Parallel Computing at DESY
Peter Wegner

Outline

- **Types of parallel computing**
- **The APE massive parallel computer**
- **PC Clusters at DESY**
- **Symbolic Computing on the Tablet PC**

Parallel Computing at DESY
Peter Wegner

Types of parallel computing :

- **Massive parallel computing**

  tightly coupled large number of special purpose CPUs and special purpose interconnects in n-Dimensions (n=2,3,4,5,6)
  Software model – special purpose tools and compilers

- **Event parallelism**

  trivial parallel processing characterized by communication independent programs which are running on large PC farms
  Software model – Only scheduling via a Batch System

Types of parallel computing cont.:

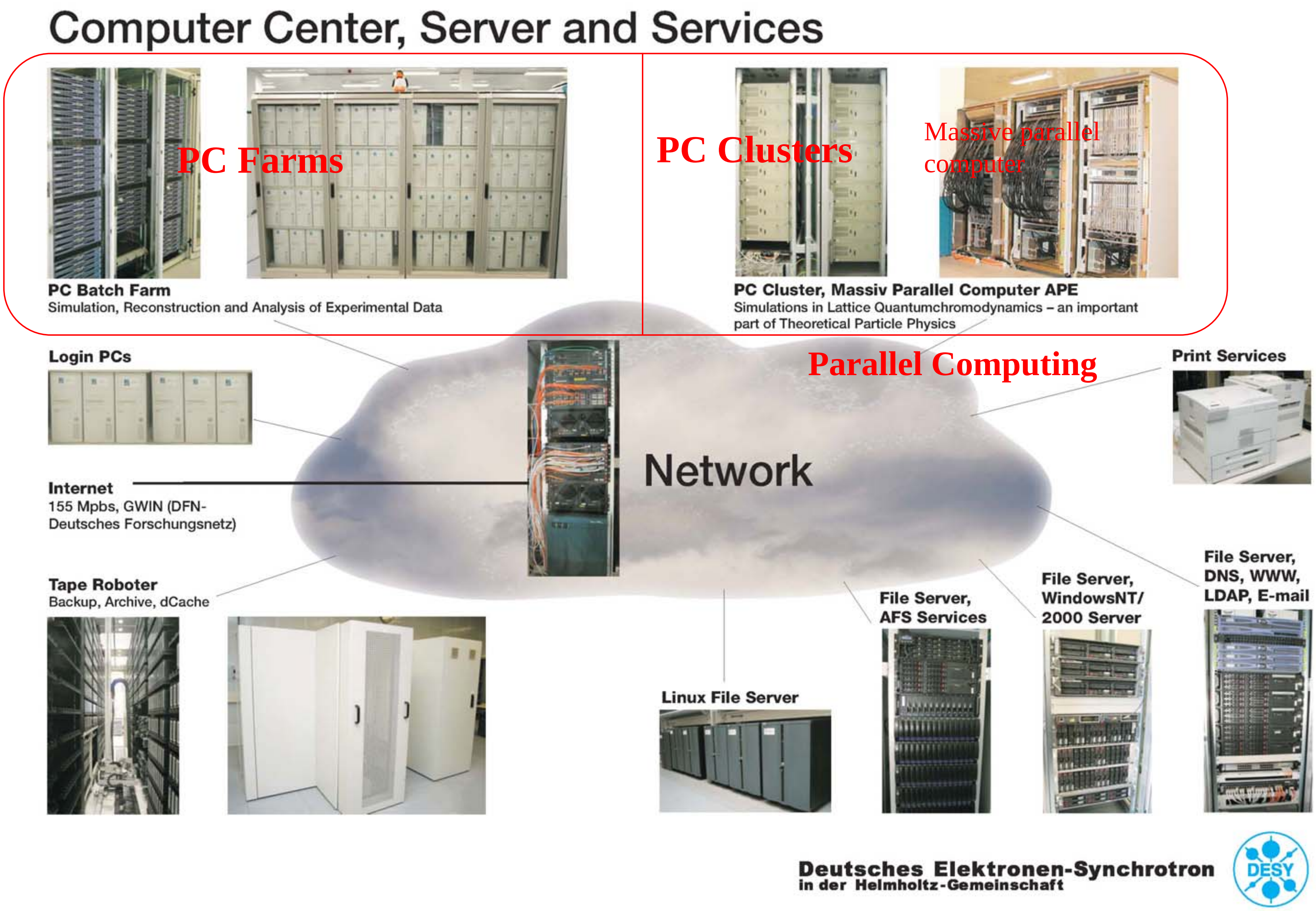
- **"Commodity " parallel computing on clusters**

  one parallel program running on a distributed PC Cluster, the cluster nodes are connected via special high speed, low latency interconnects (GBit Ethernet, Myrinet, Infiniband)
  Software model – MPI (Message Passing Interface)

- **SMP (Symmetric MultiProcessing) parallelism**

  many CPUs are sharing a global memory, one program is running on different CPUs in parallel
  Software model – OpenPM and MPI

Parallel computing at DESY: Zeuthen Computer Center



Parallel Computing at DESY

Massive parallel APE (Array Processor Experiment) - since 1994 at DESY, exclusively used for Lattice Simulations for simulations of Quantum Chromodynamics in the framework of the John von Neumann Institute of Computing (NIC, FZ Jülich, DESY)
http://www-zeuthen.desy.de/ape

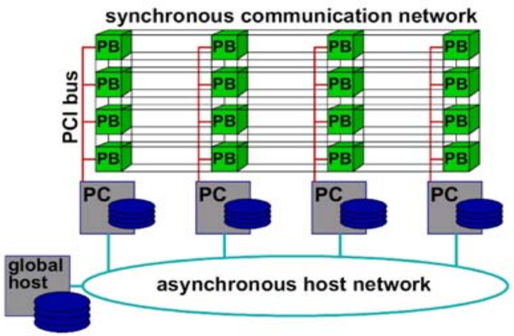
PC Cluster with fast interconnect (Myrinet, Infiniband) – since 2001, Applications: LQCD, Parform ?

