

InfiniBand Experiences of PC²

Dr. Jens Simon
simon@upb.de
Paderborn Center for Parallel Computing (PC²)
Universität Paderborn

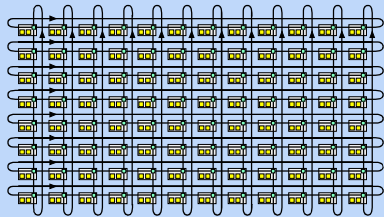
hpcLine-Infotag, 18. Mai 2004



PC² - Paderborn Center for Parallel Computing

- ▶ **structure:** scientific institute of Paderborn University
- ▶ **founded:** April 1991
- ▶ **main task:** competence center for parallel and distributed computing in NRW
- ▶ **main goal:** efficient use of parallel and distributed systems
- ▶ **overview:**
 - ▶ services:
 - ▶ HPC systems and software packages
 - ▶ user consulting
 - ▶ research:
 - ▶ HPC systems and applications
 - ▶ distributed HPC and networking

1999: Paderborn SCI Cluster (First hpcLine)



- Vendor:** Fujitsu-Siemens
System: hpcLine, 96 nodes, 48 GByte RAM, 162 GFlop/s peak
Interconnect: SCI (Scalable Coherent Interface): 500 MByte/s, 3 μ s latency
Topology: 12 x 8 torus with distributed switches
Compute node: Primergy Server (2x Intel Pentium III, 850 MHz, 512 MByte)
OS: Linux
ScaMPI: 84 MByte/s unidirectional, 5 μ s latency (ping 0 Byte)

Place 351 in the TOP 500 list Nov 1999

J. Simon

▶ 3



Itanium-Cluster

- ▶ Configuration:
 - ▶ 32 compute nodes
 - ▶ 4 visualization nodes
 - ▶ 1 front-end
 - ▶ segmented display
 - ▶ Red Hat AS 2.1
- ▶ Front-end
 - ▶ HP zx6000 (2x Intel Itanium2, 0.9 GHz, 4 GByte DRAM)
- ▶ Viz. nodes
 - ▶ HP zx6000 (2x Intel Itanium2, 1 GHz, 12 GByte DRAM, nVIDIA Quadro4 900XGL)
- ▶ Compute nodes
 - ▶ HP rx2600 (2x Intel Itanium2, 1.3 GHz, 3 MB-L3, 4 GByte DRAM)



J. Simon

▶ 4



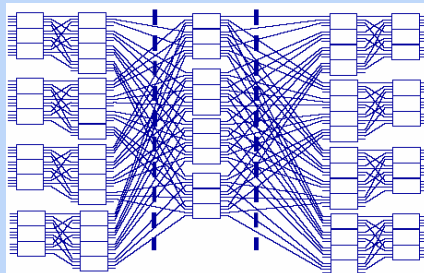
2003: Opteron InfiniBand Cluster

- System:** AMD Opteron, 6 nodes,
- Interconnect:** GigEthernet, Mellanox InfiniBand
- Topology:** switched GbE, switched IBA
- 4 Compute nodes:** Newisys 2100,
Dual AMD Opteron, 1.4 GHz,
2 – 8 GByte DRAM
- OS:** SuSE United Linux 1.0
- 2 Compute nodes:** Fujitsu-Siemens Celsius v810
Dual AMD Opteron,
2.2 GHz, 2 GByte DRAM
- OS:** RedHat AS 3

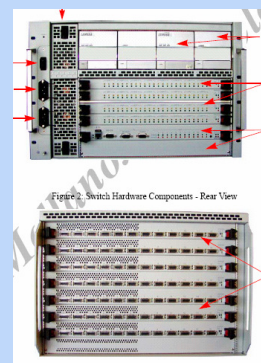


Mellanox InfiniBand: HCA + Gazelle Switch

- ▶ HCA
 - ▶ Two bidirektional Links with 4 x 2.5 Gbit/s signal rate
 - ▶ RDMA: 4.98 μ s + 725 MByte/s uni-dir.
 - ▶ MPI: 6.79 μ s + 705 MByte/s uni-dir.
- ▶ Gazelle Switch
 - ▶ Up to 96 ports
 - ▶ Full bisection-bandwidth
 - ▶ 280 ns per internal switch hop



Switch-Topologie mit 96 Ports



- Power Supply
- Spine Cards
- Mngmt Card
- Leaf Boards

Mellanox Gazelle Switch

MPI performance of Itanium2 / InfiniBand Cluster

- ▶ Mellanox MTPB23108-CE128 (fw 3.1)
- ▶ Mellanox MTS 9600-36port

MPI Version	latency [μs]	p2p bw (1 MB pack.) [MByte/s]	all2all bw (32 nodes) [MByte/s]	Barrier (32 nodes) [μs]
NCSA VMI2.0 MST b3	7.07	703	8850	60.6
OSU MVAPCIH 0.9.2	6.79	700	9060	54.9
Scali ScaMPI 3.3.0.1	7.26	685	6120*	54.2*

