ICCSA 2006

The 2006 International Conference on Computational Science and its Applications (ICCSA 2006)

An Overview of High-Performance Computing and Challenges for the Future

Jack Dongarra
University of Tennessee
and
Oak Ridge National Laboratory

Overview

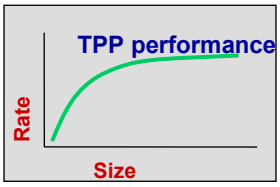
- Look at current state of high performance computing
  - Past, present and a look ahead
- Potential gains by exploiting lower precision devices
  - GPUs, Cell, SSE2, AltaVec
- New performance evaluation tools
  - HPCS – HPC Challenge

TOP500 supercomputer

H. Meuer, H. Simon, E. Strohmaier, & JD

- Listing of the 500 most powerful Computers in the World
- Yardstick: Rmax from LINPACK MPP

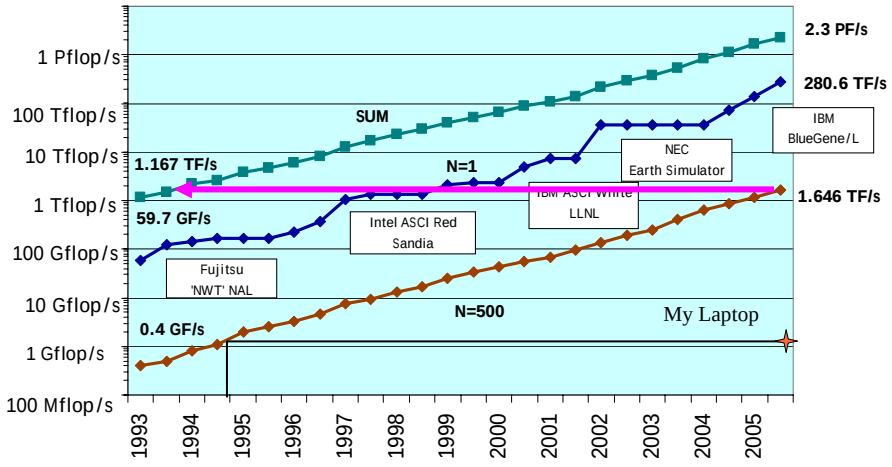
  $Ax=b$, *dense problem*



- Updated twice a year
  SC'xy in the States in November
  Meeting in Germany in June
- All data available from **www.top500.org**

3

Performance Development



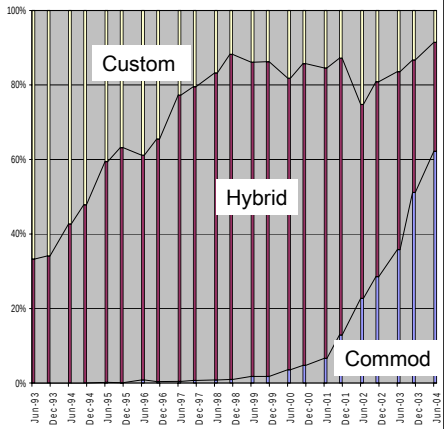
4

Architecture/Systems Continuum

Tightly Coupled

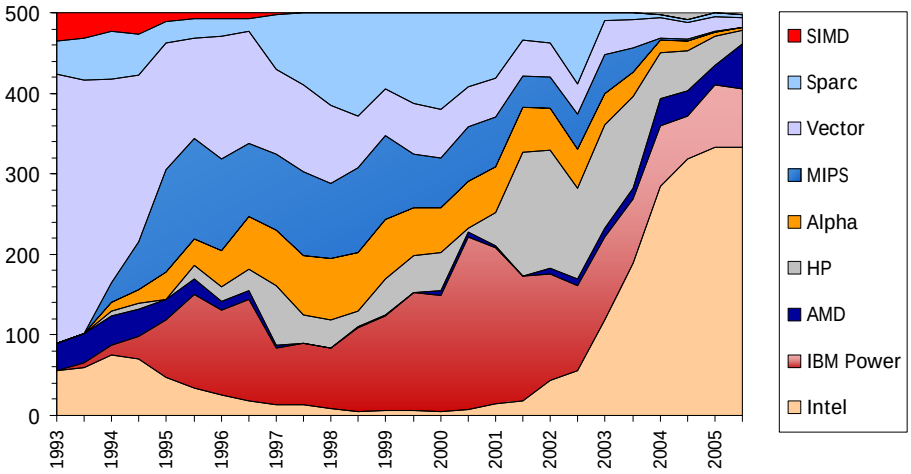
- **Custom processor with custom interconnect**
  - Cray X1
  - NEC SX-8
  - IBM Regatta
  - IBM Blue Gene/L
- **Commodity processor with custom interconnect**
  - SGI Altix
    - Intel Itanium 2
  - Cray XT3, XD1
    - AMD Opteron
- **Commodity processor with commodity interconnect**
  - Clusters
    - Pentium, Itanium, Opteron, Alpha
    - GigE, Infiniband, Myrinet, Quadrics
  - NEC TX7
  - IBM eServer
  - Dawning

