White Paper

Enabling Quality of Service With Customizable Traffic Managers

Introduction

Communications networks are changing dramatically as lines blur between traditional telecom, wireless, and cable networks. The value of bundling communication services is accelerating carrier investment in triple-play (voice, video, and data) networks. Managing these services on a common IP network leads to a dramatic reduction in operational expenses. In addition, this migration to IP networks enables carriers to offer enhanced services to drive increased revenues. These emerging services are the key to carrier growth and sustainability in the future. Supporting these future services requires a solution that meets today's stringent quality of service (QoS) requirements, yet is flexible enough to adapt to tomorrow's changing requirements.

Next Generation System Architecture Challenges

The dramatic changes in the communications market trickles down the supply chain, creating an enormous impact on the equipment suppliers. Successful communications OEMs are the ones most able to adapt to this new environment.

New Market Landscape

The emergence of low-cost competition has altered the communications original equipment manufacturer (OEM) landscape. To take advantage of lower engineering costs, development increasingly stretches across multiple geographies. To further reduce the cost and risk of development, the solutions being developed are leveraged many times across multiple platforms. Successful OEMs will continue to reduce engineering development costs while also retaining core competencies necessary to maintain long-term differentiation against low-cost competitors.

Fewer Standard Products

The contraction of the communications market has not only dramatically reduced the number of ASIC starts from equipment manufacturers, but also the number of standard product offerings targeting high-end applications.

Standard products that target applications with changing requirements, evolving standards, or product-specific needs, inevitably have deficiencies. These are likely resolved by external bridging or co-processing devices. FPGA traffic managers can reduce system cost by integrating the external bridging and co-processing functions while also differentiating the system by adding customized features.

Challenges of High-Speed Traffic Management

Memory is a major bottleneck in traffic-management applications above 2.5 Gbps. Memory is required to buffer packets, to store pointers to the buffered packets, to maintain control and state information for the queues, and to collect statistics for billing and maintenance purposes. These external memories require fast random access and response times, and often require specialized memories such as RLDRAM II. Development of efficient memory management architectures can be deceptively complex.

To interface to the external memories, high-end traffic-management solutions require a large number of pins. In applications where pin counts drive the die size, standard product solutions do not have pricing advantages over FPGA implementations.

Altera's Solution

Altera's traffic manager provides the benefits of:

■ Future-proofing your traffic manager with the inherent flexibility of the FPGA
■ Differentiating your solution using Altera's modular building blocks
■ Reducing development costs and time-to-market by leveraging proven blocks
■ Meeting 10-Gbps throughput for high performance applications
■ Reducing board space and cost by integrating external co-processing or bridging functions
■ Scaling to adapt the solution across multiple platforms or cards

Altera Traffic Manager Overview

Altera has a working 10-Gbps traffic manager solution targeting Stratix® II FPGAs. (See Figure 1 for a schematic.) The inherent programmability of the FPGA allows the traffic manager to be adapted to support changing carrier requirements or emerging services. This Traffic Manager meets guaranteed line-rate performance for Ethernet, Packet over SONET/SDH, MPLS, and ATM traffic. The design was built in a modular fashion, enabling system designers to customize partial or complete sub-blocks of the Traffic Manager.

Packet Flow: Line Side Interface Through NPU

Altera's Traffic Manager was designed for 10G line cards and interfaces to both a network processor and a fabric interface chip (FIC) on the data path. In the ingress direction, traffic flows into the line card encapsulated in Ethernet, SONET/SDH, RPR, or OTN frames. The Framer or MAC device performs the necessary Layer 1 processing and transmit the traffic to an NPU with a Layer 2 header attached. The NPU performs classification, modification, and forwarding operations as dictated by the protocols being processed. The NPU also adds additional headers to the packet for communicating information to downstream devices, including both ingress and egress Traffic Managers, as well as the egress NPU. The header for the ingress Traffic Manager contains information such as class of service (CoS), multicast, and drop precedence, which are necessary for the device to appropriately prioritize the traffic.

