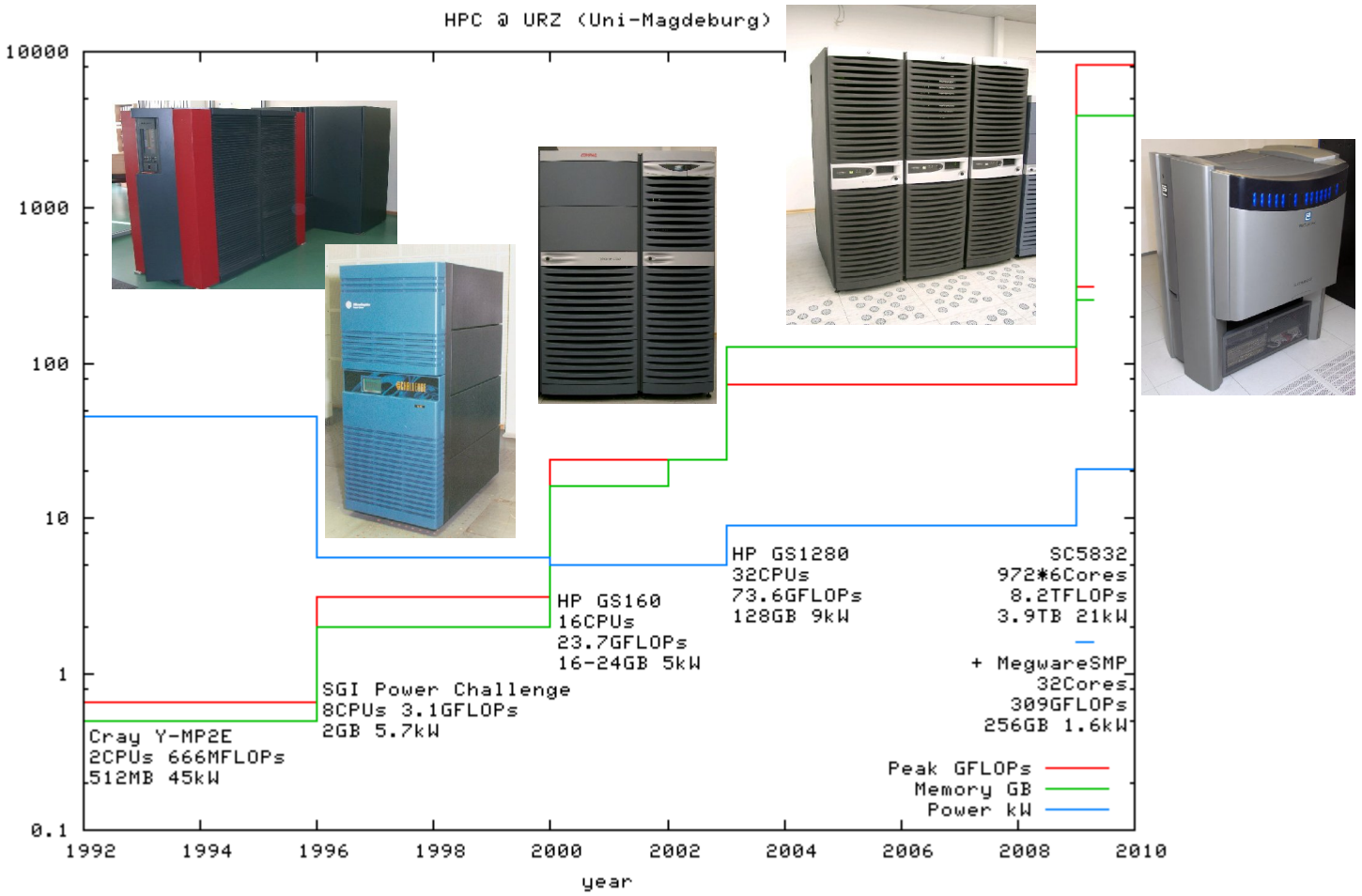
Otto-von-Guericke-Universität Magdeburg
Universitätsrechenzentrum