Loosely Coupled



5

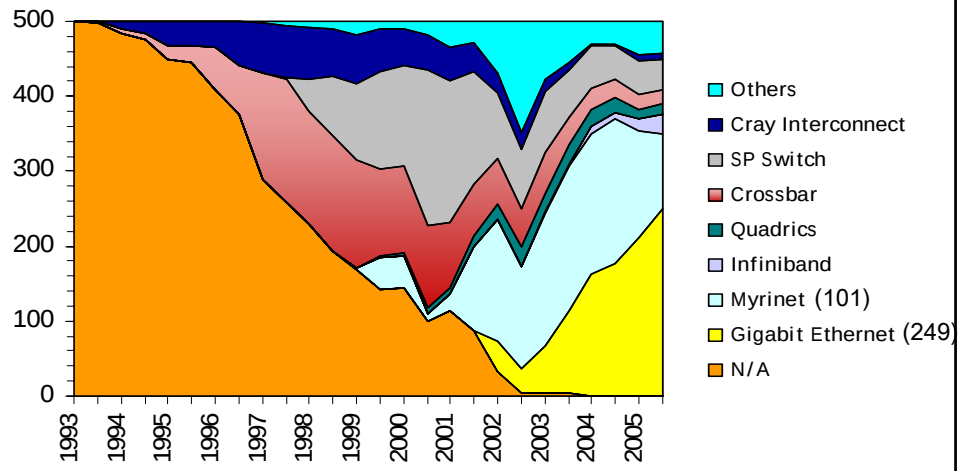
Processor Types



Intel + IBM Power PC + AMD = 91%

6

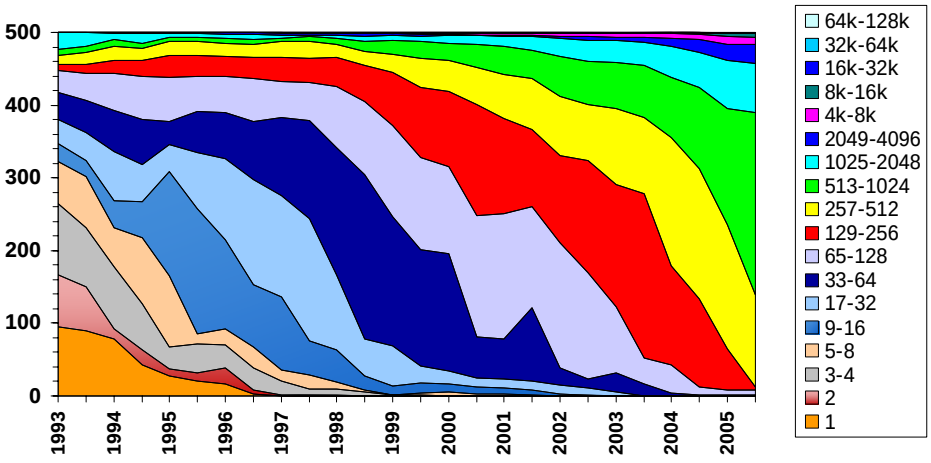
Interconnects / Systems



GigE + Myrinet = 70%

7

Parallelism in the Top500



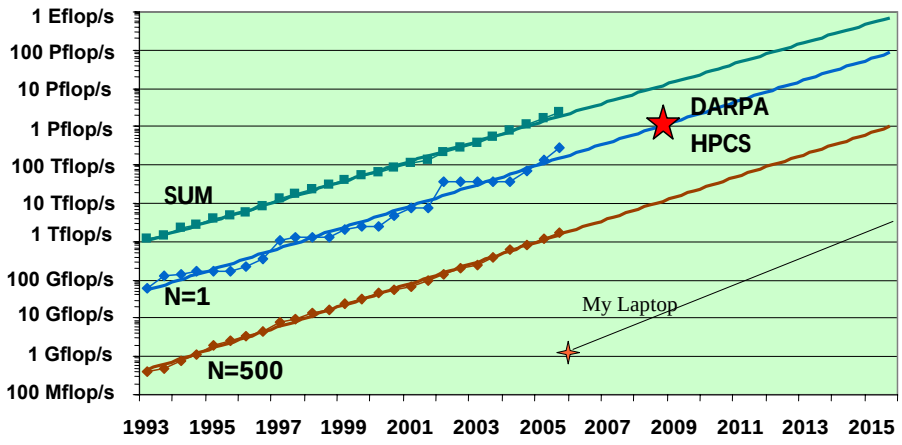
8

26th List: The TOP10

	Manufacturer	Computer	Rmax [TF/s]	Installation Site	Country	Year	#Proc
1	IBM	BlueGene/L eServer Blue Gene	280.6	DOE Lawrence Livermore Nat Lab	USA	2005 custom	131072
2	IBM	BGW eServer Blue Gene	91.29	IBM Thomas Watson Research	USA	2005 custom	40960
3	IBM	ASC Purple Power5 p575	63.39	DOE Lawrence Livermore Nat Lab	USA	2005 custom	10240
4 (3)	SGI	Columbia Altix, Itanium/Infiniband	51.87	NASA Ames	USA	2004 hybrid	10160
5	Dell	Thunderbird Pentium/Infiniband	38.27	DOE Sandia Nat Lab	USA	2005 commod	8000
6 (10)	Cray	Red Storm Cray XT3 AMD	36.19	DOE Sandia Nat Lab	USA	2005 hybrid	10880
7 (4)	NEC	Earth-Simulator SX-6	35.86	Earth Simulator Center	Japan	2002 custom	5120
8 (5)	IBM	MareNostrum PPC 970/Myrinet	27.91	Barcelona Supercomputer Center	Spain	2005 commod	4800
9 (6)	IBM	eServer Blue Gene	27.45	ASTRON University Groningen	Netherlands	2005 custom	12288
10	Cray	Jaguar Cray XT3 AMD	20.53	DOE Oak Ridge Nat Lab	USA	2005 hybrid	5200

9

Performance Projection



10

A PetaFlop Computer by the End of the Decade

- **10 Companies working on a building a Petaflop system by the end of the decade.**
  - Cray
  - IBM
  - Sun

  } HPCS

  - Dawning
  - Galactic
  - Lenovo

  } Chinese Companies

  - Hitachi
  - NEC
  - Fujitsu

  } Japanese "Life Simulator" (10 Pflop/s)

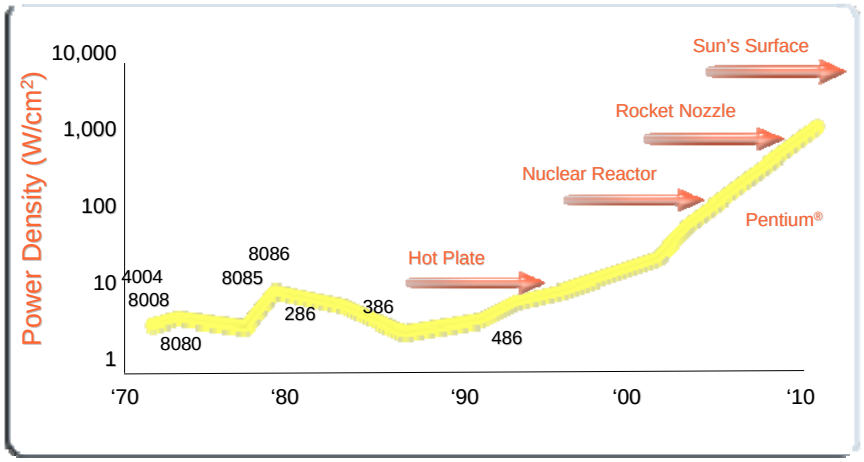
  - Bull

11

Today's CPU Architecture:

Heat becoming an unmanageable problem

Increasing the number of gates into a tight knot and decreasing the cycle time of the processor

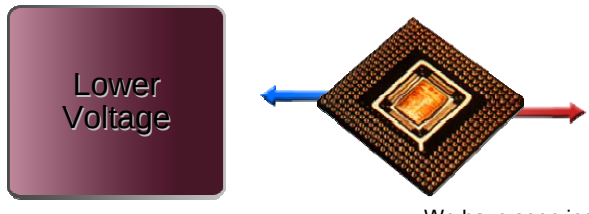


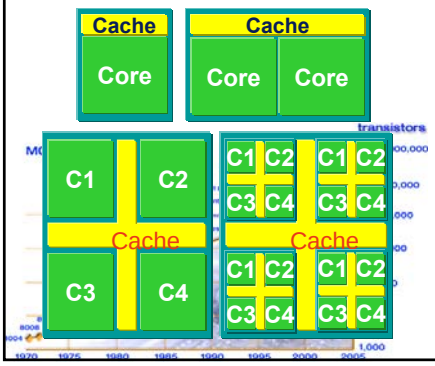
Intel Developer Forum, Spring 2004 - Pat Gelsinger (Pentium at 90 W)

Square relationship between the cycle time and power.

12

Increasing CPU Performance:
A Delicate Balancing Act





We have seen increasing number of gates on a chip and increasing clock speed.

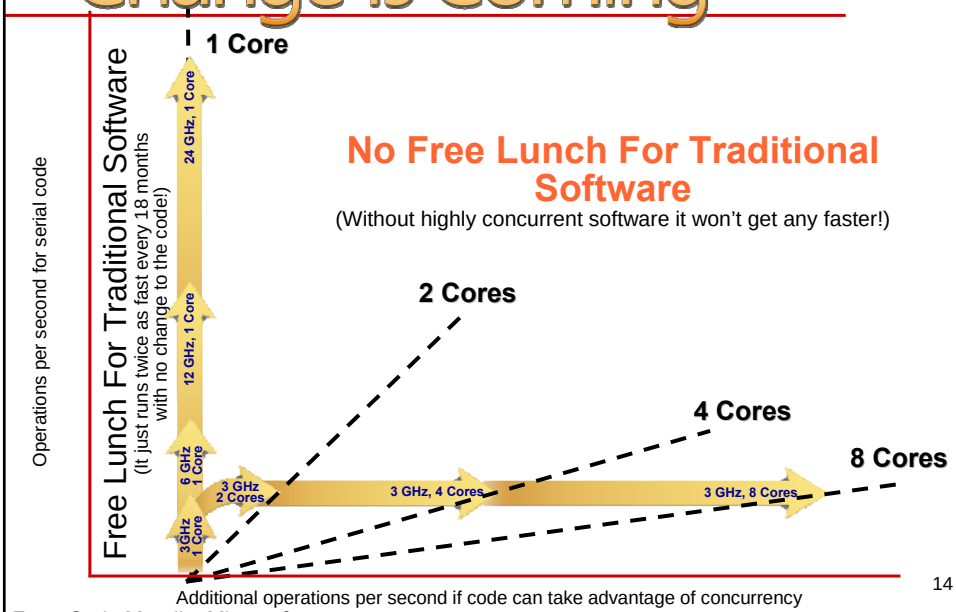
Heat becoming an unmanageable problem, Intel Processors > 100 Watts

We will not see the dramatic increases in clock speeds in the future.

However, the number of gates on a chip will continue to increase.

