Network Bandwidth Measurements and Ratio Analysis with the HPC Challenge Benchmark Suite (HPCC)

Rolf Rabenseifner, Sunil R. Tiyyagura, and Matthias Müller
rabenseifner@hlrs.de, sunil@hlrs.de, mueller@hlrs.de

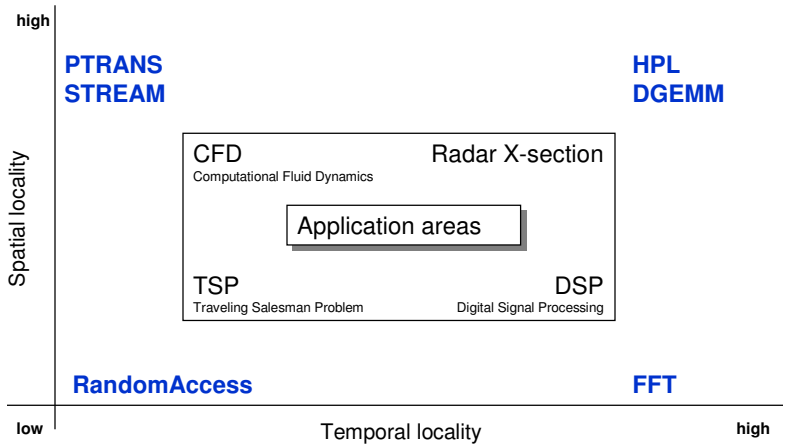
University of Stuttgart
High-Performance Computing-Center Stuttgart (HLRS)
www.hlrs.de

EuroPVM/MPI'05
Sorrento, Italy, Sep. 18-21, 2005
(HPCC data status Sep. 8, 2005)

Balance Analysis with HPC Challenge Benchmark Data

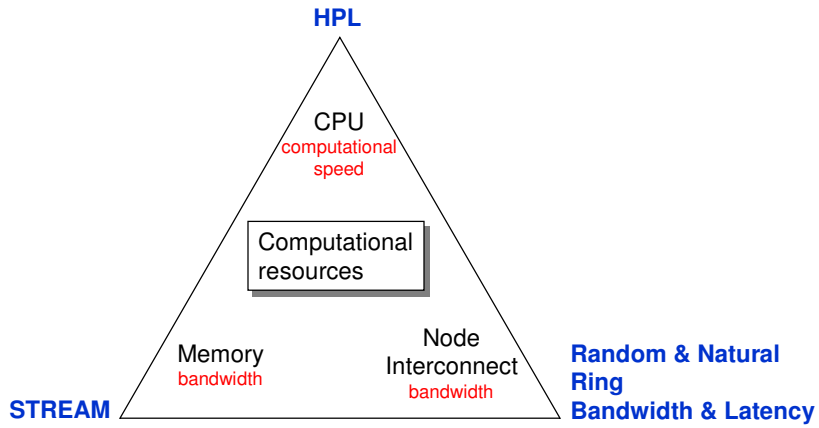
- How HPCC data can be used to analyze the balance of HPC systems
  - **Details on ring based benchmarks**
- Resource based ratios
  - **Inter-node bandwidth and**
  - **memory bandwidth**
  - **versus computational speed**
- HPCC footprint
  - **Comparing the platforms**

Application areas & HPC Challenge Benchmarks



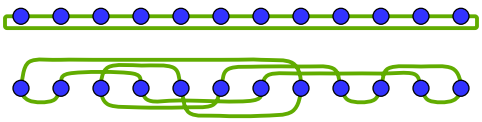
Courtesy of Jack J. Dongarra and Piotr Luszczek

Computational Resources & HPC Challenge Benchmarks



Random & natural ring bandwidth & latency

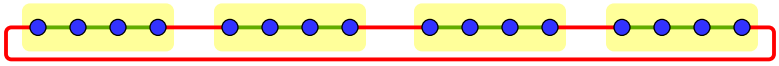
- Parallel communication pattern on all MPI processes (●)
  - Natural ring
  - Random ring



- Bandwidth per process
  - Accumulated message size / wall-clock time / number of processes
  - On each connection messages in both directions
  - With *2xMPI_Sendrecv* and *MPI non-blocking* → best result is used
  - Message size = 2,000,000 bytes
- Latency
  - Same patterns, message size = 8 bytes
  - Wall-clock time / (number of sendrecv per process)

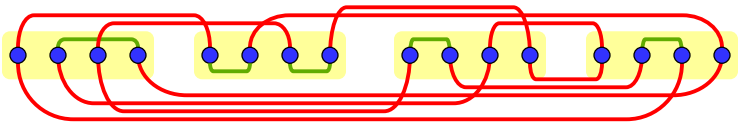
Inter-node bandwidth on clusters of SMP nodes

- Natural Ring:
  - Only one incoming and one outgoing message per node at the same time
  - Accumulated bandwidth
    := bandwidth per process x #processes
  - Reflects the communication bandwidth in the **first** dimension of a Cartesian domain decomposition
  - Does **<u>not</u>** reflect the accumulated inter-node bandwidth (nor the bi-section bandwidth) of an HPC system



