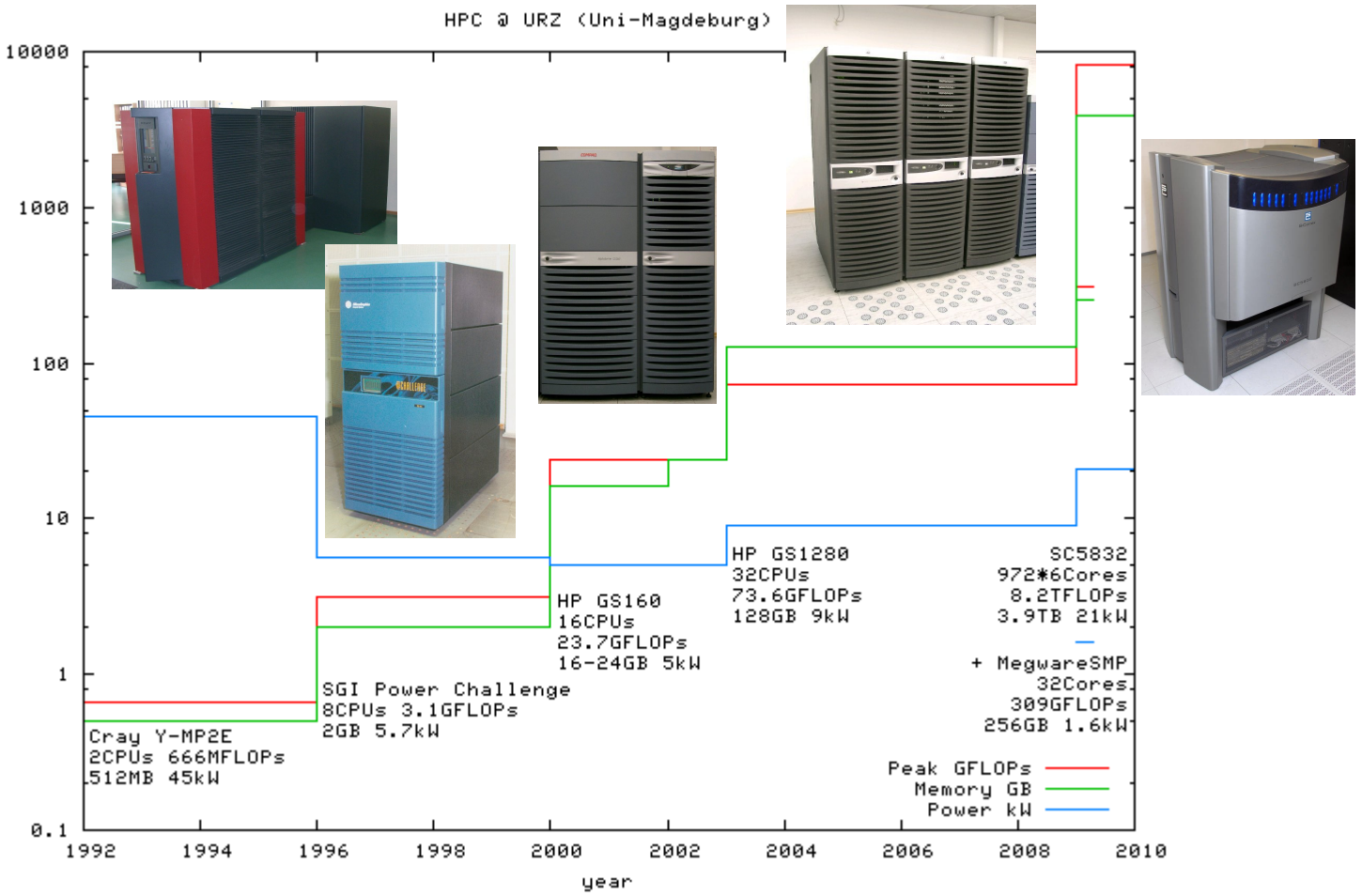
@URZ
OTTO-VON-GUERICKE-UNIVERSITÄT MAGDEBURG

Otto-von-Guericke-Universität Magdeburg
Universitätsrechenzentrum

Ausschreibung, Lieferung und erste Erfahrungen mit der SiCortex SC5832

HPC-Geschichte des URZ



Zukünftige Systemarchitektur?

