



Converged Enhanced Ethernet (CEE) and Datacenter Bridging (DCB)

Using FocalPoint Switches

White Paper

February, 2009

Introduction

As the size and density of datacenters increase, the cost, area, power and support of multiple interconnect fabric technologies cannot be tolerated. Because of this, the IEEE is developing several new standards that will enable Ethernet as the single unified fabric for data, storage and HPC traffic. The industry has coined these new initiatives Converged Enhanced Ethernet (CEE) or Datacenter Bridging (DCB). This paper will discuss several advanced FocalPoint features available in Bali that have been developed to support CEE-DCB applications.

The figure below shows an overview of several features which are used for congestion management in Bali. The way these features can be used for CEE-DCB will be described in more detail throughout the rest of this paper.

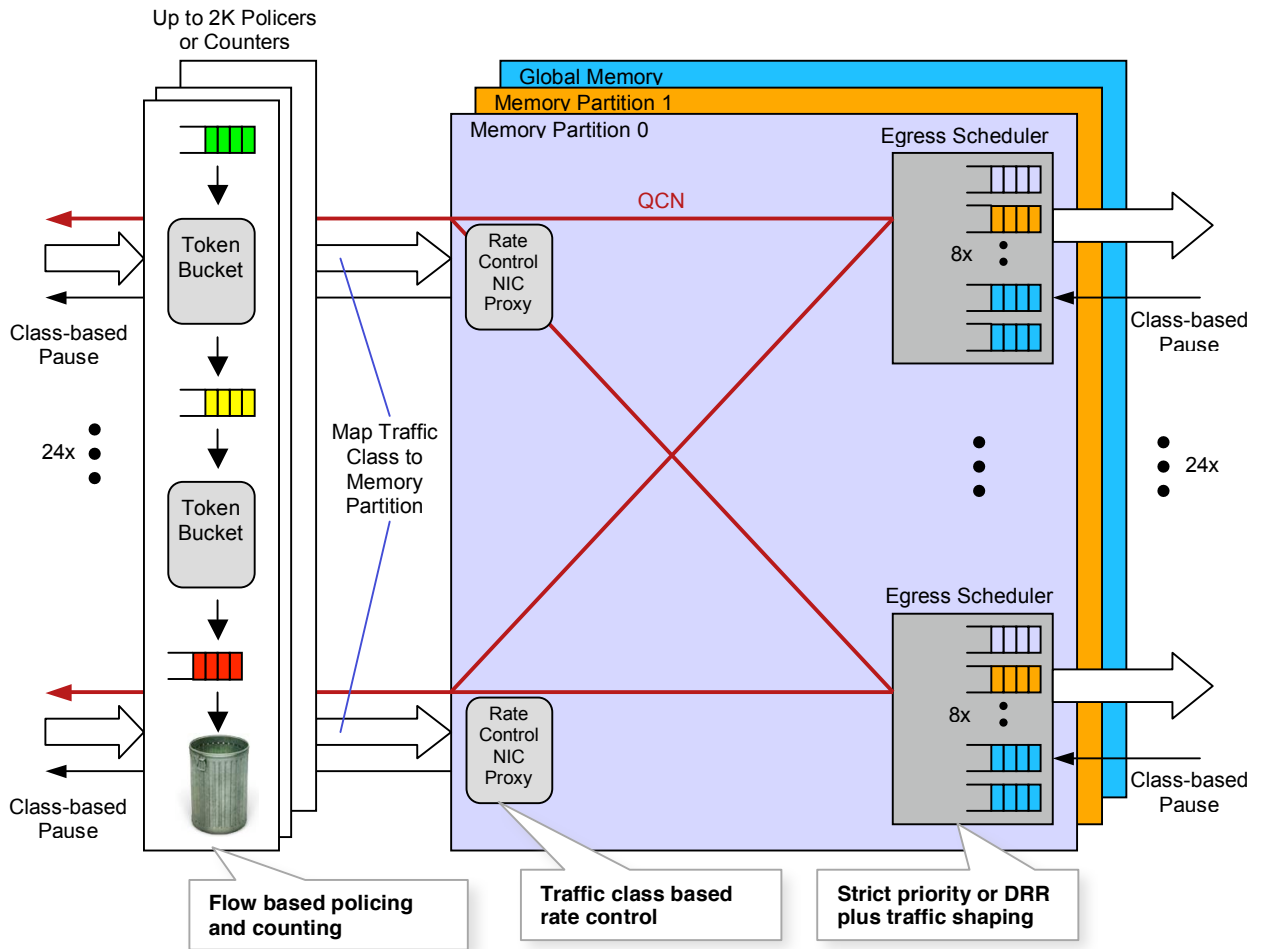


Figure 1: Bali congestion management features.

