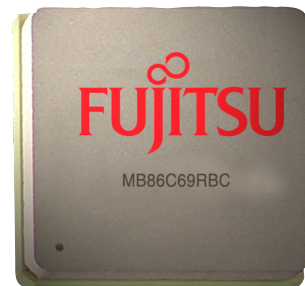


Essential Ethernet Switch Features for Data Centers

Congestion management, priority PAUSE, and other capabilities in the Fujitsu MB86C69RBC switch chip



**T E C H N O L O G Y
B A C K G R O U N D E R**

Introduction

As today's trends drive the demand for increasing bandwidth in data centers, Ethernet switches must respond with the ability to use the available bandwidth effectively. Congestion management is particularly important, along with advanced quality-of-service (QoS) features that enable the switches to maximize throughput for priority traffic.

This technology backgrounder profiles the capabilities of the Fujitsu MB86C69RBC Ethernet switch chip with a focus on data center requirements. To understand why these capabilities are vital, it is useful to begin with an overview of data center trends.

Data Center Trends Drive Bandwidth Requirements

An explosion in the popularity of applications such as IPTV, video streaming and gaming is pushing the demand for networking bandwidth and for optimizing servers using server virtualization and convergence of data fabrics. These trends are driving the need for 10-gigabit Ethernet (10 GbE).

Traditionally, servers have been dedicated to a single application such as web hosting, storage, database management or cluster computing. This usage model has led to a proliferation of physical servers, resulting in poor server utilization in the data center.

Virtualization seeks to improve server utilization by allowing multiple virtual machines to run on the same physical server. This consolidation of functionality into a single physical server has driven each server's bandwidth demands far beyond the capacity of a single 1-GbE link. The only solution is to increase the bandwidth to 10 GbE.

Also driving adoption of 10 GbE is a converged data fabric in data centers that has resulted from the adoption of blade servers. Data centers have traditionally had different fabric options optimized for specific applications. Infiniband supports high-performance computing applications, for example, while Fiber Channel works well for network storage, and Ethernet is ubiquitous in LANs. With data center complexity growing, however, managing these different interconnect technologies is becoming too costly.

Ethernet is the clear choice for a converged fabric in the data center that reduces costs, as prices for 10 GbE switch chips decrease year over year while performance and port density increase. To be viable for handling all data center tasks,

though, Ethernet needs shorter latency (to compete with Infiniband) and lossless transmission (to compete with Fiber Channel). Achieving these goals depends on overcoming the limitations of Ethernet's traditional best-effort-based protocols and use of static priority scheduling that can lead to starvation for lower-priority traffic.

Congestion management is essential to prevent congestion collapse. At the same time, the network fabric must ensure a fixed data rate for traffic types that are sensitive to delay, so Ethernet devices must guarantee a minimum QoS. Several Ethernet switch features can help meet these requirements: backward congestion notification, priority PAUSE, metering on the ingress port and shaping on the egress port.

This technology backgrounder outlines how these features work in the Fujitsu MB86C69RBC and the benefits these features offer for the data center of today and tomorrow.

Overview of MB86C69RBC Quality of Service Features

The Fujitsu MB86C69RBC Ethernet switch chip offers a variety of enhanced QoS features, including eight priority queues; programmable egress scheduling between strict and deficit round robin; ingress traffic metering; and egress traffic shaping. These features are essential for ensuring that traffic is properly prioritized while traversing the switch.

Figure 1 helps illustrate the QoS features by showing how an Ethernet packet passes through the switch. When the packet enters the switch, it goes into the input buffer for immediate classification. The input buffer allows classification in one of five scenarios – MAC based, DiffServ based, Extended VLAN based, VLAN based or port based.

Once the packet has been classified, it moves to shared memory. The transfer of the packet into shared memory is based on priority determined at the input buffer. Thus, higher-priority packets move into shared memory first, with lower-priority packets following.

Once in shared memory, the packet is scheduled to leave the switch using the egress scheduler. The classification assigned to the packet determines the packet's priority for leaving, so that higher-priority packets leave the switch first with lower-priority packets following. This scheme ensures that high-priority packets traverse the switch with minimum delay, while low-priority packets remain in the queue.