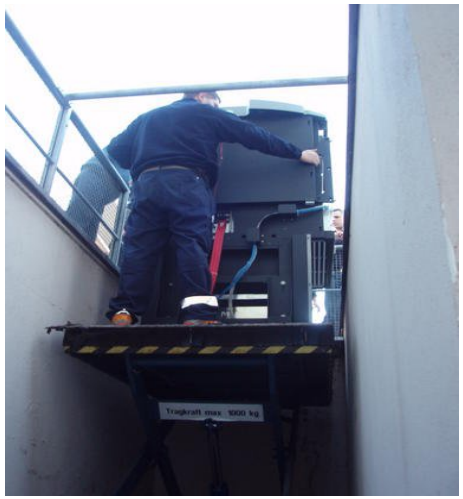
Erste Erfahrungen mit dem etwas anderen HPC-Cluster von SiCortex

HPC-Geschichte des URZ



Beschaffung 2008/2009

- traditionell sehr speicherhungrige Anwendungen (1GB : 1GFLOP) aus QPhysik
- deshalb Kommunikationsbandbreite und Latenzen zum RAM hoch gewichtet
- sanfter Wechsel SMP zu Cluster im Visier
- 2007/2008 Umstellung Code von SMP auf MPI, Tests bis 1000 Cores (vorher 32 Cores)
- Apr. 2008: AKSC Düsseldorf (SiCortex-Vortrag) – reine Neugier ... Testrechnungen
- Architekturoffene Ausschreibung mit Anwendungsbenchmark als Hauptkriterium
- SC5832 am schnellsten (+Bonus: 1/3 Stromverbrauch)



Hardware mit breiteren Flaschenhälsen:

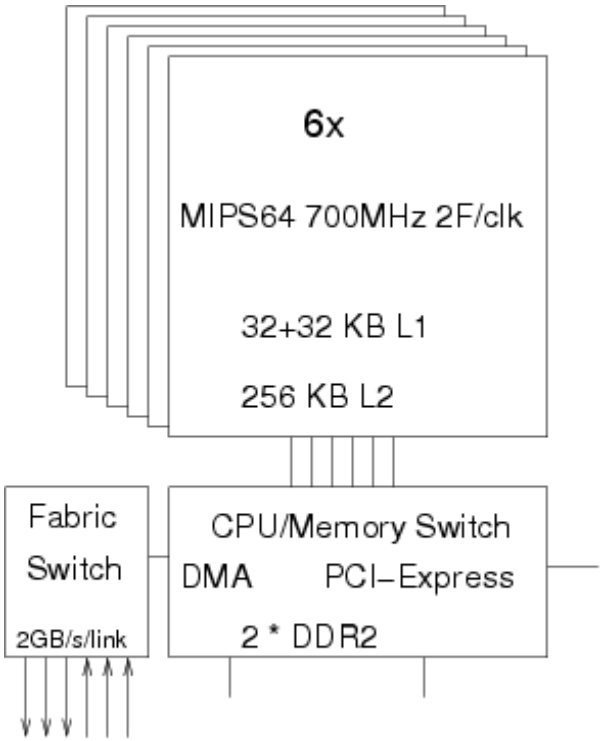
- Knoten: 6*MIPS64 700MHz 2FLOP/clk, 4GB Memory
- Kautz-Netz: 3 (In+Out)/Node, max. 6 hops, 4*3^(6-1)=972 Nodes
  - 6 lustre - 4 root - 2 head - 1 nfs server = 959 (98.7%)
  - toleriert Ausfälle von Links + Knoten!
- Sendrecv(alltoall) 52MB/s(*5748)...1.6Gb/s(*2), 1.5...7.4us
- Random Mem. (no prefetch) 6*34MB/s (256MB) (4-6*x86/peak)
- Mem. Stream 6*420MB/s (0.9-2*x86/peak)
- 21KW (0.3-0.6*(x86+IB)/peak)

Software:

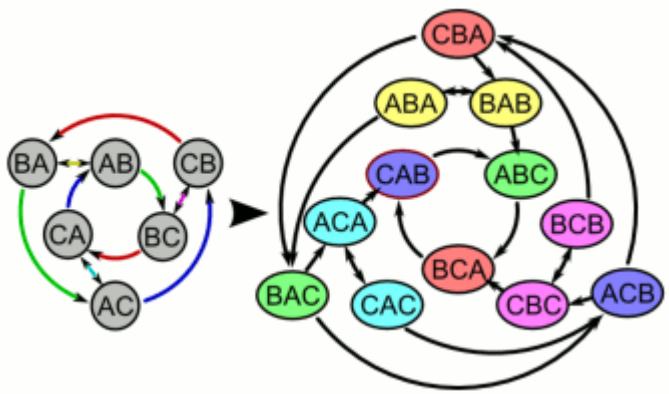
- Gentoo Linux + Slurm (srun/sbatch startet schnell)
- Compiler: PathScale + GnuC, MPICH2, HPCToolkit
- Lustre FS

Hardware / Übersicht:

- **Knoten: 6*MIPS64 700MHz 2FLOP/clk, 4GB Memory**
- **Kautz-Netz: 3 (In+Out)/Node, max. 6 hops, 4*3^(6-1)=972 Nodes**