Figure 1. Altera's 10-Gbps Traffic Manager
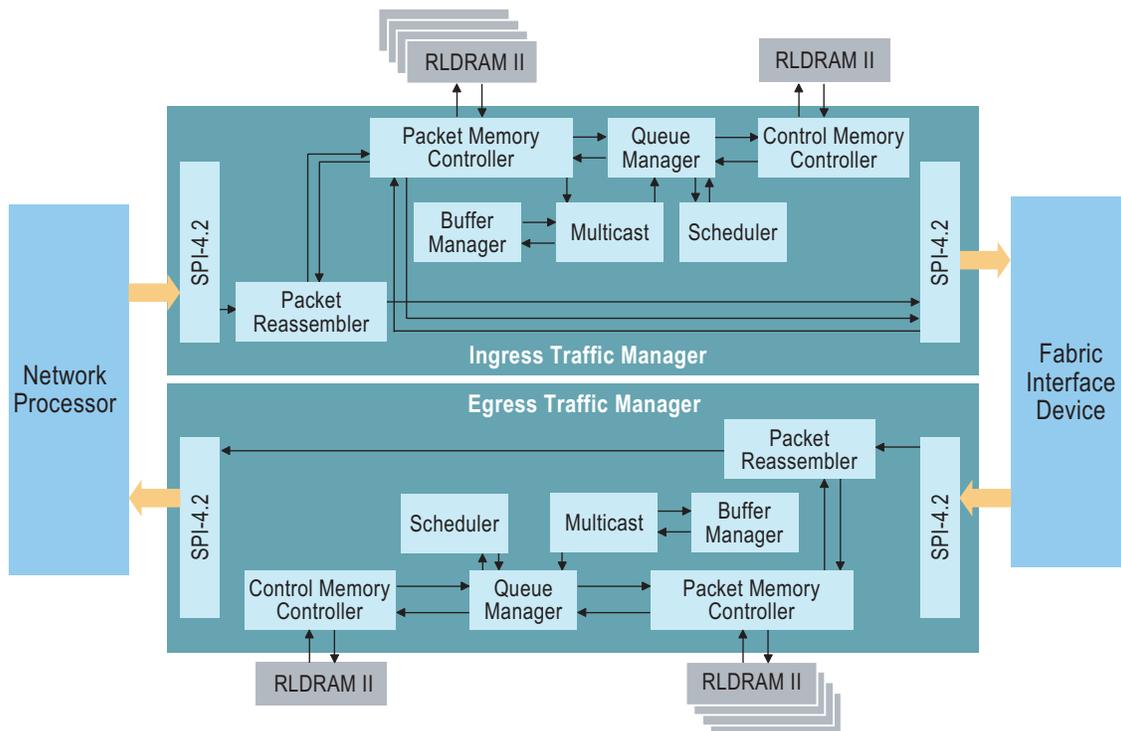
Packet Flow: Ingress Traffic Manager

The current implementation utilizes Altera's SPI-4.2 MegaCore® function for the NPU interface. This core uses the Atlantic™ interface on-chip, which enables designers to easily replace the SPI 4.2 interface with another (possibly proprietary) interface that supports the Atlantic interface on-chip.

The packet reassembler block receives data from the SPI-4.2 interface and parses the header for relevant parameters (such as CoS). The parsing can be customized to locate parameters within any location of the header. This information is stored in the control memory, along with pointers to the packets. The reassembler block converts the received data into fixed-sized cells and communicates with the buffer manager block to ensure that the configured memory partitions are not violated by packet enqueues.

The buffer manager passes the enqueue request to the queue manager block, which maintains state and pointer information about each of the queues. If the enqueue is valid, the reassembler block selects an available pointer to external memory and passes the segmented packets to the packet-memory controller for storage in external memory. The list of available pointers is maintained both off-chip in the packet memory and in an on-chip cache using the FPGA's embedded memory.

After the packet has been written to external memory, it is eligible to be transmitted by the scheduler. The scheduler examines all ports that have data to send and chooses the next port according to a hierarchical scheduling scheme. The scheduling algorithms are configurable and the entire scheduling block is customizable in order to support proprietary scheduling algorithms.

When the scheduler has selected a cell from a packet for transmission, it issues the request from the queue manager. The queue manager initiates the dequeue operation through the reassembler block.

Packet Flow: Fabric Interface to Egress Flow

Packets that have been scheduled are sent through the FIC interface. The current FIC interface utilizes Altera's SPI-4.2 MegaCore function. A header is added to each of the cells in order to enable the fabric to switch the cells to the appropriate port with the appropriate priority. The cells flow through the switch fabric, then the egress FIC converts the traffic to a SPI-4.2 interface for the egress Traffic Manager.

The egress traffic manager's data flow is similar to that of the ingress traffic manager. It also interfaces to the NPU through a SPI-4.2 interface, which determines the appropriate Layer 2 header to place on the packet for sending to the Framer or MAC.