Parallel computing at DESY: APEmille

APE Teraflop Computers for Simulations of Elementary Particle Physics

APEmille : Todays QCD Engine

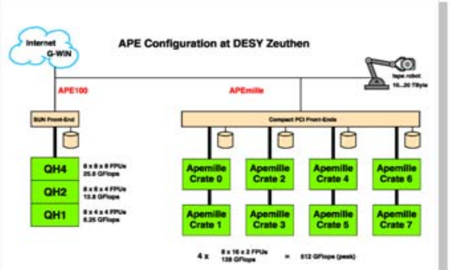
Numerical simulations are an important tool for understanding the theory of strong interactions called quantum chromodynamics (QCD), which remains one of the biggest challenges of modern physics.
For this purpose the theory has to be discritized on a space-time lattice.



APEmille system architecture (1/2 rack)

The APEmille installation at DESY Zeuthen.
Each rack hosts 256 nodes.

- 3-D communication topology
- SIMD (Single Instruction Multiple Data)
- instruction scheduling by software
- communication by direct remote access to distributed data memory
- custom developed processors with optimised floating point unit:
  - ⇒complex normal operations : a * b + c
  - ⇒large register file instead of data cache
  - ⇒simple parallel programming model



Current APEmille installations:

Zeuthen (Germany):	550 Gflops
Europe:	~2 Tflops total at 10 sites

by APE Collaboration

INFN Intituto Nazionale di Fisica Nucleare

DESY Zeuthen

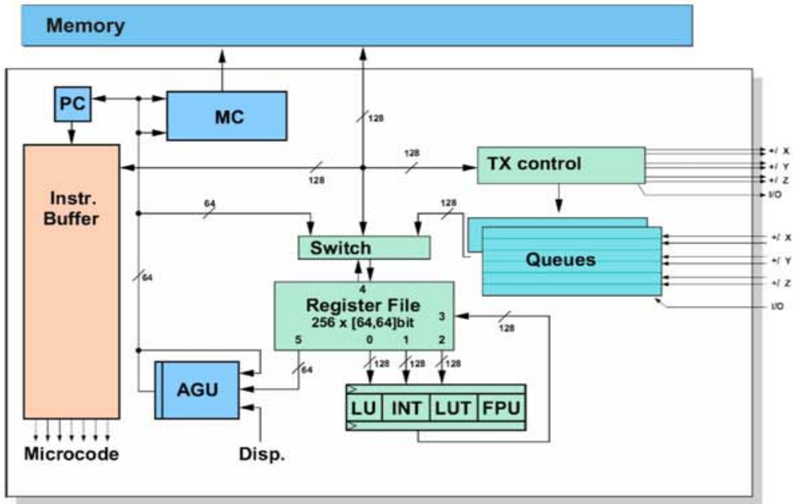
Parallel computing at DESY: apeNEXT

APE Teraflop Computers for Simulations of Elementary Particle Physics

apeNEXT : Developement for the Future

Aim : 0(3) Tflops/system peak in 2003

- SPMD architecture:
  ⇒ autonomous and asynchronous processors
  ⇒ easier technology upgrade
- prefetch queues for local and remote data
- 64bit arithmetics



possible apeNEXT Installation

	apeNEXT (in developement)
Peak perf. / rack	0.8 Tflops
Architecture	SIMD / SPMD
Communication	nearest neighbour
Bandwith / direction	ca. 200 MByte/s
Processor	1.6 Gflops peak
Arithmetics	(a * b + c) complex 64bit
Clock	200 MHz
Technology	1 custom chip, 0.18 μ
Memory	256 - 1024 Mbyte / node
Power consumption	4-5 W / Gflops
Density	~400 Gflops / m^3
Price	0.5 Euro / Mflops (peak)

by APE Collaboration

Intituto Nazionale di Fisica Nucleare

Zeuthen

UNIVERSITE PARIS SUD IX

Parallel computing at DESY: Motivation for PC Clusters

1. **Since 1999/2000 extremely performance improvement on PCs due to new (SSE) instructions, increasing memory bandwidth, increasing clock rate**

   Meanwhile:

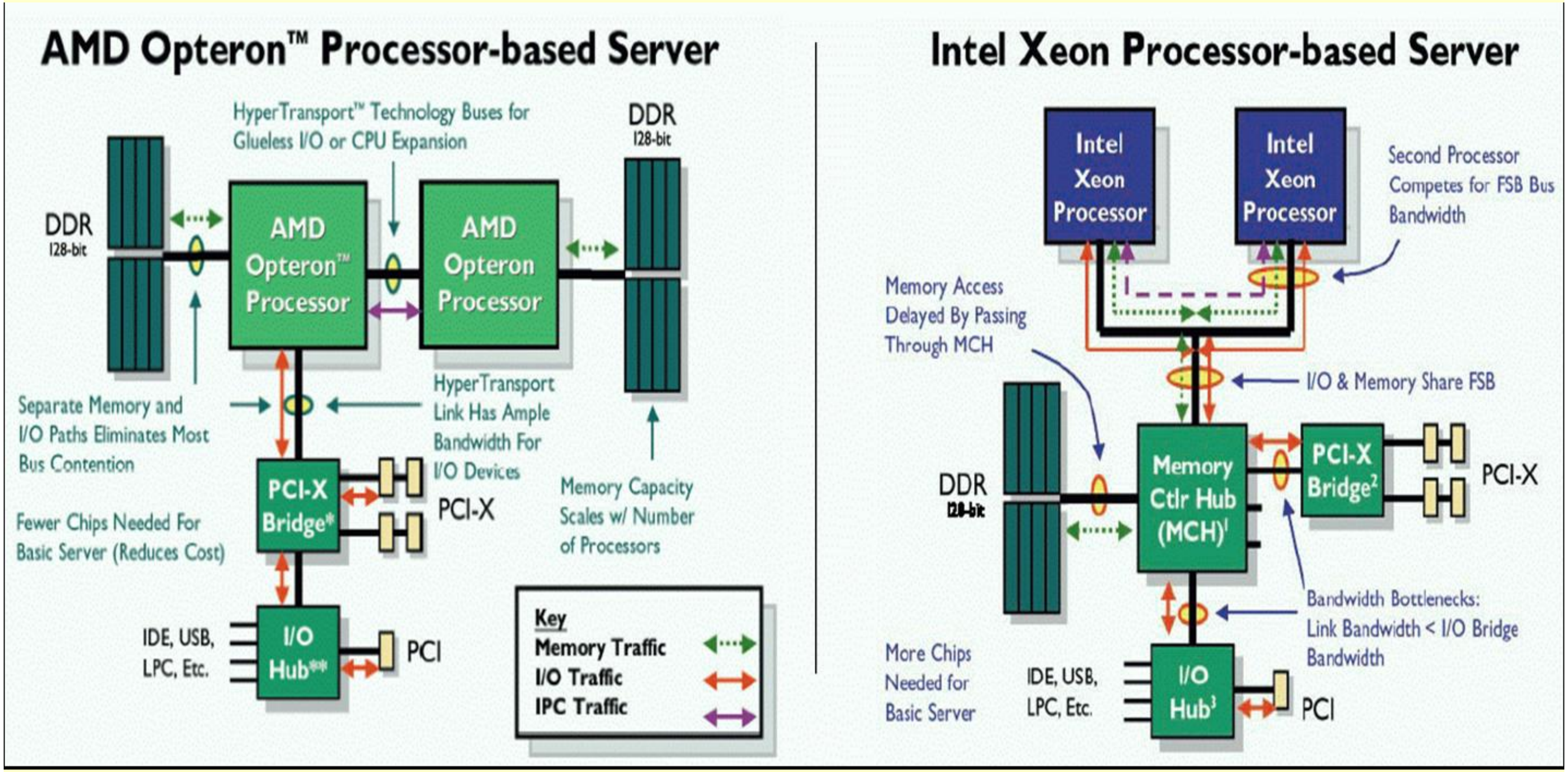
   ca 1.8 Gflops [32 bit arithmetic], 0.9 Gflops [64 bit arithmetic] sustained CPU performance.

2. **Increasing external bandwidth provided by the chipsets (PCI/PCIe) which leads to high speed interconnects (Myrinet2000, Infiniband…):**

   250… 350 MByte/s sustained bandwidth in applications

PC Cluster

PC – components most influencing the performance



PC - Cluster Hardware

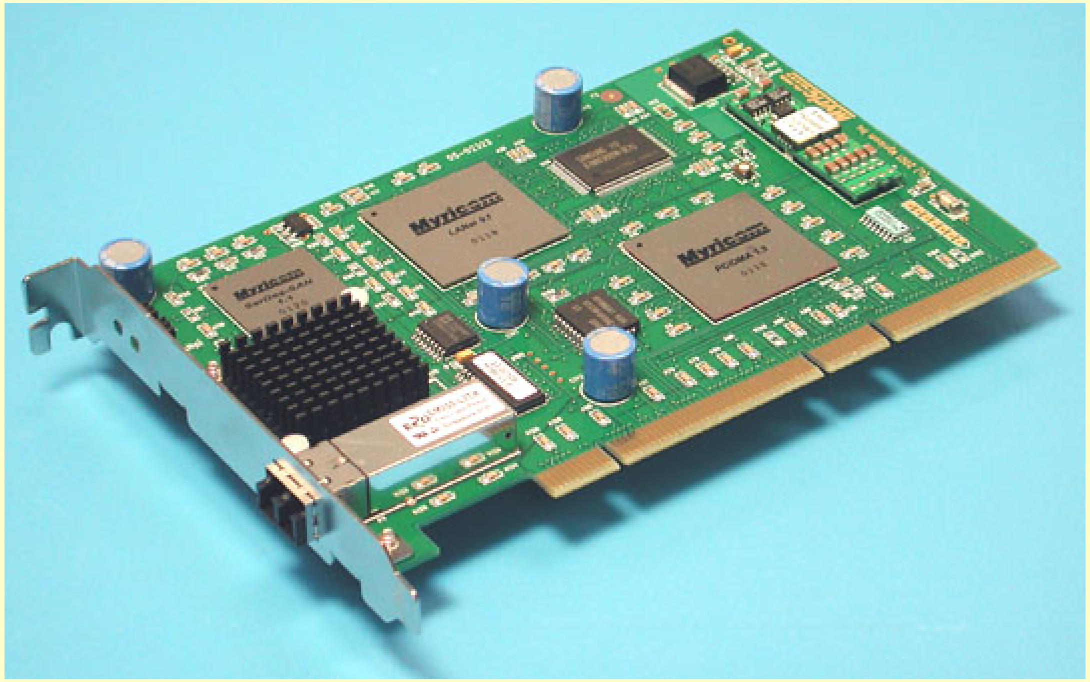
Myrinet

Nodes	**Mainboard Supermicro P4DC6** **2 x XEON P4, 1.7/2.0 GHz, 256 kByte Cache** **1 Gbyte RDRAM** **Myrinet 2000 M3F-PCI64B-2 Interface**
Myrinet Switch	**M3-E32 5 slot chassis, 2xM3-SW16 Line cards**
Installation	**Zeuthen: 16 dual CPU nodes,** **Hamburg: 32 dual CPU nodes**

