Ethernet switch technology increases performance and flexibility, lowers latency

As Gigabit Ethernet performance continues to increase beyond one gigabit per second (Gbps), another nagging issue arises: latency. Evident especially in data centers where applications are distributed among clusters of servers and storage equipment, latency and data delays cause problems with synchronization of the overall applications. Latency issues can degrade data center applications to the extent that the data center's architecture becomes disjointed, as shown in Figure 1. The compute elements of the data center use specialty switches that might use low-latency communication technologies such as InfiniBand. Storage within the data center may be connected via Fibre Channel. Finally, the enterprise is connected with the inexpensive, scalable topology of Ethernet.

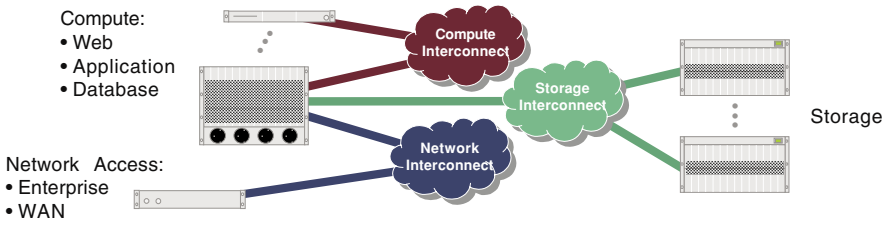

Figure 1

If you're thinking that these data center challenges have nothing to do with AdvancedTCA systems design, think again. These performance and latency issues extend beyond the data center into the heart of the Advanced Telecom Computing Architecture. AdvancedTCA backplane and switching give rise to exactly the same performance and latency issues. Many common AdvancedTCA systems have storage, compute, and network interconnect functions all using the AdvancedTCA backplane to pass data and control information. AdvancedTCA systems are specified to deliver the same kind of data center services, creating significant risk that AdvancedTCA will not be able to meet backplane switching latency requirements for these kinds of applications.

At the Network Systems Design Conference (NDSC) I ran across a new Ethernet switching technology that provides latency characteristics for one and 10 Gb systems that fall well within the latency threshold for data center applications. This new technology from Fulcrum Microsystems enables data centers to adopt a unified, 10 Gb data center architecture with latencies low enough to rival the special compute interconnect topologies of today. It also has the possibility of enhancing AdvancedTCA backplane switches to meet more stringent latency requirements between AdvancedTCA blades as multimedia applications become more prevalent in these systems.

Unified data center

Fulcrum Microsystems uses two key internally developed components to provide low latency, high performance Ethernet switch chips called Nexus and RapidArray. The technology's main objective is to enable unified data center architecture through the use of 10 Gigabit Ethernet. So, with 10 Gigabit Ethernet switching technology that meets performance and latency requirements of the data center, the infrastructure simplifies to that shown in Figure 2.

Nexus is a fully nonblocking terabit crossbar switch. The crossbar itself is benchmarked at three nanoseconds to transport data across chip, which includes arbitration, setup, and tear down of the connection through the switch. So, this fundamental building block enables system-level latency that rivals low-latency technologies such as InfiniBand, yet leverages less expensive, more widely deployed Ethernet topology.

RapidArray is high performance Static Random Access Memory (SRAM), with special features that enable it to run up to twice as fast, yet offer the same yield and density as standard SRAM. One additional feature of asynchronous RAM is that the RapidArray's overall power consumption is typically significantly lower, with consumption taking place only when there is activity through the switch. For example, under full 10 Gb traffic load through the Ethernet switch products using this technology, the products use about 150 mW per port.

Latency

So, what is the data center latency threshold? Figure 3 compares topologies with Fulcum's Nexus plus RapidArray 10 Gb Ethernet switch products called FocalPoint.