Traffic Manager Blocks

This section reviews the functionality of the individual blocks of the Traffic Manager. Each function is designed to be a module that can be customized or replaced.
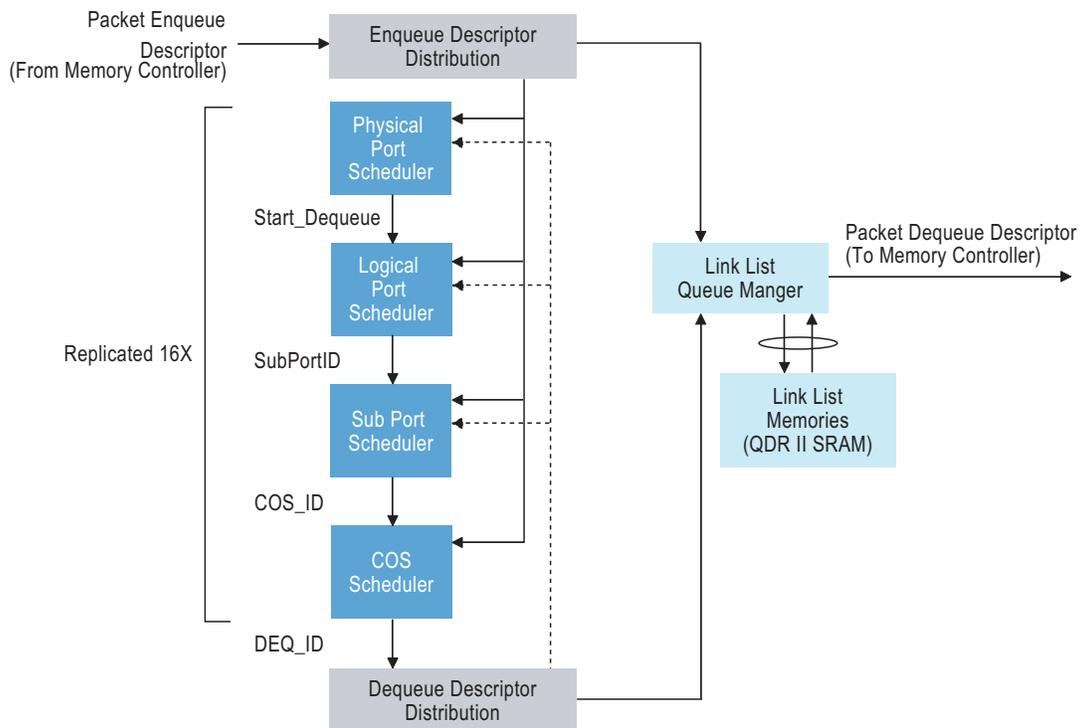
Scheduler

Traffic scheduling ensures that, during times of congestion, each port and each class of service gets its fair share of bandwidth. The scheduler interacts with the queue manager block, notifying it of scheduling events and receiving information about queue lengths.

The Altera® traffic manager's scheduler includes the following features:

■ Four-level hierarchical scheduler
■ 128K-queue scheduler with 4 classes of service, extensible to 256K queues with 8 classes of service
■ Physical port scheduler schedules 16 ports (16 x OC-3)
■ Logical port scheduler schedules 16 logical ports (16 x DS-3)
■ Sub-port scheduler schedules 128 sub-ports (128 x subscribers)
■ Class of service scheduler schedules 4 x CoS
■ Configurable scheduling algorithms
  ● Strict priority per-port scheduling (up to 4 CoS levels) ensures mission-critical control traffic is expedited
  ● Weighted fair queuing (WFQ) per-port scheduling can be used in combination with strict priority
  ● Per-port shaping supported
■ Per-port scheduling is packet-size-aware, ensuring that small packets are not unnecessarily penalized

Figure 2 shows an example of a four-level hierarchical scheduler. The physical port scheduler arbitrates among 16 unicast ports and a single multicast port. It uses a round robin (RR) scheduler to arbitrate among all unicast ports and uses a WFQ scheduler to arbitrate between the selected unicast port and the multicast port.

Figure 2. Four-Level Hierarchical Scheduler

The logical port scheduler arbitrates between 16 logical ports, using either a RR or WFQ algorithm. The sub-port scheduler arbitrates between 128 sub-ports, also using either a RR or WFQ algorithm. The CoS scheduler block is responsible for arbitrating among four classes of service to determine the next M-cell that should be sent to the queue manager. The CoS uses a 4-input WFQ scheduler or shaper, depending on the mode configured.