Inter-node bandwidth on clusters of SMP nodes

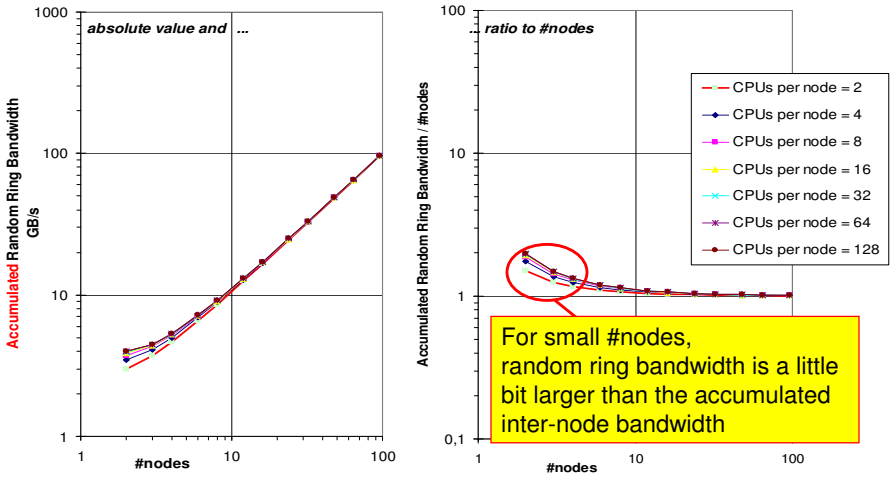
- Random Ring
  - Reflects the other dimension of a Cartesian domain decomposition and
  - Communication patterns in unstructured grids
  - Some connections are inside of the nodes
  - Most connections are inter-node
  - Depends on #nodes and #MPI processes per node



  - Accumulated bandwidth
    := bandwidth per process x #processes
  - ~= accumulated inter-node bandwidth x $(1 – 1 / \#nodes)^{-1}$

skipped

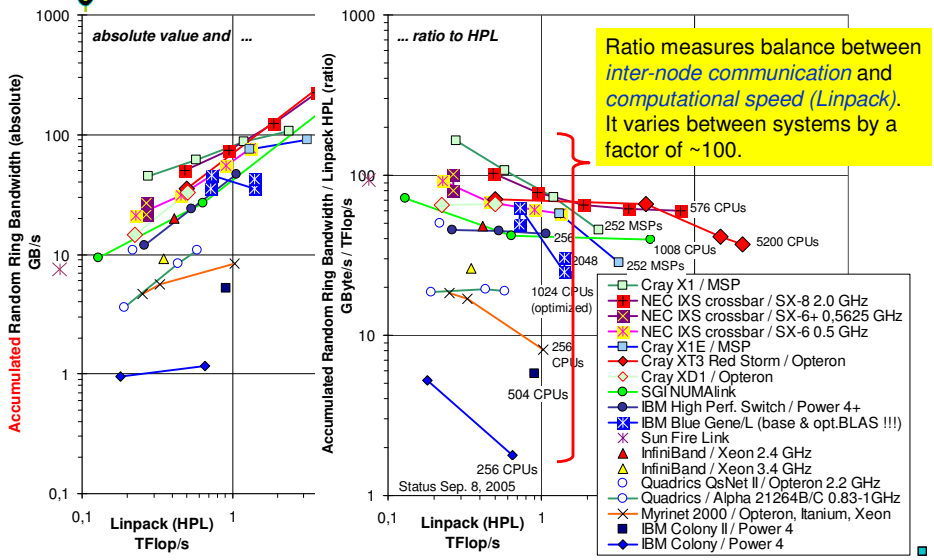
On an ideally switched cluster …



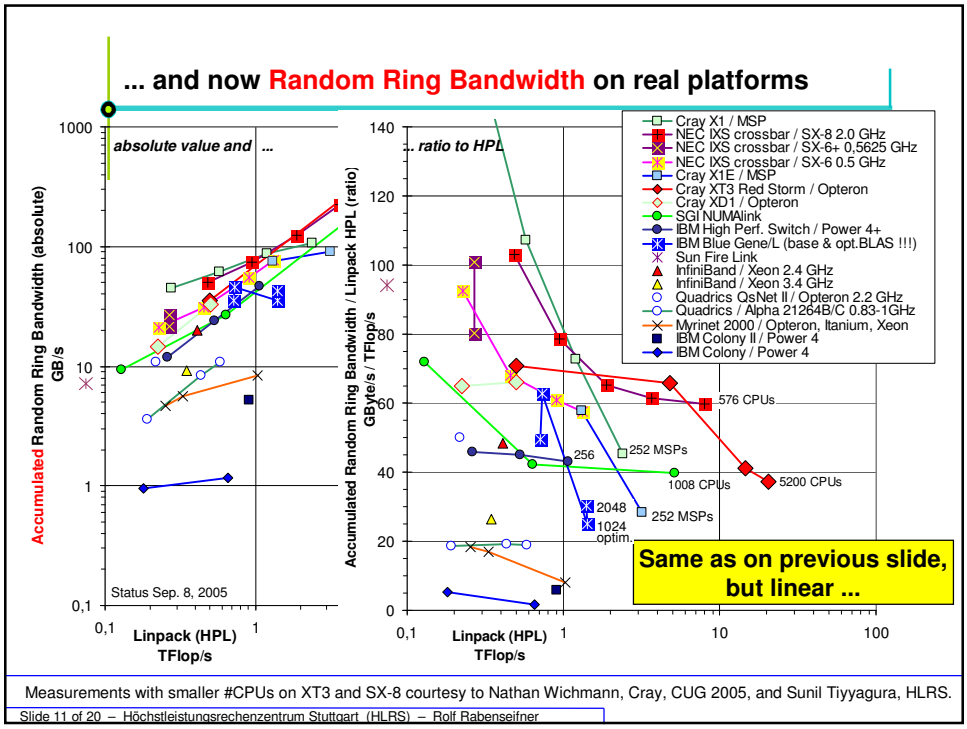
Balance Analysis with HPC Challenge Benchmark Data

- Balance can be expressed as a set of ratios
  - e.g., accumulated memory bandwidth / accumulated Tflop/s rate
- Basis
  - **Linpack (HPL)** → **Computational Speed**
  - **Random Ring Bandwidth** → **Inter-node communication**
  - **Parallel STREAM Copy or Triad** → **Memory bandwidth**
- Be careful:
  - Some data are presented for the **total system**
  - Some per **MPI process** (HPL processes)
  - i.e., balance calculation always
    - **with accumulated data on total system, or**
    - **with divided data to one MPI process**

... and now Random Ring Bandwidth on real platforms

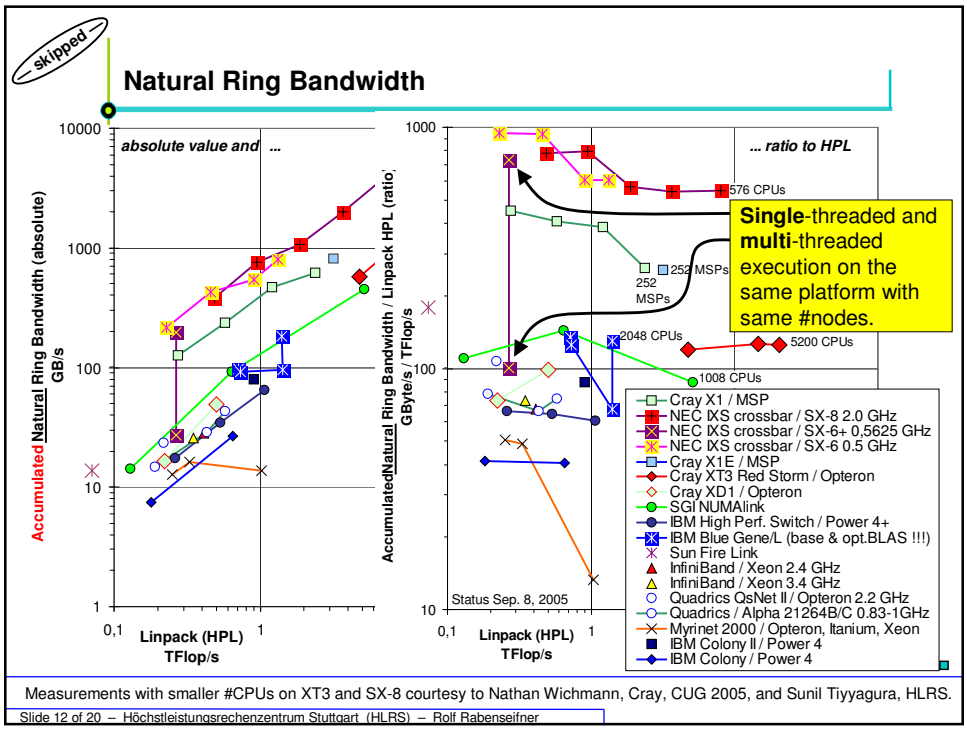


Measurements with smaller #CPUs on XT3 and SX-8 courtesy to Nathan Wichmann, Cray, CUG 2005, and Sunil Tiyyagura, HLRS.



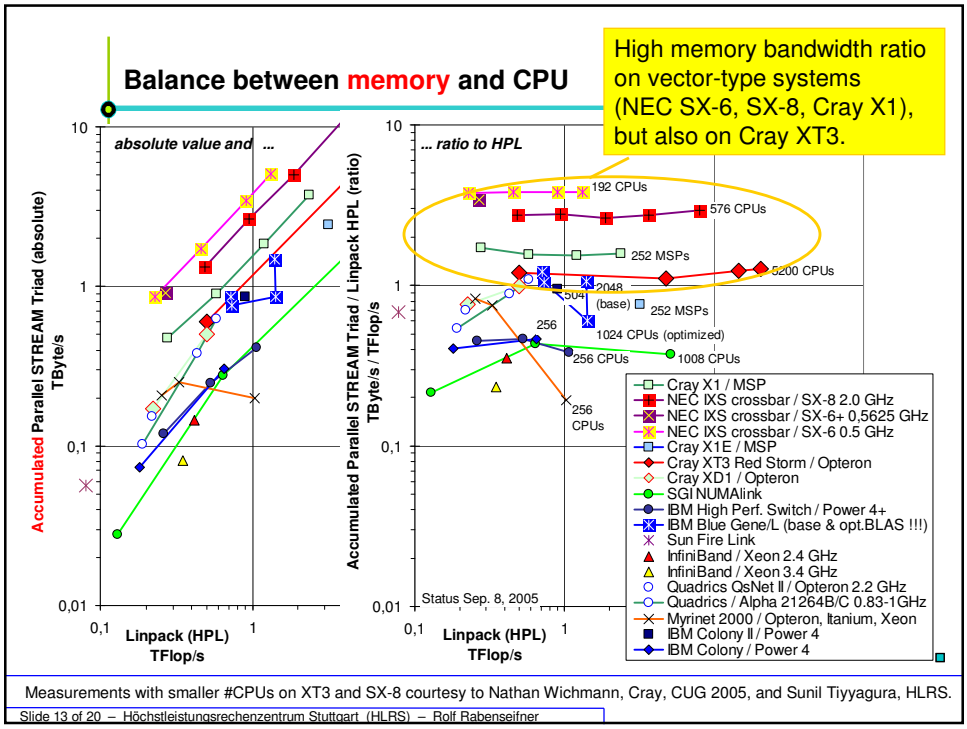
Measurements with smaller #CPUs on XT3 and SX-8 courtesy to Nathan Wichmann, Cray, CUG 2005, and Sunil Tiyyagura, HLRS.

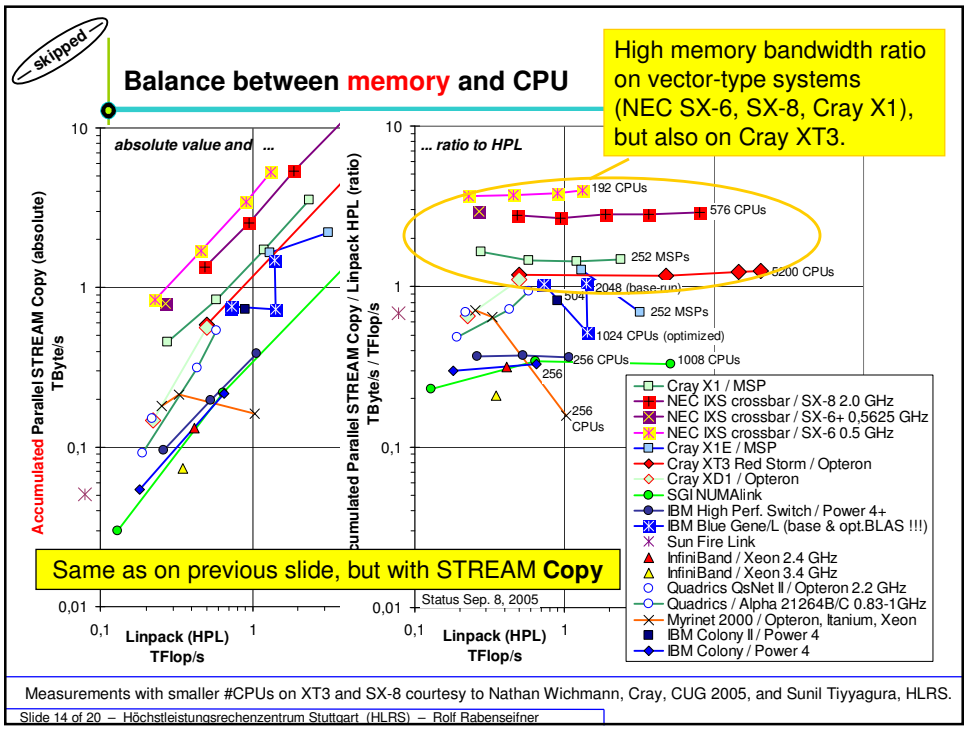
Slide 11 of 20 – Höchstleistungsrechenzentrum Stuttgart (HLRS) – Rolf Rabenseifner

