IBM Research Division

What will be IBM's HPC successes by 2009 ?

Ruud A. Haring

March 8, 2006

IBM's HPC successes by 2009

- **Disclaimers:**
  - Present speaker will not do IBM product pre-announcements.
  - Present speaker spends most of his time on chip design and chip design project management.

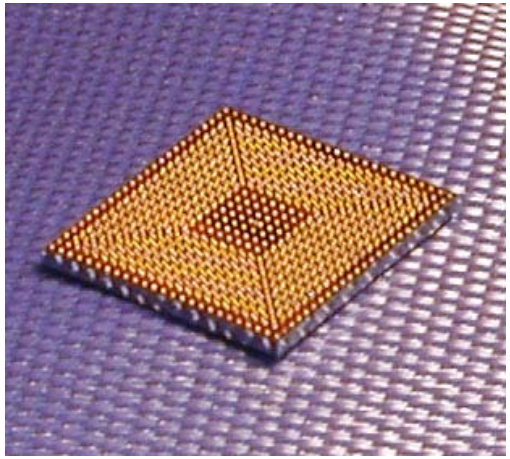
  Hence sees pessimism and paranoia as a virtue.

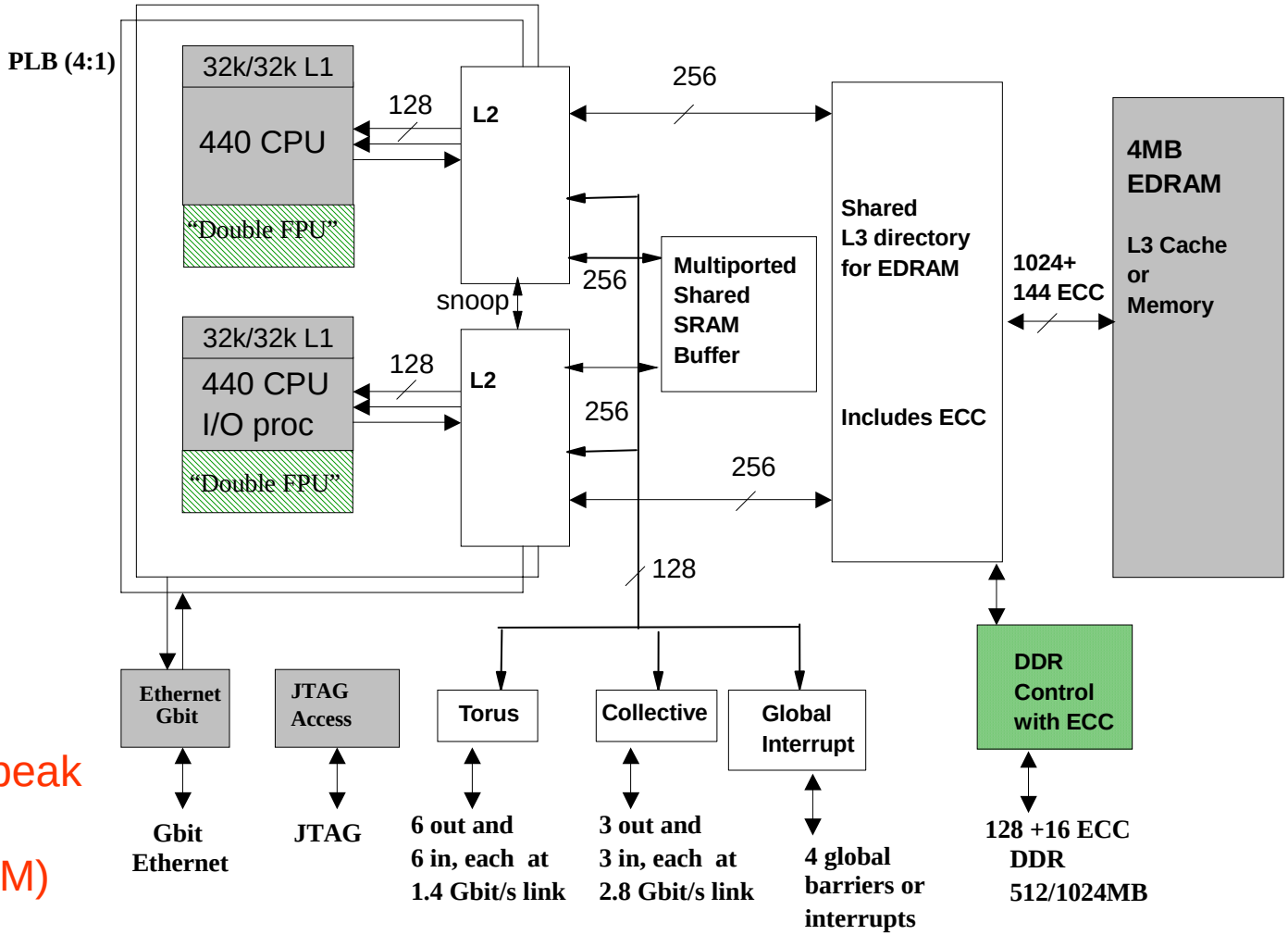
Hypothesis 1: With HPC users' relentless appetite for compute power, one could expect 1 PFLOPS peak systems ~ 2009

