DESIGN SOLUTION:

A Carrier-Grade Ethernet Switch in one **SoC**

*By **Alex Kugel**, Altera Corp.*

MEF

The Design Team:

Altera Wireline System Solutions Engineering is Altera's center of excellence for system designs in the networking and transmission space. They have built numerous reference systems and turn-key designs for some of Altera's biggest customers, including switches, routers, bridges and network interface functions with capacities ranging from a few gigabits per second to one terabit per second. They also offer complete off-the-shelf customizable OTN SoftSilicon® solutions in transport applications.

Challenge:

Create a single-chip carrier-grade Ethernet switch that can be used in small, time-critical systems or integrated with additional functions to create an intelligent network appliance.

Solution:

Using a moderately-sized 28 nm SoC FPGA, the designers implemented a 5-port, full-duplex Gbit switch that received Metro Ethernet Forum Carrier Ethernet 2.0 certification. There is adequate headroom in both the FPGA fabric and the dual CPU cores for applications to work in conjunction with the switch.

The Project

Carrier Ethernet began with a move to add features to Ethernet to meet the needs of metropolitan-level telecommunications carriers in terms of reliability, flow control, and network management functions. These added functions have made carrier Ethernet attractive in other applications as well, including data-center networks and even embedded systems where an industry-standard facility for traffic shaping, fault recovery, and time synchronization can be indispensable.

While Gbit Ethernet switch chips are commodity products, there are relatively few sources of carrier-grade hardware at moderate data rates. And none of these offer the ability to easily integrate additional hardware, such as deep-packet-inspection engines or time-critical functions, into the switch. The Altera team set out to create a single-chip device that would offer these benefits to small design teams, as well as serving as a reference design for the use of Altera Ethernet IP cores.

The Design Challenge

Individually, none of the functional blocks in a carrier Ethernet switch is a formidable design challenge for an experienced networking hardware/software design team. But integrating the blocks, providing adequate bandwidth between them—especially to external DRAM—integrating with operating and application software, and achieving Metro Ethernet Forum certification are all challenges beyond the reach of the average design team that might want to create such a part. The challenge was to create such a solution in an FPGA that would be both affordable and easy to integrate with additional functions, so that a design team without deep carrier Ethernet experience could integrate these capabilities with their own hardware and software designs.

The proper vehicle for this integration was the Arria V SoC: an FPGA with on-die dual ARM Cortex-A9 CPU cores, integrated DRAM controller, high-speed SerDes and transceiver circuitry and other supporting hardware, plus an infrastructure of embedded operating system and development tool choices.

ALTERA®
MEASURABLE ADVANTAGE™

The Design Solution

Altera's Carrier Ethernet design supports wire-speed network processing with 5Gbps of aggregated traffic. It supports Provider Bridge switching as specified by IEEE 802.1ad.The switch implements Quality of Service/traffic management features, IEEE1588v2 capabilities and supports a variety of Layer 2 Control Protocols. It fully supports MEF services such as E-LINE, E-LAN, E-TREE and E-ACCESS. The design has successfully passed MEF CE 2.0 certification to ensure that it complies with strict networking requirements of IEEE, ITU and MEF standards.

The switch has been successfully integrated with an industry leading networking software stack (Aricent ISS), running on the Arria V SoC.

The Carrier Ethernet Switch design includes the following functions and features:

- Altera MAC and PHY IP
- Transmission and reception of packets to/from Ethernet MACs
- Efficient buffering of the received frames in external DDR3 memory
- Using internal on-chip memory for configuration and statistics
- An efficient L2 and L3 classification engine
- Industry standard Ethernet forwarding engine
- Light-weight traffic management capabilities (queuing, scheduling, etc.)
- Modifying the data as it goes through the pipe-line
- IEEE1588v2 Precision Time Protocol capabilities
- Integration with the ARM hard processor system (HPS), provided by Altera SoC devices, including a complete Linux environment that runs drivers, API and control software

A full list of technical features is in Appendix I.

