively assigned. The BLRM calculates the worst case latency that the requesting flow would encounter similar to the DBTS case except that both DBTS and DB flows are included in the calculation. Furthermore, as new DB flows are requested, the BLRM must ensure that previous DB commitments are not affected. Assuming that all of these conditions are met, the BLRM may approve the request.

It is important to note that when considering a DB request for approval, the BLRM does not need to check for possible impact on DBTS flows. Since the DBTS

flows are transported on higher priority queues, they are unaffected by DB flows.

The BE and CH flows are simply assigned to queues 2 and 3, respectively. Interaction with the BLRM is not required. These flows will be guaranteed a minimum bandwidth, $B - B_d$, and will additionally consume any bandwidth unused by the DBTS and DB flows, if needed.

III. IB MECHANISMS TO SUPPORT QoS

Summarizing the previous section, a QoS solution may be constructed from three major components:

- A Bandwidth/Latency Reservation Mechanism
- A mechanism to label packets within a flow indicating a particular class of service
- A mechanism to provide forwarding behavior appropriate to the class of service requested

IBA does not specify a particular Bandwidth/Latency Reservation Mechanism. There is currently significant work being done in the IETF with respect to RSVP and SBM. It is thought that these mechanisms are likely applicable either directly or with some minor adjustments. Since these protocols are generally independent of the underlying transport, there seemed little value in IBA developing yet another method.

Before discussing the remaining two items a brief discussion of the structure of IBA communication is in order. IBA's 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 with the exception of the last packet of a message (which carries the residual). The negotiation process of path MTU is beyond the scope of this paper; suffice to say that the valid outcomes are powers of 2 between 256 bytes and 4096 bytes, inclusive. The most common path MTU's are likely to be 256 bytes and 2048 bytes.

IBA supports hierarchical packet forwarding with two layers of hierarchy. Switches perform the lower layer forwarding. A fabric 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.

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 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. The source of packets may either set the TClass field or routers may set the TClass field based on the SL field. As is the case with SL, there is not an IBA defined relationship between TClass and specific forwarding behaviors. (Note: TClass was defined as eight-bits to ease integration between IBA and IP protocols; one particular use of this is described later in this paper).

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

IBA switches may implement between one and 15 virtual lanes. A virtual lane is an independent set of receive and transmit resources (i.e. packet buffers) associated with a port. In addition to SL, the LRH contains a field (VL) 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. (The mapping also permits use of virtual lanes for functions other than QoS, e.g.

deadlock avoidance in loop topologies). Note that an implication of this is that the VL indication in a packet may change from hop-to-hop; the SL indication remains constant within a subnet (but may vary between subnets as will be discussed shortly). Note that packets within one virtual lane may pass packets in another virtual lane as they transit a switch.

The virtual lane mechanisms for IBA routers are not currently fully defined; however, it is envisioned that the mechanism will be similar to that of switches. In general routers use the TClass field (and potentially other parameters) in the GRH to determine which virtual lane to use on the outbound port. Since the meaning of SL is subnet in scope, it is envisioned that routers will reassign the SL as a packet transits from one subnet to another based on TClass. TClass becomes the invariant indication of traffic class end-to-end.

Virtual lane arbitration is the mechanism by which an output port selects from which virtual lane to transmit. IBA specifies a dual priority weighted round robin 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).

(IBA also specifies a mechanism to ensure that the high priority virtual lanes are not able to completely prevent the low priority virtual lanes from transmitting. However, this mechanism is not relevant to this discussion.

The final component is 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, communication 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.

IV. APPLYING THE IBA MECHANISMS TO THE GENERALIZED QoS APPROACH

The mapping of the IBA mechanisms to the Generalized QoS approach is straightforward. Four service levels are selected (one each for each time of service: DBTS, DB, BE, and CH). For this example, service levels zero through three respectively are used; however, the selection is arbitrary. For routing between subnets, TClass values must be similarly chosen. Again, for simplicity of the example, TClass values zero through three are also chosen.

The SL to VL mapping tables in each switch are loaded such that the packet never changes virtual lanes. Note that since the virtual lanes are abstracted from the service level, it is possible to provide a degraded level of QoS through switches that do not support four virtual lanes. For example, on a two virtual lane switch, the DBTS and DB traffic may be aggregated together and the BE and CH traffic may be aggregated together. Upon reaching a switch with four virtual lanes, the traffic may once again be separated. Performing such an aggregation may impact the latency guarantee for the DBTS traffic; however, in many cases it is likely that the end-to-end quality will remain sufficiently high to retain the usefulness of the class.

