Friday, February 15. 2008

Infiniband in the datacenter

EETimes wrote an interesting article about some statements of our top management regarding the future architecture of datecenter. In "Sun preps Infiniband for broad net role" EETimes writes:"In most data centers there are three installations often run by three different teams and using systems from three different vendors," said John Fowler, executive vice president of the systems division at Sun. "We think there are technology and economic advantages for bringing these three together," he said at a press event Wednesday (Feb. 13).
Fowler said Sun will roll out in 2008 products to run all data center traffic on Infiniband.Nice to see Andy and John sharing my opinion;) Joke beside. I talked about something similar with colleagues some month ago while driving to a Sales Kick Off somewhere in the middle of Nowhere and thought and talked about such an solution while designing the network infrastructure for a really big RfP. It´s really obvious, that Infiniband will play a bigger role in future. Andy and John have forgotten more about computer systems than i know about computer systems, so it seems my opinion isn´t completly weird.
When you look into a datacenter you have multiple fabrics. You have your SAN, you have your IP networks. They are separted. You have your SAN, you have your Ethernets. In the normal mindset TCP/IP is mentally hardwired to Ethernet or ATM and SAN storage is hardwired to FC. But it hasn´t to be this way: FC is just SCSI over a serial line. TCP/IP is inherently independent from the transport channel (just think about RFC 1149).

Now just step away from the common design principles of a datacenter for a few minutes. You don´t want to have a certain interconnect, you want to use certain protocols, SCSI and TCP/IP for example. At the end it doesn´t matter how this protocols reached your systems and storage.

All you need is a high speed low latency interconnect. High speed to fullfil all bandwith needs of all the consolidated fabrics. Low latency to have at least the latency characteristics of the fastest protocol in the fabric.

Infiniband is such an interconnect. The effective throughput of a quad link data rate infiniband (expected for this year) is 32 GBit per second (there is a roadmap up to 96 GBit per Second). More than enough to substitute some Gigabit Ethernet and Fibre Channel ports. The latency of Infiniband ist even faster than the one of Fibrechannel. And by the way, the primary usage of Infiband in HPC is just an accident, Infiniband was designed as an I/O interconnect at first.

All you have to do is to find efficient ways to encapsulate TCP/IP and FC into Infiniband. There are already such methods like SRP, the SCSI RDMA protocol to enable block device transfers over an Infiniband fabric. Solaris has already the capability to use SRP storage. You can download it on the Sun website. TCP/IP over IB is in solaris for quite a while now.

Okay, when all components are finally in place you get a high performance fabric for all jobs without loosing performance. Sure, this isn´t something for small installations but for bigger installation it´s a viable way to get rid of the multiple fabrics.

We talk about datacenter wide deployments here. So the shorter cable length of Infiband compared with Ethernet or FC isn´t much a factor. Ffor connectivity to legacy networks just use some UltraSPARC T2 system with Infiniband on one side and the legacy interconnects on the other side.

By the way, did you really thought, we designed Sun Datacenter Switch 3456 just for a few huge HPC gigs per year?

Posted by Joerg Moellenkamp in English, Oracle at 22:44

Does Infiniband have error correction? As far as controller cards are concerned, I only know about Mellanox Products, are there any other? Oh, convincing the HBA vendor to build TOE functions for IP-over-IB into the cards surely would be a nice idea. But in the end you are right, high bandwidth and low latency at a cheaper price than 10Gig Ethernet sounds promising.
Anonymous on Feb 17 2008, 19:01