Theory of Operation

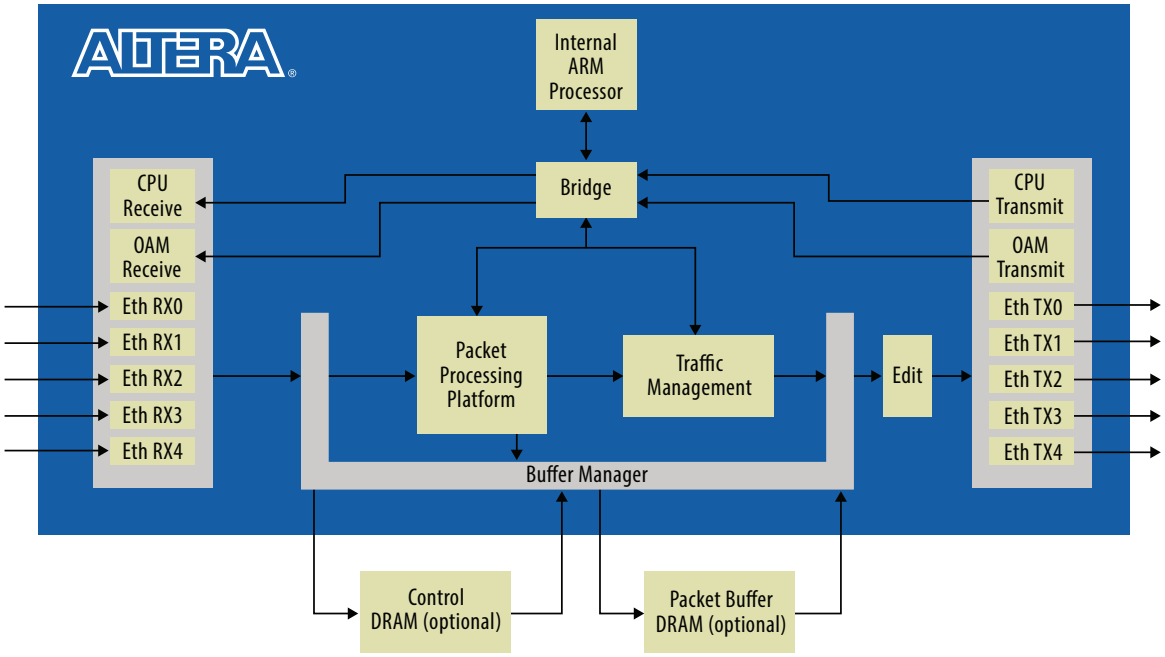
Figure 1 shows a high-level diagram of the Altera Carrier Ethernet switch SoC. The switch has five 10/100/1000 Ethernet interfaces. The switching logic essentially comprises of a buffer system, packet processing platform, traffic management and packet edit blocks.

When the traffic arrives from the five different MACs it is multiplexed into a single stream. The ingress Ethernet frames are marked with additional information such as the IEEE-1588v2 timestamp, the id of the originating port, etc.

The buffer system then stores the payload and the control information (packet length, pointers to the payload, etc.) in external DDR3 memory banks. The system also creates packet descriptors which are associated with the stored payload and control information and are used by the switch for switching decisions.

The packet processor and the traffic management process packet descriptors to classify ingress frames into flows, make switching decisions, queue the traffic and schedule the egress frames according to the switch configuration.

Figure 1. Altera Carrier Ethernet Switch SoC design



On the egress side, the traffic management/buffer system will forward the original frames with all headers and payload towards the packet edit block. The payload will be read from the external DDR3 buffers and the associated buffer memory will be released and marked as free by the buffer manager.

The packet edit function can be configured to add, remove or swap VLANs and can also update the class of service information in the frames.

The system then sends the packets to one or more of the five egress MACs, which may provide additional functionality such as IEEE1588v2 timestamping, CRC insertion, flow control operations, etc.

The design also allows traffic arriving from Ethernet ports to be sent to the SoC's ARM processor and traffic generated by the SoC's ARM processor can be sent to the Ethernet ports.

Additionally, the switch can support Ethernet OAM functionality, with optional hardware acceleration.

Implementation details of the switch are out of scope of this document, and can be made available upon request.

Triple speed Ethernet MAC with 1588 support

The design uses the Altera MAC and PHY MegaCore IP. We have used the "10/100/1000Mb Ethernet MAC with 1000Base-X/SGMII PCS" variation of the core to instantiate five MAC modules. A MAC/PCS variation can be created either using the MegaWizard method or with the QSYS tool. In this case we used the QSYS tool. The advantage of using QSYS is that the Avalon-MM management ports can be wired up easily and required address decoders are generated automatically.

Additionally, the design enables timestamping with PTP 1-step clock support for IEEE1588v2. The timestamping mechanism requires a time-of-day clock to be supplied to the design, which was included in the Carrier Ethernet Switch demo design.

DDR3 Memory

The design uses two 1024 MByte DDR3 SDRAM interfaces for very high-speed memory access. The DDR3 controllers are instances of an Altera soft IP core. One memory bank is used for storing the payload, and the second one is used to store control data. Each of the memory interfaces has 32-bit data bus (two×16 devices internally). The memory devices are running at 333MHz.

IEEE1588v2 Precision Time Protocol (PTP) Support

The design supports 1-step transparent clock operation. All received frames are timestamped by the Ethernet MACs. The switch then classifies the PTP frames, marks them for further treatment by the egress path and forwards them in the same way as all other frames, according to the switch configuration. The timestamp is ignored for all non-PTP frames.

When PTP frames exit the switch, the MAC compares the received timestamp of the frames with the current time-of-day value and calculates the delay the frame experienced in the switch. The MAC then updates the value of the correction field in the PTP frame with the measured value of the delay.

For PTP frames with UDP encapsulation, the design ensures that the IPv4/IPv6 UDP checksums are correct.

Hard Processor System (HPS)

Altera Arria V SoCs integrate a dual-core ARM® Cortex™-A9 MPCore™ processor-based hard processor system (HPS, labeled Internal ARM Processor in Figure 1) consisting of processor, peripheral and memory interfaces, connected with the FPGA fabric using a high-bandwidth interconnect backbone. The HPS also contains a hard DDR3 controller to connect to external DRAM operating at 533MHz. The design uses DMA for efficient transfer of the data between FPGA and HPS memory. The HPS hosts the embedded-software portion of the carrier Ethernet switch design.

Altera SoC Embedded Software

The design uses the Altera SoC Embedded Design Suite (EDS) to develop embedded software for SoC devices.

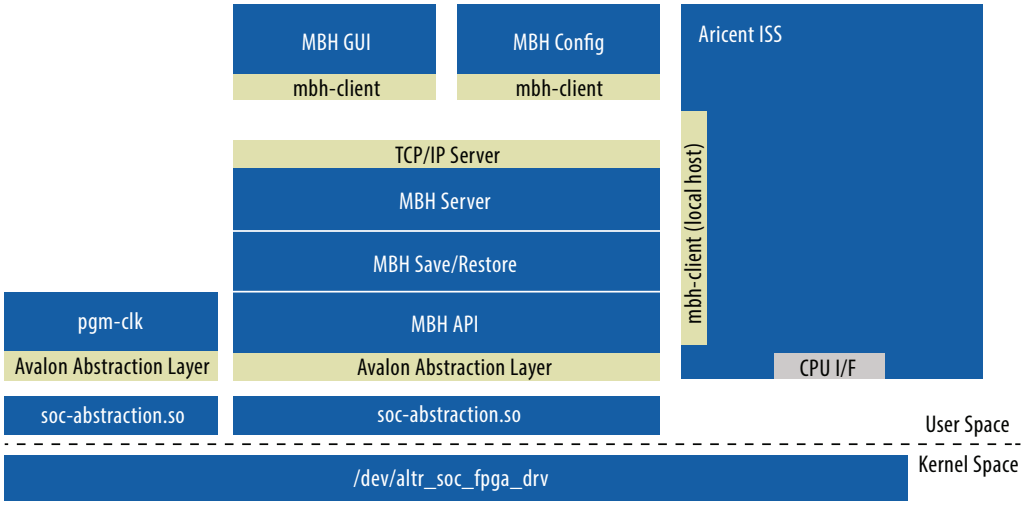
The Carrier Ethernet Switch was developed together with the software that consists of device drivers, dynamic libraries, embedded executables and shell scripts. We also developed executables on PC (running under Windows or Linux) to demonstrate how an external management system can be used to control the embedded switch.

The embedded software runs under Linux on the HPS. This software is written in a portable form and can also be compiled and executed on a generic CPU running Linux or Windows. In this case, the CPU should be connected to the system through the PCIe.

Alternatively, users may want to use a generic control CPU connected to the switch through an MII port. The generic architecture of the design allows this by replacing the FPGA-to-HPS bridge with simple PCIe or MII custom bridges.

Figure 2 shows the structure of the software that was developed for the Carrier Ethernet Switch. The most important building block of the software system is the altr_soc_fpga_drv.ko device driver that provides access to the FPGA via the HPS-FPGA AXI bridges. The driver allows DMA to be used for fast transfers of the data between the HPS memory and FPGA.

Figure 2. Software components of the Carrier Ethernet Switch design



The soc-abstraction.so dynamic library provides the hardware abstraction layer (HAL), which provides the interface between the Avalon abstraction layer and the HPS-FPGA bridges. This library is used by higher layers of the software to control the FPGA.

An important part of the software package is the MBH-server which runs on top of the soc-abstraction.so dynamic library. This provides server functionality which allows the Carrier Ethernet Switch software to be used in a typical client-server architecture. The clients can be either local (on the same HPS) or remote (on another PC, or another embedded device).