Infiniband

Nodes	**2 x AMD OPTERON Mod. 250, 2.4 GHz, 1 MB L2 Cache,** **4Gbyte PC2700 ECC DDR SDRAM** **Mellanox Infiniband HA 4X**
Infiniband Switch	**Mellanox InfiniScale III 2400 Switch 24 Port**
Installation	**Zeuthen: 8 dual CPU nodes,** **Hamburg: 10 dual CPU nodes**

Myrinet Network Card (Myricom, USA)

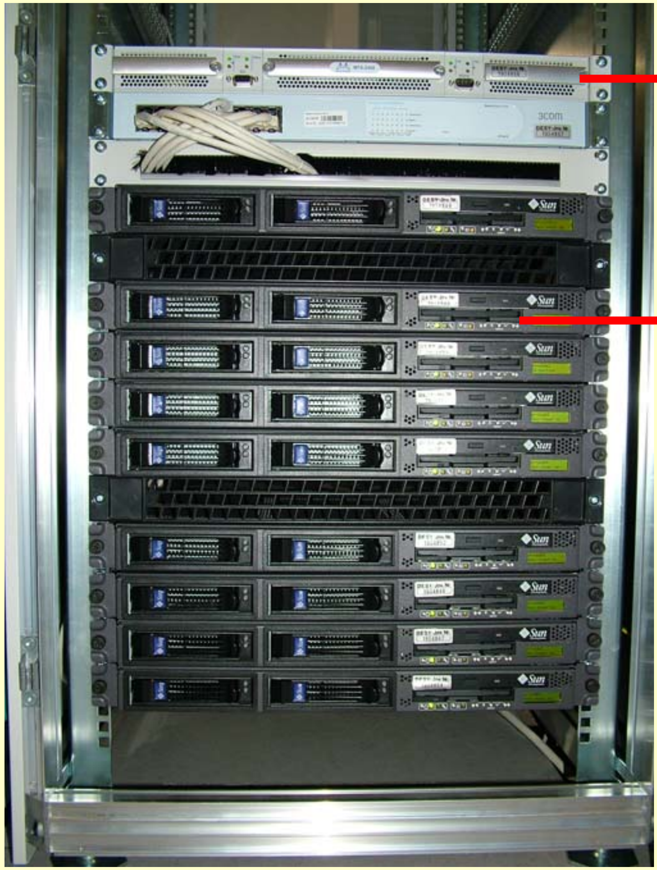


Technical details:
200 MHz Risc processor
2 MByte memory
66MHz/64-Bit PCI-connection
2.0+2.0 Gb/s optical-connection, bidirectional