The typical data center latency threshold is around 200 to 300 nanoseconds. Fibre Channel provides storage interconnect characteristics that just meet this threshold. The standard 1 and 10 Gb Ethernet products available come with latencies far outside the acceptable bounds, 1 Gb being benchmarked in the low microseconds, and 10 Gb still being more than 500 ns. This issue relegates Ethernet switching to

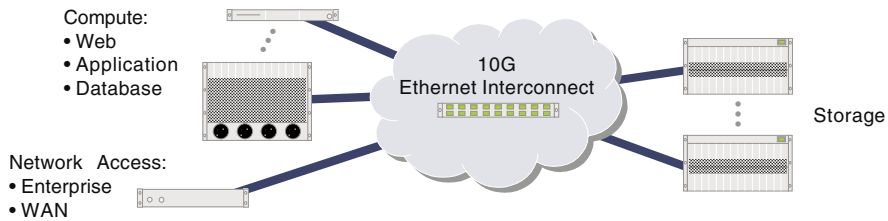

Figure 2

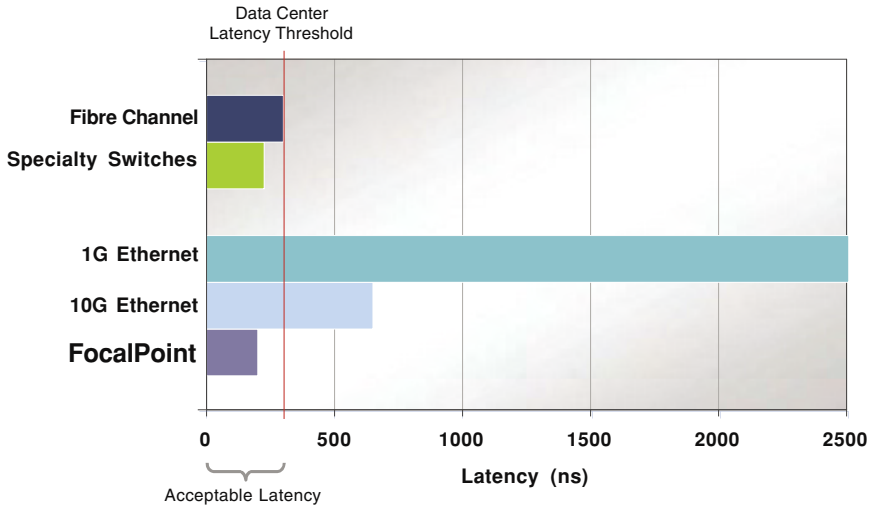

Figure 3

the network interconnect, and specialty switches using InfiniBand technology take care of the actual compute interconnect. Specialty interconnects offer comparable latency to FocalPoint, around 200 ns. The FocalPoint Ethernet switch product that incorporates the Nexus and RapidArray technologies benches well within the latency threshold, making Ethernet once again a viable solution as a compute interconnection within the data center.

Performance

Once the latency issue is resolved to enable a unified, 10 Gb interconnected data center, the architecture simply rides the performance advances of standard Ethernet technology. The simple fact is most data centers and network providers are still an order of magnitude or two from standard Gigabit Ethernet. Therefore, viable technology up to 10 Gbps provides more than enough performance road map for the vast majority of applications involving Ethernet.

AdvancedTCA applications

So what does this have to do with AdvancedTCA? Well, the entire purpose of the AdvancedTCA standards is to create standards and interoperability between components within an AdvancedTCA system that ultimately drives low cost, high reliability equipment for a wide variety of applications. But the backplane fabric interconnect remains open to a variety of standards. PICMG 3.1 through PICMG 3.5 define the use of Ethernet/Fibre Channel, InfiniBand, StarFabric, Advanced Switching, and Serial RapidIO. The reason is very similar to the data center issues already described. But this also has the potential to be the *weak link* in the interoperability chain that is AdvancedTCA. Lack of interoperability in any area of AdvancedCTA runs the risk of slowing adoption and limiting interoperability.

The reason why Ethernet isn't always the obvious choice for the high speed backplane is congestion management. Ethernet link layer flow control can be used at Layer 2 for congestion management, but there is no resolution into the flows riding on top of the link layer. For example, one particular flow may be able to be throttled, allowing all the other flows using the link layer to resume normally. The IEEE 802.3 Working Group is currently adding additional congestion management features to Ethernet in order to solve this shortcoming. In the meantime this is the main issue blocking an interoperable AdvancedTCA backplane solution.