In shaping mode, the entry at the top of the structure is checked against the shaper clock tick count. Since the structure is sorted, the entry at the top of the structure is guaranteed to be the next eligible port. If the shaper clock tick count has advanced sufficiently so that it either exceeds or is equal to the entry at the top of the structure, then the entry at the top is eligible to be scheduled.

In addition to the scheduling of traffic, the scheduler is also responsible for performing the scheduling of RLDRAM refresh cycles for the external control memory.

Congestion Management

Features

■   Weighted Random Early Detection (WRED) or Tail Drop supported
■   Depth is configurable on all queues

Packet congestion can cause severe network problems, including reduced throughput, increased delay, and increased packet loss. Congestion management can improve network congestion by intelligently dropping packets. Tail Drop is a passive technique that reacts to full queues by dropping incoming packets until space becomes available. WRED is an active congestion management scheme that allows for a reaction prior to the queue becoming full. As the queues are nearing congestion, WRED can drop a packet with a weighted probability associated to that particular queue. This selective dropping of packets prior to full congestion signals to TCP hosts to reduce the transmission until congestion is reduced.

Many other congestion management schemes exist, including proprietary algorithms. In addition, standards work (including IEEE 802.3ar) continues to define congestion management requirements for next-generation systems. FPGA-based traffic managers enable differentiated congestion management implementations, while also retaining the flexibility to support emerging standards activities.

Multicast Replication

Features

■   Multicast replication supported on ingress and egress
■   A dedicated multicast queue for each of eight classes
■   4K Multicast Groups supported
■   32K Multicast Leaves supported

As carriers add to their offerings by supporting streaming audio or video services, the need to efficiently replicate packets becomes essential. Altera's traffic manager supports multicast replication in either the ingress, egress, or both directions, depending on the application requirements.

The multicast replication block receives all packet enqueue messages (unicast and multicast) from the packet reassembler. Multicast packets are placed on a work queue (after passing a multicast partition check). When the multicast packet reaches the head of the work queue, the replication logic dequeues the packet and replicates.

Queue Manager

Features

■   10-Gbps line rate performance
■   On-chip prefetch cache reduces external memory

The queue manager block is responsible for initiating packet enqueue and dequeue requests based on notifications from the buffer manager (for enqueues) and the scheduler (for dequeues). The queue manager also maintains a prefetch FIFO buffer to cache packet descriptors and manages the head/tail pointers of each of the externally stored queues. The prefetch cache can be implemented using on-chip memory for applications with a low number of queues, or by using external QDR II SRAM for applications requiring a high number of queues.

Memory Controllers

Features

■ Uses RLDRAM II
■ Stores up to 0.5M packets in 128 Mbytes (4.32-Mbyte RLDRAM II memories)
■ RLDRAM II and QDR II are used for statistics and control memory

High-end traffic management applications require memory management that can achieve high bandwidths with reduced latency. The Altera traffic manager uses RLDRAM II memories for buffering packets and storing packet descriptors and statistics. The packet storage memory comprises four RLDRAM II devices in each direction, organized in two groups of two memory chips. The RLDRAM II interface is a bidirectional, double-data rate memory interface running up to 300 MHz. The internal core clock is edge-aligned at one-half the frequency.

To achieve the high-bandwidth requirements for 10G traffic management, the Altera traffic manager uses a time division multiplex (TDM) strategy. This scheme dedicates specific time slices for the read direction and the write direction. It also dedicates no operation (NOP) and turn-around (T_A) time slots to meet the required row-cycle times (t_{RC}) and read and write latencies required by the RLDRAM II specification.

Figure 3 shows the results of the throughput, highlighting the ability to maintain line rate performance for 10G SONET/SDH down to 40-byte packets.