(What is your company's technology to achieve this?)

- IBM's Blue Gene/L installation at Lawrence Livermore National Laboratories (LLNL) achieves 367 TFlop/s **peak**
- Thus we have to bridge a factor 2.7 to get to 1 PFlop/s peak
- We are confident that a BlueGene/L follow-up machine will do that …

BlueGene/L Compute ASIC: system-on-a-chip
integrates processors, memory sub-system and IO subsystems





- IBM CU-11, 0.13 µm
- 11 x 11 mm die size
- 700 MHz / 5.6 GFlop/s peak
- 12 Watt
- (17W with 512 MB DRAM)

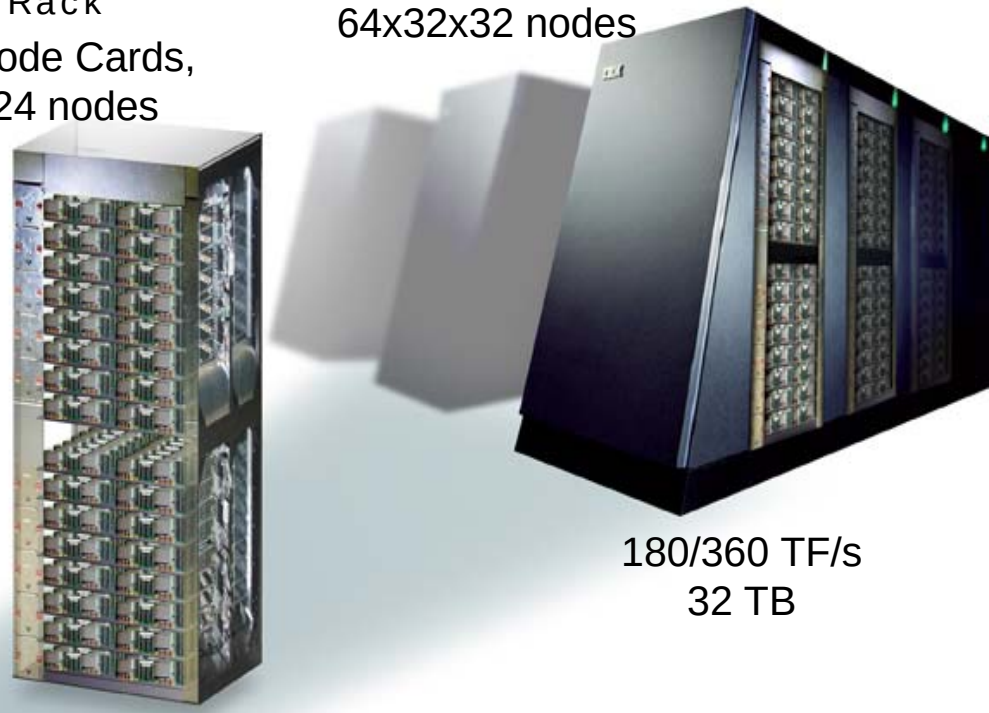
These chips are cool, pack a mean FPU capability -- and are **not** bleeding edge…

