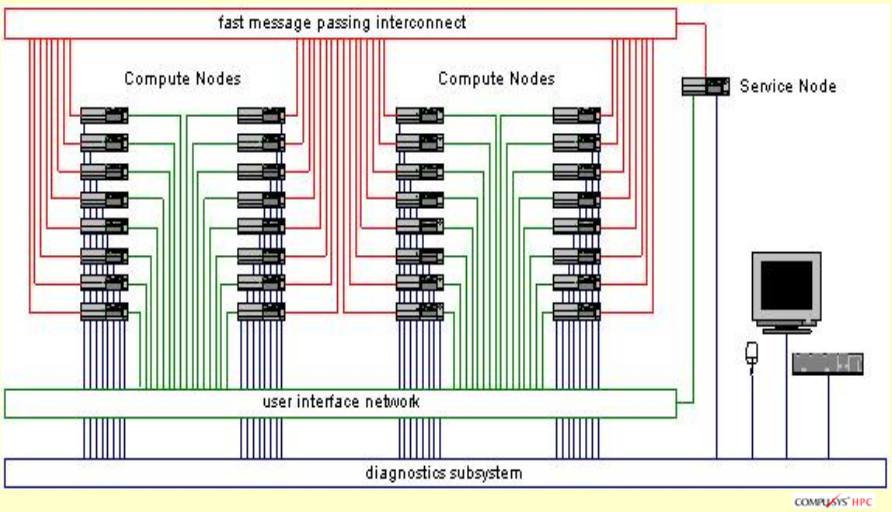
PC Cluster @ DESY
Peter Wegner

1. Motivation, History
2. Myrinet-Communication
4. Cluster Hardware
5. Cluster Software
6. Future …

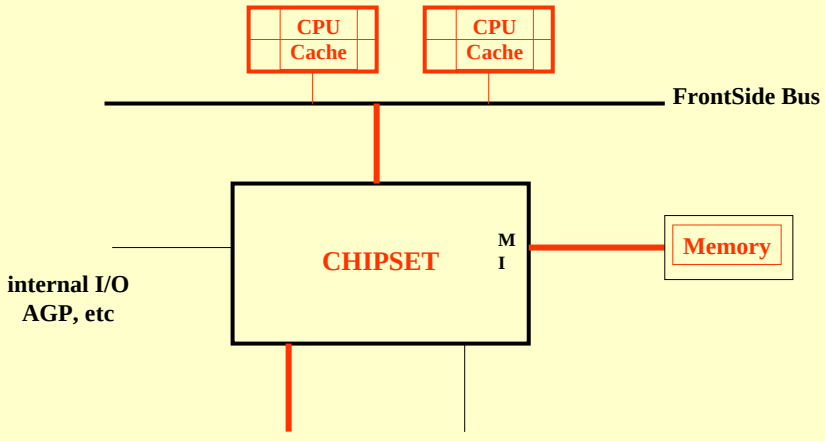
PC Cluster
“Definition 1”

Idee: Herbert Cornelius (Intel München)

PC Cluster
“Definition 2”



PC Cluster
PC – HPC related components



Motivation for PC Cluster

Motivation: LQCD, Stream benchmark，Myrinet bandwidth

32/64-bit Dirac Kernel, LQCD (Martin Lüscher, CERN):
P4, 1.4 GHz, 256 MB Rambus, using SSE1(2) instructions incl. cache pre-fetch
Time per lattice point:
0.926 micro sec (1503 Mflops [32 bit arithmetic])
1.709 micro sec (814 Mflops [64 bit arithmetic])

Stream Benchmark, Memory Bandwidth:
P4(1.4 GHz, PC800 Rambus) : 1.4 … 2.0 GB/s
PIII (800MHz, PC133 SDRAM) : 400 MB/s
PIII(400 MHz, PC133 SDRAM) : 340 MB/s

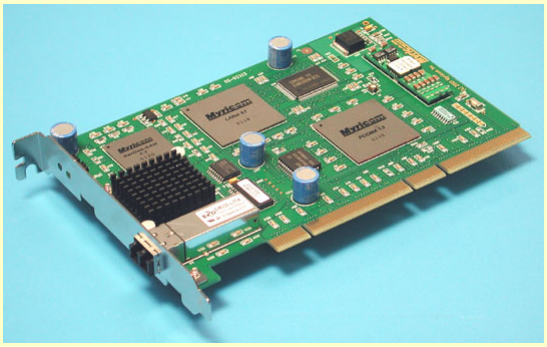
Myrinet, external Bandwidth:
2.0+2.0 Gb/s optical-connection, bidirectional, ~240 MB/s sustained