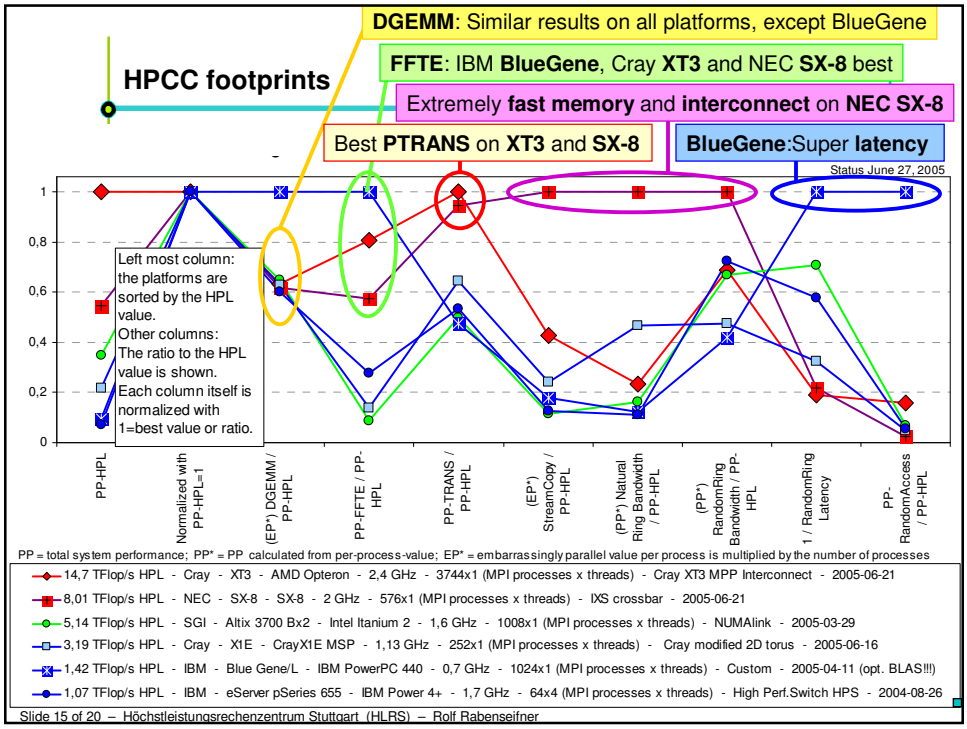


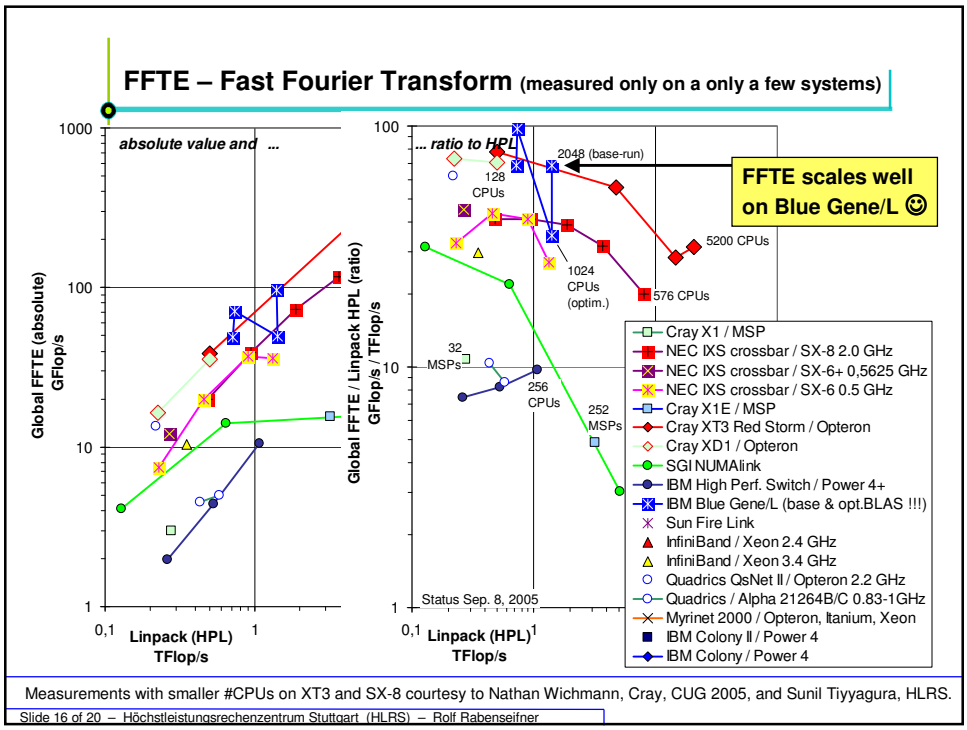
Measurements with smaller #CPUs on XT3 and SX-8 courtesy to Nathan Wichmann, Cray, CUG 2005, and Sunil Tiyyagura, HLRS.

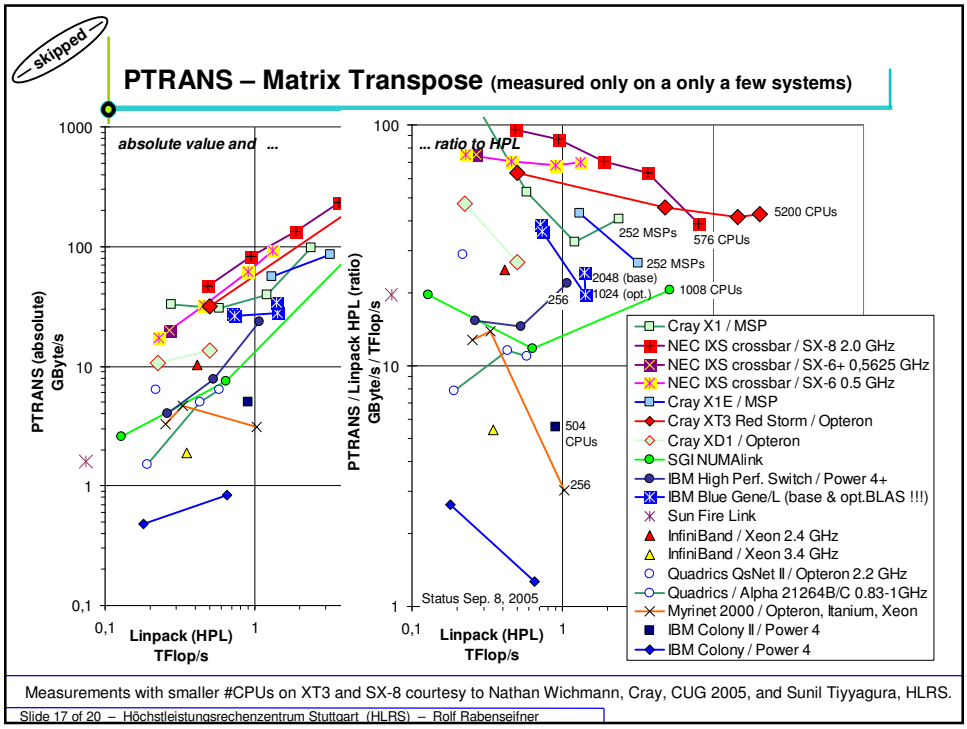
Slide 12 of 20 – Höchstleistungsrechenzentrum Stuttgart (HLRS) – Rolf Rabenseifner