BlueGene/L System Build-up

aggressive packaging,
up to the power/cooling limits
of an air-cooled rack

System
LLNL= 64 Racks,
64x32x32 nodes

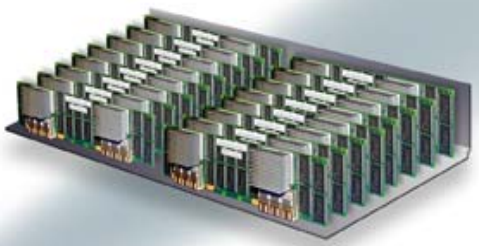
Rack
32 Node Cards,
1024 nodes



180/360 TF/s
32 TB

2.8/5.6 TF/s
512 GB

Node Card
(32 nodes 4x4x2)
16 compute, 0-2 IO cards



90/180 GF/s
16 GB

Compute Card
2 nodes,
1x2x1

5.6/11.2 GF/s
1.0 GB

Chip
2 processors

2.8/5.6 GF/s
4 MB

64-rack BlueGene/L installation at Lawrence Livermore National Laboratory

Wish list for a PetaScale computer

ref. W.J. Camp, SOS10

- ✓ **Balanced system performance**
  - BlueGene/L swept the HPC Challenge benchmarks
  - optimization allowed/encouraged!
- ✓ **Usability**
- ✓ **Scalability > 50k processors**

  -- I presume weak and strong
- ✓ **Reliability > 50hr MTBF; full RAS**
- ✓ **Upgradeability, maintainability**
- ✓ **Space, Power, Cooling**
- ✓ **Price/Performance**
  - BlueGene/L shows that the HPC community is *not* necessarily beholden to x86
  - low power ASIC has better Flop/s /W -- speed daemons are not necessarily the right building block for supercomputers
  - better Flop/s /W reduces both acquisition cost and total cost of ownership
- ✓ **True MPP**
- ✓ **Distributed memory / MIMD**
- ✓ **Fully connected 3D-mesh /Torus**
- ✓ **Function partitioned across h/w: Service / FE / IO / compute**
- ✓ **Partitioned O/S**
- ✓ **Separate RAS network (mgmt, monitoring)**
- ✓ **Source-based routing in primary network**
- ✓ **Diskless nodes**
- – **Vertical blades**
- ✓ **Passive backplane**
- – **Detailed rack dimensions**
- ✓ **Air-cooled, possibly air/water heat exchanger**
- ✓ **…**

Hypothesis 2. Increased system peak is no use without increasing performance achieved by user workloads.

(How will your company increase application scalability by then?)

- We are very happy with BlueGene/L's scaling behavior, both in weak scaling and in strong scaling.
- Customers are too!

