

Cluster Grid Interconnects

Tony Kay – Chief Architect
Enterprise Grid and Networking



Agenda



Cluster Grid Interconnects



The Upstart - Infiniband



The Empire Strikes Back - Myricom



Return of the King – 10G Gigabit



Summary

Agenda



Cluster Grid Interconnects



Cluster Grid Problems

- Performance often Interconnect determined
- 3rd party interconnects (IB, Myrinet, Quadrics)
 - Expensive (Card + Port)
- Limitations
 - ISV Support
 - OS Support
 - Scalability
- Not all address the “storage issue”
 - NAS still rules, DAS expensive

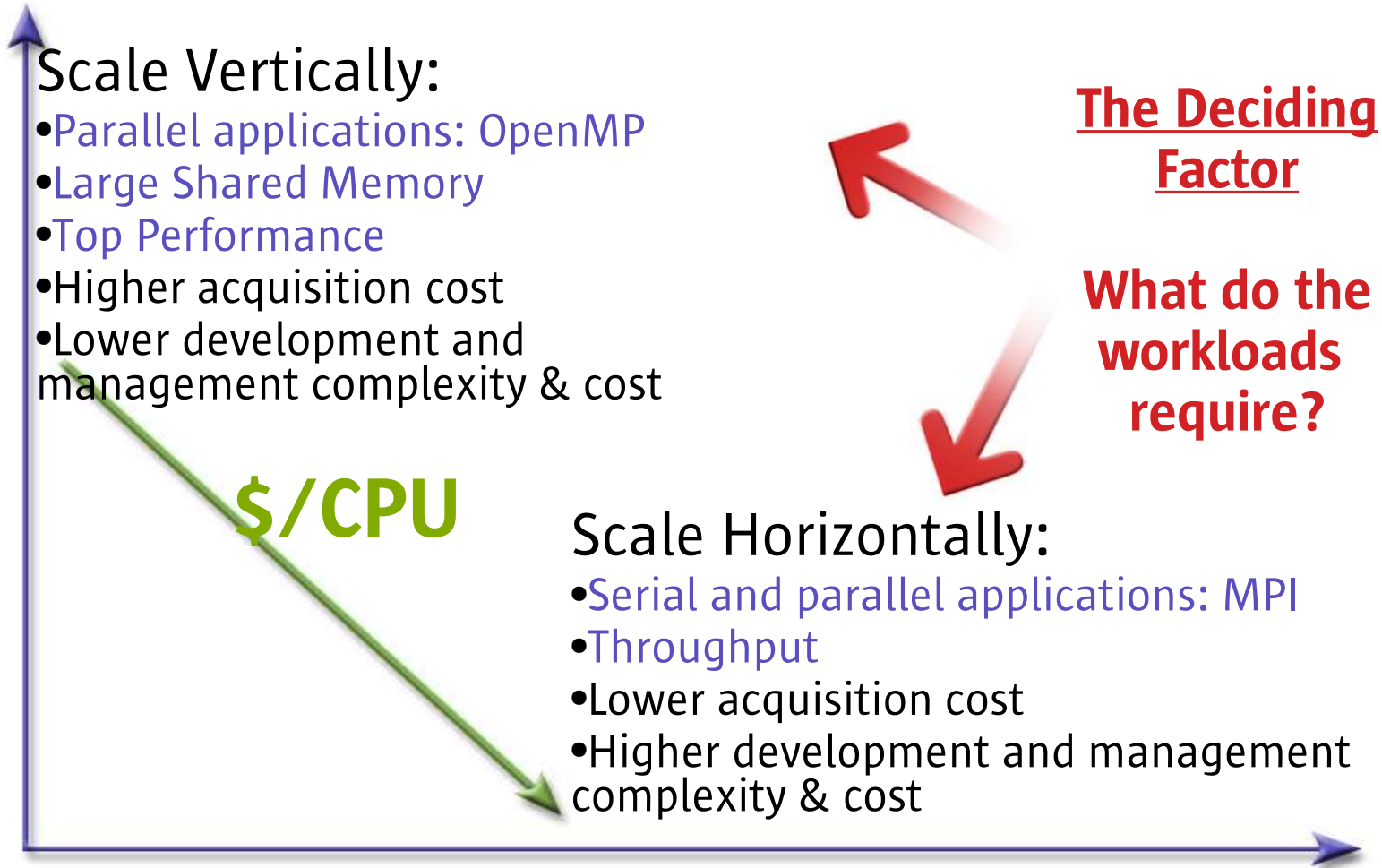
Workload Performance Factors

- Processor speed, capacity and throughput
- Memory capacity
- System interconnect latency & bandwidth
- Network and storage I/O
- Operating system scalability
- Visualization performance and quality
- Optimized applications
- Network service availability

#1 issue
for real world
cluster
performance
and scaling



The Grid Architecture Dilemma: Scale Vertically or Scale Horizontally?



