A “stretching InfiniBand cables” how-to

(& why-to)

David Southwell

June 22 2006 – OpenFabrics Paris Workshop

Contents

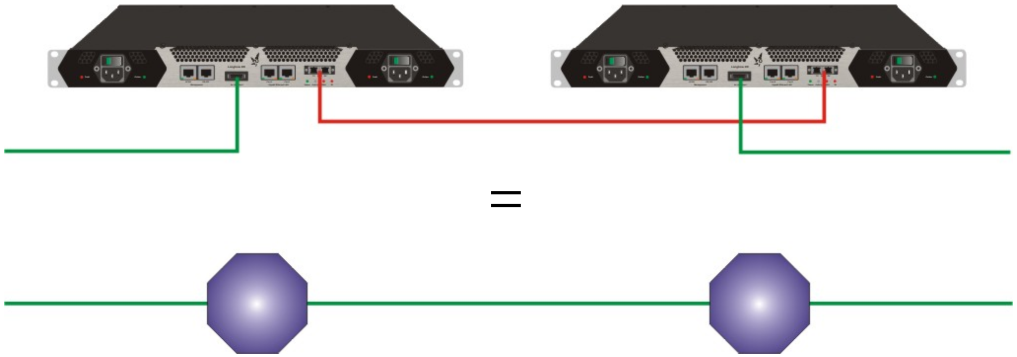
- Why does InfiniBand range extension exist?
- What does it look like?
- Optical link options?
- Applications?
- Future developments?

Why Long Haul?

- Expand InfiniBand’s footprint beyond the machine room floor – new applications
- Real performance advantages - (latency & bandwidth)
- Required for the implementation of a true Unified Fabric model
- **Customers are requesting it**

What does it look like?

- The subnet manager is allowed to conclude that the optical link is an ordinary InfiniBand cable spanning two 2-port switches...



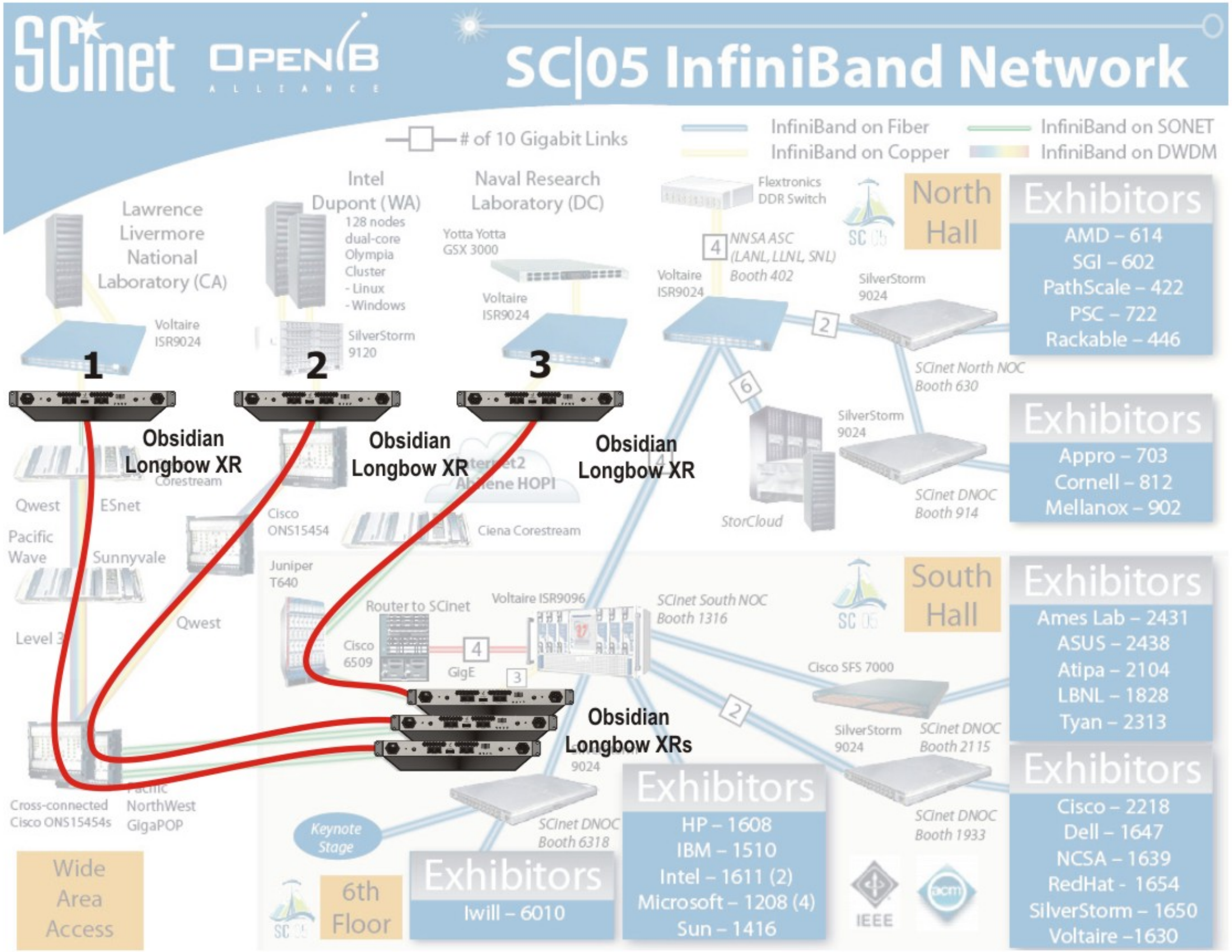
- Link two sites with a pair of boxes and an optical connection, and the sites merge their subnets
- Apart from the unavoidable optical flight latency, the InfiniBand equipment, stacks and applications see nothing unusual