Mix?

Shared Memory (OMP, MPI)
- jahrelange Erfahrung
- Physik Nutzer
- teuer, wenig Anbieter

Distributed Memory (MPI)
- wenig Erfahrung
- braucht skalierenden Code!
- preiswert, viele Anbieter

Kompromiss (?):

- traditionell Anwender (QPhysik) mit hohen Speicherbedarf
- Infiniband-Cluster von 256GB 32Core-SMPs
- mehr Flexibilität bei weniger Peak-Performance
- Problem 1: SMP-Code auf MPI anpassbar?
- Problem 2: reicht IB Bandbreite bei großen Knoten?

ca. 8 Knoten (32 Cores, 256GB)

2TFLOP : 2TB memory : 20TB disk

Beschaffung 2008/2009

- März 2007: Konzepterstellung Beschaffung bis Ende 2008
- April 2007: Nutzer-/Anbietergespräche (Orientierung: 256GB-Node-Cluster)
- Mai 2007: Angebote SUN, SGI, HP, IBM
- Dez. 2007: MPI Version SpinPack (Test bis 100 Prozesse, 100MbE, GbE)
- Feb. 2008: DFG Antrag
- März-Aug. 2008: Gutachter Rückfragen 1-3
- Apr. 2008: AKSC Düsseldorf (SiCortex-Vortrag)
- Apr-Jun. 2008: SpinPack skaliert bis 1000 Prozesse
- Aug. 2008: Testbenchmarks in Halle IB, Housten SC, SUN X4600, LRZ Altix4700
- Sept. 2008: Ausschreibung (IB-Cluster o.ä., min. 1x 256GB-Node)
- Dez. 2008: Auswahl/Bestellung Megware/SiCortex SC5832 4TB
- 19. Jan. 2009: Lieferung (Problem: 1.50m x 1.50m vs. Türen)
- 12. Febr.: Nachrüstung Storage

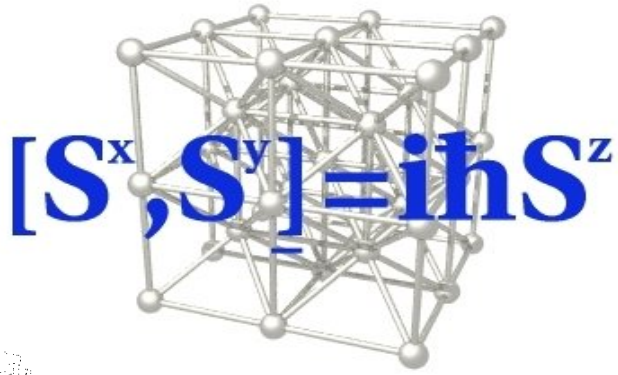


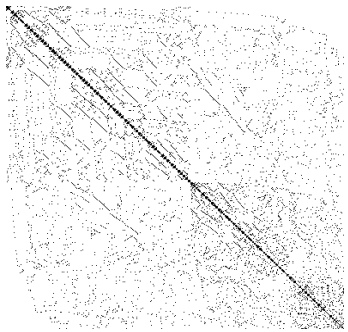
Ausschreibungseckdaten:

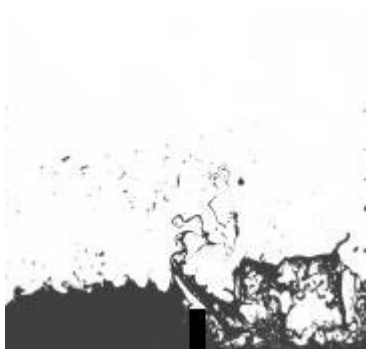
- Orientierung x86 + IB (aber Architektur offen, SMP/SC)
- 1GFLOP : 1TB Mem : 10TB Disk, IB oder vergleichbar
- Mind. Ein FatNode: x86 256GB
- Max. 750e3 EUR, max. 60kW, mind. 1TB
- Eigene Benchmarks für 50% Angebotsgröße
  (Spinpack, OpenFoam, MPI_Stress, Memspeed)
- Linux, Benchmark Compiler
- Unterscheidung muss/soll/%Wertung

Anwenderbenchmarks ...