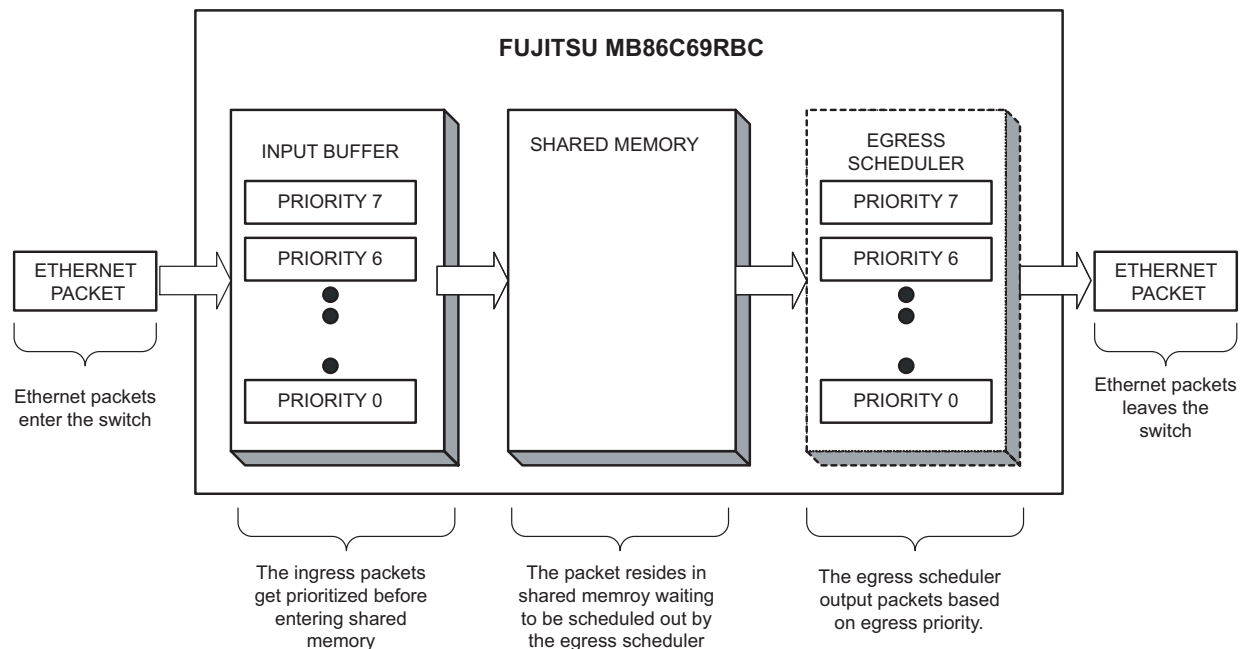


Figure 1 – Life of a Packet in the MB86C69RBC

Input Buffer

The Fujitsu MB86C69RBC provides input buffering per port, which can be divided up evenly or unevenly depending on user requirements (Figure 2). Input buffering is an important requirement for fabrics because, without proper input buffering packets that are in transmission when PAUSE, Priority PAUSE or BCN frames are sent to the endpoint will be dropped.

Each MB86C69RBC input buffer can be assigned a single priority or multiple priorities. This flexibility allows designers to allocate the appropriate input buffer for each priority.

Ingress Early Drop

As mentioned previously, today's data center often carries a wide variety of traffic, often with varying QoS requirements. The unique input queue scheme of the MB86C69RBC allows congestion avoidance in the network by enabling users to set watermarks at the input queue for low- and high-priority traffic. Thus, low-priority traffic can be dropped during times of extreme congestion while enabling the high-priority traffic to go through.

Figure 3 shows how watermarks can be set for each priority. In this example, the watermark for each priority is staggered with Priority 0 having the lowest watermark threshold and Priority 7

the highest. Thus, during extreme congestion, Priority 0 packets are dropped first, followed by Priority 1 traffic and so on until Priority 7 is reached.

The advantage of this scenario is that low-priority, best-effort traffic can be dropped before it has the chance of causing severe congestion in the network.