The Carrier Ethernet Switch software provides examples of two software clients for the MBH-Server:

- MBH-Config – a local embedded software client running on the ARM HPS system, which is used to configure the switch with a predefined sets of configurations.
- MBH-GUI – a generic visual client that can be executed on a Linux or Windows machine and provides full remote management capabilities for the switch.

The two additional SW modules shown on Figure 3 are: pgm-clk – a utility that allows programming the oscillators controlling the FPGA transceivers, and the Aricent ISS stack. Aricent Intelligent Switching Solutions (ISS) is a comprehensive, feature-rich software product for developing a wide range of intelligent Ethernet switching applications. The ISS stack was ported to the Altera Arria V SoC and is completely integrated with the drivers and with the API of the Altera Carrier Ethernet Switch design.

HW Platform

The Altera Carrier Ethernet Switch demo was implemented on the Arria V SoC Development kit. This kit is a complete design environment that includes both the hardware and software Altera customers need to develop Arria V SoC designs.

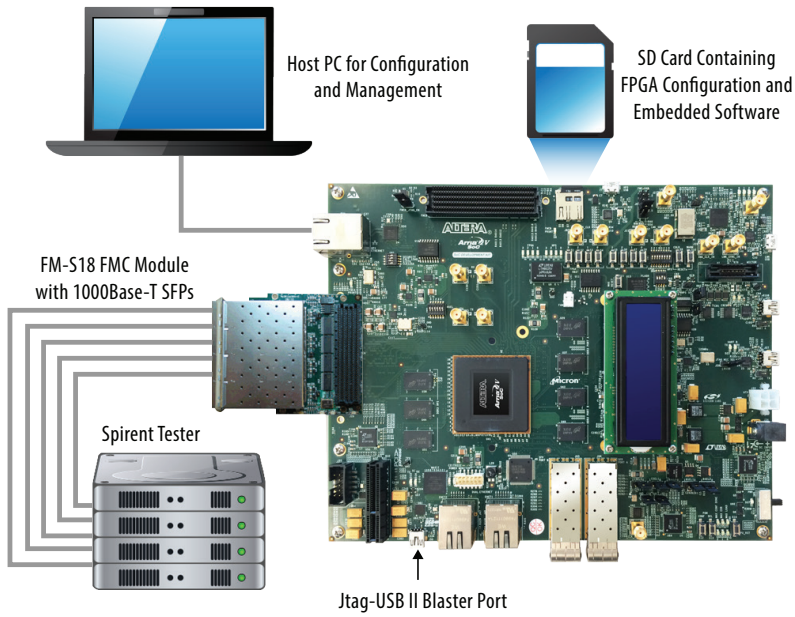
The development board features the following major component blocks:

- One Arria V SoC (5ASTFD5K3F40I3)
- Memory
  - One 1024-MB HPS DDR3 SDRAM with Error Correction Code
  - Two 1,024-MB FPGA DDR3 SDRAM
  - One Micro SD flash memory card
- Communication Ports
  - One PCI Express x4 Gen1/Gen2 socket -not used
  - Two FPGA mezzanine card (FMC) ports
  - Two SFP+ ports (not used)
  - One Gigabit Ethernet port (used for remote management)

Since the Arria V SoC development kit does not have enough Ethernet ports for the Carrier Ethernet Switch design, an FM-S18 FPGA mezzanine card with 8x1GE SFP sockets is used.

Figure 3 shows the setup of the Arria V SoC board when operating the Altera Carrier Ethernet Switch. The FM-S18 extension board connected to the FMC port of Arria V SoC development kit. These ports are the network ports of the switch and can be connected to a network tester, e.g. Spirent.

Figure 3. Carrier Ethernet Switch on Arria V SoC development platform



The system boots from the microSD flash card, which contains the FPGA configuration and embedded software. The user can then use ssh to login to the system and execute the MBH-config system that controls the switch.

Results

Arria V SoC is a flexible platform which allows efficient implementation of complete System-on-a-Chip solutions. Such solutions can combine networking switch, Ethernet ports, ARM HPS and can use external DDR3 memory for storing and processing the frames.

The EDS software environment provides all the tools which are required to rapidly develop embedded software for client-server architectures on embedded Linux platforms and embedded/remote management systems.

The Carrier Ethernet Switch design was certified by the Metro Ethernet Forum as part of the Carrier Ethernet 2.0 certification program. This demonstrates the capabilities of the Arria V SoC technology to implement such functionality and operate it alongside third party software (Aricent ISS).

