The Economics of High Speed Networking & Network Offload

Jim Pinkerton
Architect, Transports & Connectivity
Microsoft Corporation

Opinions are my own, and do not necessarily reflect that of Microsoft

Where is High Speed Networking Today?

- Network offload strategies have converged on RDMA approaches (Send, Write, Read)
- RDMA has only succeeded within HPC
  - But most HPC sites use 100 or 1000 Mb Ethernet
- RDMA is a commercial failure outside HPC
  - Performance is there, but not price/performance
  - Database benchmark specials, no revenue
  - File storage marginally more successful
- Need to focus on economics
  - Price-performance, not absolute performance

Can RDMA Go Beyond HPC High End?

- Criteria for success: **Total Cost of Ownership (TCO)**
  - **Initial cost** must be less than the cost of the CPU cycles that we're saving
  - **Management costs** must be small
    - Deployment of a 3rd network is a huge obstacle outside of HPC (3rd after Ethernet and Fibre Channel)
      - Specially trained people, parts inventory, integration with existing network management
  - **Reliability/Availability/Servicability**
    - Must be robust enough for five 9's deployment

What is Affordable Link Cost?

- High End Example - Where RDMA shines today
  - $10,000/CPU + support circuitry costs, network load for real application uses 50% of the CPU, 50% of the time (compute, communicate). Thus on average 25% of the CPU is saved.
  - $2,500 link max cost - NIC, cable, switch port.
- Mid-End Example
  - $1000/CPU + circuitry costs, same load,
  - $250 link cost (max)
- Low End Example - Why 100baseT shines today
  - $100/CPU + circuitry costs, same load
  - $25 link cost (max)

Rough RDMA fabric costs
(*informal poll of list price on SC2003 show floor*)

- **Economics of all of today's RDMA fabrics relegate it to the high end**
  - **Costs below could be halved with same result**
  - **Quadrix link cost (8.5 gigabit, ~2 usec)**
    - **$2900 – Adapter $1200, cable $200, switch port $1500**
  - **Topspin link cost (8 gigabit InfiniBand, ~6 usec)**
    - **$1600 - Adap ~$1000, cable ~$100?, switch port $500**
  - **Myrinet link cost (2 gigabit, ?? usec)**
    - **$1245 - Adapter $695, cable $50???, switch port $400**

- Caveats: no haggling list price on show room floor, small switch configs, not based on quotes, probably some vendors gave me more aggressive pricing than list, I probably wrote some of the prices down incorrectly, perf numbers above are from vendors…
- ??? Means wild guess, **your mileage will vary…**
- Above is a snap-shot – pricing always comes down…
- **Apologies to Dolphin, ran out of time…**

Can Network Offload Break Out of High End?

- Economics for high end appear sustainable
- Must reduce TCO to move out of high end
  - Management costs
    - Leverage existing fabric (most likely Ethernet)
    - Leverage existing management infrastructure
  - RAS
    - Ethernet (with TCP/IP) infrastructure is already at five 9's reliability
  - Pricing
    - Potential to leverage Ethernet volumes (not a given)

Leveraging Ethernet

- Ethernet standards status
  - RDMA Consortium Specs are done – handed off to IETF
    - Decreased InfiniBand Verbs overhead by 40-50%
    - InfiniBand moving to support new verbs
    - Includes Block Storage, Sockets Direct Protocol
  - IETF Standard is closing – maybe 2H'04
  - Expect implementations late '04 and '05
- Pricing is strongly dependent on volume, volume is dependent on applications
  - Get off the local subnet (w/ congestion control)
  - API that doesn't require a rocket scientist (well, a-bomb scientist) – i.e. more than just MPI or VIA or uDAPL