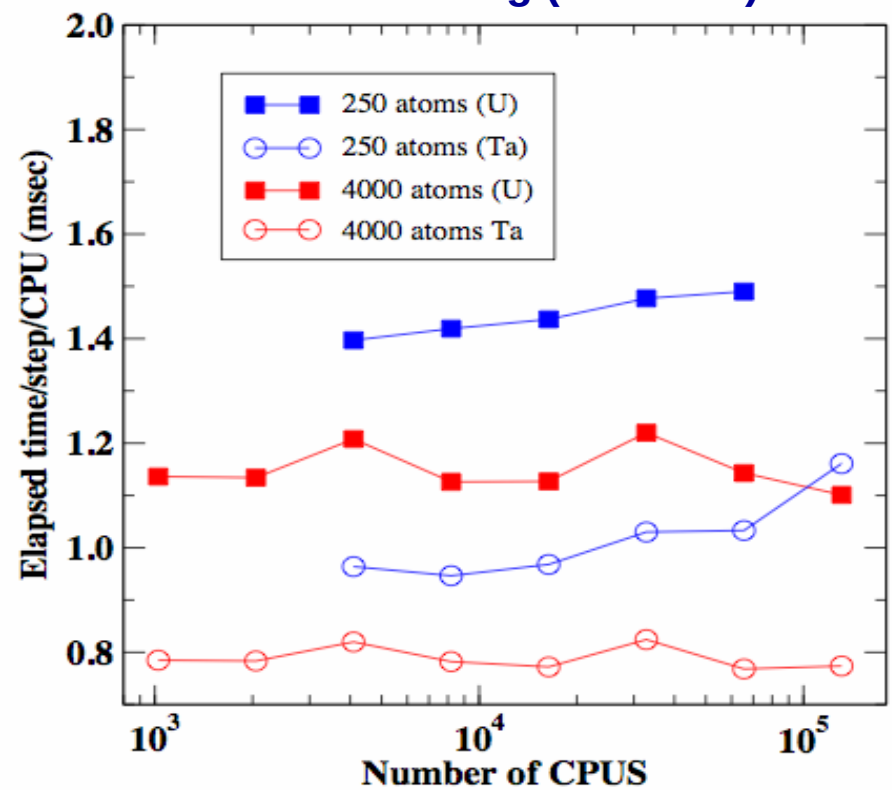
Classical MD – ddcMD (LLNL) -- 2005 Gordon Bell Prize Winner

524 million atom simulations on 64K nodes achieved 101.5 TF/s sustained.

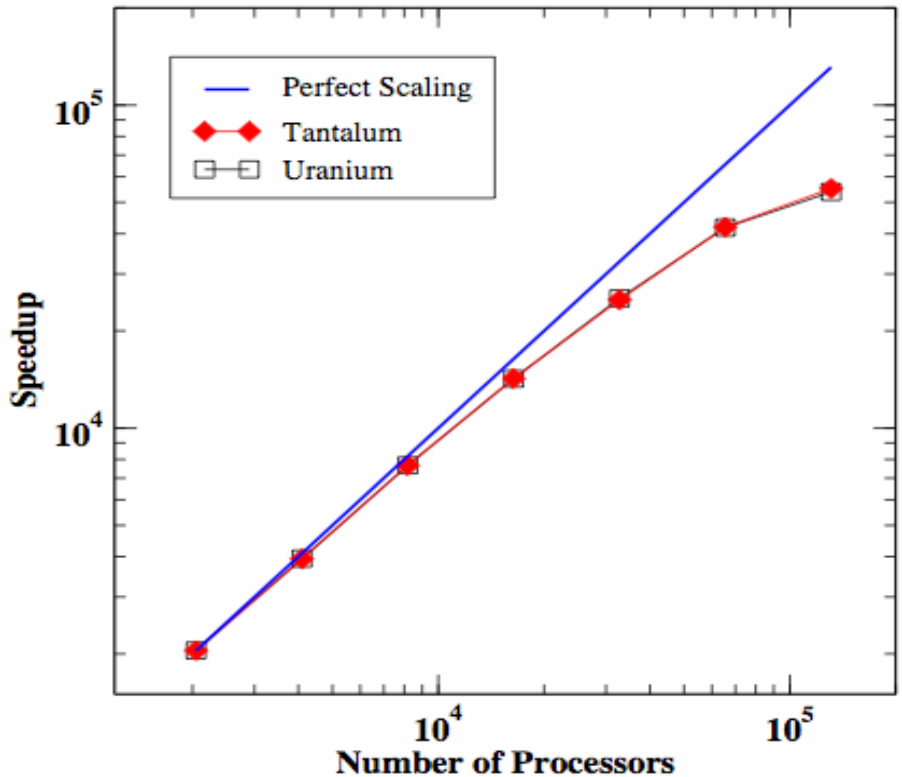
"… unprecedented scaling of size or time"