Motivation for PC Cluster, History

March 2001 Pentium4 systems (SSE instructions, Rambus memory,
66MHz 64-bit PCI) available,
Dual Pentium4 systems, XEON expected for May 2001,
First systems on CeBit (under non-disclosure)
Official announcement end of May 2001: Intel XEON processor, i860 chipset,
Supermicro motherboard P4DC6 –the only combination available
BA (Ausschreibung) July 2001
First information about the i860 problem
Dual XEON test system delivered August 2001

Final decision end of August 2001(Lattice2001 in Berlin)
Installation:
December 2001 in Zeuthen, January 2002 in Hamburg

PC cluster interconnect - Myrinet Network Card (Myricom, USA)



Technical details:
200 MHz Risc processor
2 MByte memory
66MHz/64-Bit PCI-connection
2.0+2.0 Gb/s optical-connection, bidirectional

Myrinet2000 M3F-PCI64B PCI card with optical connector

Sustained bandwidth: 200 ... 240 MByte/sec

PC cluster interconnect - Myrinet Switch

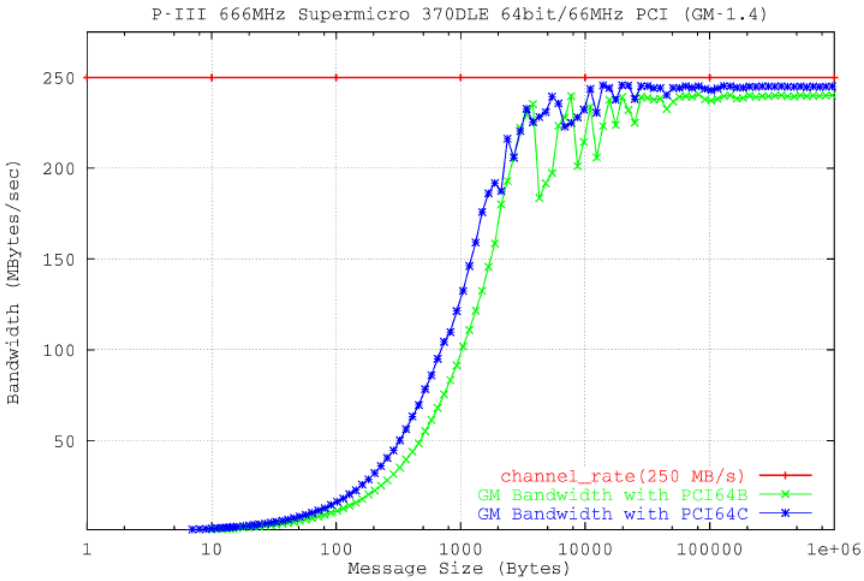
Technical details:
200 MHz Risc processor, 2 MByte memory
66MHz/64-Bit PCI-connection 2.0+2.0 Gb/s optical-connection, bidirectional