13

Change Is Coming

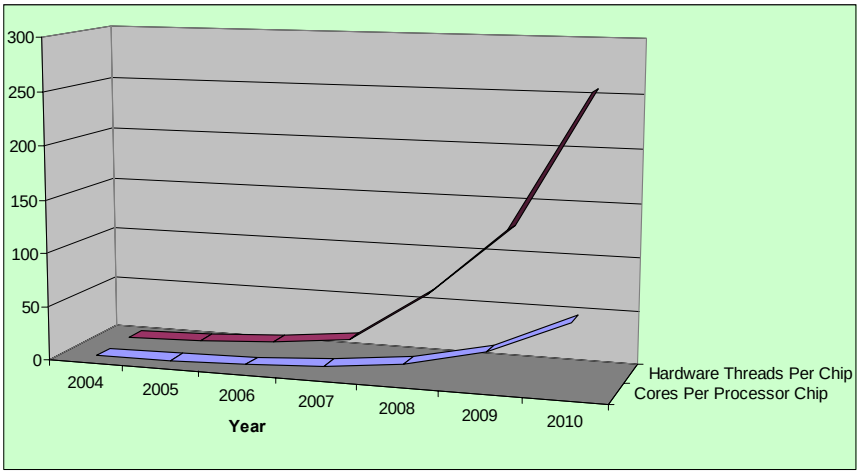


14

From Craig Mundie, Microsoft

CPU Desktop Trends – Change is Coming

- Relative processing power will continue to double every 18 months
- 256 logical processors per chip in late 2010



15

Commodity Processor Trends

Bandwidth/Latency is the Critical Issue, not FLOPS

Got Bandwidth?

	Annual increase	Typical value in 2006
Single-chip floating-point performance	59%	4 GFLOP/s
Front-side bus bandwidth	23%	1 GWord/s = 0.25 word/flop
DRAM latency	(5.5%)	70 ns = 280 FP ops = 70 loads

Source: *Getting Up to Speed: The Future of Supercomputing*, National Research Council, 222 pages, 2004, National Academies Press, Washington DC, ISBN 0-309-09502-6.

16

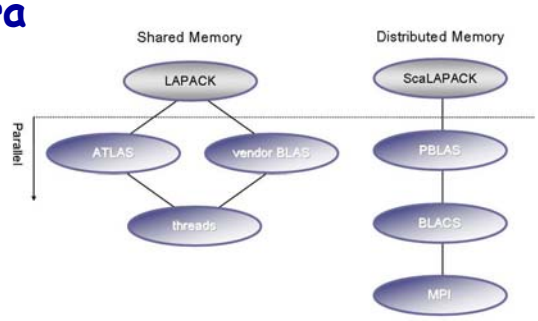
That Was the Good News

- Bad news: the effect of the hardware change on the existing software base
- Must rethink the design of our software
  - Another disruptive technology
  - Rethink and rewrite the applications, algorithms, and software

17

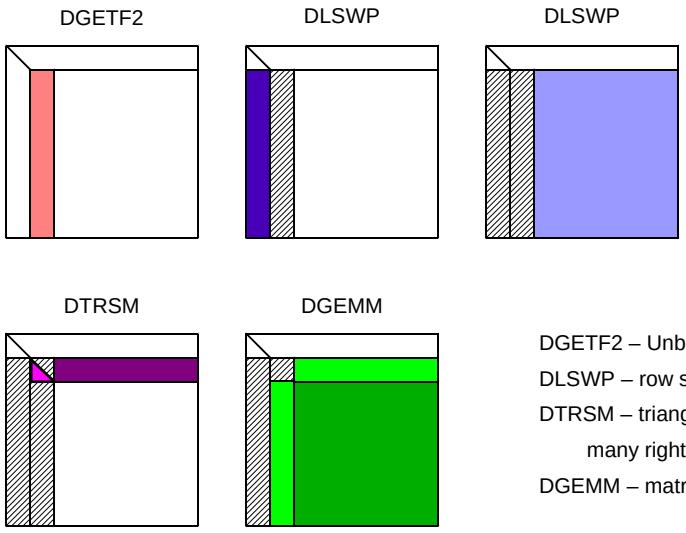
LAPACK - ScaLAPACK

- Numerical libraries for linear algebra
- LAPACK
  - Late 1980's
  - Sequential and SMPs
- ScaLAPACK
  - Early 1990's
  - Message passing systems



18

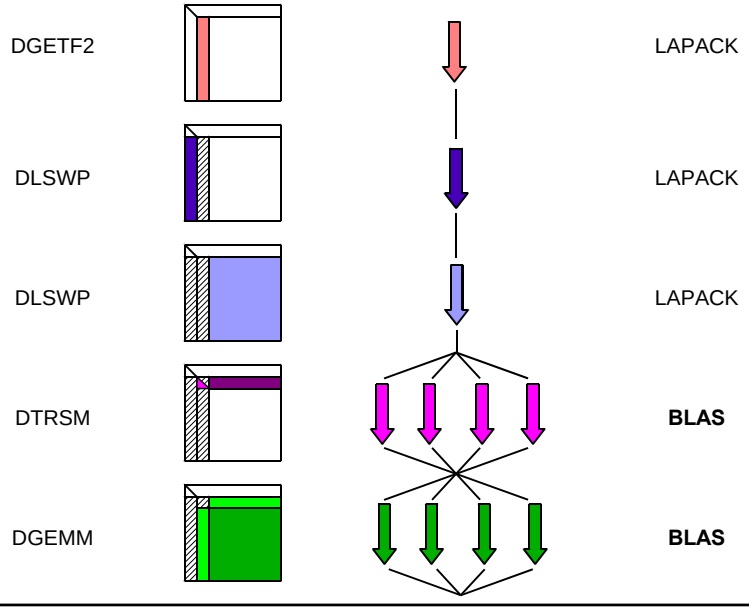
Right-Looking LU factorization (LAPACK)



DGETF2 – Unblocked LU
DLSWP – row swaps
DTRSM – triangular solve with many right-hand sides
DGEMM – matrix-matrix multiply