- **Weak scaling is virtually flat across the entire machine - enables simulation of tens of billions of atoms (roughly a cubic micron of material)**
- **Strong scaling shows speedup down to 8 atoms/CPU - enables simulations involving millions of steps (typically ns of simulated time)**

Weak Scaling (Ta and U)

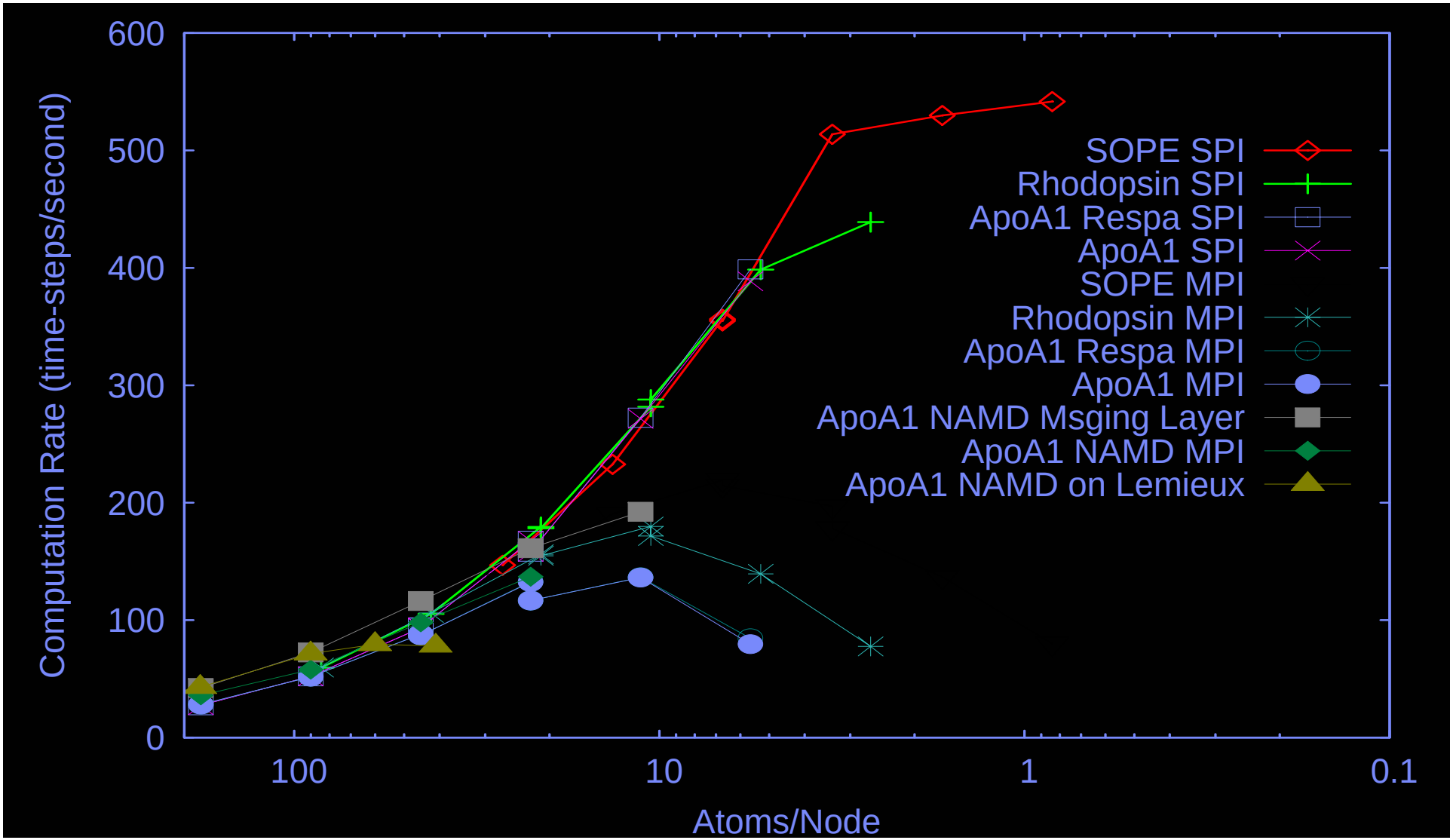


Strong Scaling (Ta and U)