Myrinet2000 M3F-PCI64B PCI card with optical connector

Sustained bandwidth: 200 ... 240 MByte/sec

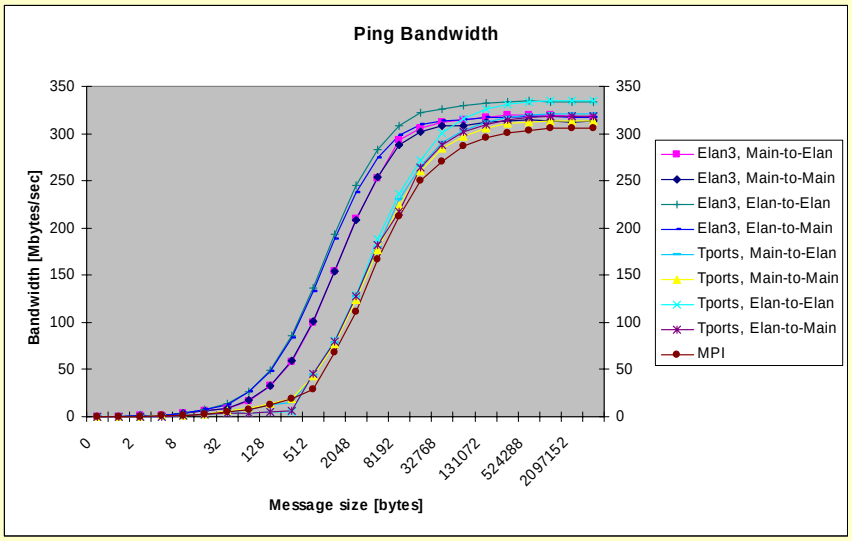
PC - Cluster interconnect, performance

Myrinet performance

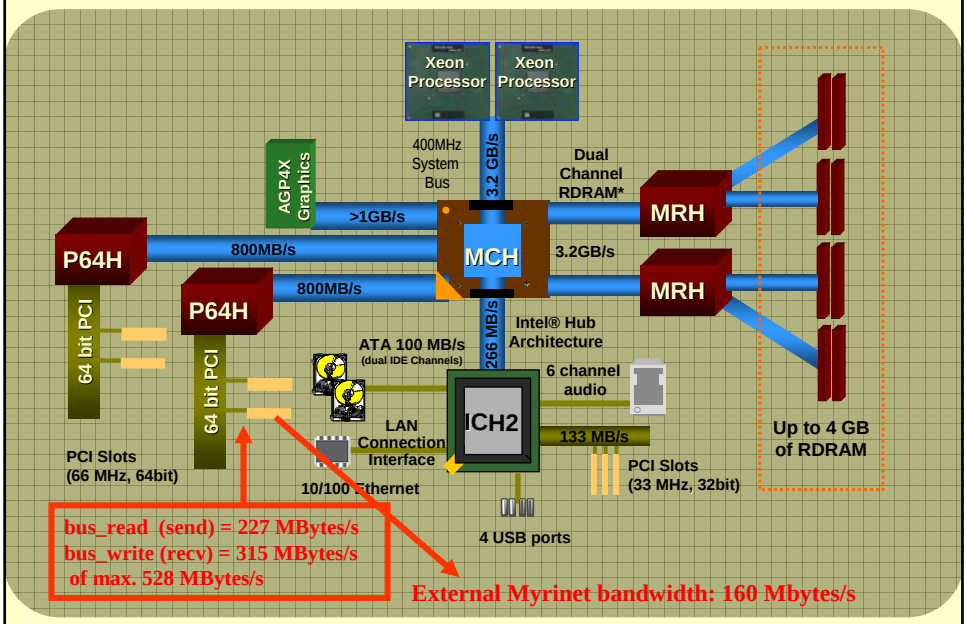


PC - Cluster interconnect, performance

QSnet performance (Quadrics Supercomputer World)



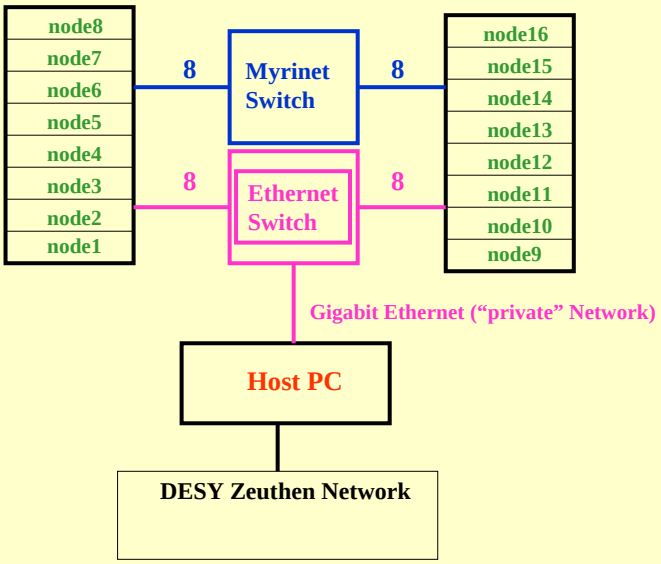
PC Cluster, the i860 chipset problem



PC - Cluster Hardware

Nodes	**Mainboard Supermicro P4DC6** **2(1) x XEON P4, 1.7 GHz, 256 kByte Cache** **1 Gbyte (4x 256 Mbyte) RDRAM** **IBM 18.3 GB DDYS-T18350 U160 3.5” SCSI disk** **Myrinet 2000 M3F-PCI64B-2 Interface**
Network	**Fast Ethernet Switch Gigaline 2024M,** **48x100BaseTX ports + GIGAline2024 1000BaseSX-SC**
Myrinet Fast Interconnect	**M3-E32 5 slot chassis, 2xM3-SW16 Line cards**
Installation	**Zeuthen: 16 dual CPU nodes,** **Hamburg: 32 single CPU nodes**