19

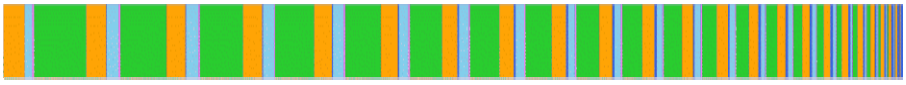
Steps in the LAPACK LU



20

LU Timing Profile

LAPACK + BLAS threads



Time for each component

1D decomposition and SGI Origin

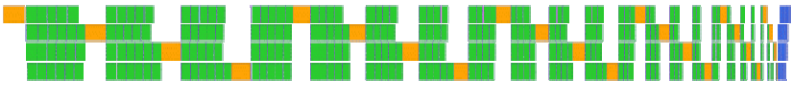
DGETF2
DLASWP(L)
DLASWP(R)
DTRSM
DGEMM

LU Timing Profile

LAPACK + BLAS threads

Time for each component

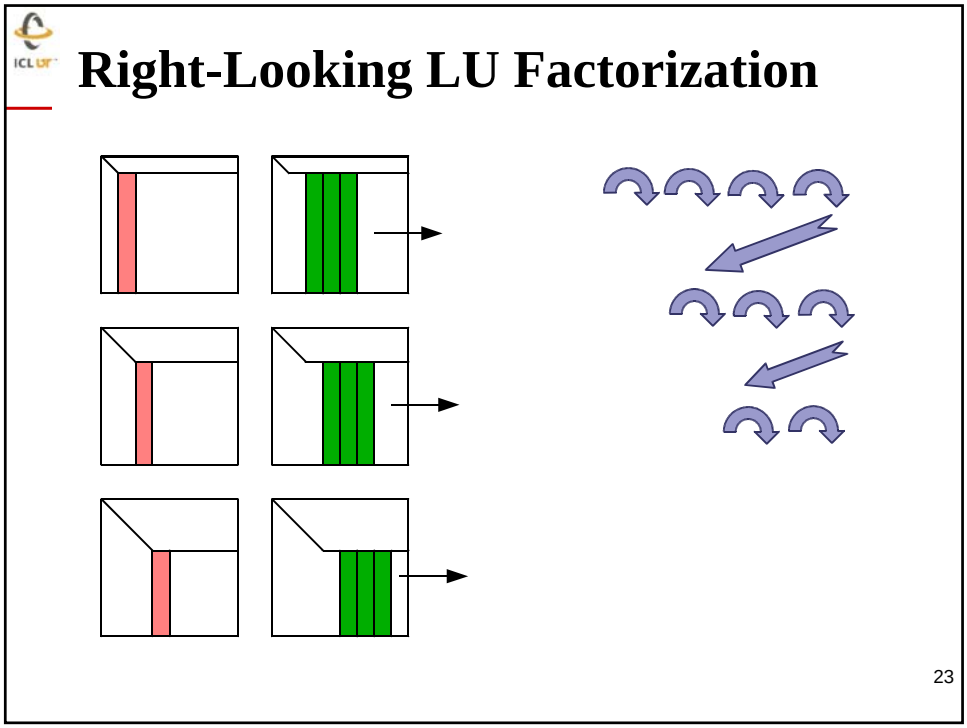
Threads – no lookahead

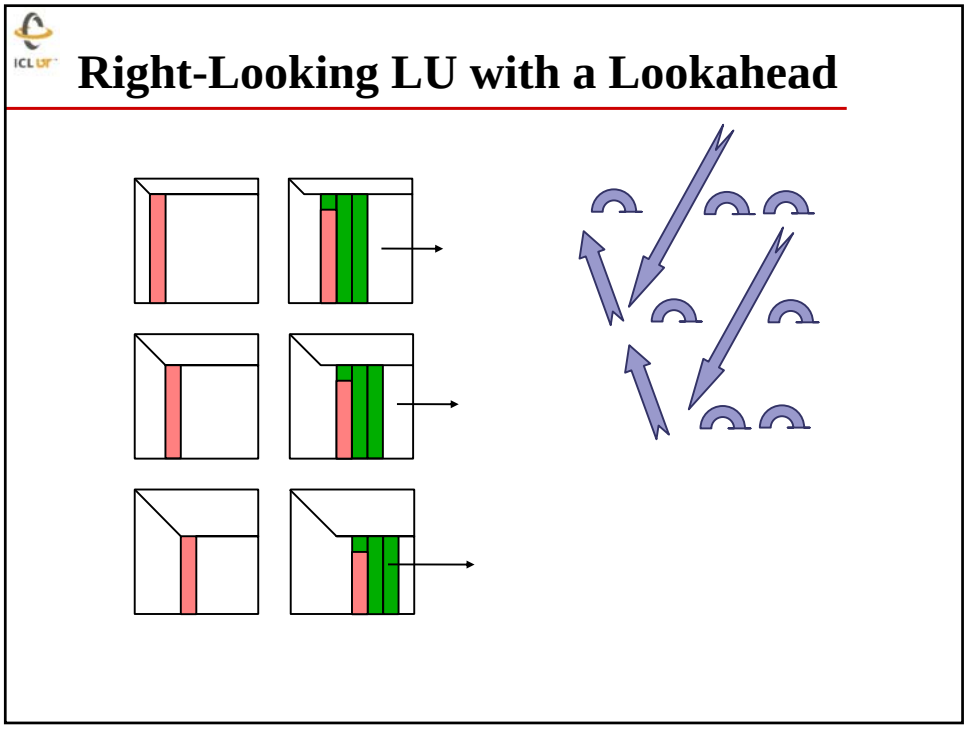


In this case the performance difference comes from parallelizing row exchanges (DLASWP) and threads in the LU algorithm.

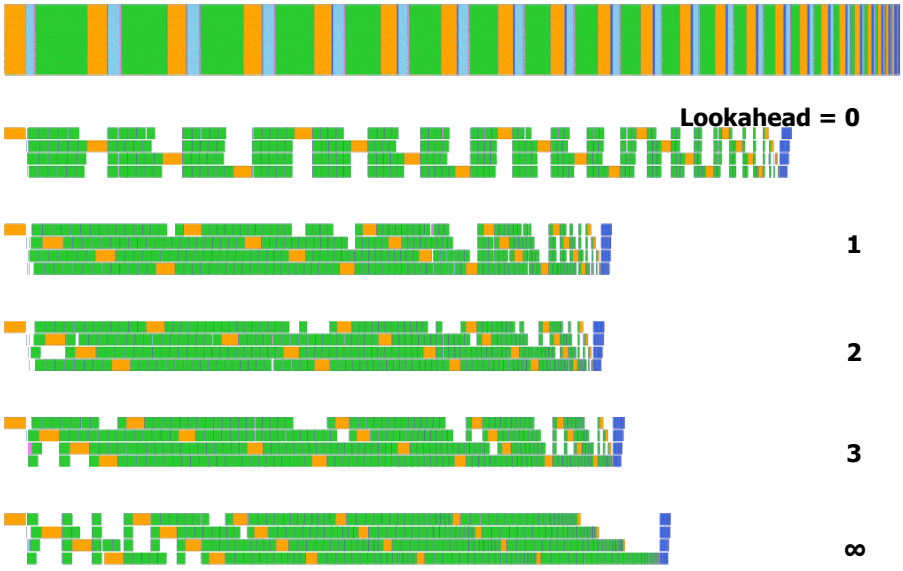
1D decomposition and SGI Origin

DGETF2
DLASWP(L)
DLASWP(R)
DTRSM
DGEMM


Right-Looking LU Factorization
23

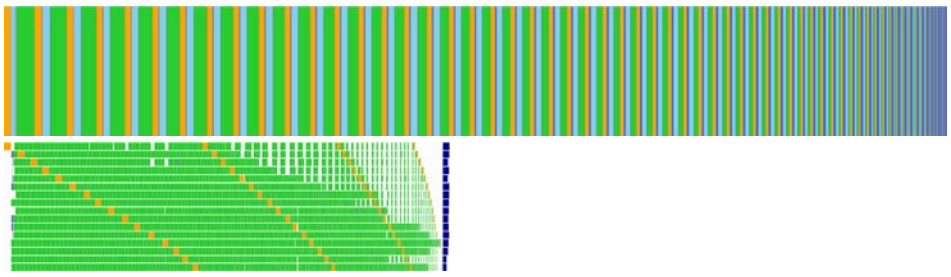

Right-Looking LU with a Lookahead

Pivot Rearrangement and Lookahead
4 Processor runs



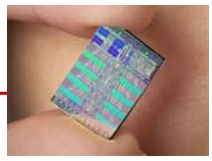
25

Pivot Rearrangement and Lookahead
16 SMP runs

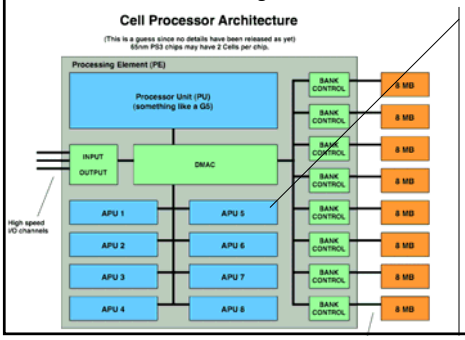


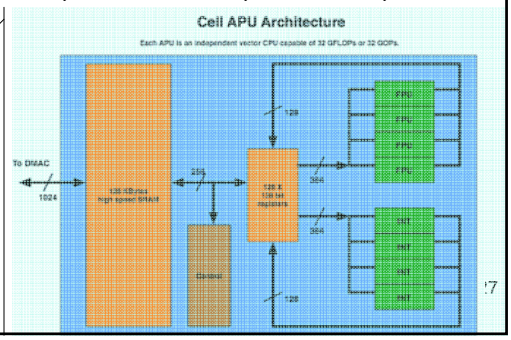
26

Motivated by…



- The PlayStation 3's CPU based on a chip codenamed "Cell"
- Each Cell contains 8 APUs.
  - An APU is a self contained vector processor which acts independently from the others.
  - 4 floating point units capable of a total of 32 Gflop/s (8 Gflop/s each)
  - 256 Gflop/s peak! 32 bit floating point; 64 bit floating point at 25 Gflop/s.
  - IEEE format, but only rounds toward zero in 32 bit, overflow set to largest
    - According to IBM, the SPE's double precision unit is fully IEEE854 compliant.





17

GPU Performance

GPU Vendor	NVIDIA 	NVIDIA 	ATI 
Model	6800Ultra	7800GTX	X1900XTX
Release Year	2004	2005	2006
32-bit Performance	60 GFLOPS	200 GFLOPS	400 GFLOPS
64-bit Performance	must be emulated in software		

28

Idea Something Like This…

- Exploit 32 bit floating point as much as possible.
  - Especially for the bulk of the computation
- Correct or update the solution with selective use of 64 bit floating point to provide a refined results
- Intuitively:
  - Compute a 32 bit result,
  - Calculate a correction to 32 bit result using selected higher precision and,
  - Perform the update of the 32 bit results with the correction using high precision.

