Slide 1:

Hello everyone, I am Kallol. On behalf of NucleoDyne Systems I would like to welcome you all to the talk on Ethernet Over PCIe Express Bus.

Slide2: Outline of presentation

Here is the outline of today’s presentation. At first I will provide an over view then we will go over, #1 the highlights of PCI Express protocol , #2 System to system communication using PCIe protocols, #3 Advantages & Issues and finally the Application areas.

Slide 3: Overview

Traditionally Ethernet devices are used to establish communication between two systems. On the other hand PCI Express bus is used as a motherboard-level interconnect to link motherboard mounted peripherals or add-in devices.

We explored PCI Express switch capabilities to interconnect multiple computers to establish communication like Ethernet. The new use model relies on making remote system’s memory available to local system for load/store CPU instructions and DMA operations. This has been captured in the next two slides

Slide 4: This slide depicts PCI express switches connect multiple computer systems at system bus level.

Slide 5: In this slide we have shown that a user space process directly reads or writes information to the address space of a remote process that runs on a different computer system.

Slide 6:

As I have explained in the previous slides the new way of inter process communication across two systems use direct load/store or DMA from one system to another. However, there are many existing applications written using socket programs and TCP/IP protocol. So we experimented with sending Ethernet frames over PCI Express link. A communication thus established over the PCI Express switch based interconnect demonstrated favorable metrics, like lower latency, higher throughput , lower cost, lower power consumption over the traditional TCP/IP over Ethernet.

Slide 7: Highlights of PCI Express Protocol

* PCIe is a packet based Serial Bus Protocol
* It is point to point and bi-directional; data can flow in both directions at the same time
* Depending on version of the protocol on each lane the effective data rate 2, 4 or 8gbs
* Two or more data lanes can be aggregated between two devices, adding up the bandwidth
* PCIe implements Layer architecture
  + The Software layer is typically a PCIe Driver
  + The transaction layer primarily creates PCI Express requests and completion transactions (TLP). It supports optional end-to-end CRC (ECRC) to ensure content of a TLP is properly conveyed from one end to destination even when it passes through a PCI Express Switch.
  + The data Link layer provides data Integrity across link – main responsibilities are error correction, detection and reliable transmission across link as well as support link management functions. The layer implements sequence number, LCRC, Retries, Flow control and power management support.
  + The physical layer is responsible for scrambling, encoding, framing and electrical Signalling
* PCIe has path based routing and globally flat addressable fabric
* It supports QoS

Slide 8: Two PCs are connected via PCIe switches and external PCIe cable as shown in the picture. The switches are mounted on add-on PCIe adapter boards.

Slide 9: Let’s go over a step-by-step illustration of how data transfer takes place between two nodes based on our proposed use model.

We have two computer systems A and B. These are connected using PCIe switches and an external cable.

Suppose an application on system A wants to send 4 bytes of data to an application on system B.

The Application will issue a store instruction say

store r3, < to an address of the application on system B>.

The root complex on system A will turn this write into a transaction layer packet and forward to adjacent switch.

The switch will decide that the TLP has to go out through an egress port to the partner switch on system B over external PCIe bus.

When the PCI express switch on system B will receive the TLP, and determine that the transaction is destined to local system’s DDR memory. It will modify the TLP header and re-route towards the root complex on local system.

The root complex will receive the TLP, interpret and convert it to a 4 byte write to DDR.

Slide 10: In the previous slides an application wrote data bytes directly to another remote application’s address space. If the applications wish to establish communication through TCP/IP stack we can do so using PCI express switches as shown in the picture. We have implemented this model and obtained tests results running network benchmarks.

Now let us look at comparative results.

Slide 11: Throughput

In PIO mode, with 4 bytes write at a time we have achieved throughput of 9.97 Gigabits per second over 4 lane Gen2 PCIe link.

The CPU utilization number is higher as we have used PIO to move data. DMA should bring down the utilization number.

Slide 12: Latency

As you can see in this graph the latency of data access is 35% better than an Ethernet adapter, i.e. with PCIe driver we could complete 35% more request response transactions.

Slide 13: Advantages

PCIe switches consume less power than Ethernet controllers. Typical power consumption on a 16 lane Gen2 device is approximately 2.5W.

The switches cost much less than Ethernet controllers. The cost is around $1 per lane, so an 8 lane device costs 8 dollars. The big savings is in ecosystem, i.e. in clocks, connectors, cables and test equipments.

Slide 14: We do see promising prospect in our PCIe switch based interconnect model. However we also feel that that at this moment it has some limitations which need to be overcome before it can be deployed as widely as the Ethernet based network. Two predominant limitations are:

1. Limitation on the physical spread of the network

&

1. Present upper cap of 64k nodes

Currently the maximum length of the available external cables is approximately 5 meters. There are optical connectors that extend the length upto 100 meters and also at most 64k devices can be interconnected in a fabric. However, a Gen2 cluster will not scale much beyond 32 nodes without penalizing performance. In Gen3 upper cap is much larger.

Slide # 15: Let’s look at the applications where the PCIe switch based model can be deployed for added advantages:

We propose following applications:

1. As a replacement of parallel buses like VME (Versa Modular Eurocard bus) buses: A 4 lane PCIe switch shows higher data transfer rate than an existing VME parallel bus even in source synchronous mode (9.97 Gbps Gbps over 2.56 Gbps). Also higher bandwidth may be attained by aggregating more lanes.
2. In a multi-node cluster development: PCIe switch based interconnect can be used to successfully build a multi-node cluster where lower latency is required and the nodes are not physically far apart.
3. In NVRAM mirroring: Many of the enterprise class Filesystems mirror NVRAM writes to a partner nvram. If DDR memory is battery backed up, PCIe switch based interconnects can successfully replace such NVRAMs.
4. In High frequency trading systems and banking industry: A PCIe switch based distributed processor network can improve the performance in a high frequency trading system because of lower latency.

Slide #16: Conclusion

We have proposed a new way on communication between applications running on different systems, where a PCIe switch based interconnect can be used with advantages over the traditional Ethernet based network. We also went over test results and potential applications.

Slide 17: Who we are

Now please allow me to introduce ourselves to you.

We are a system software services company that is located in Cupertino, California. We develop low level system software, kernel modules and device driver. We also have expanded into developing end to end embedded systems starting from processor core modification/development, system board development, OS porting or new OS development to custom system oriented application development.

Slide 18:

On behalf of Nucleodyne, I would like to thank you all. Thank you for listening and for your time.

With that let’s go over to question and answer session: