Progress in Supercomputing: The Top Three Breakthroughs of the Last 20 Years and the Top Three Challenges for the Next 20 Years

Horst Simon
Associate Laboratory Director
Lawrence Berkeley National Laboratory

ISC 2005 Heidelberg
June 22, 2005
Cray-2
- 244 MHz (4.1 nsec)
- 4 processors
- 1.95 Gflop/s peak
- 2 GB memory (256 MW)
- 1.2 Gflop/s LINPACK R_max
- 1.6 m² floor space
- 0.2 MW power
IBM BG/L @ LLNL

- 700 MHz (x 2.86)
- 65,536 nodes (x 16,384)
- 180 (360) Tflop/s peak (x 92,307)
- 32 TB memory (x 16,000)
- 135 Tflop/s LINPACK (x 110,000)
- 250 m² floor space (x 156)
- 1.8 MW power (x 9)
1985 versus 2005

- custom built vector mainframes
- 30 Mflops sustained is good performance
- vector Fortran
- proprietary operating system
- remote batch only
- no visualization
- no tools, hand tuning only
- dumb terminals
- remote access via 9600 baud
- single software developer, develops and codes everything
- serial, vectorized algorithms

- commodity massively parallel platforms
- 1 Tflops sustained is good performance
- Fortan/C with MPI, object orientation
- Unix, Linux
- interactive use
- visualization
- parallel debugger, development tools
- high performance desktop
- remote access via 10 Gb/s; grid tools
- large group developed software, code share and reuse
- parallel algorithms
The Top 10 Major Accomplishments in Supercomputing 1985 – 2005

• My own personal opinion
• Selected by “impact” and “change in perspective”

10) The TOP500 list
9) NAS Parallel Benchmark
8) The “grid”
7) Hierarchical algorithms: multigrid and fast multipole methods
6) HPCC initiative and Grand Challenge applications
#5 The “Attack of the Killer Micros”

- First used by Eugene Brooks (LLNL) at Supercomputing 89
- Became a catchy shorthand expression for the technology change from custom ECL to commodity CMOS
- Commodity CMOS micro processors did change the face of supercomputing, but they were neither inevitable, nor the only technology choice

Intel I860XP - a killer micro at its time
#4 Beowulf Clusters

• Thomas Sterling et al. established vision of low cost, high end computing

• Demonstrated effectiveness of PC clusters for some (not all) classes of applications

• Provided software and conveyed findings to broad community (great PR) through tutorials and book (1999)

• Made parallel computing accessible to large community worldwide; broadened and democratized HPC

• However effectively stopped HPC architecture innovation for at least a decade; narrower market for custom systems
#3 Scientific Visualization


- Change in point of view: transformed computer graphics from a technology driven subfield of computer science into a medium for communication

- Added artistic element

- The role of visualization is “to reveal concepts that are otherwise invisible” (Krytstof Lenk)
Before Scientific Visualization (1985)

Computer graphics typical of the time:
- 2 dimensional
- line drawings
- black and white
- “vectors” used to display vector field

Images from a CFD report at Boeing (1985).
After scientific visualization (1992)

The impact of scientific visualization seven years later:
– 3 dimensional
– use of “ribbons” and “tracers” to visualize flow field
– color used to characterize updraft and downdraft

Images from “Supercomputing and the Transformation of Science” by Kauffman and Smarr, 1992; visualization by NCSA; simulation by Bob Wilhelmson, NCSA

Orange ribbons represent tracers that rise through the depth of the storm in the updraft and blue ribbons represent tracers that eventually fall to the ground in the downdraft in this image tracing particle paths from approximately 75 minutes after the start of the simulation to 2 hours and 18 minutes after.

Tracer ribbons illustrate the development of streamwise vorticity during a 20-minute period. The relative magnitude of the vorticity is largest in the lower part of the storm, where the ribbon is most tightly coiled.
#2 Message Passing Interface (MPI)
Parallel Programming 1988

- At the 1988 “Salishan” conference there was a bake-off of parallel programming languages trying to solve five scientific problems

- The “Salishan Problems” (ed. John Feo, published 1992) investigated four programming languages
  - Sisal, Haskel, Unity, LGDF

- Significant research activity at the time

- The early work on parallel languages is all but forgotten today
The availability of real parallel machines moved the discussion from the domain of theoretical CS to the pragmatic application area.

In this presentation (ca. 1990) Jack Dongarra lists seven approaches to parallel processing:

1. Automatic parallelizing compiler - low level language
2. Compiler directives
3. Add-on to existing sequential languages (monitors, send/receive, SCHEDULE, macros, subroutines)
4. Special parallel languages - SISAL, Linda, Strand
5. Standard parallel features in existing languages - PFC
6. High-level languages - Lisp (pure), Prolog (pure)
7. New Programming models - spreadsheet, neural networks

Note that message passing libraries are a sub-item on 2)
Parallel Programming 1994

Evolution in Programming Models

1988
- single architecture
- single programming model

1993
- single architecture
- multiple programming models

199?
- multiple architectures
- multiple programming models
Parallel Programming 1996

• At NERSC in 1996 among scientific users
  – 30% used PVM
  – 30% used MPI
  – 30% used HPF
  – 10% used SHMEM or CRAFT

• It was only in about 1998 that it was clear that MPI was the choice model for parallel programming
Until 2010: A New Parallel Programming Methodology? - NOT

The software challenge: overcoming the MPI barrier

- MPI created finally a standard for applications development in the HPC community
- Standards are always a barrier to further development
- The MPI standard is a least common denominator building on mid-80s technology

Programming Model reflects hardware!

From a presentation by HDS in Heidelberg, 2001

“I am not sure how I will program a Petaflops computer, but I am sure that I will need MPI somewhere” – HDS 2001
DEVELOPMENT OF PARALLEL METHODS FOR A 1024-PROCESSOR HYPERCUBE

JOHN L. GUSTAFSON†, GARY R. MONTRY†, AND ROBERT E. BENNER†

Abstract. We have developed highly efficient parallel solutions for three practical, full-scale scientific problems: wave mechanics, fluid dynamics, and structural analysis. Several algorithmic techniques are used to keep communication and serial overhead small as both problem size and number of processors are varied. A new parameter, operation efficiency, is introduced that quantifies the tradeoff between communication and redundant computation. A 1024-processor MIMD ensemble is measured to be 502 to 637 times as fast as a single processor when problem size for the ensemble is fixed, and 1009 to 1020 times as fast as a single processor when problem size per processor is fixed. The latter measure, denoted scaled speedup, is developed and contrasted with the traditional measure of parallel speedup. The scaled-problem paradigm better reveals the capabilities of large ensembles, and permits detection of subtle hardware-induced load imbalances (such as error correction and data-dependent MFLOPS rates) that may become increasingly important as parallel processors increase in node count. Sustained performance for the applications is 70 to 130 MFLOPS, validating the massively parallel ensemble approach as a practical alternative to more conventional processing methods. The techniques presented appear extensible to even higher levels of parallelism than the 1024-processor level explored here.

Key words. fluid dynamics, hypercubes, MIMD machines, multiprocessor performance, parallel computing, structural analysis, supercomputing, wave mechanics
The argument against massive parallelism (ca. 1988)

From a presentation by Jack Worlton, LANL, at NCSA 1989
The argument against massive parallelism (ca. 1988)

Massively-Parallel Approaches (continued)

For continued performance improvement —

Case 1 \[ \sigma(p) = c \log_2 p \quad p \leq 6931 \quad S_p \leq 704 \]

Case 2 \[ \sigma(p) = \frac{c}{2} p \quad p \leq 141 \quad S_p \leq 70 \]

Case 3 \[ \sigma(p) = \frac{c}{2} p \log_2 p \quad p \leq 53 \quad S_p \leq 29 \]

where the normalization coefficient \( c = 0.0001 \).

From a presentation by Jim Hack at DOE

Salishan High Speed Computing, 1988
Scaled Speed Up

Helped the community to overcome a conceptual barrier

- Exposed the fallacy of the fixed size speed-up
- Focused back on the fact that we want to use larger computers in order to solve larger problems
- Opened the way for parallel applications development

Without this change in our understanding of parallel computing the early successes in adapting parallel computers would not have happened
Three Challenges for 2005 - 2025

• These are my personal opinion
• The challenges are in chronological order
  ▪ 2005 - 2012
  ▪ 2010 - 2018
  ▪ 2015 - 2025
Challenge 2005 - 2010: Scaling Applications to Petascale Systems

(Assume nominal Petaflop/s system with 100,000 commodity processors of 10 Gflop/s each)

Three major issues:

• Scaling to 100,000 processors and multi-core processors
• Topology sensitive interconnection network
• Memory Wall
Application Status in 2005

- A few Teraflop/s sustained performance
- Scaled to 512 - 1024 processors
Parallelism has Stagnated for a Decade

Number of processors in the most highly parallel system in the TOP500

![Graph showing the number of processors over time with Intel Paragon XP, ASCI RED, and IBM BG/L highlighted.](image-url)
Integrated Performance Monitoring (IPM)

- brings together multiple sources of performance metrics into a single profile that characterizes the overall performance and resource usage of the application
- maintains low overhead by using a unique hashing approach which allows a fixed memory footprint and minimal CPU usage
- open source, relies on portable software technologies and is scalable to thousands of tasks
- developed by David Skinner at NERSC (see http://www.nersc.gov/projects/ipm/)
Scaling Portability: Profoundly Interesting

A high level description of the performance of cosmology code MADCAP on four well known architectures.

Source: David Skinner, NERSC, IPM project
http://www.nersc.gov/projects/ipm/
16 Way for 4 seconds

(About 20 timestamps per second per task) *(1...4 contextual variables)
64 way for 12 seconds
256 Way for 36 Seconds
Applications on Petascale Systems will need to deal with

(Assume nominal Petaflop/s system with 100,000 commodity processors of 10 Gflop/s each)

Three major issues:

• Scaling to 100,000 processors and multi-core processors
• Topology sensitive interconnection network
• Memory Wall
• Currently most applications work with a flat MPI model, this is already a simplification
• More processors means more complex interconnects and topology sensitivity
• Example: BG/L
  – five different interconnection networks
  – latency dependent on distance
Even today’s machines are interconnect topology sensitive

Four (16 processor) IBM Power 3 nodes with Colony switch
If the interconnect is topology sensitive, mapping will become an issue (again)

“Characterizing Ultra-Scale Applications Communications Requirements”, by John Shalf et al., submitted to SC05
Interconnect Topology BG/L
Applications on Petascale Systems will need to deal with

(Assume nominal Petaflop/s system with 100,000 commodity processors of 10 Gflop/s each)

Three major issues:

• Scaling to 100,000 processors and multi-core processors

• Topology sensitive interconnection network

• Memory Wall
The Memory Wall

Source: “Getting up to speed: The Future of Supercomputing”, NRC, 2004
Characterizing Memory Access

Memory Access Patterns/Locality

HPCS Challenge Points
HPCchallenge Benchmarks

High
Temporal Locality
Low

Low
Spatial Locality
High

Mission Partner Applications

RandomAccess

FFT

HPL
PTRANS
STREAM

Source: David Koester, MITRE
Apex-MAP characterizes architectures through a synthetic benchmark

Temporal Locality

1=Low

"Global Streams"

"Short indirect"

1/Re-use

"HPL"

"Small working set"

0 = High

0 = High

1/L

1=Low

Spatial Locality

Source: Erich Strohmaier, NERSC, LBNL
Apex-Map Sequential

Seaborg Sequential

Source: Erich Strohmaier, NERSC, LBNL
Apex-Map Sequential

Source: Erich Strohmaier, NERSC, LBNL
Apex-Map Sequential

Source: Erich Strohmaier, NERSC, LBNL
Challenge 2010 - scaling to Petaflops level

• Applications will face (at least) three challenges
  – Scaling to 100,000s of processors
  – Interconnect topology
  – Memory access

• We have yet to scale to the 100,000 processor level
  – Algorithms
  – Tools
  – System Software
Challenge 2010 - 2018: Developing a New Ecosystem for HPC

From the NRC Report on "The Future of Supercomputing":

- Platforms, software, institutions, applications, and people who solve supercomputing applications can be thought of collectively as an ecosystem.

- Research investment in HPC should be informed by the ecosystem point of view - progress must come on a broad front of interrelated technologies, rather than in the form of individual breakthroughs.

Pond ecosystem image from http://www.tpwd.state.tx.us/expltx /eft/txwild/pond.htm
Supercomputing Ecosystem (1988)

Cold War and Big Oil spending in the 1980s

Powerful Vector Supercomputers

20 years of Fortran applications base in physics codes and third party apps
Supercomputing Ecosystem (until about 1988)

Cold War and Big Oil spending in the 1980s

Powerful Vector Supercomputers

20 years of Fortran applications base in physics codes and third party apps
Supercomputing Ecosystem (2005)

Commercial Off The Shelf technology (COTS)

“Clusters” 12 years of legacy MPI applications base
Supercomputing Ecosystem (2005)

Commercial Off The Shelf technology (COTS)

“Clusters”

12 years of legacy MPI applications base
How Did We Make the Change?

• Massive R&D Investment
  – HPCC in the US
  – Vigorous computer science experimentation in languages, tools, system software
  – Development of Grand Challenge applications

• External Driver
  – Industry transition to CMOS micros

• All changes happened virtually at once
  – Ecosystem change
Observations on the 2005 Ecosystem

• It is very stable
  – attempts of re-introducing old species failed (X1)
  – attempts of introducing new species failed (mutation of Blue Gene 1999 to BG/L 2005)

• It works well
  – just look around the room

• So why isn’t everybody happy and content?
The supercomputing community is aware that the current situation is suboptimal for HPC

- “divergence problem” and Blue Planet at NERSC
- concern about the “right” benchmarks

The current ecosystem will become untenable after about 2010 in the face of the architectural and software challenges

How are we going to change the ecosystem?