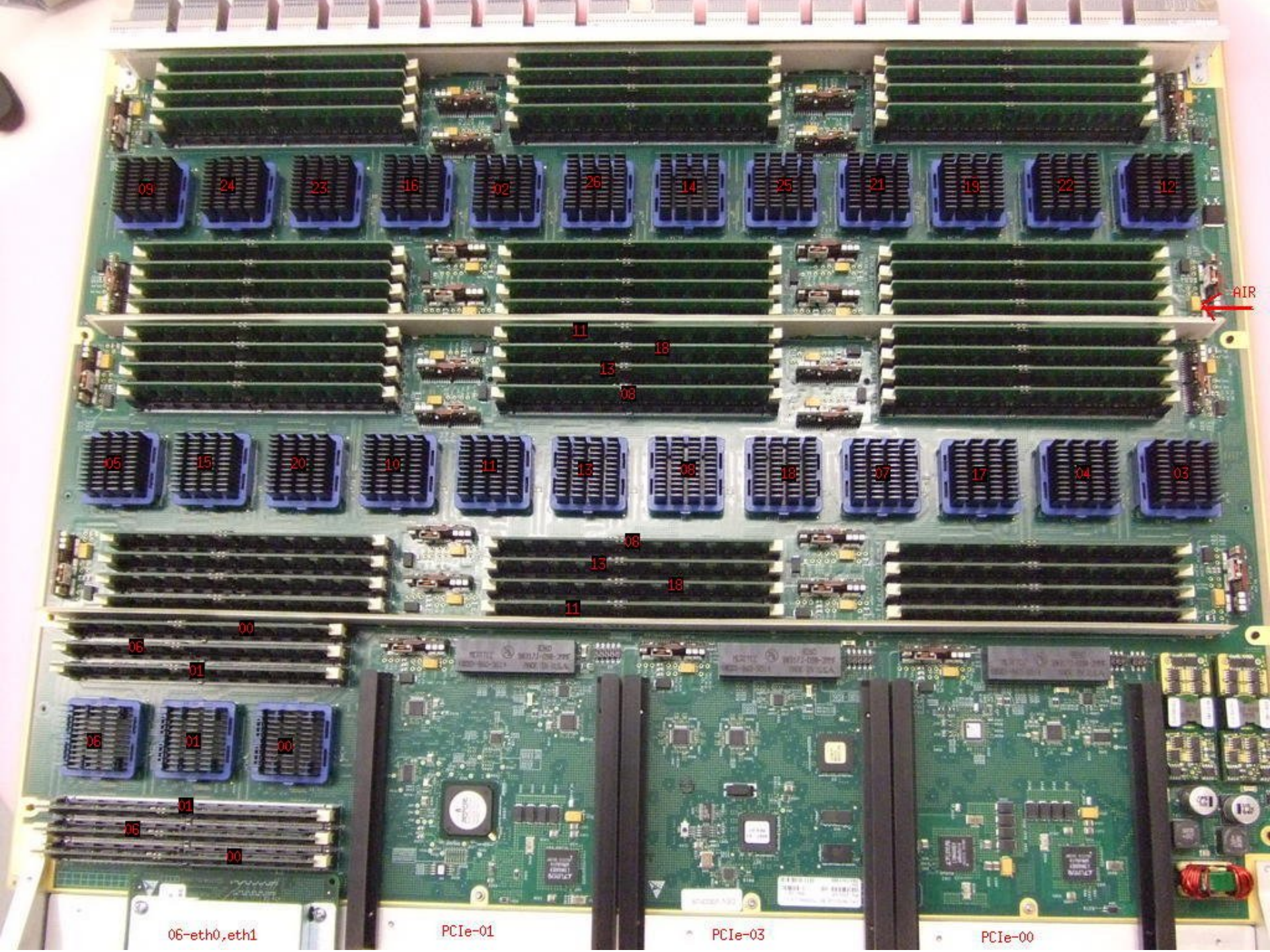
90nm, 170mm^2, 1152pins

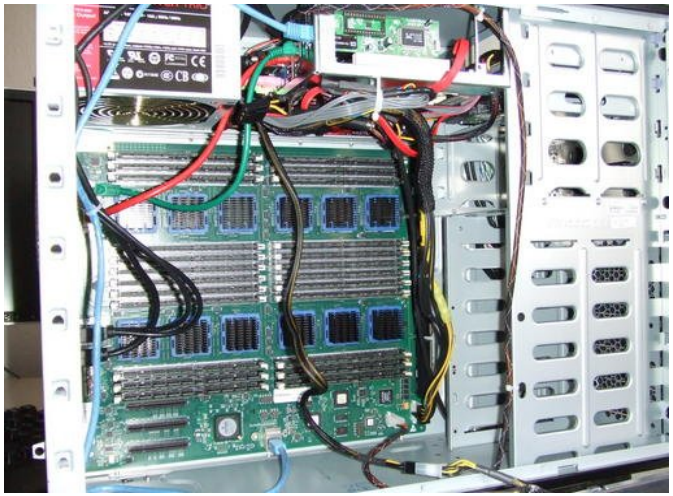


Kautzgraph: 2 Links, 2 or 3 hops

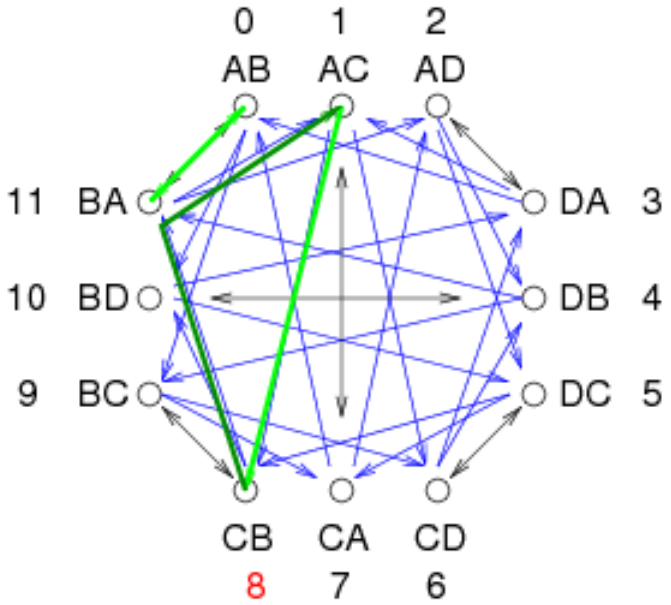
Nodes = (Links+1)Links^(hops-1)

Abb: C.Rocchini, en.wikipedia.org


09
24
23
16
02
26
14
25
21
19
22
12
AIR
11
18
13
08
05
15
20
10
11
13
08
18
07
17
04
03
08
13
18
11
00
06
01
06
01
00
01
06
00
06-eth0,eth1
PCIe-01
PCIe-03
PCIe-00