*Figure 3. RLDRAM II Throughput for Packet Over SONET/SDH in Stratix II Devices*
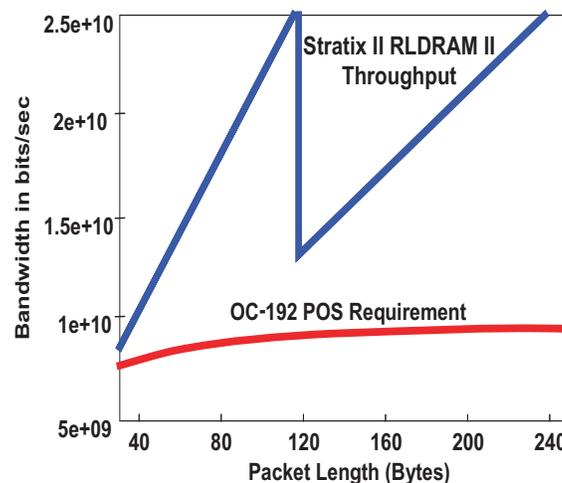
The flexibility of the FPGA implementation supports a variety of cadences for the TDM scheme, which allows system designers to optimize the memory management for a particular application.

A script was written to generate the memory bandwidth for each of the TDM structures and for various frequencies. The results of the throughput analysis were then graphed to determine if the memory bandwidth was sufficient to sustain the packet throughput for each packet size.

Figure 4 shows the 10-Gbps Traffic Manager Demonstration Board.

Figure 4. Altera's 10-Gbps Traffic Manager Demonstration Board
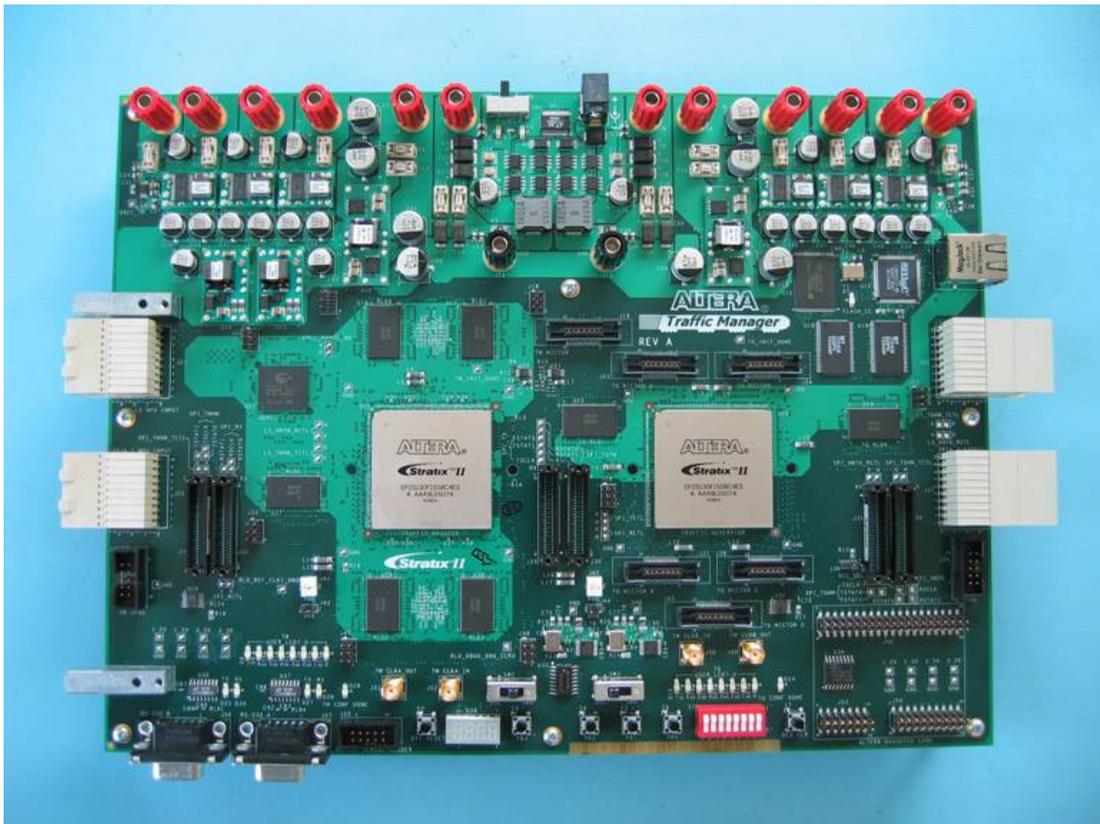
Conclusion

Altera's 10-Gbps traffic manager solution meets the demands of next-generation networks by supporting high-speed throughput in a solution that can adapt to the changing market. The solution can be demonstrated in a hardware environment using Altera's demonstration board. For more information on this solution, contact your Altera salesperson.

101 Innovation Drive
San Jose, CA 95134
(408) 544-7000
http://www.altera.com