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







Hardware mit breiteren Flaschenhälsen:

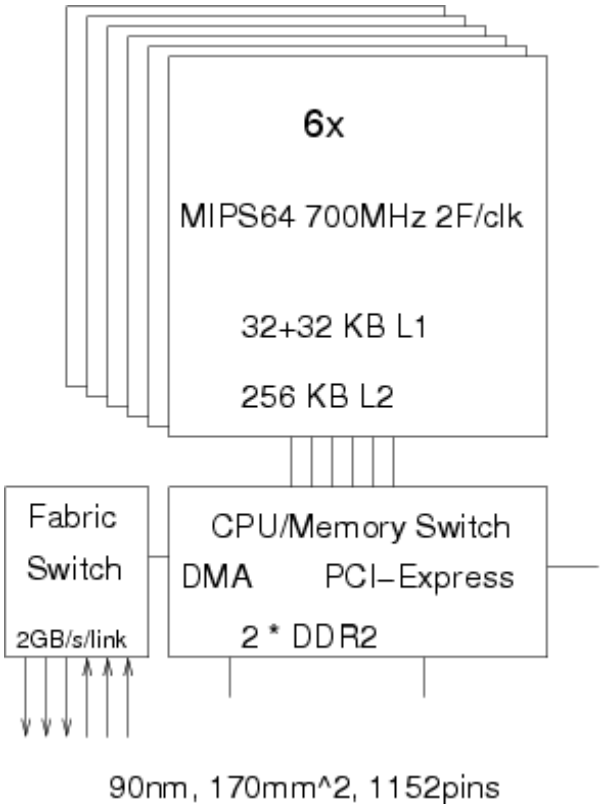
- Knoten: 6*MIPS64 700MHz 2FLOP/clk, 4GB Memory
- Kautz-Netz: 3 (In+Out)/Node, max. 6 hops, 4*3^(6-1)=972 Nodes
  - 6 lustre - 4 root - 2 head - 1 nfs server = 959 (98.7%)
- Sendrecv(alltoall) 52MB/s(*5748)...1.6Gb/s(*2), 1.5...7.4us
- Random Mem. (no prefetch) 6*34MB/s (256MB) (4-6*x86/peak)
- Mem.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

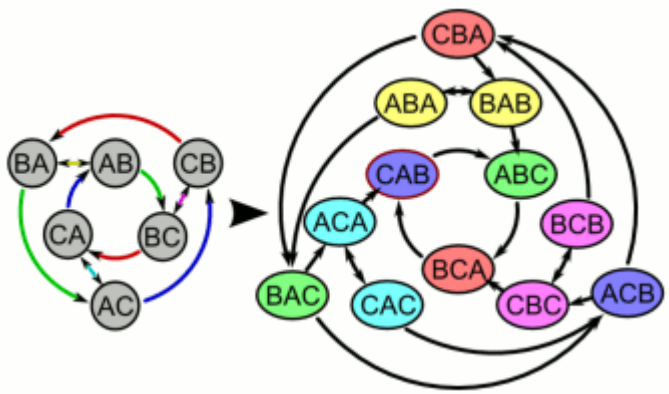
Software:

- Gentoo Linux (kernel+driver genehmigt sich z.Z. 400MB)
- 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**