29

32 and 64 Bit Floating Point Arithmetic

- Iterative refinement for dense systems can work this way.

Solve $Ax = b$ in lower precision, save the factorization ($L*U = A*P$); $O(n^3)$

Compute in higher precision $r = b - A*x$; $O(n^2)$

Requires the original data A (stored in high precision)

Solve $Az = r$; using the lower precision factorization; $O(n^2)$

Update solution $x_+ = x + z$ using high precision; $O(n)$

Iterate until converged.

- Wilkinson, Moler, Stewart, & Higham provide error bound for SP fl pt results when using DP fl pt.
- We can show using this approach that we can compute the solution to 64-bit floating point precision.

Requires extra storage, total is 1.5 times normal;
$O(n^3)$ work is done in lower precision
$O(n^2)$ work is done in high precision

Problems if the matrix is ill-conditioned in sp; $O(10^8)$

30

On the Way to Understanding How to Use the Cell Something Else Happened …

- Realized have the similar situation on our commodity processors.
  - That is, SP is 2X as fast as DP on many systems
- The Intel Pentium and AMD Opteron have SSE2
  - 2 flops/cycle DP
  - 4 flops/cycle SP
- IBM PowerPC has AltiVec
  - 8 flops/cycle SP
  - 4 flops/cycle DP
    - No DP on AltiVec

Processor and BLAS Library	SGEMM (GFlop/s)	DGEMM (GFlop/s)	Speedup SP/DP
Pentium III Katmai (0.6GHz) Goto BLAS	0.98	0.46	2.13
Pentium III CopperMine (0.9GHz) Goto BLAS	1.59	0.79	2.01
Pentium Xeon Northwood (2.4GHz) Goto BLAS	7.68	3.88	1.98
Pentium Xeon Prescott (3.2GHz) Goto BLAS	10.54	5.15	2.05
Pentium IV Prescott (3.4GHz) Goto BLAS	11.09	5.61	1.98
AMD Opteron 240 (1.4GHz) Goto BLAS	4.89	2.48	1.97
PowerPC G5 (2.7GHz) AltiVec	18.28	9.98	1.83

31

Performance of single precision and double precision matrix multiply (SGEMM and DGEMM) with n=m=k=1000

Speedups (Ratio of Times)

Architecture (BLAS)	*n*	DGEMM /SGEMM	DP Solve /SP Solve	DP Solve /Iter Ref	# iter
Intel Pentium IV-M Northwood (Goto)	4000	2.02	1.98	1.54	5
Intel Pentium III Katmai (Goto)	3000	2.12	2.11	1.79	4
Intel Pentium III Coppermine (Goto)	3500	2.10	2.24	1.92	4
Intel Pentium IV Prescott (Goto)	4000	2.00	1.86	1.57	5
AMD Opteron (Goto)	4000	1.98	1.93	1.53	5
Sun UltraSPARC IIe (Sunperf)	3000	1.45	1.79	1.58	4
IBM Power PC G5 (2.7 GHz) (VecLib)	5000	2.29	2.05	1.24	5
Cray X1 (libsci)	4000	1.68	1.57	1.32	7
Compaq Alpha EV6 (CXML)	3000	0.99	1.08	1.01	4
IBM SP Power3 (ESSL)	3000	1.03	1.13	1.00	3
SGI Octane (ATLAS)	2000	1.08	1.13	0.91	4

Architecture (BLAS-MPI)	# procs	*n*	DP Solve /SP Solve	DP Solve /Iter Ref	# iter
AMD Opteron (Goto – OpenMPI MX)	32	22627	1.85	1.79	6
AMD Opteron (Goto – OpenMPI MX)	64	32000	1.90	1.83	6

32

Refinement Technique Using Single/Double Precision

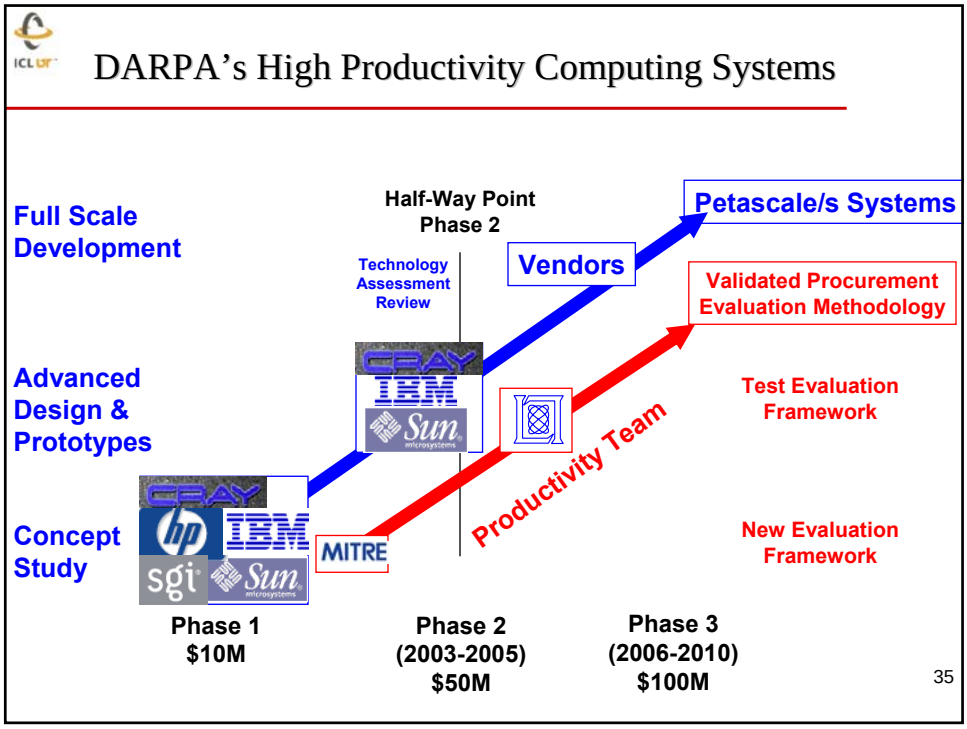
- Linear Systems
  - LU (dense and sparse)
  - Cholesky
  - QR Factorization
- Eigenvalue
  - Symmetric eigenvalue problem
  - SVD
  - Same idea as with dense systems,
    - Reduce to tridiagonal/bi-diagonal in lower precision, retain original data and improve with iterative technique using the lower precision to solve systems and use higher precision to calculate residual with original data.
    - $O(n^2)$ per value/vector
- Iterative Linear System
  - Relaxed GMRES
  - Inner/outer scheme

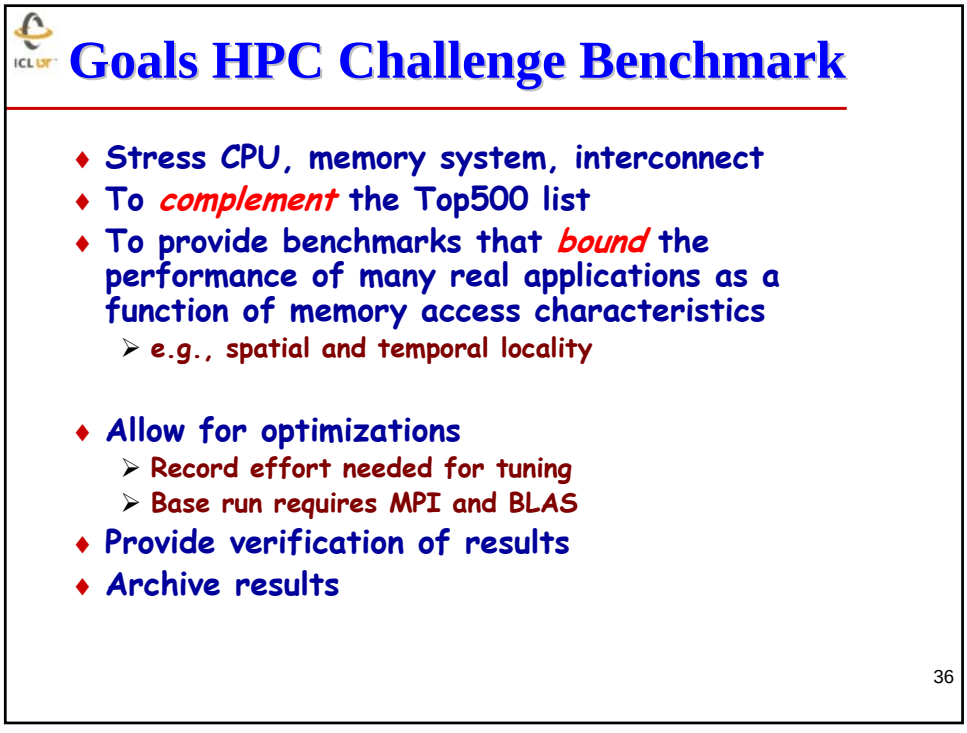
Motivation for Additional Benchmarks

Linpack Benchmark

- Good
  - One number
  - Simple to define & easy to rank
  - Allows problem size to change with machine and over time
- Bad
  - Emphasizes only "peak" CPU speed and number of CPUs
  - Does not stress local bandwidth
  - Does not stress the network
  - Does not test gather/scatter
  - Ignores Amdahl's Law (Only does weak scaling)
  - …
- Ugly
  - Benchmarketeering hype