Right Interconnect for the job

- 900 dual CPU Nodes:
 - 7 x 256 CPU clusters
- 3 x 256 CPU I/O intense
 - Gigabit (dual)
- 4 x 256 CPU MPI intense
 - Gigabit (TCP/IP)
 - Myrinet (MPI)

- Choose right interconnect for the job



Sun's CRS build 38 racks containing 1,800 CPU cluster nodes (6 lorry loads)

Final Delivery: 1 lorry load

Good News

- Interconnects are getting:
 - Faster
 - Lighter
 - Greater throughput per %CPU
 - Cheaper
 - Multi-protocol support
 - Infiniband – FCAL, Ethernet friendly
 - Myrinet – Ethernet friendly

Bad News – what to buy in 2005?

- No clear winner in 2005
- Myrinet
 - 'De facto' legacy standard?
- IB
 - The promise of a 'single fabric'
- 10Gigabit Ethernet
 - Undisputed **Networking** World champion?
 - 1995 Token Ring
 - 199x FDDI
 - 199x ATM
 - Myrinet & Infiniband - **TBA?**

Agenda



The Upstart - Infiniband

VOLTAIRE

InfiniCon
systems

 **TOPSPIN**

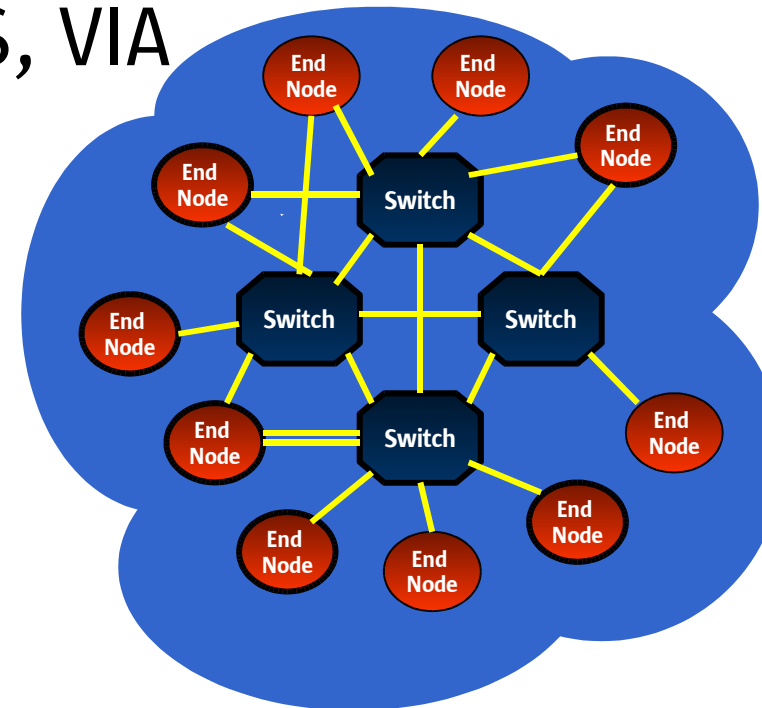


Why Infiniband

- Open 'Standard'
- Price dropping quickly
- Number of players:
 - Mellanox, Infinicon, TopSpin, Voltaire
- Optimised; Low Latency, High Bandwidth
 - Dual port 10GBps PCI-X now (850MB/s sustained)
 - 30GB/s PCI Express coming
- 1 Port, 3 roles:
 - Cluster (MPI)
 - LAN TCP/IP
 - Storage

Infiniband Protocols - Rich

- Host to Host - SDP (Sockets Direct Protocol)
- Storage – SRP (SCSI RDMA protocol for IB and IB to FC bridges)
- Network - TCP termination and IPoverIB
- Also used for MPI, DAFS, NDIS, VIA
- Rich ISV support for API's
- Only 'Open' RDMA