In the generalized approach, absolute priority was assumed as the method of arbitration between the virtual lanes. IBA does not provide a mechanism to support four absolute levels of priority; however, the dual priority weighted round robin scheme not only is sufficient but also offers certain advantages. To implement the generalized approach, virtual lane zero (the virtual lane carrying the DBTS traffic) is assigned to the higher priority while all other virtual lanes are assigned to the lower priority. Weights are assigned to virtual lanes one (DB traffic) and two (BE traffic) in proportion to the bandwidth that is to be reserved for DB traffic compared to that for BE traffic. A small weight is assigned to virtual lane three for the challenged traffic. With this assignment, one can make the following observations:

- DBTS traffic will at worst wait for one packet to be transmitted from non-DBTS flows thereby ensuring the BLRM's ability to commit to a latency guarantee.
- DB traffic may need to wait for several BE and CH packets to be transmitted; however, the delay is bounded and easily calculated given the weights

assigned to each virtual lane. Again, this enables the BLRM to commit to a latency guarantee for DB traffic (although this latency may be greater than that achievable with a strict priority scheme).

- Even if the DB traffic sources abuse their bandwidth reservations attempt to inject data in excess of their reservations, the BE traffic is guaranteed to receive a minimum amount of bandwidth based on the relative weights between the DB and BE traffic. This is not achievable with the strict prioritization arbitration scheme.
- Finally, it is as possible to adjust just how "challenged" the CH class really is.

This example covers a very basic QoS implementation. IBA defined mechanisms to enable QoS, but specifically did not define the policies for utilizing those mechanisms. Given these mechanisms, it is possible to develop a rich set of QoS strategies to support a wide variety of applications.

V. IBA AND DIFFERENTIATED SERVICES

The Internet Engineering Task Force (IETF) is currently in the process of developing an architecture for providing QoS in the internet. This effort is referred to as Differentiated Services [2]. The mechanisms provided by IBA map well into this effort.

DiffServ's operation is largely analogous to the generalized approach presented above. DiffServ assumes the presence of a BLRM but does not specify one. An obvious candidate is ReSerVation Protocol (RSVP) [1] including subsequent related work.

IP datagrams are assigned to traffic classes referred to as Behavior Aggregates (BA). The DS field in the IPv4 and IPv6 header is used to label each datagram with the appropriate BA. (The DS field replaces the TOS field in IPv4 and the Traffic Class field in IPv6 per [5]).

A forwarding behavior is defined for each BA (there are currently two forwarding behaviors, in addition to best effort under consideration: Assured Forwarding [7] and Expedited Forwarding [8]). Each of these behaviors is referred to as Per-Hop Behavior (PHB). The PHB is mapped into the capability of the forwarding device.

Support for DiffServ over IBA is simple and direct. The Subnet Bandwidth Manager (SBM) component of RSVP is directly applicable to IBA subnets providing

the BLRM functionality. The DS value is mapped on a per-subnet basis to a service level. Like IP, the DS value is transported directly in the GRH TClass field. The service levels are mapped (via the SL to VL mapping table in each switch) at each hop to the virtual lane that is to handle the corresponding BA. The VL arbitration tables (assigning weights to virtual lanes and virtual lanes to one of the two priorities) are loaded to provide the appropriate PHB. Finally, routers examine the DS field directly to determine the appropriate PHB.

VI. SUMMARY

Provision of QoS in data communication networks involves the segregation of traffic into groups of flows requiring similar delivery characteristics. Of particular importance are bandwidth and latency guarantees. Once segregated and appropriately identified, the forwarding elements of a network may implement forwarding mechanisms that ensure the QoS commitments are sustained.

IBA provides these mechanisms. These include:

- The ability to mark individual packets with a value to indicate a class of service requirement
- The ability of forwarding elements (routers and switches) to identify the class of service and provide appropriate forwarding services.
- The ability, through link level flow control, to ensure that packets are not dropped due to congestion.

These mechanisms are sufficient to implement a wide variety of QoS strategies including strategies being developed as industry standards such as Differentiated Services.

REFERENCES

- [1] Braden, B., Zhand, L., Berson, S., Herzog, S. and S. Jamin, "Resource ReSerVation Protocol (RSVP) – Version 1 Functional Specification", RFC 2205, September 1997.
- [2] ISO/IEC Final CD 15802-3 Information technology – Telecommunications and information exchange between systems – Local and metropolitan area networks – Common specifications – Part 3: Media Access Control (MAC) bridges.
- [3] Almquist, P., "Type of Service in the Internet Protocol Suite", RFC 1349, July 1992.
- [4] Deering, S., and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", December 1998.
- [5] Nichols, K., Blake, S., Baker, F. and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.
- [6] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.
- [7] Heinanen, J, Baker, F., Weiss, W., and J. Wroclawski, "Assured Forwarding PHB Group", RFC 2597, June 1999.
- [8] Jacobson, V., Nichols, K., and K. Poduri, "An Expedited Forwarding PHB", RFC 2598, June 1999.