- 2 hops: 4*3^(2-1)=12 Nodes
- MPI speed per Node:
  3*2GB/s*8/10*16/19 /(1..2) = 2..4GB/s
  phys. code pckt hops
- stress: 1.2..1.7GB/s/Node 1.2..2.4us

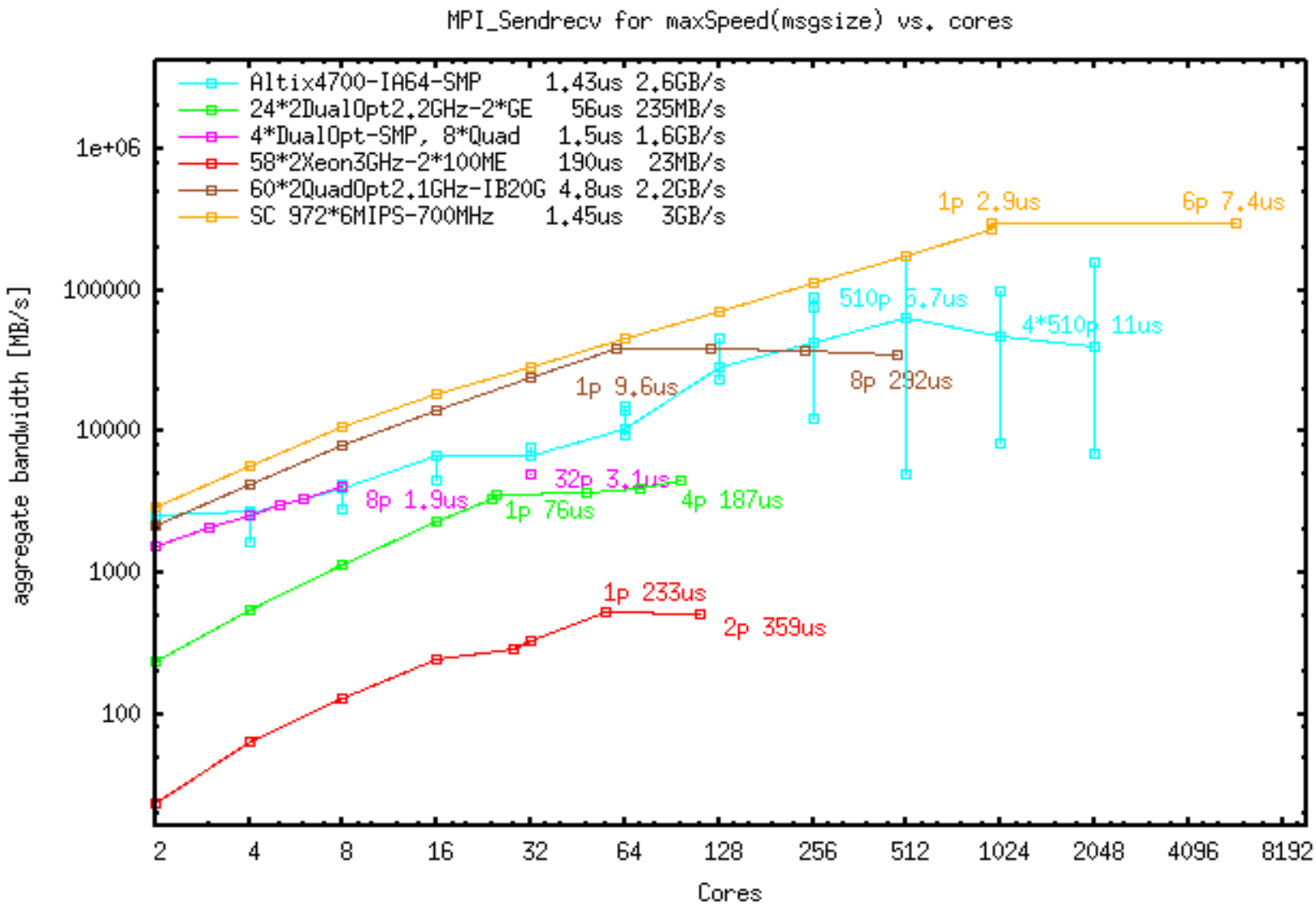


MPI_stress.c:

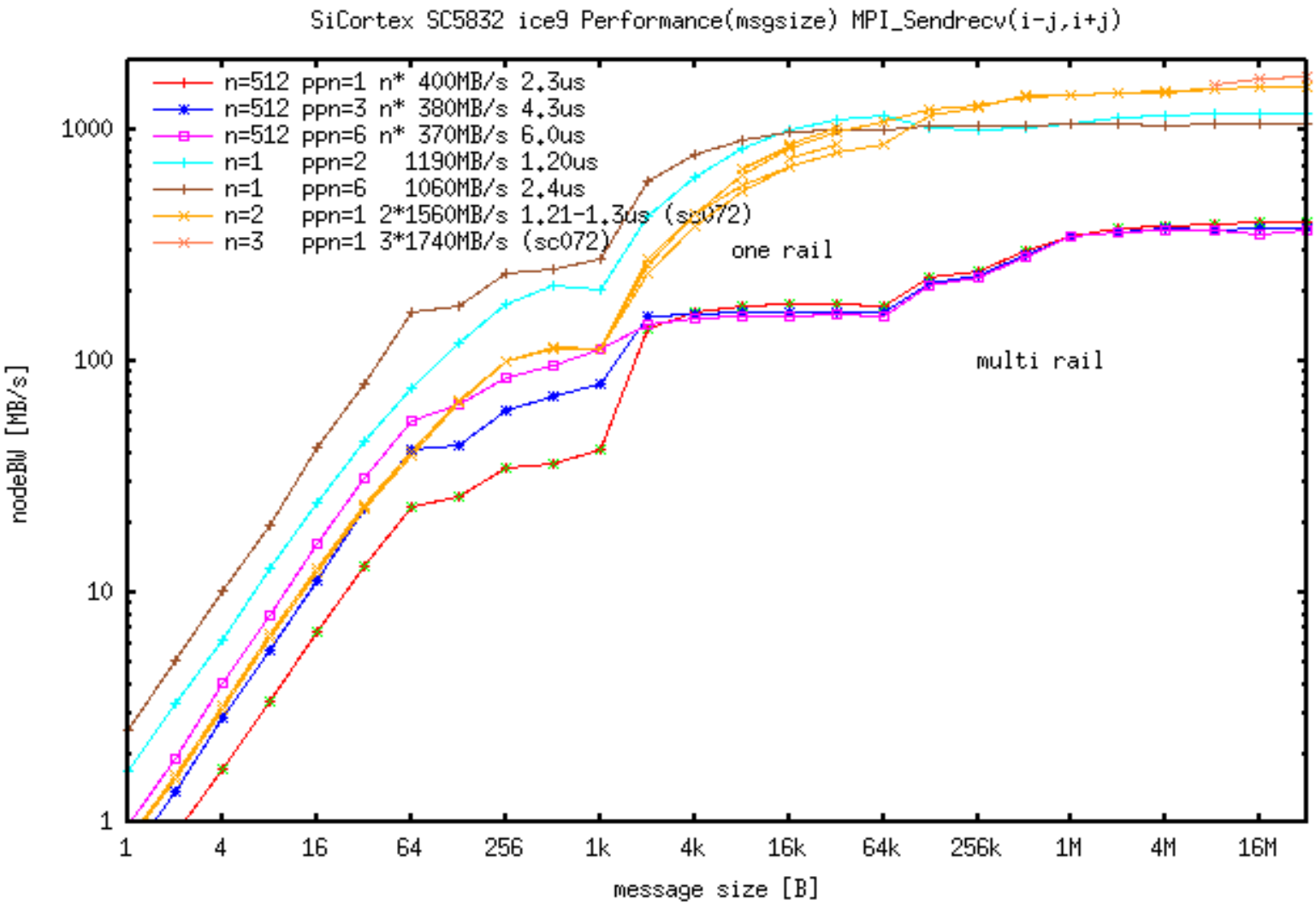
- all-to-all data distribution of spinpack (here same length)
- src = i – j, dst = i + j

```
for (;(i2<<1)<i1;i2++) {
  MPI_Sendrecv(buf1,len,MPI_BYTE,node_dst,0,
               buf2,len,MPI_BYTE,node_src,0,...);
  node_src=next_src[node_src];
  node_dst=next_dst[node_dst];
  /* 2nd call for NUMA systems */
  MPI_Sendrecv(buf2,len,MPI_BYTE,node_dst,0,
               buf1,len,MPI_BYTE,node_src,0,...);
  node_src=next_src[node_src];
  node_dst=next_dst[node_dst];
}
```