What does it look like?

SCinet's OpenIB subnet spills from the exhibition halls to remote InfiniBand installations in three other cities:

1. LLNL, CA	**1200**	**km**
2. Intel, Dupont WA	**100**	**km**
3. NRL, DC	**5000**	**km**

(DWDM and SONET)



Optical link options?

Global	Encapsulation over SONET/ SDH, ATM, 10GbE
Metro	Latency optimised SONET/ SDH
Campus	Dark-fibre (Single-/ Legacy Multi-Mode Fibre)

- Pretending to be an InfiniBand cable :
  - Model is broken if the link is shared
  - Error rates must be very low (FEC if necessary)
  - A dedicated lambda is preferred (WDM – no problem)

- Failover can be handled in the InfiniBand or optical domains (may be much faster e.g. SONET/ SDH)

Application: Bulk data transport

- Streaming huge datasets across large distances as quickly as possible
  - Technology streams at wire speed; very low CPU loading, with little or no stack tweakage
  - Streams of high definition media or science data
  - Streams of disaster tolerance backup/ restore data

Application: Low latency messaging

- Moving small time-sensitive data chunks from one InfiniBand cluster to another
  - Avoid forcing the application to speak a second protocol to cross distances
  - Userspace-to-Userspace latencies compare very favourably to other solutions in Metro Area Networks

Application: Visualisation

- Remote InfiniBand powered visualisation clusters, workstations or personal supercomputers can tap into a compute cluster’s native InfiniBand fabric directly
  - Crisper, smoother visualisation experiences
  - Distributes access to valuable compute resources
  - Simpler coding of applications (one stack)

Application: Storage

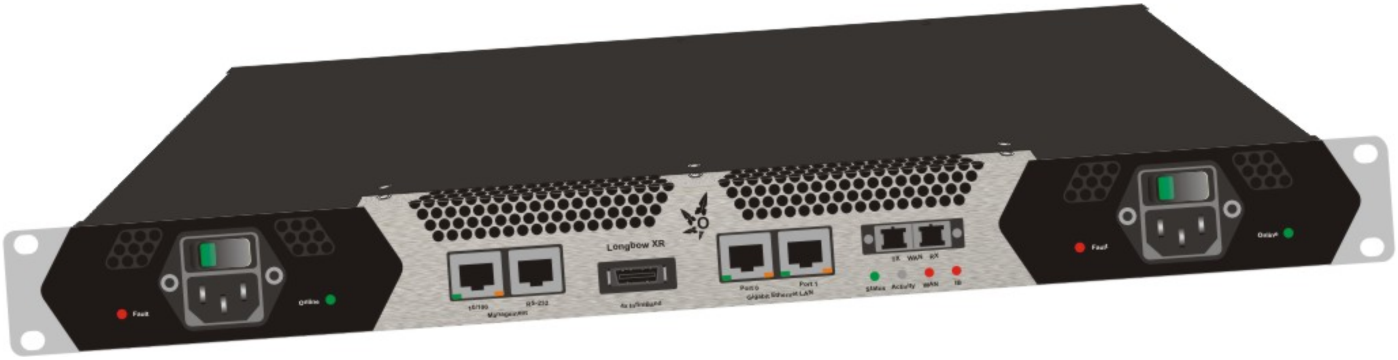
- Native InfiniBand storage can be tapped directly across large distances
  - Optimises rapid replication applications
  - Allows centralised InfiniBand storage deployments in a campus/ metro/ WAN environment

Application: SuperClusters

- Parallel long haul links between clusters within a campus or research park allows clusters to be quickly aggregated into larger subnets
  - Flexible and efficient use of clustered compute farms
  - Same infrastructure could serve visualisation/ replication roles
  - Little or no application changes required to assimilate and exploit (not very) remote InfiniBand clusters

Future Developments

- InfiniBand router mode – it is not always helpful to unify remote subnets
- Support faster InfiniBand speeds (@ OC-768+)
- Tighter integration with optics-side infrastructure



Thank you