PC - Cluster Zeuthen schematic



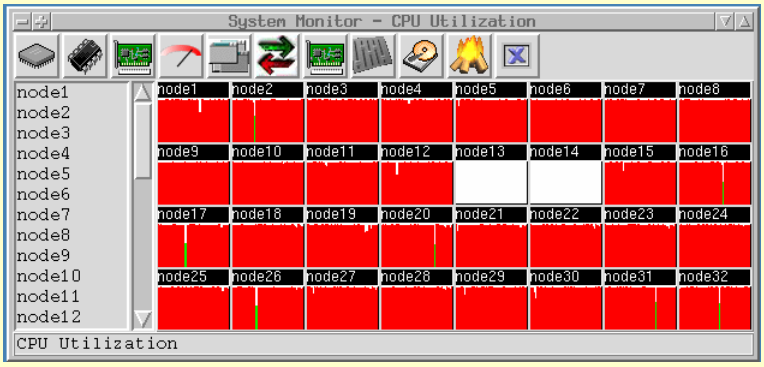
PC - Cluster Software

Operating system	**Linux (z.B. SuSE 7.2)**
Cluster tools:	**Clustware (Megware) Monitoring of temperature, fan rpm, cpu usage,**
Communication software:	**MPI - Message passing interface based on GM (Myricom low level communication library)**
Compiler:	**GNU, Portland Group, KAI, Intel Compiler**
Batch system:	**PBS (OpenPBS)**
Cluster management:	**Clustware, SCORE**

PC - Cluster Software, Monitoring Tools

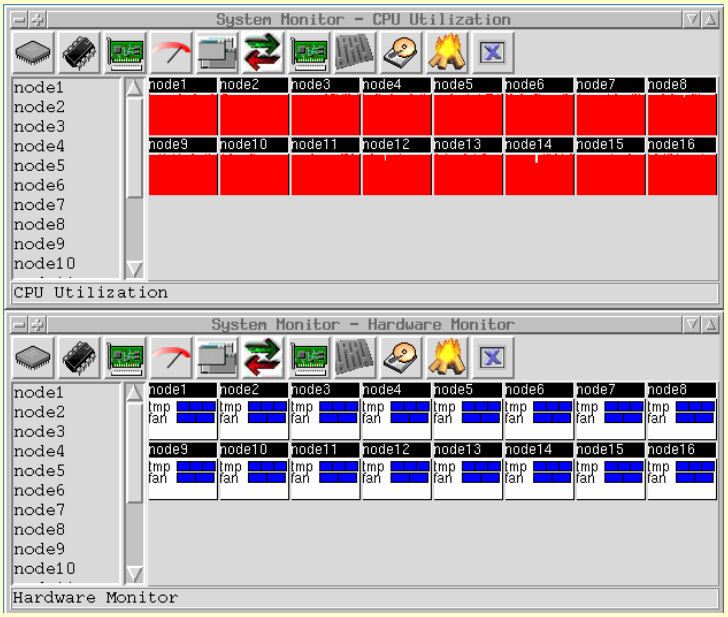
Clustware from Megware

Monitoring example: CPU utilization DESY HH



PC - Cluster Software, Monitoring Tools

Monitoring example: CPU Utilization, Temperature, Fan speed, DESY Zeuthen



PC - Cluster Software: MPI

```
...
if (myid == numprocs-1)
      next = 0;
   else
      next = myid+1;

   if (myid == 0)
   {
      printf("%d sending '%s' \n",myid,buffer);
      MPI_Send(buffer, strlen(buffer)+1, MPI_CHAR, next, 99, MPI_COMM_WORLD);
      printf("%d receiving \n",myid);
      MPI_Recv(buffer, BUFLEN, MPI_CHAR, MPI_ANY_SOURCE, 99, MPI_COMM_WORLD,
            &status);
      printf("%d received '%s' \n",myid,buffer);
      /* mpdprintf(001,"%d receiving \n",myid); */
   }
   else
   {
      printf("%d receiving  \n",myid);
      MPI_Recv(buffer, BUFLEN, MPI_CHAR, MPI_ANY_SOURCE, 99, MPI_COMM_WORLD,
            &status);
      printf("%d received '%s' \n",myid,buffer);
      /* mpdprintf(001,"%d receiving \n",myid); */
      MPI_Send(buffer, strlen(buffer)+1, MPI_CHAR, next, 99, MPI_COMM_WORLD);
      printf("%d sent '%s' \n",myid,buffer);
   }
...
```