Metering

As a result of the converged fabric, multiple data streams will be moving over a single physical link. Thus, next-generation Ethernet switches will need to meter and shape traffic flows based on these different data streams, with each stream containing its own QoS requirements. The Fujitsu MB86C69RBC has the ability to meter ingress traffic moving from the input buffer to shared memory.

Referring to the packet flow illustrated in Figure 1, once the packet has been properly classified in the input buffer, it moves to shared memory. This transfer of the packet into shared memory can be metered on a priority basis, which allows higher-priority, latency-sensitive traffic to move into shared memory before lower-priority, latency-insensitive traffic.

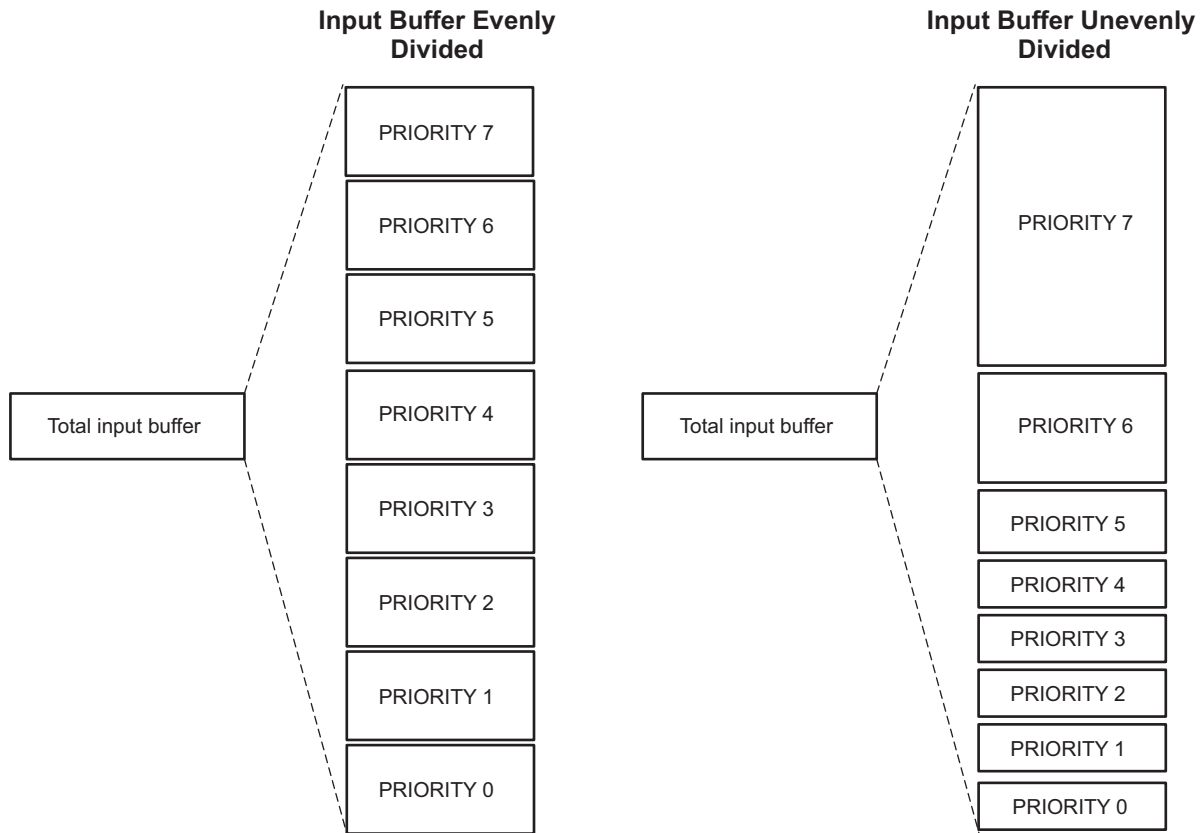


Figure 2 – MB86C69RBC Input Buffer

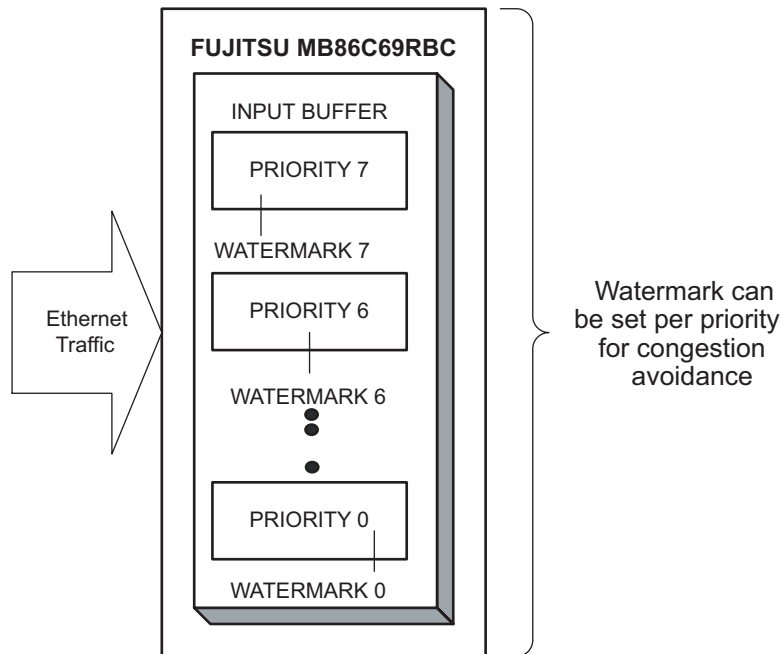


Figure 3 – Setting Watermarks on a Per Priority Basis on the MB86C69RBC

Figure 4 illustrates a scenario in which Priority 6 and Priority 0 traffic is metered at 75 and 25 percent levels, respectively, while Priority 7 traffic is not metered and allowed to move through the switch unimpeded. Metering the traffic guarantees that customer service-level agreements are met.

Egress Scheduler

While metering handles traffic at the ingress port, the scheduler shapes traffic at the egress port. Both are required to maintain a minimum QoS for the customer.

In the MB86C69RBC, packets leave the input buffer and then stay in shared memory until being serviced by the egress scheduler. This servicing of the packets depends on each packet's priority and the egress scheduling scheme chosen by the designer.