Myrinet2000 M3F-PCI64B PCI card with optical connector

Sustained bandwidth: 200 ... 240 MByte/sec

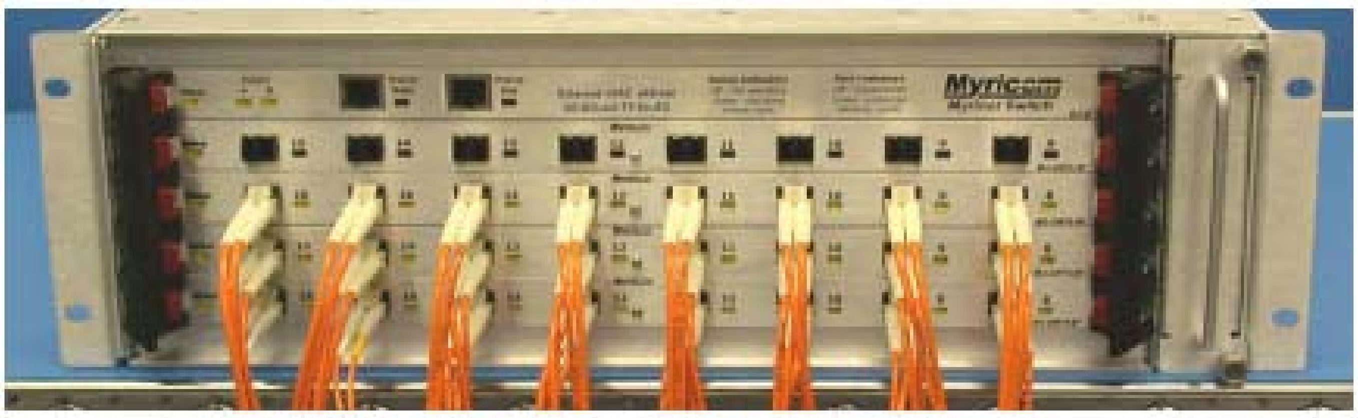
Infiniband Opteron Cluster



Infiniband
Switch

Dual Opteron
Server

Myrinet Switch



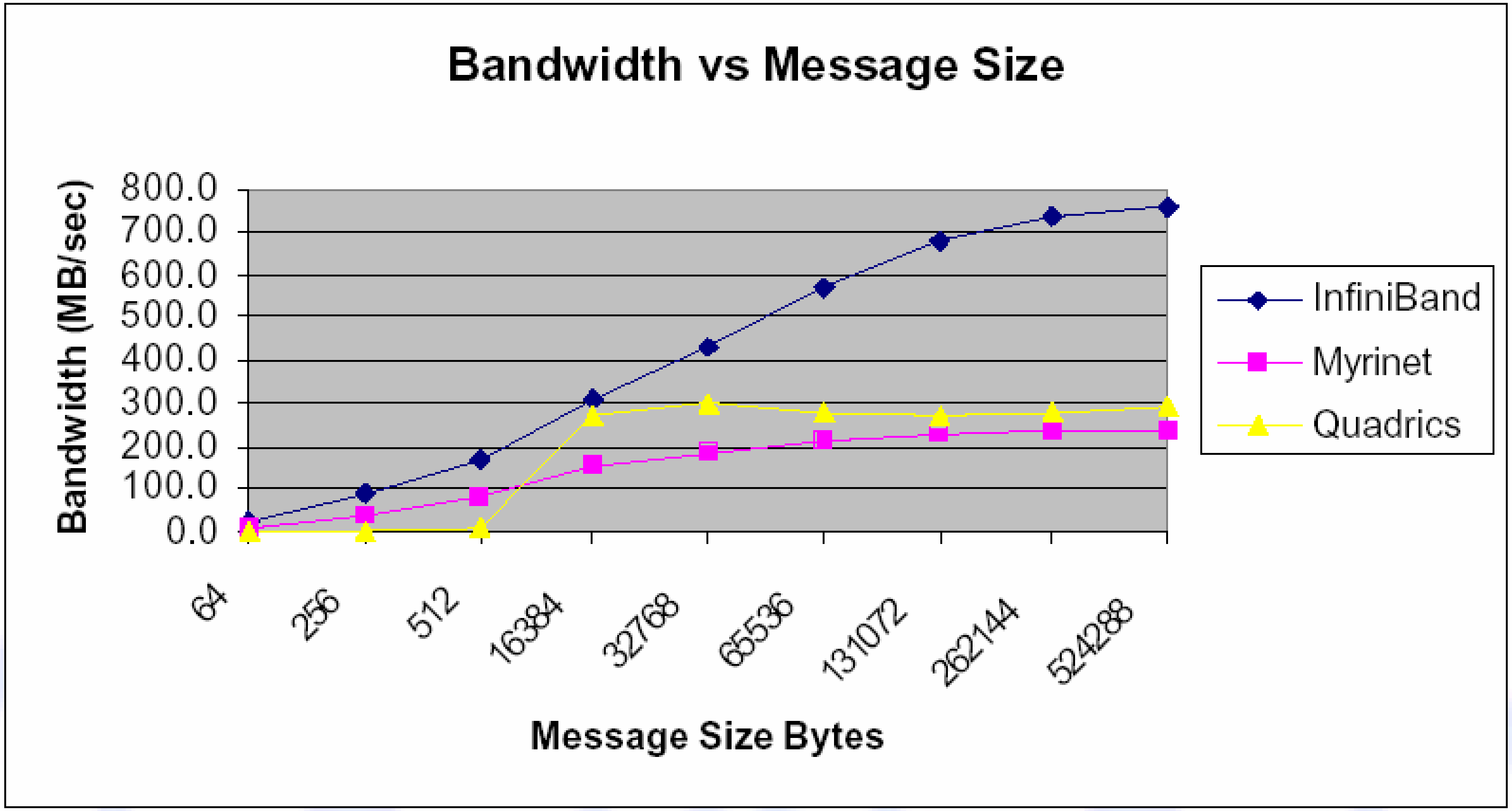
Technical details:
200 MHz Risc processor, 2 MByte memory
66MHz/64-Bit PCI-connection 2.0+2.0 Gb/s optical-connection, bidirectional