Lossless Operation using PFC

Storage protocols require deterministic, bounded latency to ensure that time-out values are not exceeded. Traditional Ethernet switches drop packets during periods of high congestion, requiring retransmission at a higher layer. This added latency cannot be tolerated in storage applications. Standard IEEE pause frames will pause all traffic during periods of congestion. So with large enough buffers, traffic will not be dropped, but unacceptable latency times may still be seen. Because of this, the IEEE is standardizing Priority Flow Control (PFC), which can differentiate traffic classes, providing bounded latency for storage traffic.

Bali contains a Frame Forwarding Unit that can be used for deep header inspection including identification of traffic types such as storage or HPC. It can also assign these traffic types to one of 8 internal traffic classes. These traffic classes can also be assigned to one of 2 shared memory partitions. When a partition fills past a watermark, pause frames will be generated for all traffic classes assign to that memory partition. For example, storage traffic can be assigned to one memory partition and data traffic can be assigned to another memory partition. In this case, the switch will not pause storage traffic if data traffic causes congestion in the switch.

The Bali device supports PFC. This capability is compliant with the latest version of draft IEEE specification: IEEE P802.1Qbb/D0.2. This capability was also designed in compliance with the initial Cisco specification made public through the IEEE in May, 2007 and later updated in proposals from Cisco. In addition, interoperability has been successfully tested at Fulcrum's lab with several industry leading NIC vendors.

Bali may be configured at ingress and at egress to support PFC. At the switch ingress, a PFC pause frame can be generated from the Bali device to the upstream link partner based on memory partition watermark settings that are per port. When a watermark is exceeded, a PFC pause frame is generated, in accordance with the specification.

At the switch egress, a PFC pause frame can be received by the Bali device and interpreted to pause a particular traffic class. When this happens, the egress scheduler will not schedule frames of that priority to that egress port, but may schedule frames from other priority queues to that egress port. All PFC features in the Bali device can perform at line rate with no impact to the latency or overall packet forwarding performance.

Flow Optimization Using ETS

In any switch fabric, traffic from multiple ingress ports can compete for the limited bandwidth in a single egress port. For example, bursts of data traffic could reduce the bandwidth available to storage traffic, causing congestion and increased latency for the storage traffic. To solve this problem, the IEEE is developing Enhanced Transmission Selection (ETS) which provides advanced traffic scheduling including features such as guarantee minimum bandwidth for certain traffic classes like storage or HPC.

The Bali device supports ETS and is compliant with the most current version of the IEEE specification, IEEE P802.1Qaz/D0.1. The ETS support in Bali enables the definition of priority groups at each egress port of the switch. Priority queues are combined into groups, and deficit round robin scheduling algorithm can be applied to the groups. All of the queues in a particular group can have different priority, but the switch is able to apply similar congestion control capabilities to the multiple queues within a group. The switch is also able to establish some priority groups to be associated with PFC and some to be configured without PFC. It is important that both capabilities are configurable. The IEEE allows for the details of the scheduling algorithm to be vendor specific, provided that the general capability of the behavior on the wire can be met.

White Paper: CEE and DCB using FocalPoint

In addition to the capability defined by the IEEE, the Bali device supports numerous enhancements in this area. Bali is capable of implementing a hierarchical scheduling algorithm between the priority groups and the priority queues, which enables strict priority between the priority queues within a priority group. Bali also supports shaping between the priority queues. Bali supports 8 priority queues per egress port, which is contemplated by the IEEE but not required.

The following sections describe some of the mechanisms and features available in Bali's multi-level egress scheduler.

Traffic Classes

Bali contains a Frame Forwarding Unit that can be used for deep header inspection including identification of traffic types such as storage or HPC. It can also assign these traffic types to one of 8 internal traffic classes. These traffic classes are used for ingress flow control and egress scheduling.

Strict Priority

Scheduler groups can be defined to provide either strict priority or deficit round robin. For strict priority, higher number traffic classes are always scheduled before lower number traffic classes unless that traffic class has been bandwidth limited or is being flow controlled by a downstream device. Consecutive traffic class numbers can also be assigned to the same strict priority group such that any traffic class in the group with a frame queued will always be scheduled before lower number traffic classes unless all traffic classes in the group have been bandwidth limited or are being flow controlled by a downstream device.