The MB86C69RBC has a large number of virtual output queues per port which gives designers the flexibility to configure how the egress scheduler serves each output queue group. The groups can be served in either a strict or deficit round robin, or some combination of the two scheduling schemes. By selecting deficit round robin, users can change the quantum value, thus shaping the traffic flows based on priority. In a typical scenario, higher-priority traffic would have more bandwidth reserved than lower-priority traffic, so the higher-priority traffic would leave the switch before the lower-priority traffic.

In addition to being able to shape the traffic at the egress port using deficit round robin, users can set an output-queue congestion drop threshold. This capability makes it possible to reserve bandwidth for high-priority traffic and drop lower-priority traffic in a congestion situation.

Similarly, the MB86C69RBC can detect congestion at the egress port and notify the endpoint that is responsible for the congestion in the network to back off traffic into the switch. The next section provides details of this capability and other congestion-management features.

Congestion Management

Congestion management is an essential feature for data-center Ethernet. Without it, the data centers are susceptible to congestion, which can result in a collapse. The Fujitsu MB86C69RBC has specific features to handle and prevent congestion in the network, including backward congestion notification and priority PAUSE.

Native Ethernet has traditionally left congestion management to the realm of higher-level protocols such as TCP, but this approach cannot provide a high-performance response to network congestion events. The transmission behavior of TCP and similar protocols can even exacerbate congestion in the network. Moreover, endpoints that run TCP drop their data rate according to when they receive a packet loss event, despite the fact that they may not be responsible for congestion in the network.