MPI-Stress Benchmark

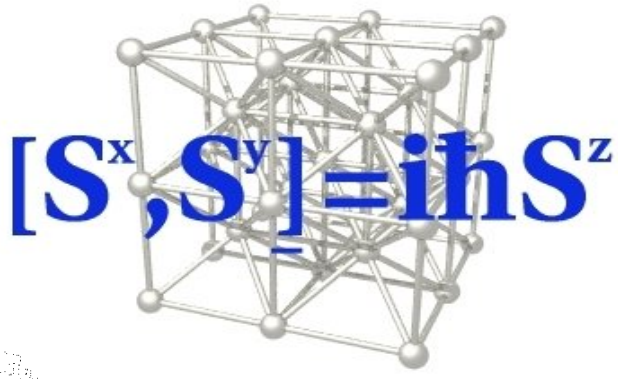


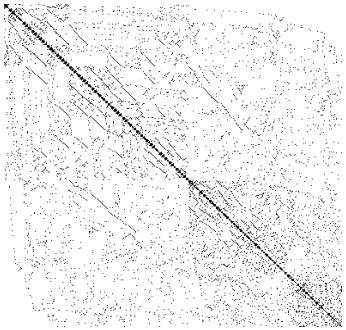
MPI-Stress Benchmark

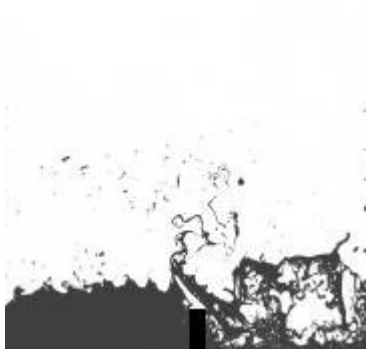


Anwenderbenchmarks ...