Myrinet2000 M3F-PCI64B PCI card with optical connector

Sustained bandwidth: 200 ... 240 MByte/sec

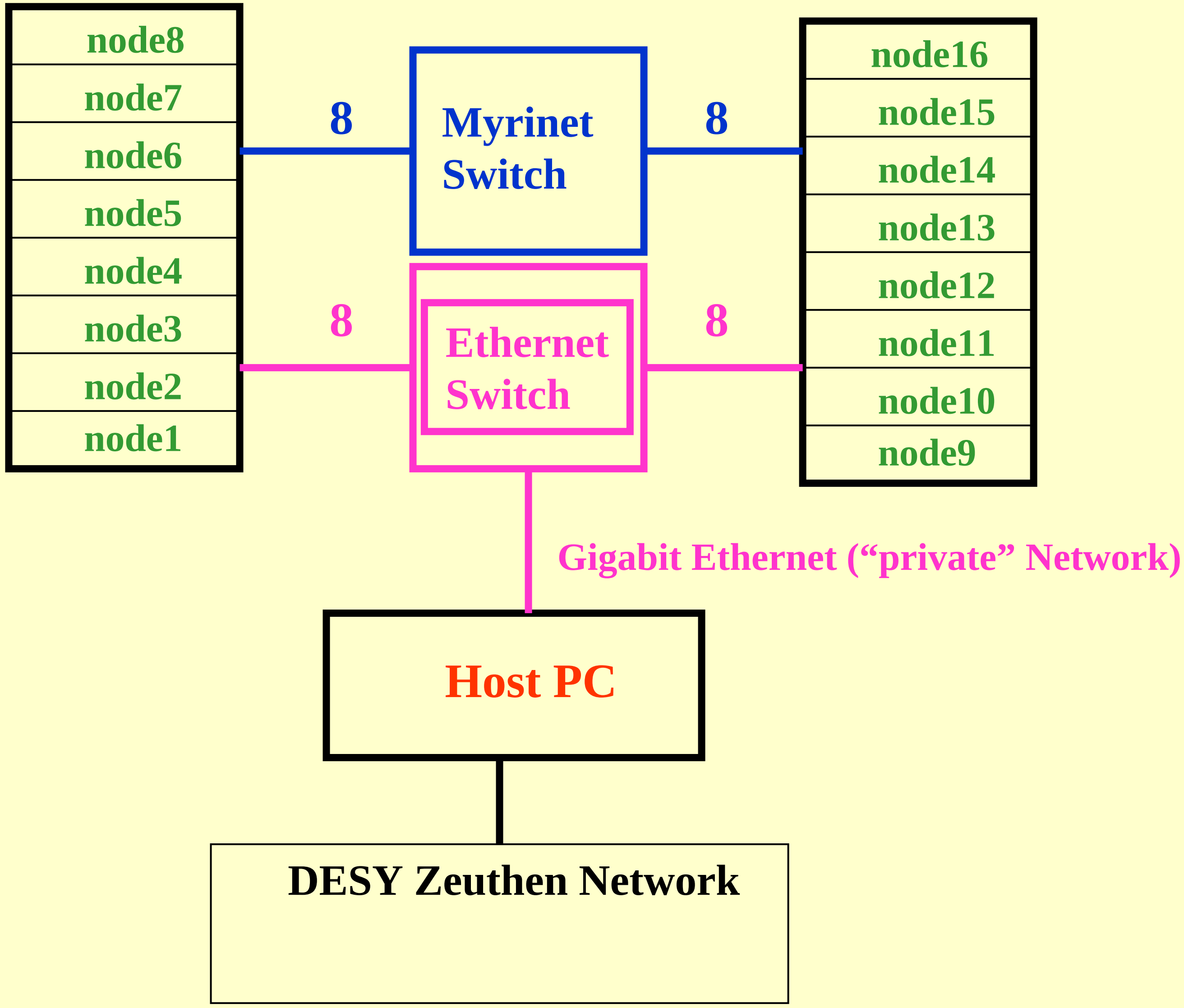
Myrinet – Infiniband Bandwidth

InfiniBand MPI Throughput Comparison



Source: Ohio State University, Xeon 2.2 GHz up processor platform

PC - Cluster Zeuthen schematic



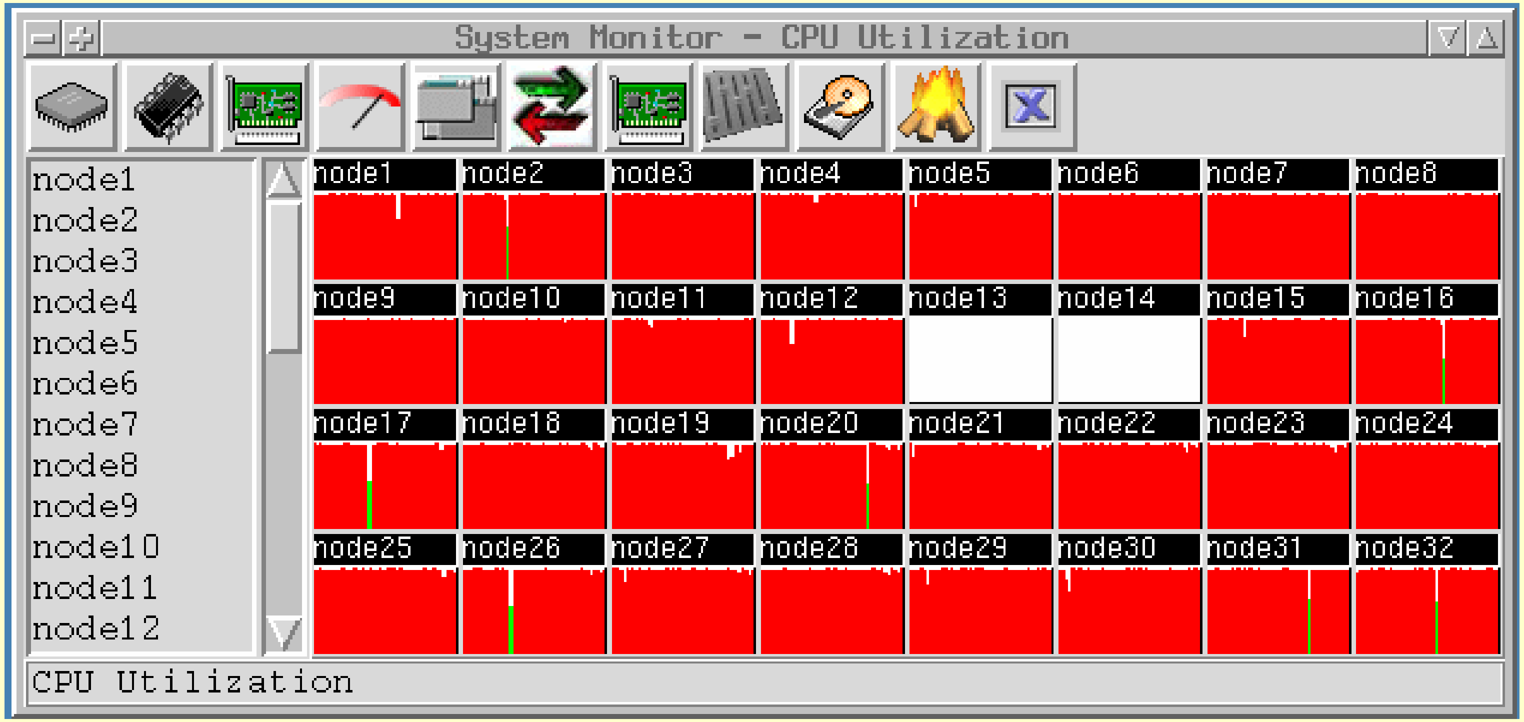
PC - Cluster Software

Operating system:	**Linux (z.B. SuSE 7.2, Scientific Linux)**
Cluster tools:	**Monitoring of temperature, fan rpm, cpu usage,**
Communication software:	**MPI - Message Passing Interface**
Compiler:	**GNU, Portland Group, Intel Compiler**
Batch system:	**PBS (OpenPBS), Sun Gridengine**