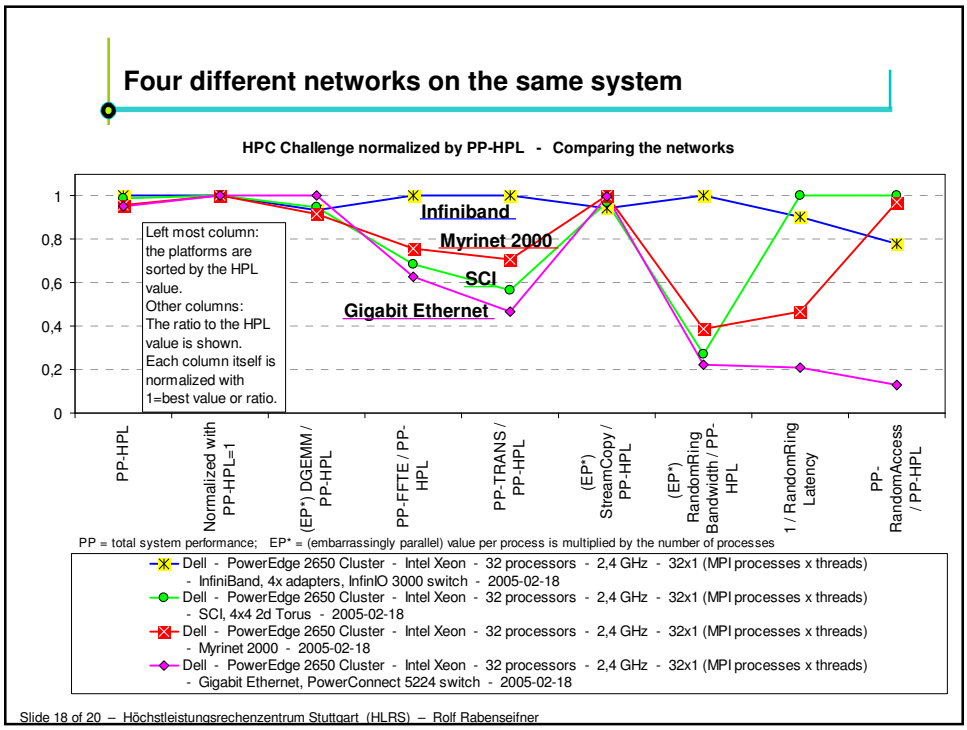

Balance between memory and CPU
High memory bandwidth ratio on vector-type systems (NEC SX-6, SX-8, Cray X1), but also on Cray XT3.
absolute value and ...
... ratio to HPL
Accumulated Parallel STREAM Triad (absolute) TByte/s
Accumulated Parallel STREAM Triad / Linpack HPL (ratio) TByte/s / TFlop/s
Linpack (HPL) TFlop/s
192 CPUs
576 CPUs
252 MSPs
5200 CPUs
2048 (base)
504
252 MSPs
1024 CPUs (optimized)
256
256 CPUs
1008 CPUs
256 CPUs
Status Sep. 8, 2005
Cray X1 / MSP
NEC IXS crossbar / SX-8 2.0 GHz
NEC IXS crossbar / SX-6+ 0,5625 GHz
NEC IXS crossbar / SX-6 0.5 GHz
Cray X1E / MSP
Cray XT3 Red Storm / Opteron
Cray XD1 / Opteron
SGI NUMAlink
IBM High Perf. Switch / Power 4+
IBM Blue Gene/L (base & opt.BLAS !!!)
Sun Fire Link
InfiniBand / Xeon 2.4 GHz
InfiniBand / Xeon 3.4 GHz
Quadrics QsNet II / Opteron 2.2 GHz
Quadrics / Alpha 21264B/C 0.83-1GHz
Myrinet 2000 / Opteron, Itanium, Xeon
IBM Colony II / Power 4
IBM Colony / Power 4
Measurements with smaller #CPUs on XT3 and SX-8 courtesy to Nathan Wichmann, Cray, CUG 2005, and Sunil Tiyyagura, HLRS.
Slide 13 of 20 – Höchstleistungsrechenzentrum Stuttgart (HLRS) – Rolf Rabenseifner