Deficit Round Robin (DRR)

Deficit round robin gives a minimum bandwidth guarantee to a traffic class, which can also be used to guarantee maximum latency. Consecutive traffic class numbers can also be grouped such that a group of traffic classes can be given a minimum bandwidth guarantee. For example, bandwidth can be allocated to various traffic classes within a priority group. There cannot be a strict priority group between DRR scheduler groups. A traffic class or classes within a DRR group will be scheduled at a minimum bandwidth rate assuming that there is traffic to be scheduled, that there is no higher priority traffic eligible that is consuming the bandwidth and there is no flow control for that traffic class.

Traffic Shaping

Traffic shaping is used to create an upper bound on the bandwidth for a traffic class in order to reduce jitter. If DRR is used, it is expected the maximum shaping bandwidth will be set higher than the minimum DRR bandwidth. Consecutive traffic class numbers can be in the same shaping group such that the aggregate bandwidth from that group does not exceed a maximum value. Traffic shaping in combination with DRR provides an excellent means to optimize traffic flows through the fabric.

Example Implementation

As you can see by the discussion above, there are a wide variety of scheduling configurations that can be implemented in each Bali egress port. The figure below is an example of how an egress scheduler could be used.

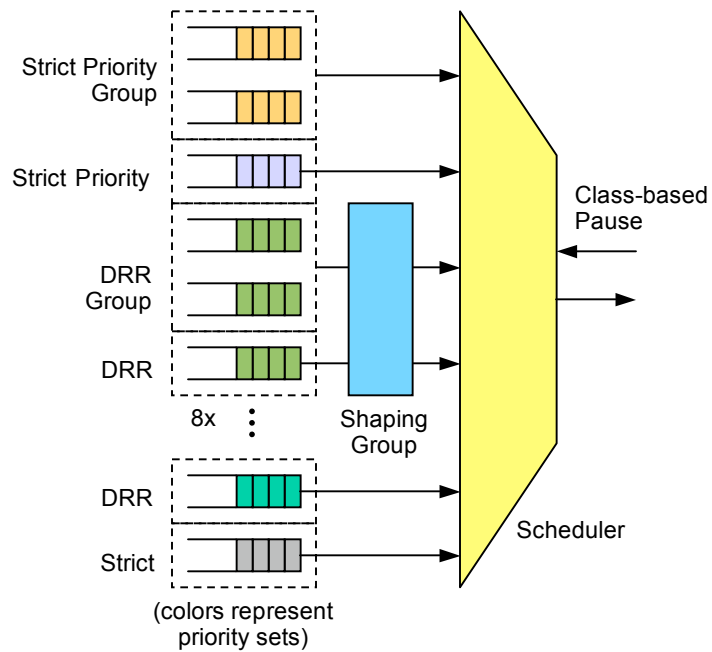


Figure 2: Bali multi-level scheduler example.

Efficient Fabric Scaling

Scalability is important in data center switches. But scalability cannot come at the expense of lower performance due to factors such as blocking. The Bali architecture provides several features to alleviate blocking in multi-stage datacenter fabrics. The main feature is a large number of 10G ports per switch element. This allows the efficient creation of a multi-tiered switch architecture commonly called a Fat Tree, which provides constant bandwidth between switching layers and forms a non-blocking, scalable switch as shown in figure 3.

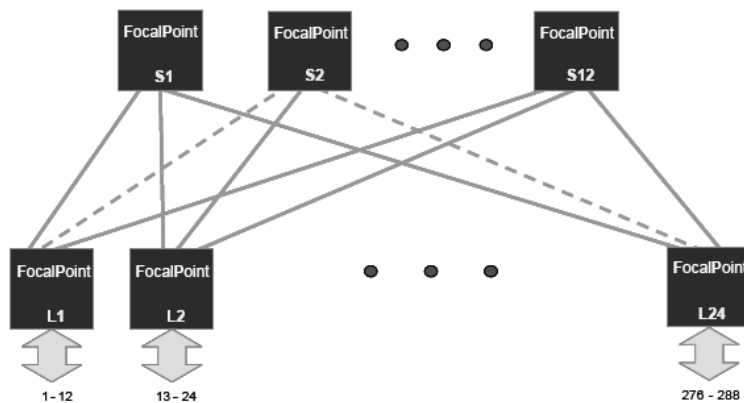


Figure 3: Multi-tiered switch architecture.