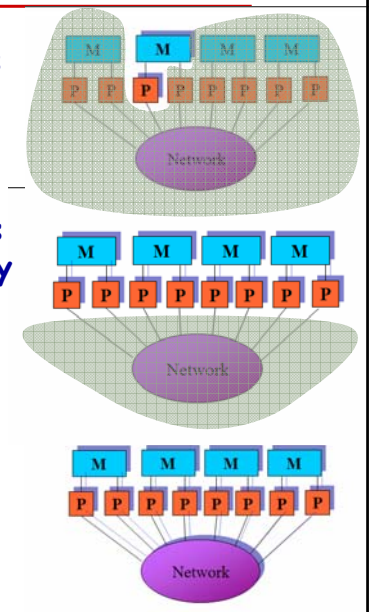
- From Linpack Benchmark and Top500: "no single number can reflect overall performance"
- Clearly need something more than Linpack
- HPC Challenge Benchmark
  - Test suite stresses not only the processors, but the memory system and the interconnect.
  - The real utility of the HPCC benchmarks are that architectures can be described with a wider range of metrics than just Flop/s from Linpack.


DARPA's High Productivity Computing Systems
Full Scale Development
Advanced Design & Prototypes
Concept Study
Half-Way Point Phase 2
Technology Assessment Review
Vendors
Petascale/s Systems
Validated Procurement Evaluation Methodology
Test Evaluation Framework
New Evaluation Framework
Productivity Team
MITRE
Phase 1 $10M
Phase 2 (2003-2005) $50M
Phase 3 (2006-2010) $100M
35


Goals HPC Challenge Benchmark
♦ Stress CPU, memory system, interconnect
♦ To complement the Top500 list
♦ To provide benchmarks that bound the performance of many real applications as a function of memory access characteristics
➢ e.g., spatial and temporal locality
♦ Allow for optimizations
➢ Record effort needed for tuning
➢ Base run requires MPI and BLAS
♦ Provide verification of results
♦ Archive results
36

Tests on Single Processor and System

- Local - only a single processor is performing computations.
- Embarrassingly Parallel - each processor in the entire system is performing computations but they do no communicate with each other explicitly.
- Global - all processors in the system are performing computations and they explicitly communicate with each other.



HPC Challenge Benchmark

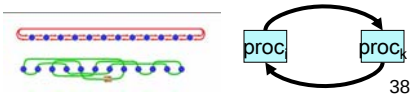
Consists of basically 7 benchmarks;

- Think of it as a framework or harness for adding benchmarks of interest.

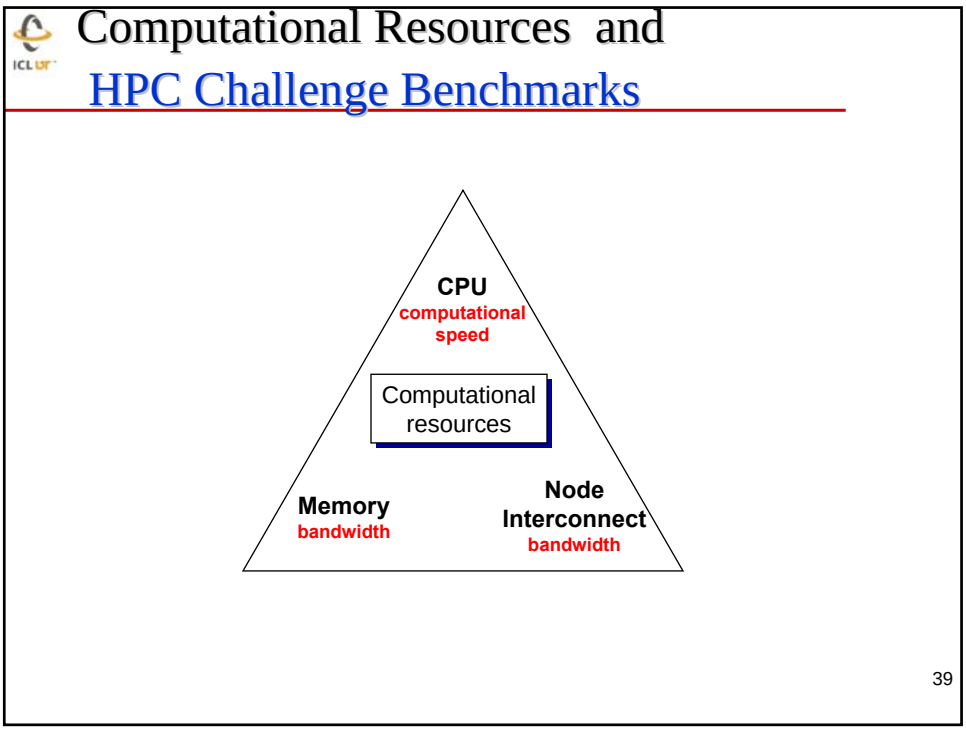
1. HPL (LINPACK) — MPI Global (Ax = b)
2. STREAM — Local; single CPU
   *STREAM — Embarrassingly parallel
3. PTRANS (A ← A + B^T) — MPI Global
4. RandomAccess — Local; single CPU
   *RandomAccess — Embarrassingly parallel
   RandomAccess — MPI Global
5. BW and Latency – MPI
6. FFT - Global, single CPU, and EP
7. Matrix Multiply – single CPU and EP

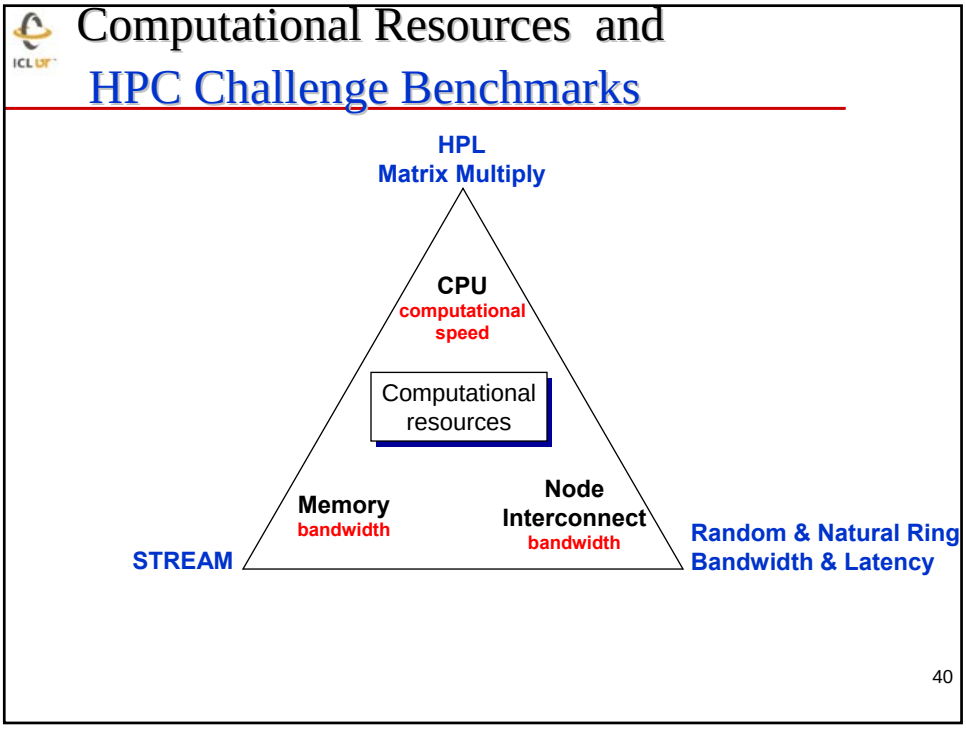
name	kernel	bytes/iter	FLOPS/iter
COPY:	a(i) = b(i)	16	0
SCALE:	a(i) = q*b(i)	16	1
SUM:	a(i) = b(i) + c(i)	24	1
TRIAD:	a(i) = b(i) + q*c(i)	24	2