Table 1 summarizes the resource usage by the Alter Carrier Ethernet switch design.

Table 1. Resource usage by the Carrier Ethernet Switch on Arria V SoC

Module	ALMs	Memory M10K
CPU Subsystem	2389	30
Switch Subsystem	44558	829
5x Ethernet MAC (with 1588)	5x3752	5x7
DDR3 for Packet payload	5368	43
DDR3 for Packet Control	4339	39
Total	~77000 ALMs	~1000 M10Ks

The total power consumption during the boot was ~5.6W when measured in a controlled indoor environment. The total power consumption when the switch was fully loaded with Ethernet traffic (all five ports receiving and sending 5x1Gbps of traffic) was measured as ~6.6W.

The measured power consumption in both cases included ~1.6W of power used by the ARM HPS system. Excluding the HPS power from the measurements, the switch consumed ~5W, with power consumption of the FPGA logic, transceivers and memory I/Os included into measurements.

Appendix I - Carrier Ethernet Switch specifications

Capacity

- >5Gbps traffic capacity (full duplex)

Ethernet Interfaces

- Five 10/100/1000 Ethernet MACs
  - Based on Altera's Triple-Speed Ethernet MegaCore function
  - 10/100/1000-Mbps MAC, PCS, and PMA
- Two additional dedicated interfaces to the CPU for control and OAM traffic
- Many external Ethernet interface options
  - MII (10/100 Mbps), GMII, RGMII, and SGMII (10/100/1000 Mbps), 1000BASE-X, and TBI (1 Gbps)
- Jumbo frames (9.6k)
- Pause frames (802.3x)
- MetroEthernetForum UNI & NNI support

Timing Recovery

- IEEE1588v2 transparent clock (1 step)
  - Highly accurate hardware time-stamping unit within the MACs
- Synchronous Ethernet ready

Switching

- Ethernet switch with learning and aging
  - 128 L2 addresses, including learned, static and multicast
  - MAC address pinning, MAC address filtering
- VLAN (802.1Q), Provider Bridging (QinQ / 802.1ad)
  - 128 VLANs
  - VLAN tag removal, addition, swap
- Metro Ethernet Forum E-Line, E-LAN and E-Tree support
- Full L2 control protocol handling as specified by MEF
- IGMP snooping

Protection

- Support for Spanning Tree Protocol (STP, RSTP, MSTP)
- Link Aggregation
- Hardware accelerated protection switching suitable for Ethernet linear protection (G.8031) and Ethernet ring protection (G.8032)

Classification

- Customisable classification engine
- Traffic flow can be determined based on port, outer VLAN ID, MAC SA/DA, IP SA/DA and Ethertype
- Class of service within a flow can additionally be determined based on VLAN PCP bits and/or IP DSCP bits

Policing

- 128 policers
- Three levels of policing – per port, per flow and per CoS
- srTCM, trTCM and MEF8 modes, colour aware and colour blind
- Egress colour marking via VLAN DEI or PCP bits

Traffic Management

- Single layer scheduler - eight queues per port
- Traffic shaping per queue and per port
- SP and WFQ scheduling
- Congestion control unit
- Over 50ms of traffic buffering per port

Security

- Rate limit different types of traffic on a per port / EVC basis (unicast, configured multicast, broadcast and flooded traffic)
- L2 address learn limits per port / EVC
- L2 address pinning, blocking, white-list

Altera Corporation
101 Innovation Drive
San Jose, CA 95134
USA
Telephone: (408) 544 7000
www.altera.com

Altera European Headquarters
Holmers Farm Way
High Wycombe
Buckinghamshire
HP12 4XF
United Kingdom
Telephone: (44) 1 494 602 000

Altera European Trading Company Ltd.
Building 2100
Cork Airport Business Park,
Cork, Republic of Ireland
Telephone: +353 21 454 7500

Altera Japan Ltd.
Shinjuku i-Land Tower 32F
6-5-1, Nishi Shinjuku
Shinjuku-ku, Tokyo 163-1332
Japan
Telephone: (81) 3 3340 9480
www.altera.co.jp

Altera International Ltd.
Unit 11- 18, 9/F
Millennium City 1, Tower 1
388 Kwun Tong Road
Kwun Tong
Kowloon, Hong Kong
Telephone: (852) 2945 7000
www.altera.com.cn

Altera Corporation Technology Center
Plot 6, Bayan Lepas Technoplex
Medan Bayan Lepas
11900 Bayan Lepas
Penang, Malaysia
Telephone: 604 636 6100