* Not optimized

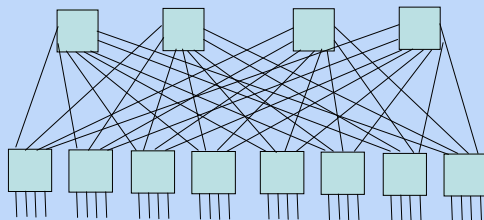
J. Simon

▶ 7



InfiniCon InfiniBand: InfinIO 3032 Switch

- ▶ InfinIO 3032 Switch
 - ▶ 32 ports
 - ▶ full-bisectional bandwidth
 - ▶ 130 ns per internal switch hop
 - ▶ 1 U height



J. Simon

▶ 8



MPI performance of Itanium2 / InfiniBand Cluster

- ▶ Mellanox MTPB23108-CE128 (fw 3.1)
- ▶ InfiniCon InfinIO 3032

MPI Version	latency [μ s]	p2p bw (1 MB pack.) [MByte/s]	all2all bw (32 nodes) [MByte/s]	Barrier (32 nodes) [μ s]
NCSA VMI2.0 MST b3	6.88	700	9500	65.4
OSU MVAPCIH 0.9.2	6.61	698	11500	52.2
Scali ScaMPI 3.3.0.1	7.26	687	11700*	43.1

* Optimized version

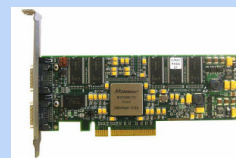
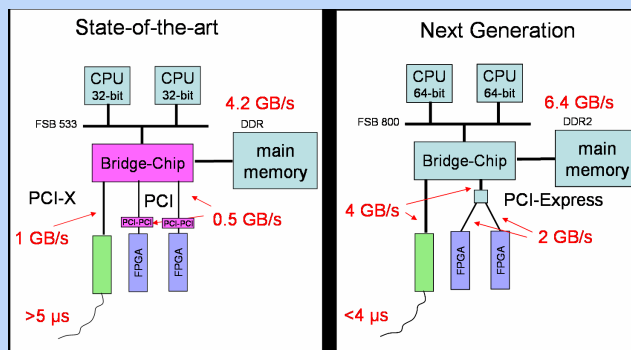
J. Simon

▶ 9



Next InfiniBand HW

- ▶ 24 port switch chips
 - ▶ Reduces internal switch hops (lower latencies)
- ▶ HCA PCI-Express 8x
 - ▶ Enables full InfiniBand 4x bandwidth (> 3 GByte/s)
 - ▶ Reduces latencies of communication



J. Simon

▶ 10



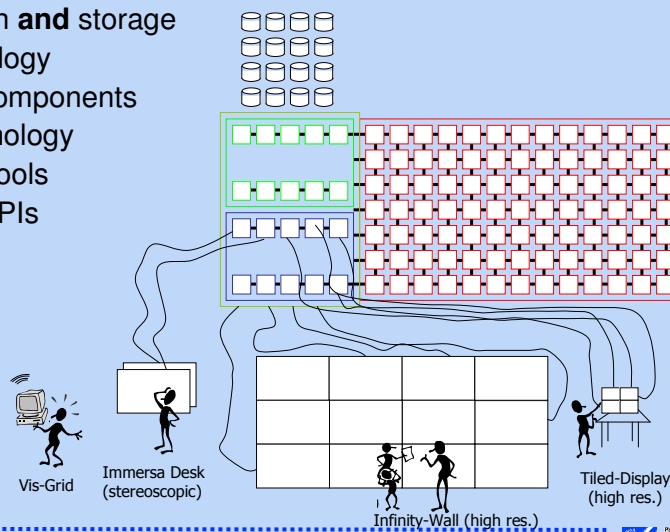
Clusters of PC²

1998	FSC hpcLine	96 Dual Pentium3, 850 MHz, 0.5 GByte	SCI-Netzwerk 7 μ s, 84 MByte/s 2.5 GByte/s all2all BW	163 GFLOPS 340 SpecFPPrate
2002	HP SFB-Cluster	4 Dual Itanium2, 1 GHz, 12 GByte	InfiniBand-Netzwerk 7.8 μ s, 420 MByte/s	32 GFLOPS 120 SpecFPPrate
2003	HP Frauenheim- Cluster	32 Dual Itanium2+, 1.3 GHz, 4 GByte	InfiniBand-Netzwerk 6.6 μ s, 705 MByte/s 11.7 GByte/s all2all BW	332 GFLOPS 1100 SpecFPPrate
2003	AMD Opteron Cluster	4 Dual Opteron, 1.4 GHz, 2 GByte	InfiniBand-Netzwerk 6.25 μ s, 687 MByte/s	22 GFLOPS 100 SpecFPPrate
2004	FSC v810 Cluster	2 Dual Opteron, 2.2 GHz, 2 GByte	InfiniBand-Netzwerk 5.18 μ s, 770 MByte/s	16 GFLOPS ~64 SpecFPPrate

All systems are available for performance tests

New Development: Comp & Viz Cluster

- ▶ Integration of services
 - ▶ Computation **and** visualization **and** storage
- ▶ Cluster technology
 - ▶ Standard components
- ▶ Software technology
 - ▶ High level tools
 - ▶ Standard APIs



PC² Benchmark Center

- ▶ For actual performance results of different compute nodes and communication networks have a look at

<http://www.upb.de/StaffWeb/jens/Projekte/Benchmarks/>