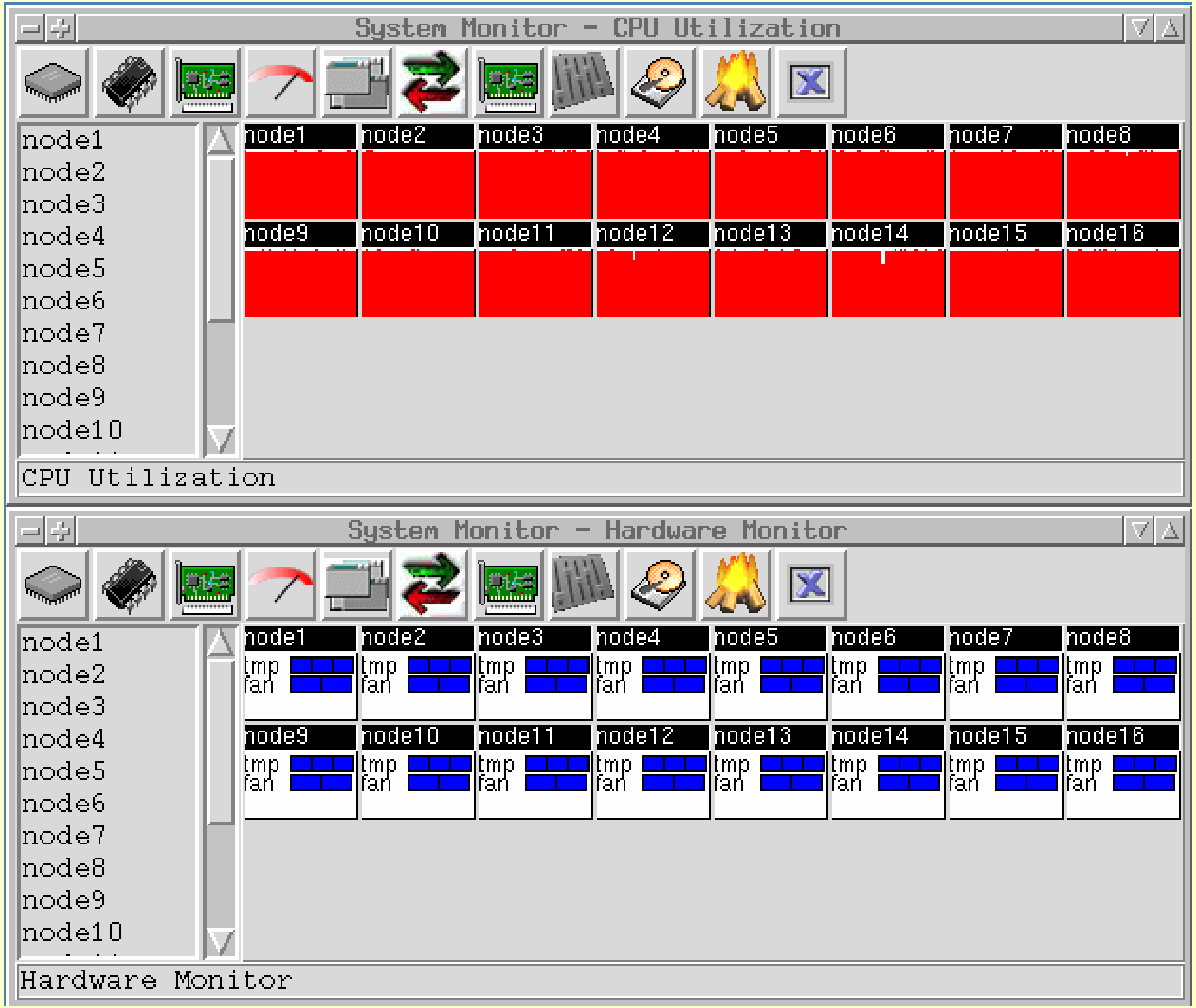
PC - Cluster Software, Monitoring Tools

Clustware from Megware

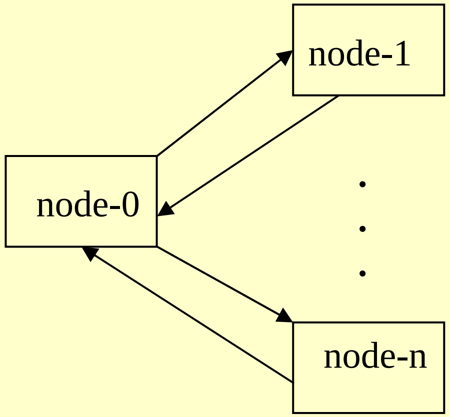
Monitoring example: CPU utilization DESY HH