White Paper: CEE and DCB using FocalPoint

This configuration is another example of the importance of low-latency switch elements. With 300ns Bali switch elements, the maximum latency between any two CPU blades in any two racks is less than 1uS, whereas with an alternative switch element, the latency can be well over 10uS for the same configuration. Adding 10uS latency to virtually any application – even simple enterprise automation applications – will meaningfully impact performance and efficiency. A 2-tier fat tree, built from top-of-rack switches and blade switches, is shown in Figure 4.

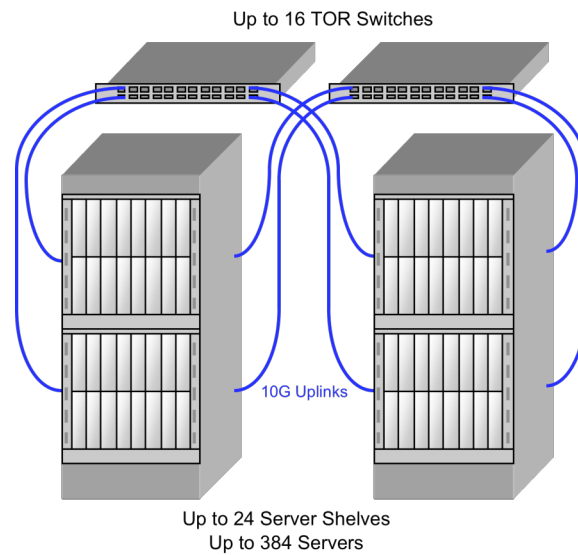


Figure 4: Example multi-tiered system using TOR switches.

The number of network ports presented by a Fat Tree architecture scales exponentially with the number of tiers in the switch. The FocalPoint switches were designed with a set of features optimized for the effective implementation of this architecture, where a 2-tier Bali-based system provides up to 288 10G ports in a non-blocking configuration and a 3-tier system up to 3,456 non-blocking 10G ports (greater density can be achieved by introducing over-subscription in the system). By using highly integrated Bali switches in this architecture and by using standard Ethernet switching throughout the fabric, scalability is achieved that is more cost-, space- and power-effective than the standard “Big Iron” approach.

FocalPoint provides additional features to avoid blocking in multi-tiered fabrics such as an output queued shared memory architecture, separate memory partitions for different traffic types, efficient load distribution across second tier switch devices and congestion feedback mechanisms such as class-based pause frames and QCN.

Bali Multi-path Support

The Bali devices support multi-pathing. This capability allows one to define multiple output ports from the switch and enables the switches to hash between the output ports to load-balance the traffic between multiple second level switches. Unlike link aggregation, there is no limitation with multi-pathing to require both ends of the wire to terminate on the same set of switches as in link aggregation. Multi-pathing is used to define Clos

switches in Ethernet, similar to Clos switches in Infiniband applications. In conjunction with DCE features, and shared memory packet storage, the performance of multi-pathing can be considerably higher than in Infiniband. Bali supports layer 2 multi-pathing with an ISL tag and layer 3 multi-pathing using ECMP. When the ISL tag is used, the multi-pathing configuration can be modified on the fly to accommodate changes in link topology for resiliency. These changes are implemented in a low-level mapping table sparing the switch from doing a spanning tree reconfiguration

Cut-Through Architecture

In order to achieve low latency independent of packet size, FocalPoint switches employ a cut-through architecture. Traditional Ethernet switches use store and forward mechanisms where the latency can be many microseconds per stage depending on the packet size, which is considered unacceptable for high performance datacenter applications. FocalPoint latency compared to other Ethernet switch products is shown in Figure 5 below:

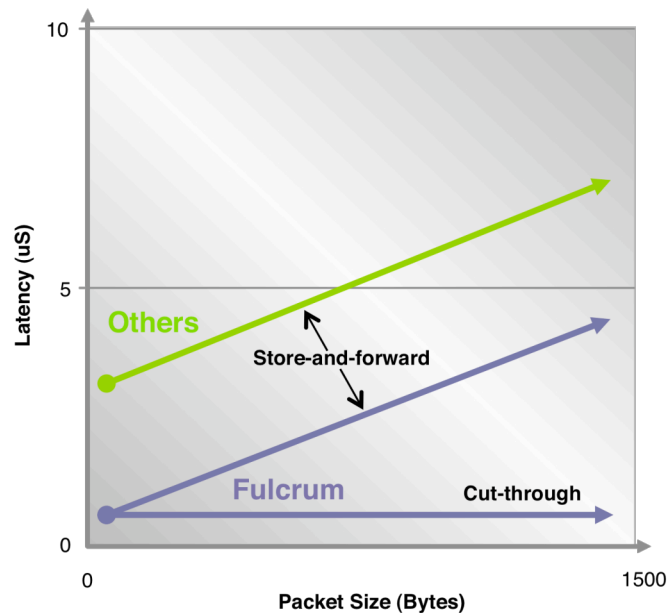


Figure 5: Single stage switch latency vs. packet size.