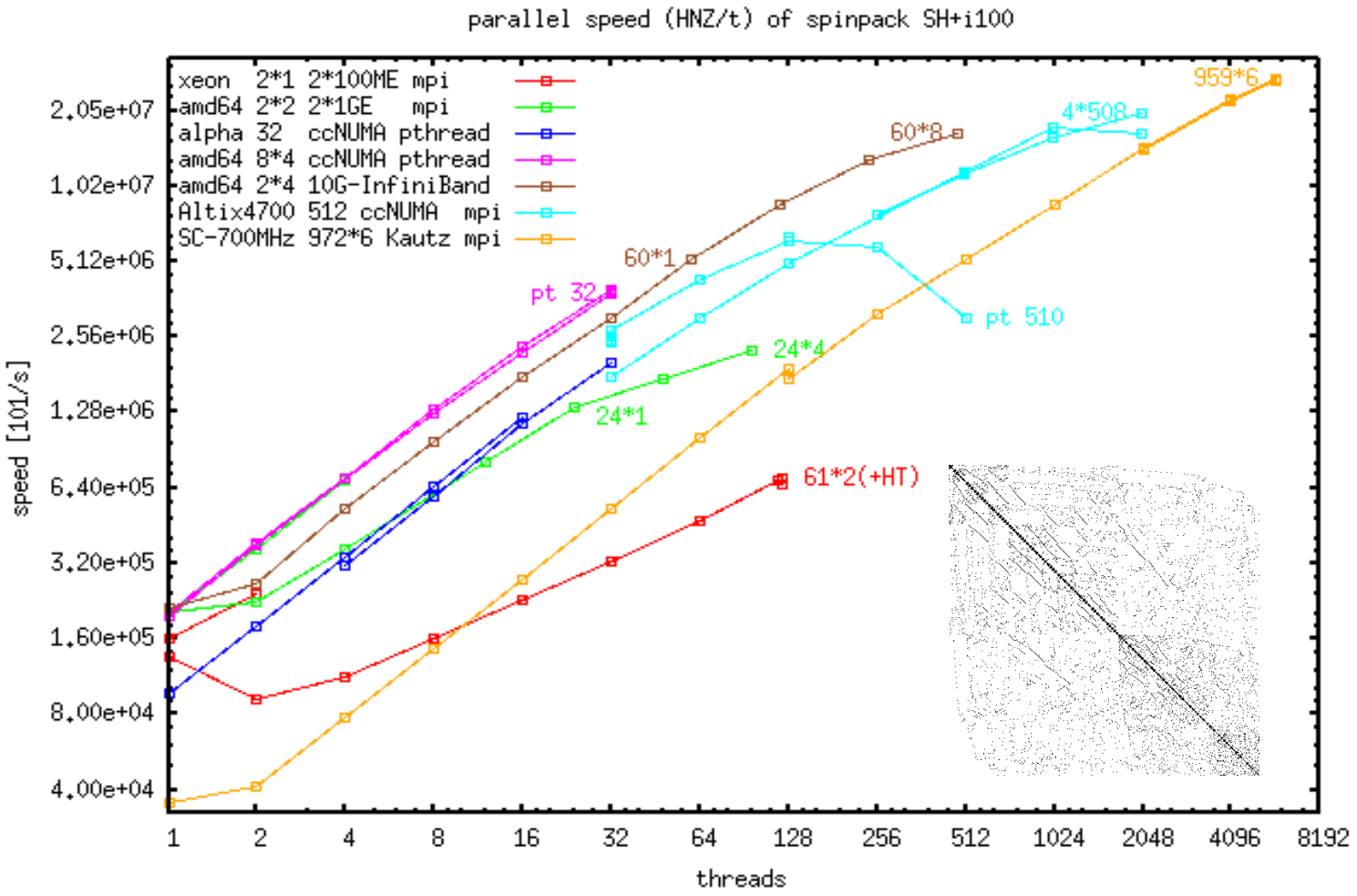


Kautzgraph: 2 Links, 2 or 3 hops

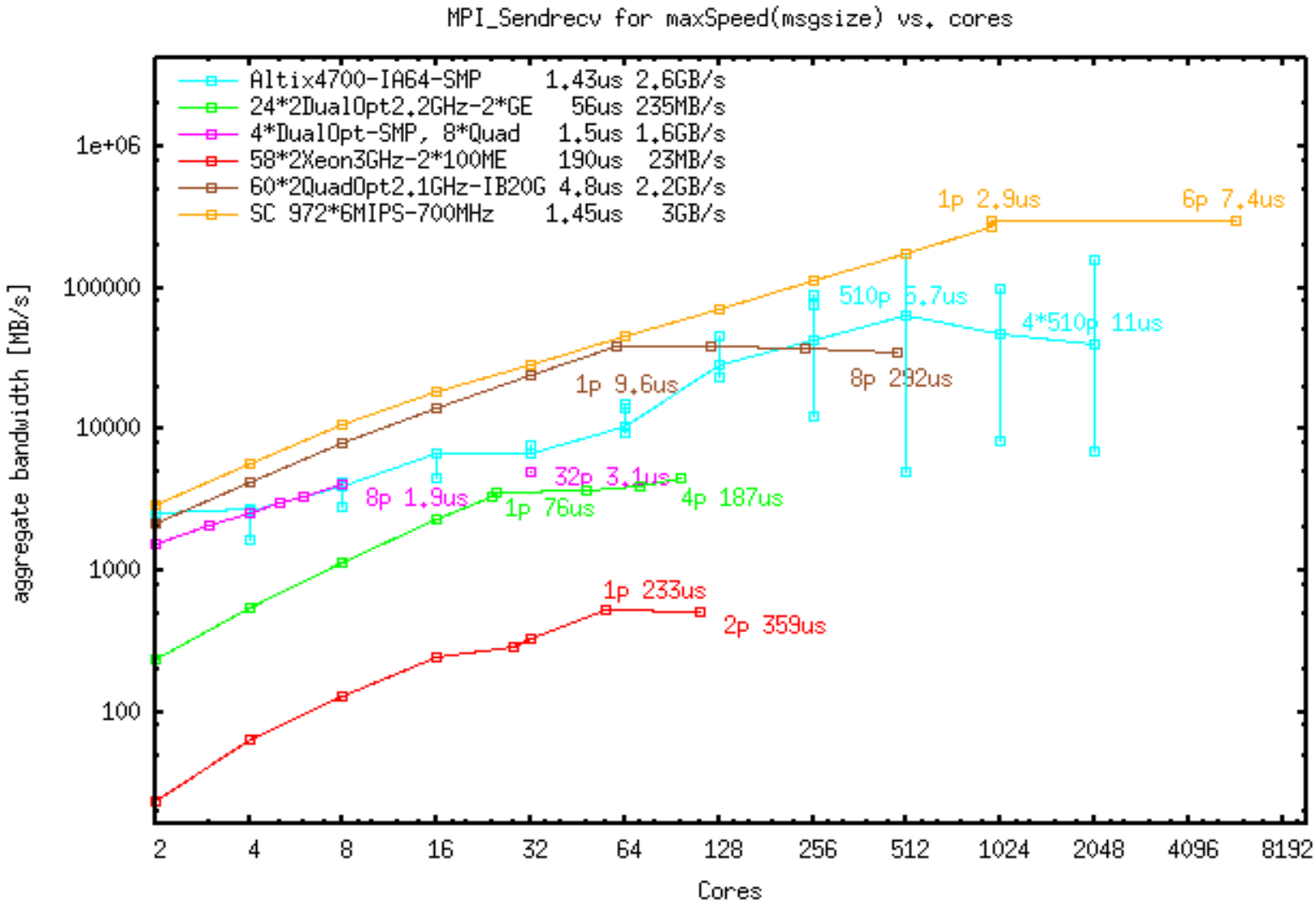
Nodes = (Links+1)Links^(hops-1)

Abb: C.Rocchini, en.wikipedia.org

Spinpack Benchmark



MPI-Stress Benchmark



Distributed Memory: SiCortex SC5832, 8TFLOP, 4TB

- 5754 Cores auf 959 Knoten für Einzeljob (sinnvoll) nutzbar
- excellente Performance (Kautz) + Stabilität (2 Monate)
- Leichtes Handling für Nutzer und Admin (slurm)
- mehr Tuningaufwand (App. speedup ab 20*GS1280)
- Genügsam (21kW, 1.5m x 1.5m)

Shared Memory: Megware, 8 Quad-Opteron, 256GB

- Pflegeleichter x86 mit 2-4 facher Leistung (zu GS1280)
- Flexibel nutzbar (ab ersten Tag gefragt)
- 1.6kW (1/5 * GS1280, 1/10 Raum, 1/10 Preis)