"Excellent scaling of ddcMD on BG/L supports solidification understanding"

BGW: Strong Scaling Results



Improved user productivity…

- While typical Linpack benchmarks score 70-80% FPU utilization…
  - Note that the 100+ TFlop/s ddcMD result utilizes only 28% of the 360 TFlop/s peak FPU capability
  - Most other user applications score less! Communication bound apps (FFT) much less!
- Thus the challenge is to work on software / communications overhead / compilers / libraries
- IBM supports customers with programming expertise, software development and training and will support users with questions on porting and optimization.
- IBM supports the *BlueGene/L Consortium* :

  "a community of BG/L expertise that will foster the rapid scientific adoption of BG/L and develop experience in order to provide critical feedback to the architects and designers of the BG follow-on system". Led by Argonne National Labs.
  http://www-fp.mcs.anl.gov/bgconsortium/

- IBM provides commercial access to its BlueGene/L facilities via an "on-demand" process
- IBM provides non-commercial access via
  - the *DOE Incite* program
  - *BGW Days* workshops sponsored by the BlueGene/L Consortium.

Hypothesis 3. The whole IT market may continue to grow, but HPC markets may decrease as a proportion of this

What is your company's view of the future size of the HPC market? How do you think HPC customers can realistically help you?

- IBM evaluates a business case for each of its offerings, including its continued offerings to the HPC market.
- We foresee continued growth in demand for HPC CPU cycles, branching out from scientific markets into commercial markets.
- However, the size of the market cannot sustain a big development budget – so we plan to keep it simple, relatively cheap, power efficient and modular.
- IBM predicts that the future of HPC will be in massively parallel systems built from tightly coupled, tightly packaged, power efficient nodes … such as exemplified in the BlueGene architecture.
- Customers can help by criticizing or endorsing this view.
  - In the latter case, please buy a few racks…

Hypothesis 4. HPC achieves increasing benefits to users through vendor competition.

Who do you expect to be your closest competitors in 2009? Where are you watching for new competitors to come from?

- Somehow supercomputers play into national pride. Japanese and Chinese governments have separately announced PetaFlop/s-scale supercomputer projects.
- We expect our US-based competitors to vigorously strive to be there.
- It is IBM's policy to sell products and services on their merits.

Outlook

- Classical CMOS power-law scaling is coming to an end
- For super computers, this will mean “wild west”, with lots of architectural experimentation.
- It is critical that the user community help us -- the vendors.
  - Keep the competition honest by insisting on real achieved performance on real applications. The HPC Challenge is a good beginning.
  - The worst machine that a user can buy is just a Linpack stunt…
  - Amdahl’s law rules… if your application is communication-bound, then Linpack performance becomes (almost) meaningless.

HPC Challenge

- **HPL** - the Linpack TPP benchmark which measures the floating point rate of execution for solving a linear system of equations.
- **RandomAccess** - measures the rate of integer random updates of memory (GUPS).
- **FFTE** - measures the floating point rate of execution of double precision complex one-dimensional Discrete Fourier Transform (DFT).
- **STREAM** - a simple synthetic benchmark program that measures sustainable memory bandwidth (in GB/s) and the corresponding computation rate for simple vector kernel.

Benchmark	64-rack BG (optimized)	16-rack BG (optimized)
HPL (TFlop/s)	259.213	67.11
RandomAccess (GUP/s)	35.46	17.29
FFT (GFlop/s)	2311.09	988.18
STREAM Triad (GB/s)	160,064	39,991

←Only ~1 TFlop/s, still best in class!

We find that for protein simulation -- mix of short range (direct space) & long range (k-space) interactions -- we get about twice the FFT rate (~ 1.85 TFlop/s on 16 racks).

So that is the appropriate figure of merit for that particular problem!