PC - Cluster Operating

DESY Zeuthen: Dan Pop (DV), Peter Wegner (DV)
DESY Hamburg: Hartmut Wittig (Theorie), Andreas Gellrich (DV)

Maintenance contract with MEGWARE
Software: Linux system, Compiler, MPI/GM, (SCORE)
Hardware: 1 reserve node + various components

MTBF: O(weeks)

Uptime of the nodes (28.05.2002):
Zeuthen – 38 days, node8 … node16 4 days break for line card replacement
Hamburg – 42 days

Problems:
Hardware failures of Ethernet Switch, node, SCSI disks, Myrinet card
All components were replaced relatively soon.
KAI compiler not running together with MPI/GM, (RedHat-SuSE Linux problem)

PC – Cluster world wide: Examples

Cluster	Site	Nodes/ Processors	Peak (Gflop)	CPU	Interconnect
1. Locus	Locus, USA	960/1920	1920	PIII/1000	FastEth.
2. HELICS	Heidelberg	256/512	1434	AMD MP/1400	Myrinet
3. EMPIRE	Mississippi	519/1038	1157	PIII/1000	Myrinet
4. RHIC	Brookhaven	706/1412	1083	PIII/PII.	FastEth.
5. Biopentium	Inpharmatica	800/1220	1061	PIII/700+	FastEth.
6. Genesis	Shell, NL	1030/1038	1037	PIII/1000	GigabitE
7. Platinum	NCSA	516/1032	1032	PIII/1000	Myrinet
8. CBRC Magi	AIST,Japan	520/1040	967	PIII/933	Myrinet
9. Score III	RWCP, Japan	512/1024	955	PIII/933	Myrinet
10. ICE Box	Utah, USA	303/406	915	PIII + AMD	FastEth.

PC – Cluster: Ongoing Future

CPUs: XEON 2.4 GHz …, AMD Athlon™ XP Processor 2000+

Chipsets: Intel E7500, ServerWorks GC…, AMD-760™ MPX Chipset – full PCI bandwidth

Mainboards: Mainboard Supermicro P4DP6

I/O interfaces: PCI-X, PCI Express

Fast Network: Myrinet, QsNet, Inifiniband(?), …

Intel® Xeon™ Processor-based Platform
with the Intel® E7500 Chipset

- Dual Intel® Xeon™ 2.4GHz Processor
- 512KB L2 cache on-die
- Hyper-Threading enabled
- 400MHz Bus (3.2GB/s)
- Dual-Channel DDR Memory (16GB)
- 3.2GB/s Memory Bandwidth
- 3.2GB/s I/O Bandwidth
- 64-bit PCI/PCI-X I/O support
- Optional SCSI and RAID support
- GbE support
- 1u and 2u dense packaging

PC Cluster – new chipset
Intel E7500

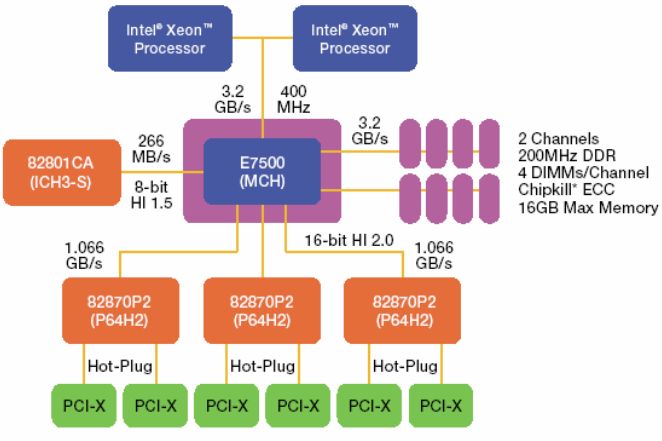
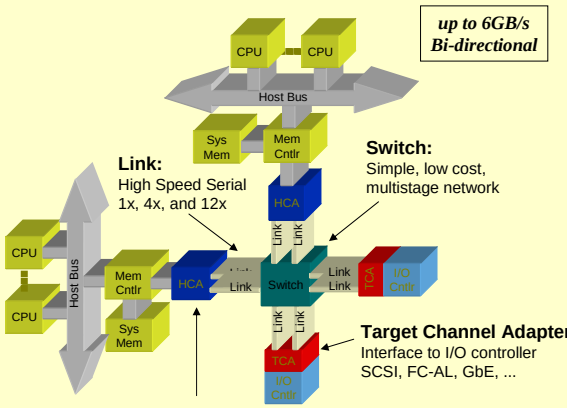