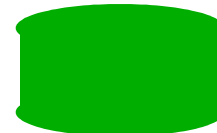
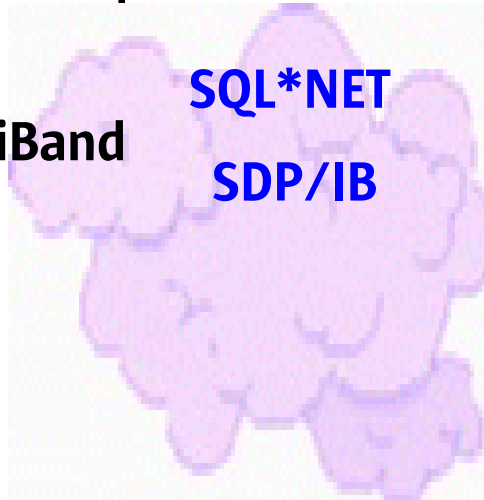
Infiniband and Oracle

- 10G RAC release 1
- Currently Red Hat Enterprise Linux
- Solaris port underway
- Topspin and Voltaire port underway



Application Servers

InfiniBand



DB Server

Cluster
interconnect
uDAPL

Agenda



The Empire Strikes Back - Myricom

Myricom[®]

Myrinet

- Dominant MPI Interconnect – today
- Good OS and ISV support
- Good performance
- Light on host operating HW/SW
- Improving road-map
- Infiniband and threat of 10G forcing:
 - Innovation
 - Big cost reduction

Myricom[®]

Myricom – Keeping itself Relevant?

- Myrinet 10G
 - 'Look and feel' like 10 Gigabit Ethernet (10 Gbe)
 - But lower latency
 - Lower host CPU overhead
 - Protocol offload
 - Same physical infrastructure
 - 10Gbit/s full duplex
 - Compatible with Myrinet-2000
 - Same APIs and application support
 - Similar switches, different data rate

Myrinet Protocol Support

- **Low-level APIs**
 - GM 1 (legacy), GM 2 (current standard), MX (new), 3rd party (e.g., SCore PM)
- **TCP/IP & UDP/IP**
 - Ethernet emulation, included in all GM and MX releases
 - 1.95 Gb/s TCP/IP netperf benchmarks on Linux (2.4, 2.6)
- **MPICH-GM, MPICH-MX**
 - An implementation of the Argonne MPICH directly over GM or MX.
 - Third-party MPI implementations over Myrinet are also available.
- **VI-GM, VI-MX**
 - An implementation of the VI Architecture API directly over GM or MX.
- **Sockets-GM, Sockets-MX**
 - An implementation of UNIX or Windows sockets (or DCOM) over GM or MX. Completely transparent to application programs. Use the same binaries!
- **Additional middleware in development**
 - e.g., uDAPL and kDAPL (for distributed data bases and file systems DB2, Oracle 10g, ...)

Agenda



Return of the King – 10G Gigabit

10G Ethernet (10 Gbe)

- Gigabit Ethernet – Ubiquitous
- Track record for 'killing' all comers:
 - Token Ring, FDDI, ATM
 - Myrinet next?
 - Infiniband TBA?
- Faster – obviously!
- Lower latencies
- More Capable
- Start-up activity
- Storage capabilities

10Gbe Faster

- 10x throughput
- Latencies falling e.g (figures changing)
 - Gigabit
 - 40-60us MPI over TCP/IP for simpler Linux stack
 - 16us MPI over TCP/IP with user-mode stack
 - 10 Gbe
 - Over TCP/IP, 9usec
 - Over MPI 7 usec
- Lots of work around:
 - TOE, RDMA, Etherfabric

10Gbe Momentum

- Lots of start-ups
 - Level 5 networks
 - Ammasso
 - Chelsio
 - S2IO
- Plus Myricom 'support'
- Impact of iSCSI?

Agenda



Summary

Summary

- Good news
 - Everything's getting faster
 - And Cheaper (both cost and also impact on host)
 - More flexible
 - Better interoperability
 - Richer protocols
- Advice
 - Look to the future – what do you need
 - Buy what best supports your applications
- Look at the new features
 - Can you take advantage of these?

Cluster Grid Interconnects

Tony.Kay@Sun.Com