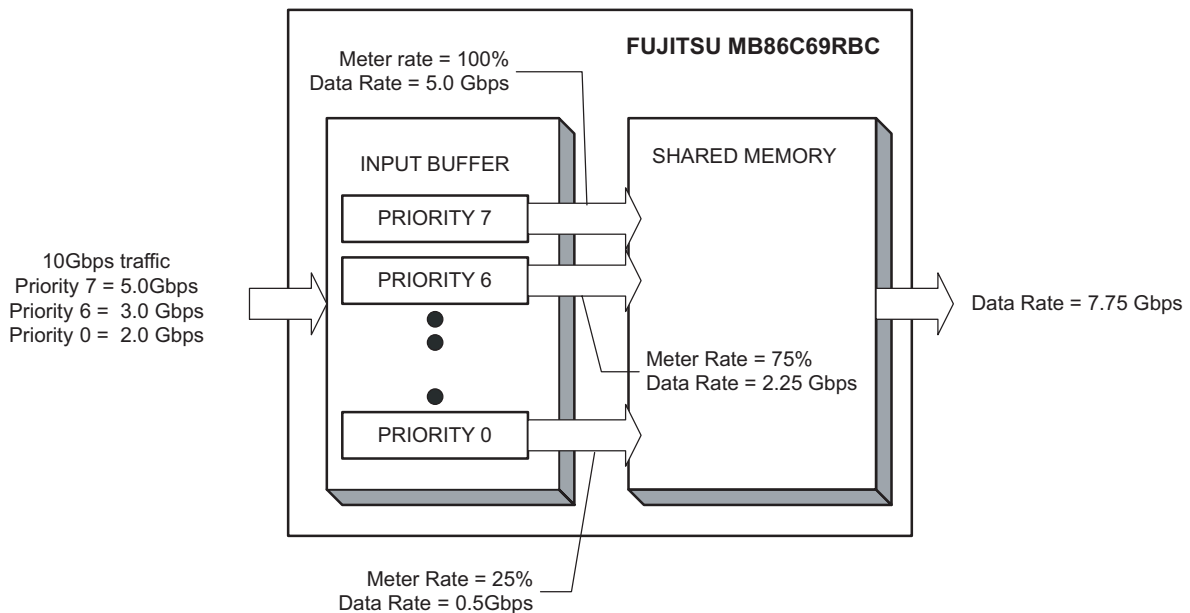


Figure 4 – Metering of Traffic in the MB86C69RBC

The solution to these limitations is to implement congestion management in the Ethernet Data Link layer. This approach improves responsiveness since congestion events can be handled immediately in the Data Link layer as opposed to being passed up the protocol stack. Figure 5 illustrates the inherent advantage of implementing congestion management in the Data Link layer.

The Fujitsu MB86C69RBC is the first chip in the industry to implement congestion management in the Ethernet Data Link layer. During a congestion event, the chip can send a backward congestion notification to the source endpoint causing the congestion, thus mitigating or eliminating the congestion in the network altogether.

Figure 5 illustrates a scenario in which the MB86C69RBC observes congestion at its egress port and sends a BCN message back to the source port responsible for the congestion. Upon receiving the BCN, the SQL server can respond by limiting traffic into the network and immediately reduce or eliminate congestion.

For this approach to work, a NIC must be able to respond appropriately to the congestion message sent by the MB86C69RBC. While the MB86C69RBC has the ability to generate different BCN message formats, the NIC still must have the ability to recognize and react to the BCN frame.

To avoid limitations imposed by NICs that do not have this ability, an alternative solution is to redirect the BCN message to the host processor connected to the MB86C69RBC. The host processor can then identify the ingress port and MAC address causing the congestion and take actions to mitigate the congestion. These actions may include assigning a lower priority to the source MAC address that is causing the congestion and then either metering that priority or dropping the traffic on that priority.

This method does not require third-party NIC vendors to recognize the BCN format. When congestion occurs, the necessary action can take place within the system integrator's own solution.