skipped
Balance between memory and CPU
High memory bandwidth ratio on vector-type systems (NEC SX-6, SX-8, Cray X1), but also on Cray XT3.
absolute value and ...
... ratio to HPL
Accumulated Parallel STREAM Copy (absolute) TByte/s
Accumulated Parallel STREAM Copy / Linpack HPL (ratio) TByte/s / TFlop/s
Linpack (HPL) TFlop/s
192 CPUs
576 CPUs
252 MSPs
5200 CPUs
2048 (base-run)
504
252 MSPs
1024 CPUs (optimized)
256 CPUs
1008 CPUs
256
256 CPUs
Same as on previous slide, but with STREAM Copy
Status Sep. 8, 2005
Cray X1 / MSP
NEC IXS crossbar / SX-8 2.0 GHz
NEC IXS crossbar / SX-6+ 0,5625 GHz
NEC IXS crossbar / SX-6 0.5 GHz
Cray X1E / MSP
Cray XT3 Red Storm / Opteron
Cray XD1 / Opteron
SGI NUMAlink
IBM High Perf. Switch / Power 4+
IBM Blue Gene/L (base & opt.BLAS !!!)
Sun Fire Link
InfiniBand / Xeon 2.4 GHz
InfiniBand / Xeon 3.4 GHz
Quadrics QsNet II / Opteron 2.2 GHz
Quadrics / Alpha 21264B/C 0.83-1GHz
Myrinet 2000 / Opteron, Itanium, Xeon
IBM Colony II / Power 4
IBM Colony / Power 4
Measurements with smaller #CPUs on XT3 and SX-8 courtesy to Nathan Wichmann, Cray, CUG 2005, and Sunil Tiyyagura, HLRS.
Slide 14 of 20 – Höchstleistungsrechenzentrum Stuttgart (HLRS) – Rolf Rabenseifner


HPCC footprints
DGEMM: Similar results on all platforms, except BlueGene
FFTE: IBM BlueGene, Cray XT3 and NEC SX-8 best
Extremely fast memory and interconnect on NEC SX-8
Best PTRANS on XT3 and SX-8
BlueGene:Super latency
Status June 27, 2005
Left most column: the platforms are sorted by the HPL value.
Other columns: The ratio to the HPL value is shown.
Each column itself is normalized with 1=best value or ratio.
PP-HPL
Normalized with PP-HPL=1
(EP*) DGEMM / PP-HPL
PP-FFTE / PP-HPL
PP-TRANS / PP-HPL
(EP*) StreamCopy / PP-HPL
(PP*) Natural Ring Bandwidth / PP-HPL
(PP*) RandomRing Bandwidth / PP-HPL
1 / RandomRing Latency
PP-RandomAccess / PP-HPL
PP = total system performance; PP* = PP calculated from per-process-value; EP* = embarrassingly parallel value per process is multiplied by the number of processes
14,7 TFlop/s HPL - Cray - XT3 - AMD Opteron - 2,4 GHz - 3744x1 (MPI processes x threads) - Cray XT3 MPP Interconnect - 2005-06-21
8,01 TFlop/s HPL - NEC - SX-8 - SX-8 - 2 GHz - 576x1 (MPI processes x threads) - IXS crossbar - 2005-06-21
5,14 TFlop/s HPL - SGI - Altix 3700 Bx2 - Intel Itanium 2 - 1,6 GHz - 1008x1 (MPI processes x threads) - NUMAlink - 2005-03-29
3,19 TFlop/s HPL - Cray - X1E - CrayX1E MSP - 1,13 GHz - 252x1 (MPI processes x threads) - Cray modified 2D torus - 2005-06-16
1,42 TFlop/s HPL - IBM - Blue Gene/L - IBM PowerPC 440 - 0,7 GHz - 1024x1 (MPI processes x threads) - Custom - 2005-04-11 (opt. BLAS!!!)
1,07 TFlop/s HPL - IBM - eServer pSeries 655 - IBM Power 4+ - 1,7 GHz - 64x4 (MPI processes x threads) - High Perf.Switch HPS - 2004-08-26
Slide 15 of 20 – Höchstleistungsrechenzentrum Stuttgart (HLRS) – Rolf Rabenseifner