Fulcrum's Ethernet chip has some capability to identify unique flows within the first 16 bytes of the packet. By identifying these flows, the chip can use its flexible port logic to implement some degree of throttling lower priority data on a link in congestion management situations. Furthermore, the flexible port logic allows the ability for the chip to switch based on this unique 16 bytes as programmed. So, these chips provide the ability to add flexibility by including congestion management and proprietary header switching within an AdvancedTCA system.

Programming the device

Software that enables easy integration into the platform and OS used is key to successfully incorporating sophisticated switch chips that support port switching and configuration. Figure 4 shows Fulcrum Microsystems' software architecture for the device.

As Figure 4 depicts, a device driver implements calls to control and configure the device. A device access API exposes all the services available from the chip. From there, the device would typically be integrated with a Layer 2 switching protocol stack. Fulcrum has used the device access API to create the adaptation layer to the LVL7 switch stack. Fulcrum has also implemented the software using an operating system abstraction layer to facilitate porting to other operating systems. Thus far, the software environment has been ported to Linux and VxWorks running on the PowerPC processor, although this model can very easily be migrated to other processor platforms as needed.

The following services are provided through the device access API:

- Initialization: Device initialization and configuration for operation
- Switch management: Setting switch attributes, setting switch to active/inactive, and getting a list of switches
- Virtual Local Area Network (VLAN) management: Configuration and management of port-based VLAN

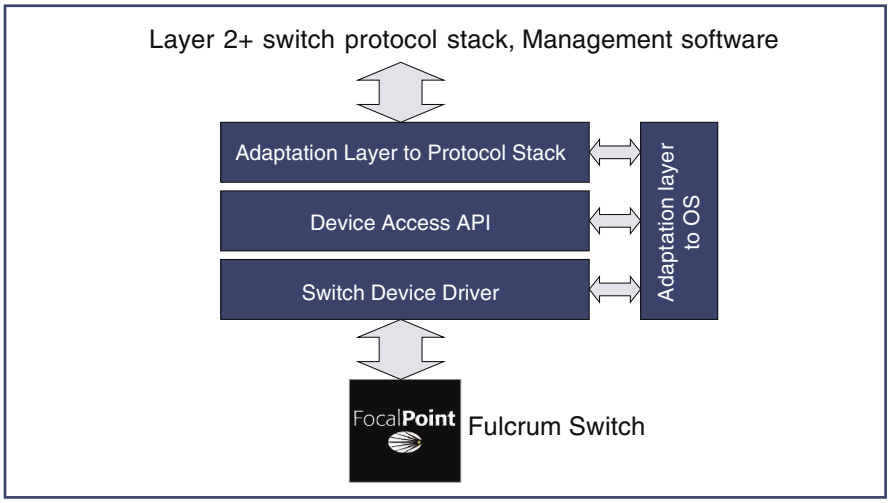

Figure 4

- Port management: Port enable/disable and attributes
- Address management: MAC address insertion, removal, and set table aging characteristics
- Link aggregation management: Setting up link aggregation groups from a set of ports and managing characteristics of those port groups
- Quality of Service (QoS): QoS global management for the device, per port, per VLAN, or per MAC address
- Statistics: Access statistics at multiple levels, the switch chip, per VLAN, and per port
- Packet transmission and reception characteristics: Customize the header switching characteristics of the device
- Event reporting: Reports for anomaly situations or activity through the device including switch and per port state, VLAN access violations, and transmission and reception characteristics

The Fulcrum driver is also implemented using a buffer management API that expects to be able to get and free a buffer, set a buffer length, and set the content offset in the buffer. These are standard services provided by UNIX *mbuf* and Linux *zbuf* facilities, so implementing the buffer management API is straightforward.

For a flavor of the API, Figure 5 shows an API call in the Fulcrum services API that gets the attributes from the switch chip.