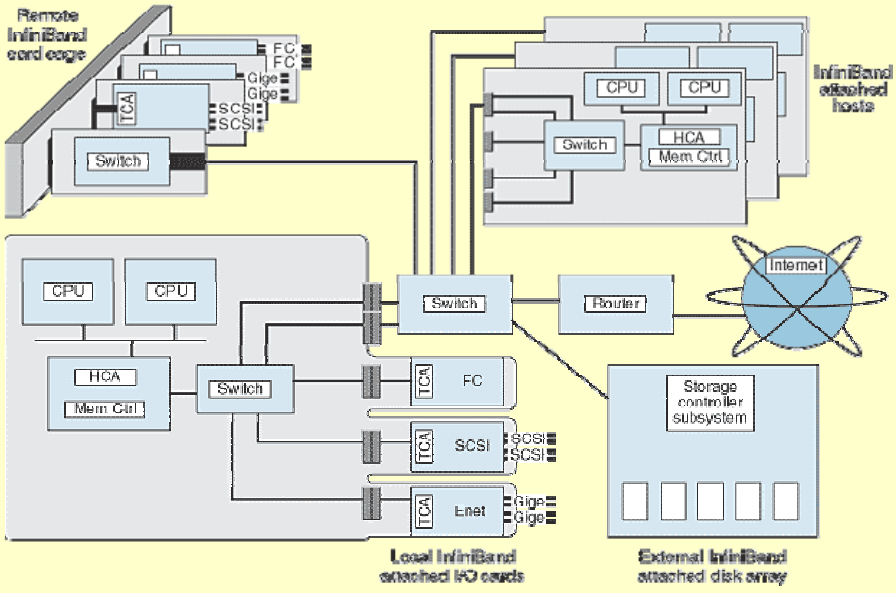
Figure 1. Intel® Xeon™ Processor and the Intel® E7500 Chipset Dual Processor Server Block Diagram

PC Cluster: Future interconnect
Infiniband Concept

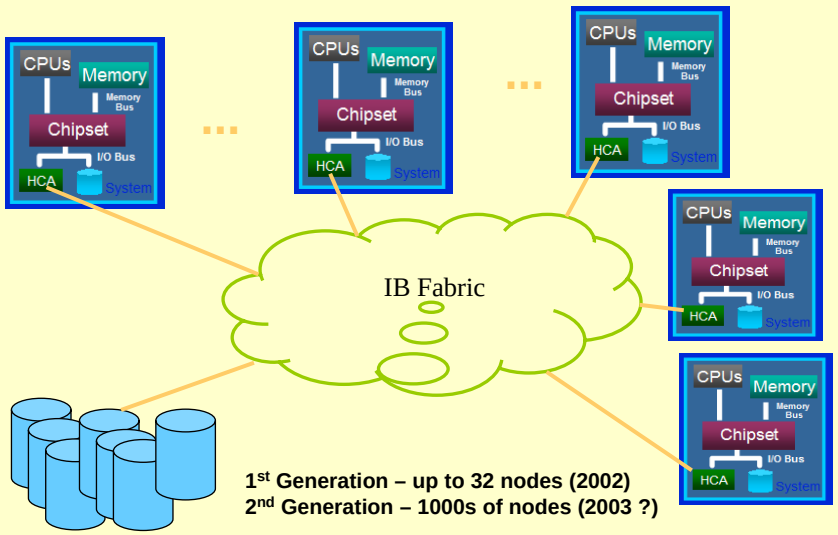


http://www.infinibandta.org

PC Cluster: Future interconnect
Infiniband Concept (IBM)



PC Cluster: Future interconnect (?)
Infiniband Cluster



1st Generation – up to 32 nodes (2002)
2nd Generation – 1000s of nodes (2003 ?)