PC - Cluster Software, Monitoring Tools



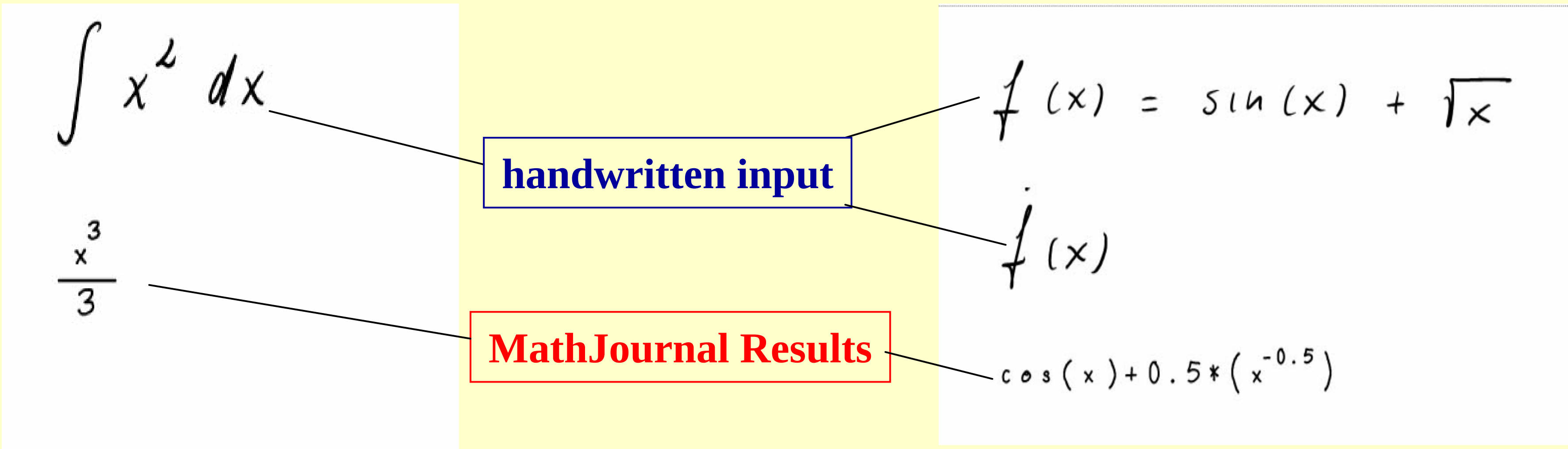
PC - Cluster Software: MPI



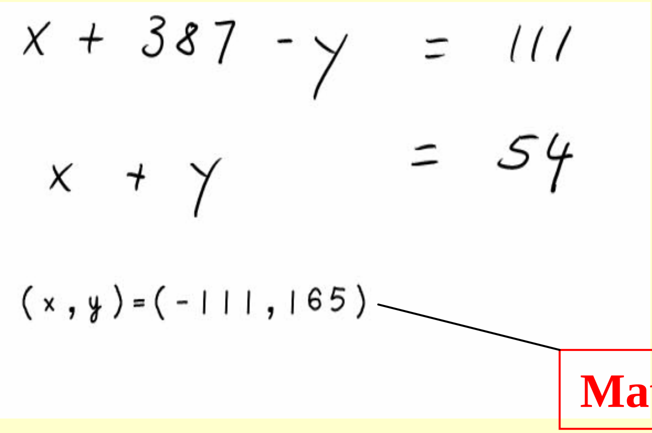

```
...
if (myid == numprocs-1)
      next = 0;
   else
      next = myid+1;

   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);
   }
...
```

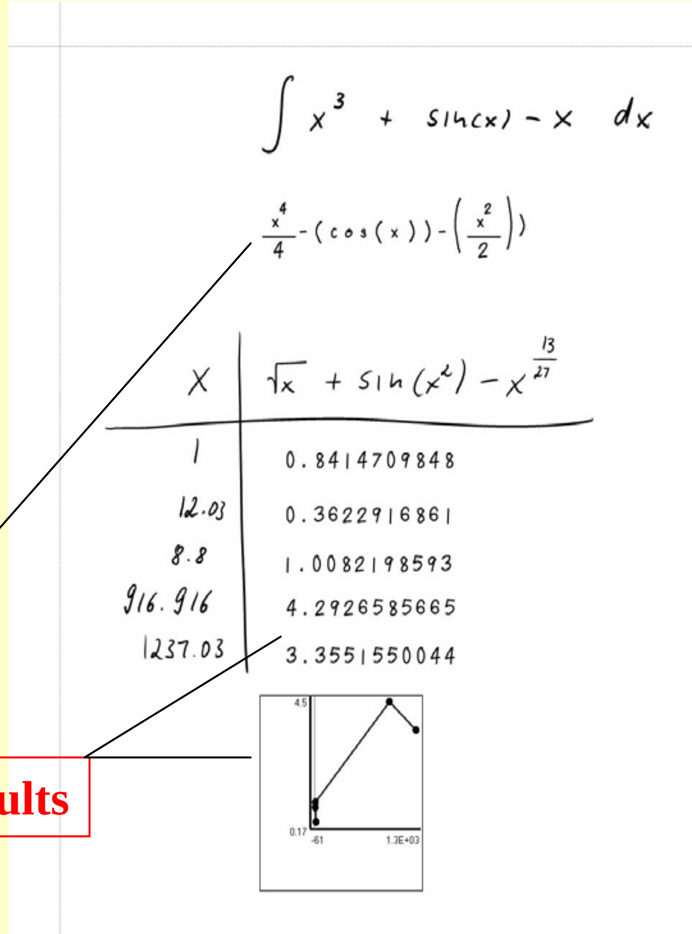
Symbolic Computing on the TabletPC – Recognizing HandwrittenMathematical Formulars and Equations (www.xthink.com)



Symbolic Computing on the TabletPC – Recognizing HandwrittenMathematical Formulars and Equations (www.xthink.com)



MathJournal Results



Symbolic Computing on the TabletPC – Recognizing HandwrittenMathematical Formulars and Equations (www.xthink.com)

MathJournal Plotting Capabilities