Random integer read; update; & write

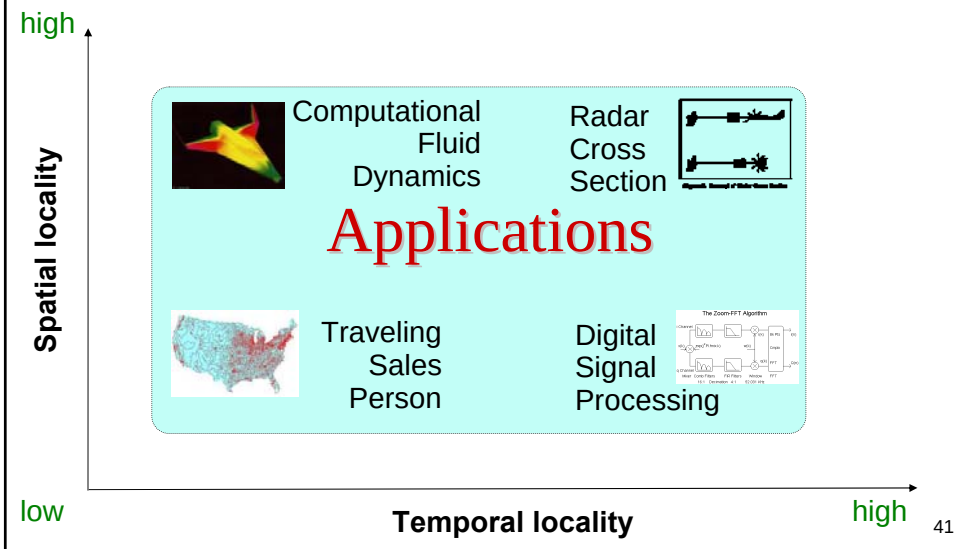


38


Computational Resources and
HPC Challenge Benchmarks
CPU
computational
speed
Computational
resources
Memory
bandwidth
Node
Interconnect
bandwidth
39

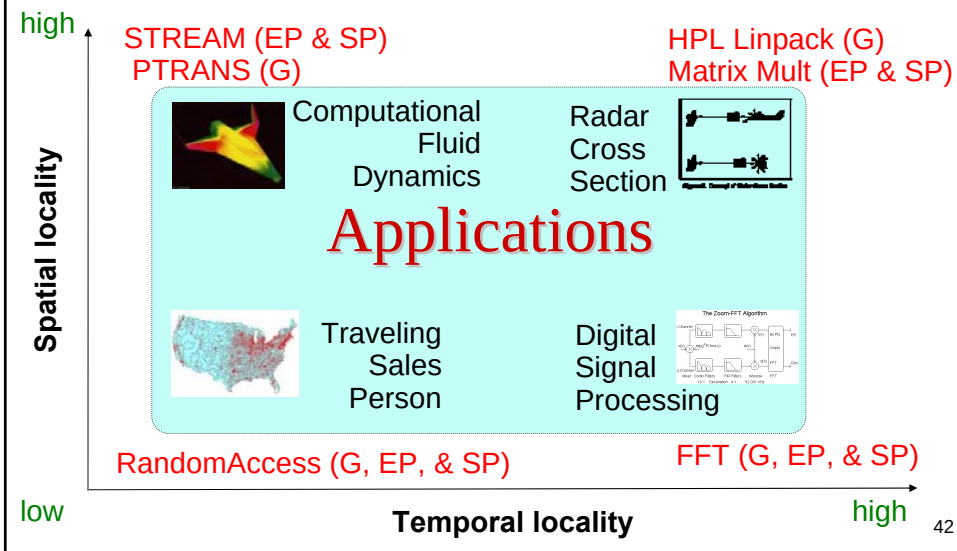

Computational Resources and
HPC Challenge Benchmarks
HPL
Matrix Multiply
CPU
computational
speed
Computational
resources
Memory
bandwidth
Node
Interconnect
bandwidth
STREAM
Random & Natural Ring
Bandwidth & Latency
40

Memory Access Patterns



41

Memory Access Patterns



42

http://icl.cs.utk.edu/hpcc/ web

HPC Challenge

Home
Rules
News
Download
FAQ
Links
Collaborators
Sponsors
Upload
Kiviat Diagram
Results

HPC Challenge Benchmark

The HPC Challenge benchmark consists of basically 7 benchmarks:

1. HPL - the Linpack TPP benchmark which measures the floating point rate of execution for solving a linear system of equations.
2. DGEMM - measures the floating point rate of execution of double precision real matrix-matrix multiplication.
3. STREAM - a simple synthetic benchmark program that measures sustainable memory bandwidth (in GB/s) and the corresponding computation rate for simple vector kernel.
4. PTRANS (parallel matrix transpose) - exercises the communications where pairs of processors communicate with each other simultaneously. It is a useful test of the total communications capacity of the network.
5. RandomAccess - measures the rate of integer random updates of memory (GUPS).
6. FFTE - measures the floating point rate of execution of double precision complex one-dimensional Discrete Fourier Transform (DFT).
7. Communication bandwidth and latency - a set of tests to measure latency and bandwidth of a number of simultaneous communication patterns; based on b_eff (effective bandwidth benchmark).

43

HPCCHALLENGE

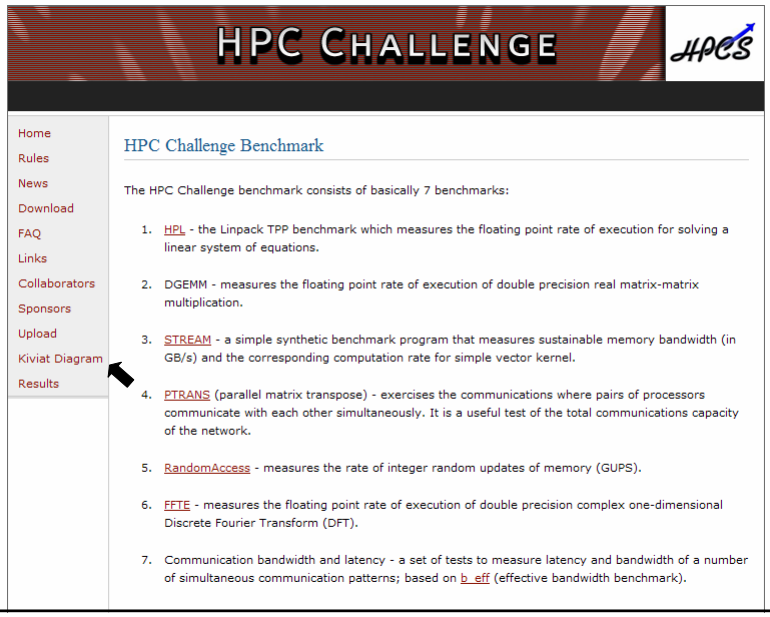
Home | Rules | News | Download | FAQ | Links | Collaborators | Sponsors | Upload | Results

Condensed Results - Base Runs Only - 102 Systems - Generated on Sun Apr 30 08:15:43 2006