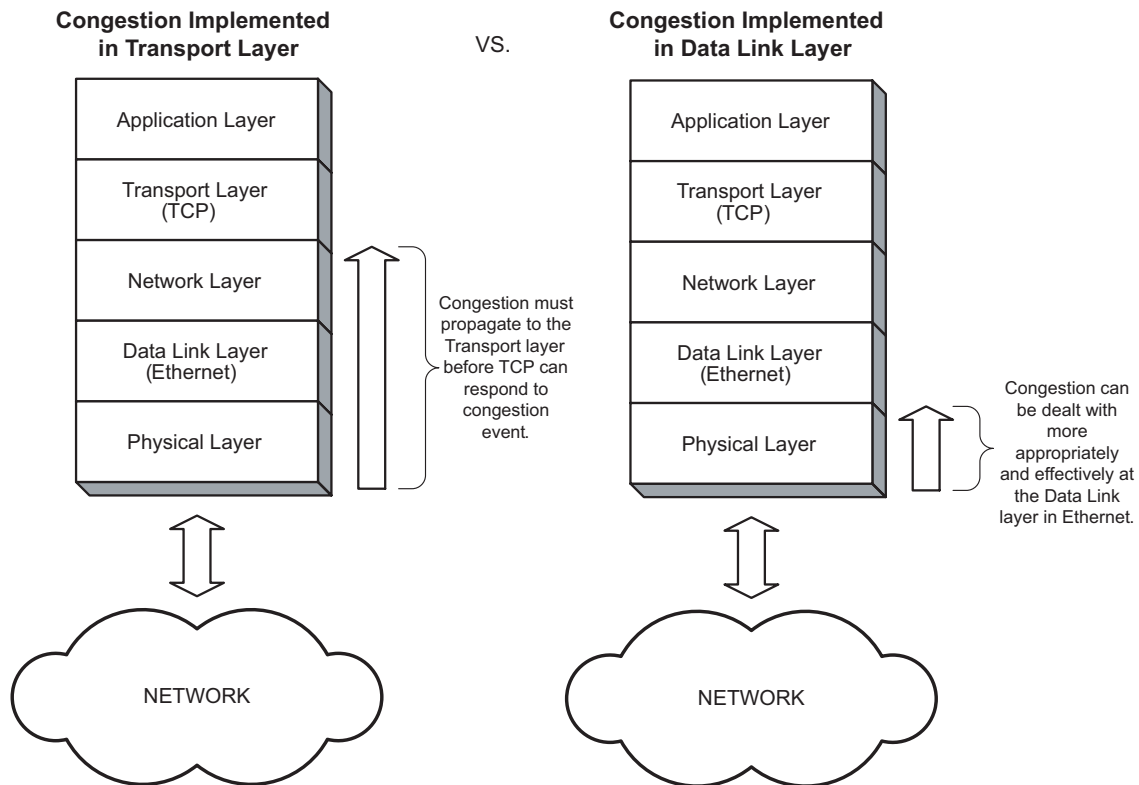


Figure 5 – Advantage of Implementing BCN in the Data Link Layer versus the Transport Layer

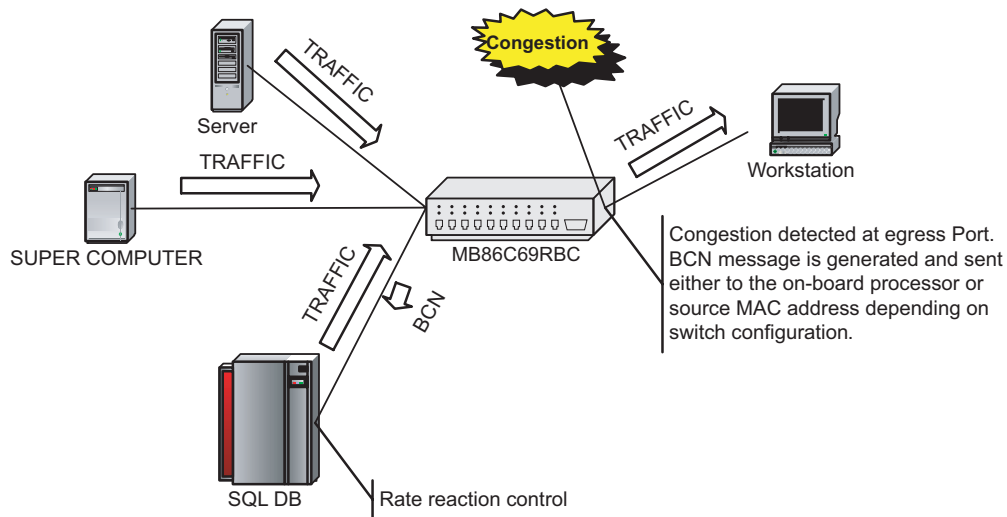


Figure 6 – Example of Backward Congestion Notification (BCN) Using the Fujitsu MB86C69RBC