Using the port aggregation API, users can set up and control transmission of a set of input ports through the switch chip. Link aggregation groups can be configured to have packets from the group dropped, sent to the CPU, or passed through to an output port.

The QoS and queue management API provides a very robust set of facilities that allow per port and global management of receive and transmit port queues. High and low watermarks can be set up globally and/or on a per port basis to implement up to eight queue priorities per port for the implementation of IEEE 802.1P.

Products

The Fulcrum Microsystems product line consists of the PivotPoint family of SPI-4.2 line card switch chips and the newly announced FocalPoint device, a backplane switch chip specifically targeted for applications such as AdvancedTCA backplane switching. The FocalPoint FM2112 is an eight port by 10 Gb plus 16 port by 1 Gigabit Ethernet switch. The eight ports are XAUI interfaces. Each interface supports quad Serializer/Deserializer (SERDES) and single SERDES modes. Each 10 Gb port can be overclocked up to 12.5 Gbps. The FocalPoint FM2224 is a 24 port by 10 Gigabit Ethernet XAUI switch.

Both FocalPoint devices include eight priority level queue management with programmable Weighted Random Early Discard (WRED) algorithms to support IEEE 802.1P priority queuing specifications. The part can be programmed for either strict priority or weighted round robin scheduling algorithms. Jumbo packet support is also included. MAC tables within the part are 16K entries and provide configurable hardware aging and the ability to lock entries in the table. For switch fabrics passing Ethernet on top of proprietary headers, the parts can also be configured to bypass the proprietary header by adjusting an offset into the packet or use the proprietary header to make switching decisions. In addition, the parts gather multiple statistics for RMON and SNMP instrumentation and incorporate VLAN recognition for port-based VLAN processing.

Fulcrum Microsystems also offers the FocalPoint FM2224-EP 10 Gigabit Ethernet interconnect evaluation platform. The platform has options for a variety of connectivity options, an embedded Linux operating system, and commercial Layer 2 control plane software from LVL7. The part itself comes with drivers and software to ease integration with operating systems and control plane software. The API provides a simple interface for controlling the flexible port logic within the chip, and the adaptation layer makes it possible to port the software to a number of operating systems.

The current PivotPoint products include the FM1010 and FM1020. These parts feature a SPI-4.2 interface that is quickly becoming the de facto standard for communications chip interfaces. SPI-4 is used by multiple NPU manufacturers, security coprocessor, and interface chip developers. The FM1010 features six SPI-4.2 interfaces; the FM1020 provides three SPI-4.2 interfaces in a smaller package.

All these parts are manufactured using a 0.13 micron process for a high level of

```
fmDriverGetSwitchInfo ()
Synopsis:     fm_status fmDriverGetSwitchInfo (fm_int switch, fm_switch_info *argv);
Description:  Get information about this switch.
Inputs:       switch: Switch number
              argv: Pointer to a location to store the information
Outputs:      *argv: Information about this switch
Returns:      FM_OK: Success
              FM_FAIL: Cannot access the switch

The fm_switch_info structure is:
struct {
    fm_int model;           // The model of this switch
    fm_int env;             // The environment for this switch
    fm_int32 info[32];      // Information about the environment
                            // around this switch:
                            // info[0] = Control CPU clock
                            // info[1...8] = EPL clocks
fm_int physicalPorts[FM_MAX_PORT];
                            // Map of logical port to physical port
                            // physicalPort[0] = 0 (CPU)
                            // physicalPort[1] = X
                            // ..etc..
} fm_switch_info
```

Figure 5

integration in a small footprint. For example, the FM1010 product is a 1232-ball Ball Grid Array (BGA) geometry with the FM1020 being half that size.

Conclusion

Standardization and ubiquity in data centers and AdvancedTCA alike drive easier to maintain, lower cost architectures for a wide variety of applications. Ethernet switching products that address latency and congestion management issues look like enablers to extend Ethernet topologies far into the data center and into the heart of the AdvancedTCA backplane.

For more information, contact Curt by e-mail at cschwaderer@opensystems-publishing.com.