FocalPoint switches have been designed for the datacenter using cut-through operation that can achieve less than 1uS latency through three fabric stages. In addition, this latency can be achieved with all L2/L3/L4 features enabled. This makes Ethernet a compelling unified fabric solution for the datacenter.

The Bali devices also support parallel multicast, which is a much-needed feature in many datacenter applications. The switch can saturate all of its output ports simultaneously with layer 2 or layer 3 multicast. In an uncongested switch, there is less than 70 nS of min-max delay between the first and last multicast packet in a 23-way multicast tree. This capability is unique to Fulcrum switches and is considered highly desirable in market trading and other transaction applications that require low-latency synchronization between nodes. It is also very useful in HPC environments for MPI collective operations.

Power Efficiency

In large multi-stage switch configurations, the power per switch has a significant impact on overall system power. Bali incorporates patented low-power Fulcrum technology, which cannot be matched by standard Ethernet devices. In addition, unused interfaces can be disabled to consume no power and the core power scales directly with the level of activity.

Bandwidth Optimization using QCN

In multi-stage fabrics, congestion hot spots can occur which can cause congestion spreading. To combat this, a new IEEE work group has been formed to develop the Quantized Congestion Notification (QCN) standard.

The Bali device was built with several features to support the QCN standard. The Bali reference design, Monaco, has implemented QCN features compatible with P802.1Qau/D1.1. In the Monaco reference design, a small, low-cost, low-power FPGA (Altera Cyclone III) is used between the switch and the CPU. This FPGA monitors Bali's queues and modifies the configuration of Bali to make it compatible with the new QCN algorithm. The vast majority of the processing necessary for QCN is in the Bali device itself, so the performance of the algorithm is not determined by the FPGA.

NIC Proxy Mechanism

Due to the fact that NICs may not implement the QCN standard for some time, Bali has been developed with a NIC proxy feature, which implements the logic that would normally go into a compliant NIC device. This allows the use of QCN before there is widespread adoption of the algorithm by adaptor vendors. Support for the configuration of the QCN algorithm is implemented in the FocalPoint API.

NIC Proxy is a feature used in QCN and supported in Bali to allow the entire algorithm to be implemented within the switch. The NIC proxy feature does this by providing the QCN rate limiters in the switch device itself. The switch is capable of trapping the QCN frames from a distant congested egress port. It then adjusts the token bucket rate limiters until the congestion point is satisfied with the aggregate flow rates.

The trick is to convert the result of rate limiters into a format the NICs can interpret. This is done through a PFC pause pacing function. The rate limiter will cause the switch ingress port to send occasional pause messages (PFC messages) until the upstream NIC achieves the desired rate required. Because of this, the NIC only needs to support PFC.

Fabric Management with DCBX

Management of a datacenter bridging fabric requires the exchange of parameters between switches. The IEEE is developing the Datacenter Bridging Exchange Protocol (DCBX) to support this.

Bali supports the DCBX protocol also described in IEEE P802.1Qaz/D0.1. Support for the DCBX protocol in the silicon is determined by the support of LLDP, and the underlying support for the capabilities that DCBX enables. Fulcrum currently supports the necessary features for DCBX packet handling in the FocalPoint API. Fulcrum also has DCBX software in development for other specialized environments. As a result of this work, Bali has features for the discovery of DCB peer capability, detection of DCB mis-configuration and DCB peer configuration.

Conclusion

Emerging datacenters will require converged fabrics in order to minimize cost, size, power and support. Because of this, the IEEE has several new initiatives, which are known collectively as Converged Enhanced Ethernet (CEE) or Datacenter Bridging (DCB). These initiatives include PFC, ETS, QCN and DCBX. The Bali switch products support all of these features while also maintaining industry-leading latency in large scalable fabric configurations.

Fulcrum Microsystems, Inc.
26630 Agoura Road
Calabasas, CA 91302
818.871.8100
www.fulcrummicro.com