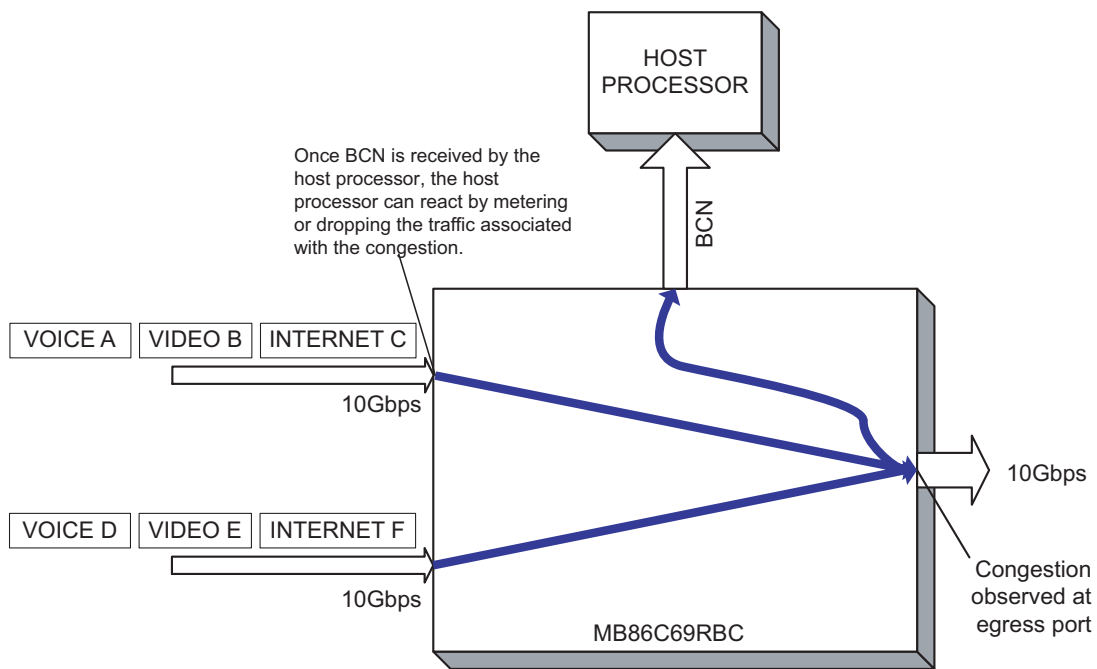


Figure 7 – Redirecting the BCN Message to the Processor

PAUSE and Priority PAUSE

PAUSE is a Link Layer flow-control mechanism in full-duplex Ethernet. When an endpoint is overwhelmed by the amount of incoming traffic, it sends a PAUSE frame asking the sender to halt transmission for a specified period of time.

Unfortunately, this PAUSE mechanism provides no granularity because it does not distinguish which data traffic may be

overwhelming the endpoint. The endpoint may be congested by incoming traffic from a single source, but the PAUSE frame halts all incoming traffic for a specified duration. Alternatively, the endpoint can ignore the fact that it is overwhelmed by incoming traffic, which ultimately leads to dropped frames. In both cases, all traffic is dropped, which is unnecessary if a single traffic flow is responsible for the congestion.