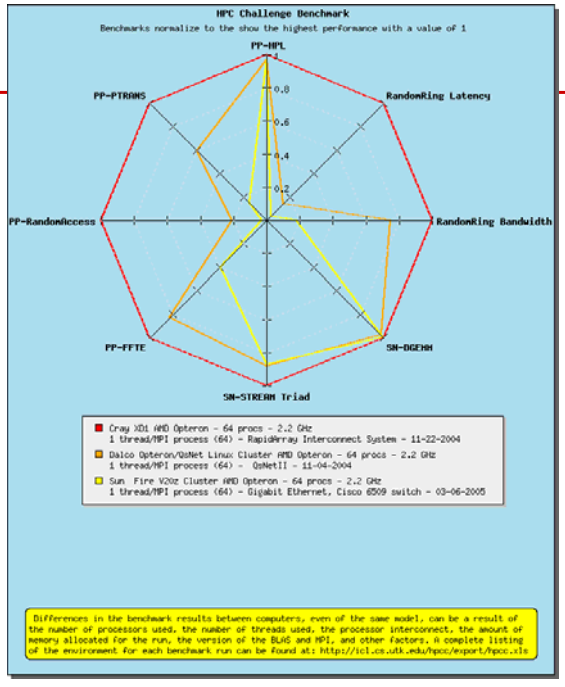
System Information System - Processor - Speed - Count - Threads - Processes					G-HPL	G-PTRANS	G-Random Access	G-FFTE	EP-STREAM Sys	EP-STREAM Triad	EP-DGEMM	RandomRing Bandwidth	RandomRing Latency
MA/PT/PS/PC/TH/PR/CM/CS/IC/IA/SD					TFlop/s	GB/s	Gup/s	GFlop/s	GB/s	GB/s	GFlop/s	GB/s	usec
Abiqa Conquest cluster AMD Opteron	1.4GHz	128	1	128	0.2526110	3.2471			208.525	1.6291		0.03627	23.68
Clustervision BV Beastie AMD Opteron	2.4GHz	32	1	32	0.1037640	0.8159	0.0002350	2.1470	106.951	3.3422	4.19493	0.02648	53.23
Cray X1 MSP	0.8GHz	64	1	64	0.5215600	3.2288			959.334	14.9896		0.94074	20.34
Cray X1 MSP	0.8GHz	60	1	60	0.5777790	30.4315			898.446	14.9741		1.03291	20.83
Cray X1 MSP	0.8GHz	120	1	120	1.0609700	2.4603			1019.519	8.4960		0.83014	20.12
Cray T3E Alpha 21164	0.6GHz	1024	1	1024	0.0481695	10.2765			529.242	0.5168		0.03174	12.09
Cray X1 MSP	0.8GHz	252	1	252	2.3847300	97.4076			3738.404	14.9143		0.42899	22.27
Cray X1 MSP	0.8GHz	124	1	124	1.2054200	39.5252			1856.664	14.9731		0.70857	20.15
Cray X1 MSP	0.8GHz	60	1	60	0.5087430	1.6342	0.0030750	3.1444	894.114	14.9019	10.91520	1.16779	14.66
Cray T3E Alpha 21164	0.675GHz	512	1	512	0.2221810	9.7741	0.0289464	15.4774	272.186	0.5316	0.66077	0.03571	8.14
Cray XD1 AMD Opteron	2.2GHz	64	1	64	0.2238980	10.5924	0.0223966	16.3611	169.955	2.6555	4.03375	0.22697	1.63
Cray X1 MSP	0.8GHz	32	1	32	0.2767140	32.6606	0.0016620	2.9649	473.846	14.8702	8.25848	1.41269	14.94
Cray XT3 AMD Opteron	2.6GHz	1100	1	1100	4.7823400	217.9230	0.1370020	266.6600	5274.696	4.7952	4.81050	0.28638	25.94
Cray XD1 AMD Opteron	2.4GHz	128	1	128	0.5020760	13.5135	0.0666722	35.3172	500.063	3.9068	4.22433	0.23919	2.06
Cray X1E X1E MSP	1.13GHz	252	1	252	2.1940900	85.2040	0.0140684	15.5352	2429.983	9.6825	14.18470	0.36024	14.92
Cray XT3 AMD Opteron	2.4GHz	3744	1	3744	14.7040000	608.5060	0.2202960	417.1720	18146.382	4.8468	4.41330	0.16164	25.32

System Information System - Processor - Speed - Count - Threads - Processes					G-HPL	G-PTRANS	G-Random Access	G-FFTE	EP-STREAM Sys	EP-STREAM Triad	EP-DGEMM	RandomRing Bandwidth	RandomRing Latency
MA/PT/PS/PC/TH/PR/CM/CS/IC/IA/SD					TFlop/s	GB/s	Gup/s	GFlop/s	GB/s	GB/s	GFlop/s	GB/s	usec
Cray XT3 AMD Opteron	2.4GHz	5200	1	5200	20.5270000	874.8990	0.2685830	644.7300	26020.800	5.0040	4.39335	0.14682	25.80
Cray xt3 AMD Opteron	2.4GHz	32	1	32	0.1387810	7.3764	0.0606017	9.3683	156.424	4.8883	4.77641	0.57281	8.74
Cray X1E	1.13GHz	32	4	32	0.3376360	18.9199	0.0089684	5.2027	307.565	9.6114	11.60560	1.40487	13.21
Cray XT3 AMD Opteron	2.6GHz	4096	1	4096	16.9752000	302.9790	0.5330720	905.5690	20656.456	5.0431	4.79166	0.16896	9.44
Cray XT3 AMD Opteron	2.6GHz	1100	1	1100	4.7276400	253.3460	0.3035680	328.2860	5161.134	4.6919	4.77440	0.39964	7.29
Cray Inc XT3 AMD Opteron	2.4GHz	5200	1	5200	20.4086000	944.2270	0.6724120	761.7290	24268.447	4.6598	4.41173	0.20636	9.20
Cray Inc XT3 AMD Opteron	2GHz	10350	1	10350	32.9065000	1813.0600	1.0176500	1118.2900	43581.780	4.2108	3.66719	0.16188	10.32
Cray Inc. X1 Cray E	1.13GHz	1008	1	1008	12.0263000	108.0190	0.0851199	82.3884	13522.091	13.3989	14.50000	0.13667	16.30
Cray Inc. XT3 AMD Opteron	2.6GHz	4128	1	4128	16.6421000	674.7860	0.6767580	821.6770	19295.676	4.6742	4.75946	0.22245	8.23
Dalco Opteron/Qaltel Linux Cluster AMD Opteron	2.2GHz	64	1	64	0.2180410	6.3195	0.0047003	13.5481	153.394	2.3968	3.87865	0.17003	11.46
Dalco Gonzales AMD Opteron	2.4GHz	64	1	64	0.2574150	9.2555	0.0399258	14.0039	224.767	3.5120	4.55878	0.17353	4.89

http://icl.cs.utk.edu/hpcc/ web



45



46

Summary of Current Unmet Needs

- Performance / Portability
- Fault tolerance
- Memory bandwidth/Latency
- Adaptability: Some degree of autonomy to self optimize, test, or monitor.
  - Able to change mode of operation: static or dynamic
- Better programming models
  - Global shared address space
  - Visible locality
- Maybe coming soon (incremental, yet offering real benefits):
  - Global Address Space (GAS) languages: UPC, Co-Array Fortran, Titanium, Chapel, X10, Fortress)
    - "Minor" extensions to existing languages
    - More convenient than MPI
    - Have performance transparency via explicit remote memory references
- What's needed is a long-term, balanced investment in hardware, software, algorithms and applications in the HPC Ecosystem.

47

Real Crisis With HPC Is With The Software

- Our ability to configure a hardware system capable of 1 PetaFlop (10^{15} ops/s) is without question just a matter of time and $$.
- A supercomputer application and software are usually much more long-lived than a hardware
  - Hardware life typically five years at most…. Apps 20-30 years
  - Fortran and C are the main programming models (still!!)
- The REAL CHALLENGE is Software
  - Programming hasn't changed since the 70's
  - HUGE manpower investment
    - MPI… is that all there is?
  - Often requires HERO programming
  - Investments in the entire software stack is required (OS, libs, etc.)
- Software is a major cost component of modern technologies.
  - The tradition in HPC system procurement is to assume that the software is free… SOFTWARE COSTS (over and over)
- What's needed is a long-term, balanced investment in the HPC Ecosystem: hardware, software, algorithms and applications.

48

Collaborators / Support

- Top500 Team
  - Erich Strohmaier, NERSC
  - Hans Meuer, Mannheim
  - Horst Simon, NERSC
- Sca/LAPACK
  - Julien Langou
  - Jakub Kurzak
  - Piotr Luszczek
  - Stan Tomov
  - Julie Langou