FFTE – Fast Fourier Transform (measured only on a only a few systems)
absolute value and ...
... ratio to HPL
Global FFTE (absolute) GFlop/s
Global FFTE / Linpack HPL (ratio) GFlop/s / TFlop/s
Linpack (HPL) TFlop/s
2048 (base-run)
128 CPUs
5200 CPUs
1024 CPUs (optim.)
576 CPUs
32 MSPs
256 CPUs
252 MSPs
FFTE scales well on Blue Gene/L ☺
Cray X1 / MSP
NEC IXS crossbar / SX-8 2.0 GHz
NEC IXS crossbar / SX-6+ 0,5625 GHz
NEC IXS crossbar / SX-6 0.5 GHz
Cray X1E / MSP
Cray XT3 Red Storm / Opteron
Cray XD1 / Opteron
SGI NUMAlink
IBM High Perf. Switch / Power 4+
IBM Blue Gene/L (base & opt.BLAS !!!)
Sun Fire Link
InfiniBand / Xeon 2.4 GHz
InfiniBand / Xeon 3.4 GHz
Quadrics QsNet II / Opteron 2.2 GHz
Quadrics / Alpha 21264B/C 0.83-1GHz
Myrinet 2000 / Opteron, Itanium, Xeon
IBM Colony II / Power 4
IBM Colony / Power 4
Status Sep. 8, 2005
Measurements with smaller #CPUs on XT3 and SX-8 courtesy to Nathan Wichmann, Cray, CUG 2005, and Sunil Tiyyagura, HLRS.
Slide 16 of 20 – Höchstleistungsrechenzentrum Stuttgart (HLRS) – Rolf Rabenseifner


skipped
PTRANS – Matrix Transpose (measured only on a only a few systems)
absolute value and ...
... ratio to HPL
PTRANS (absolute) GByte/s
PTRANS / Linpack HPL (ratio) GByte/s / TFlop/s
Linpack (HPL) TFlop/s
5200 CPUs
576 CPUs
252 MSPs
252 MSPs
2048 (base)
1024 (opt.)
1008 CPUs
256
504 CPUs
256
Status Sep. 8, 2005
Cray X1 / MSP
NEC IXS crossbar / SX-8 2.0 GHz
NEC IXS crossbar / SX-6+ 0,5625 GHz
NEC IXS crossbar / SX-6 0.5 GHz
Cray X1E / MSP
Cray XT3 Red Storm / Opteron
Cray XD1 / Opteron
SGI NUMAlink
IBM High Perf. Switch / Power 4+
IBM Blue Gene/L (base & opt.BLAS !!!)
Sun Fire Link
InfiniBand / Xeon 2.4 GHz
InfiniBand / Xeon 3.4 GHz
Quadrics QsNet II / Opteron 2.2 GHz
Quadrics / Alpha 21264B/C 0.83-1GHz
Myrinet 2000 / Opteron, Itanium, Xeon
IBM Colony II / Power 4
IBM Colony / Power 4
Measurements with smaller #CPUs on XT3 and SX-8 courtesy to Nathan Wichmann, Cray, CUG 2005, and Sunil Tiyyagura, HLRS.
Slide 17 of 20 – Höchstleistungsrechenzentrum Stuttgart (HLRS) – Rolf Rabenseifner


Four different networks on the same system
HPC Challenge normalized by PP-HPL - Comparing the networks
Infiniband
Myrinet 2000
SCI
Gigabit Ethernet
Left most column: the platforms are sorted by the HPL value.
Other columns: The ratio to the HPL value is shown.
Each column itself is normalized with 1=best value or ratio.
PP-HPL
Normalized with PP-HPL=1
(EP*) DGEMM / PP-HPL
PP-FFTE / PP-HPL
PP-TRANS / PP-HPL
(EP*) StreamCopy / PP-HPL
(EP*) RandomRing Bandwidth / PP-HPL
1 / RandomRing Latency
PP-RandomAccess / PP-HPL
PP = total system performance; EP* = (embarrassingly parallel) value per process is multiplied by the number of processes
Dell - PowerEdge 2650 Cluster - Intel Xeon - 32 processors - 2,4 GHz - 32x1 (MPI processes x threads) - InfiniBand, 4x adapters, InfinIO 3000 switch - 2005-02-18
Dell - PowerEdge 2650 Cluster - Intel Xeon - 32 processors - 2,4 GHz - 32x1 (MPI processes x threads) - SCI, 4x4 2d Torus - 2005-02-18
Dell - PowerEdge 2650 Cluster - Intel Xeon - 32 processors - 2,4 GHz - 32x1 (MPI processes x threads) - Myrinet 2000 - 2005-02-18
Dell - PowerEdge 2650 Cluster - Intel Xeon - 32 processors - 2,4 GHz - 32x1 (MPI processes x threads) - Gigabit Ethernet, PowerConnect 5224 switch - 2005-02-18
Slide 18 of 20 – Höchstleistungsrechenzentrum Stuttgart (HLRS) – Rolf Rabenseifner

Acknowledgments

- Thanks to
  - all persons and institutions that have uploaded HPCC results.
  - Jack Dongarra and Piotr Luszczek
    for inviting me into the HPCC development team.
  - Matthias Müller, Sunil Tiyyagura and Holger Berger
    for benchmarking on the SX-8 and SX-6 and discussions on HPCC.
  - Nathan Wichmann from Cray for Cray XT3 and X1E data.
  - David Koester
    for his helpful remarks on the HPCC Kiviat diagrams.

Conclusions

- HPCC is an interesting basis for
  - benchmarking computational resources
  - analyzing the balance of a system
  - scaling with the number of processors
  - with respect to application needs
- HPCC helps to show the strength of
  - vector systems
  - cluster networks

See also HPCC award at SC'06
http://icl.cs.utk.edu/hpcc/
→ Award