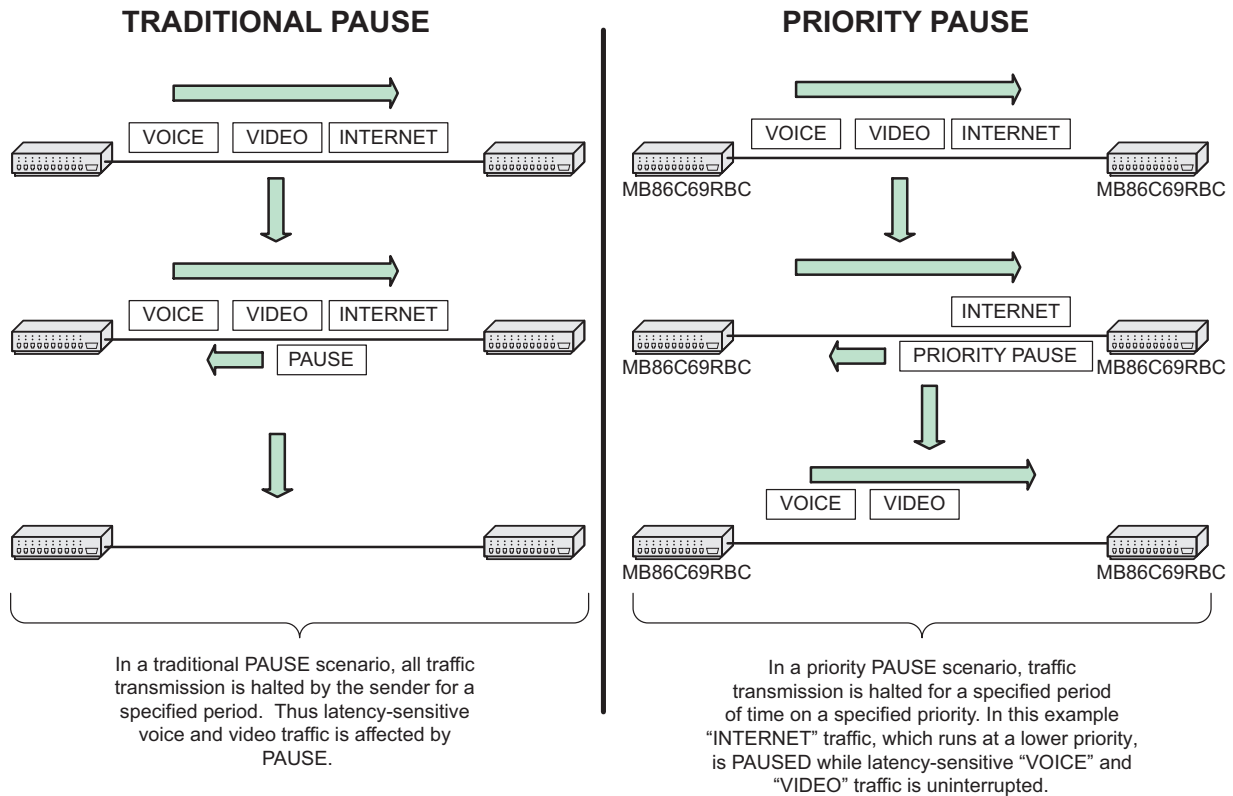


Figure 8 – An Illustrative Example of the Advantages of Priority PAUSE over PAUSE

The solution is to extend the Ethernet PAUSE mechanism so that it differentiates among the traffic flows. The Fujitsu MB86C69RBC implements this solution and is the first chip in the industry to provide priority PAUSE between link segments. The MB86C69RBC can send a PAUSE frame for a specific priority on the same physical link. As a result, traffic can be halted for a specific traffic flow without affecting other traffic on the same medium.

Figure 7 shows examples of PAUSE and priority PAUSE operations. In the latter example, the sender has halted the Internet traffic while the voice and video traffic streams remain uninterrupted.

Summary

The Fujitsu MB86C69RBC, 20-port, 10 GbE switch chip provides the features required to meet evolving data center requirements. It is the first chip in the industry to incorporate the advanced features of backward congestion notification, priority PAUSE and congestion avoidance, which are essential in today's and tomorrow's data center.

For more information

For more information about the Fujitsu MB86C69RBC and supporting products, please visit the company website at <http://us.fujitsu.com/micro/10gethernet> or send e-mail to inquiry@fma.fujitsu.com

FUJITSU MICROELECTRONICS AMERICA, INC.

Corporate Headquarters
 1250 East Arques Avenue, M/S 333, Sunnyvale, California 94085-5401
 Tel: (800) 866-8608 Fax: (408) 737-5999
 E-mail: inquiry@fma.fujitsu.com Web Site: <http://us.fujitsu.com/micro>

©2008 Fujitsu Microelectronics America, Inc. All rights reserved.
 All company and product names are trademarks or registered trademarks of their respective owners.
 Printed in U.S.A. 10GE-TB-21293-3/2008