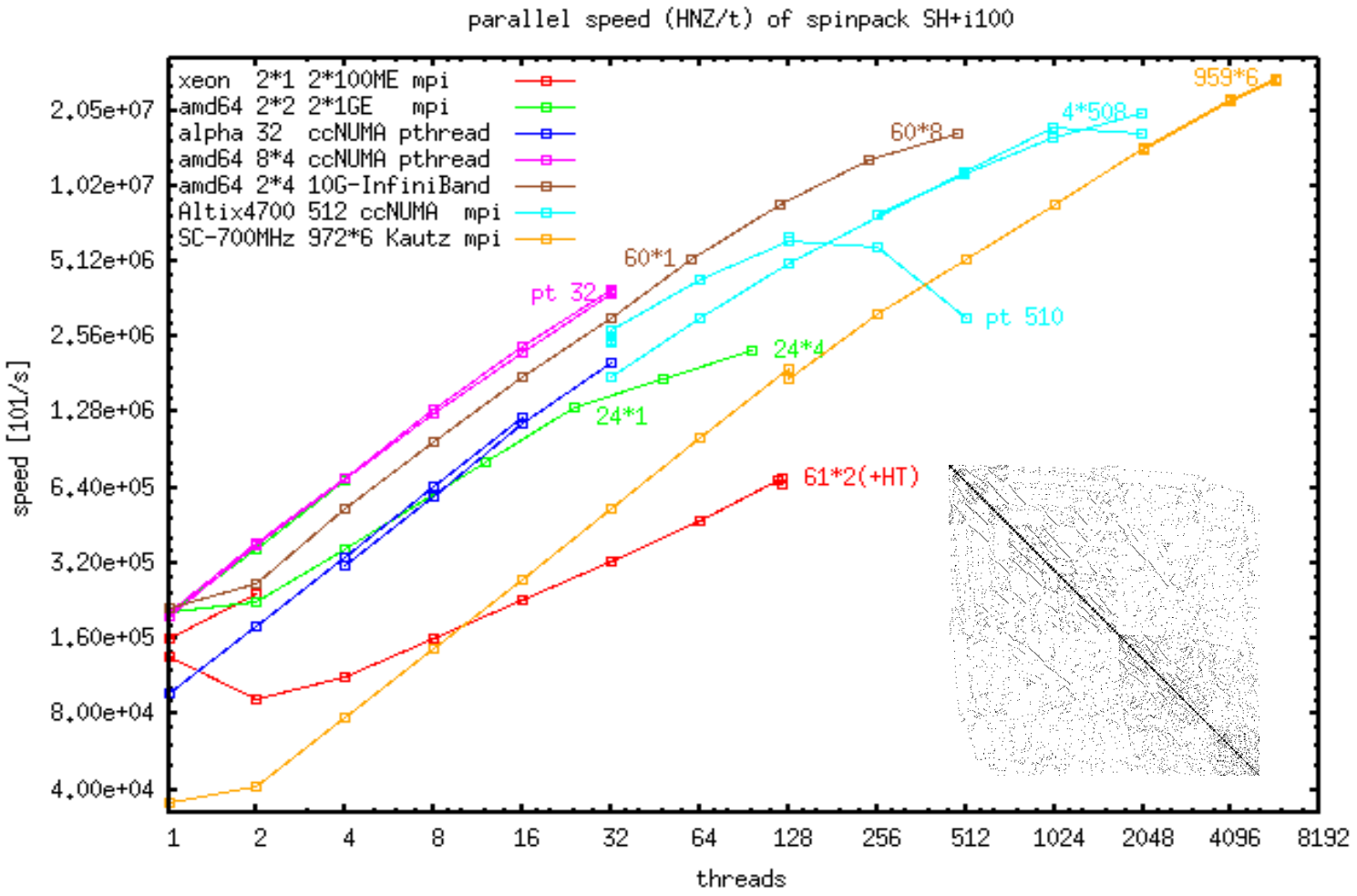
- Spinpack (Quantenphysik) = dünn besetzte Matrix (bis 5000)
- OpenFOAM (CFD, bis 100 cores getestet)







Spinpack Benchmark



```
srun -p sca -N 12 -n 72 papiex ./spin  # 0.8MFLOPs
man hpctoolkit(1)
srun -p sca -N 12-n 72 hpcex  ./spin  # statistical profiling
hpcprof -e spin ./spin.PAPI_TOT_CYC.sca-m0n0.scsystem.5960.0x

hpcex -e PAPI_L2_TCM              # ... Level 2 cache misses

Function Summary:
  36.0%  83.0% <<spinpack/exe/spin>>hamilton_geth_block
   5.7%   6.7% <<spinpack/exe/spin>>b2i
  47.7%   0.0% <<spinpack/exe/spin>>b_smallest_inlined
  ...
```

Details und Potential noch zu ergründen ...

Distributed Memory: SiCortex SC5832, 8TFLOP, 4TB

- 5754 Cores auf 959 Knoten für Einzeljob (sinnvoll) nutzbar
- excellente Performance (Kautz) + Stabilität (2 Monate)
- Besonderes Netzwerk (hohe BW, kurze Latenzen)
- Leichtes Handling für Nutzer und Admin (slurm)
- mehr Tuningaufwand durch mehr Cores
- Genügsam (max. 21kW, 1.5m x 1.5m)
- Probleme: Lustre? ein HW-Ausfall
- mehr Infos: http://www.ovgu.de/urzs/kautz/