Market Potential for 100 Gb/s and 40 Gb/s Ethernet in HPC Applications

Alan Benner, Petar Pepeljugoski

IBM

The HPC Applications of Higher-Speed Ethernet

- HPC applications for higher-speed networks include roughly 3 major uses:
  - Cluster/MPP: inside the cluster/parallel system
    - for executing the parallel applications
  - Storage: between compute & storage nodes
    - for loading data into apps & storing results
  - LAN: between compute nodes and external machines
    - for visualizing results or FTP to other machines.

Higher-speed Ethernet for Cluster/MPP Networks

- Steady drive to larger aggregates of more-powerful chips, more cores
  - Top500 data: steady 2x/year (100% CAGR) in aggregate system speed
    - Contributions from both processors (Moore's law), plus larger & faster networks
    - No signs of slowing down - several 1 PF-capable machines planned for '08-'09
- Faster 40 & 100Gbps PMDs needed just to match node performance
- A simple design example – a 2008-2010 Peta-scale machine:
  - 2 PF system, with 100 GF/node & 0.1 Byte/FLOP BW ratio, would need ~20,000 nodes with 100 Gbps ea. ➔ at least 40,000 links of 100G
  - Note: improved protocols (latency & overhead) also needed, to match bandwidth & packet-rate (packets/sec) of 40 & 100 Gbps links

Higher-speed Ethernet for HPC Storage & LAN Networks

- Requires fast, widely-interoperable, robust standard links
  - Large overlap with data center requirements, at higher bitrates
- 100Gb/s (& 40Gb/s) Ethernet would be best option, if available in a cost-effective implementation
  - Aggregation through switches allows speed-matching to use 100G or 40G, as the traffic needs arise
  - Note: Protocols will vary by traffic type
    - Storage: iSCSI or Fibre-Channel-over-Ethernet protocols or similar
    - LAN: TCP or iWARP: distributes results of supercomputing runs to other machines through FTP-like transfers

HPC Applications: 100 Gb/s vs. 40Gb/s

- 40G better matched for NIC link rate (at least until ~2011-2013)
  - Effective BW is limited by I/O buses, host SW, & memory BW to well below 100 Gb/s
  - Mainstream I/O buses (PCIe x16 Gen2) have ~(50+50) Gbps of effective usable BW
    - Matches well to a dual-ported (40+40) Gb/s adapter, not well to 100G

..however..

- 100G better matched for inter-switch & for uplinks from blade chassis
  - HPC & data centers are increasingly using Blade-style packaging, with 1st-level switch integrated inside chassis, cable uplinks carry aggregated traffic
    - Traffic aggregation in chassis switch allows increased rate for uplinks
    - Density of cables & connectors will motivate for 100G over 40G
  - A chassis full of blades w/ 40G links to chassis switches would be well-served by 100G uplinks to central switch boxes

- Net: Both 40 Gbps & 100 Gbps are needed by servers in HPC & data centers

Summary

- Both 100 Gb/s & 40 Gb/s links are critical for linking servers and switches.
  - Server⇔Switch: mostly 40Gb/s (SW & bus limitations)
  - Switch⇔Switch: 100Gb/s (BW to match increasing node perf.)
- Large numbers of 40G & 100G links are needed for Peta-scale systems in near future (2009-2011)
  - >10K-core systems, w/ >>10K-links, are steadily more common
  - For higher-level stack, may use Ethernet, a variation, or alternate with the same 40G & 100G PMDs.
  - Cable PMDs needed: mostly copper (<~10m) and short-range optical (likely parallel shortwave VCSEL, OM3 ribbon fiber)
    - Copper / optic crossover length may be shorter or longer than ~10m, depending on technology & cost evolution