What are we going to change it into?

DARPA HPCS is on the right track combining requirements for new architecture, new languages, and insistence on commercialization by vendors

But, will a $150M program be enough to change a $6 - 8B industry?
Some time between 2015 and 2025 the continued performance growth of semiconductor based microprocessors will end.

for more details see presentation by Erik DeBenedictis, Sandia, at www.zettaflops.org
Vanishing Electrons (2016)

Electrons per device

Source: Joel Birnbaum, HP, Lecture at APS Centennial, Atlanta, 1999
Driving away from FM transmitter $\rightarrow$ less signal
Noise from electrons $\rightarrow$ no change

Increasing numbers of gates $\rightarrow$ less signal power
Noise from electrons $\rightarrow$ no change
## Semiconductor Roadmap

<table>
<thead>
<tr>
<th></th>
<th>2010</th>
<th>2013</th>
<th>2016</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>YEAR OF PRODUCTION</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DRAM 1/2 PITCH (nm)</td>
<td>45</td>
<td>32</td>
<td>22</td>
</tr>
<tr>
<td>MPU / ASIC 1/2 PITCH (nm)</td>
<td>50</td>
<td>35</td>
<td>25</td>
</tr>
<tr>
<td>MPU PRINTED GATE LENGTH (nm)</td>
<td>25</td>
<td>18</td>
<td>13</td>
</tr>
<tr>
<td>MPU PHYSICAL GATE LENGTH (nm)</td>
<td>18</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>Physical gate length high-performance (HP) (nm) [1]</td>
<td>18</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>Equivalent physical oxide thickness for high-performance $T_{ox}$ (EOT) (nm) [2]</td>
<td>0.5-0.8</td>
<td>0.4-0.6</td>
<td>0.4-0.5</td>
</tr>
<tr>
<td>Gate depletion and quantum effects electrical thickness adjustment factor (nm) [3]</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$T_{ox}$ electrical equivalent (nm) [4]</td>
<td>1.2</td>
<td>1.0</td>
<td>0.9</td>
</tr>
<tr>
<td>Nominal power supply voltage ($V_{dd}$) (V) [5]</td>
<td>0.6</td>
<td>0.5</td>
<td>0.4</td>
</tr>
<tr>
<td>Nominal high-performance NMOS sub threshold leakage current, $I_{sub}$ (at 25°C) (μA/μm) [6]</td>
<td>3</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>Nominal high-performance NMOS saturation drive current, $I_{dd}$ (at $V_{dd}$, at 25°C) (μA/μm) [7]</td>
<td>1200</td>
<td>1500</td>
<td>1500</td>
</tr>
<tr>
<td>Required percent current-drive &quot;mobility/transconductance improvement&quot; [8]</td>
<td>30%</td>
<td>70%</td>
<td>100%</td>
</tr>
<tr>
<td>Parasitic source/drain resistance (Rsd) (ohm-μm) [9]</td>
<td>110</td>
<td>90</td>
<td>80</td>
</tr>
<tr>
<td>Parasitic source/drain resistance (Rsd) percent of ideal channel resistance ($V_{dd}$/$I_{dd}$) [10]</td>
<td>25%</td>
<td>30%</td>
<td>35%</td>
</tr>
<tr>
<td>Parasitic capacitance percent of ideal gate capacitance [11]</td>
<td>31%</td>
<td>36%</td>
<td>42%</td>
</tr>
<tr>
<td>High-performance NMOS device $\tau$ ($C_{gate}$ * $V_{dd}$ / $I_{sat}$-NMOS)/(ps) [12]</td>
<td>0.39</td>
<td>0.22</td>
<td>0.15</td>
</tr>
<tr>
<td>Relative device performance [13]</td>
<td>4.3</td>
<td>7.2</td>
<td>10.7</td>
</tr>
<tr>
<td>Energy per ($W/L_{gate}=3$) device switching transition ($C_{gate}$ * ($3$L_{gate}) * $V^2$) (fJ/Device) [14]</td>
<td>0.015</td>
<td>0.007</td>
<td>0.002</td>
</tr>
<tr>
<td>Static power dissipation per ($W/L_{gate}=3$) device (Watts/Device) [15]</td>
<td>9.7E-08</td>
<td>1.4E-07</td>
<td>1.1E-07</td>
</tr>
</tbody>
</table>

*White—Manufacturable Solutions Exist, and Are Being Optimized*
*Yellow—Manufacturable Solutions are Known*
*Red—Manufacturable Solutions are NOT Known*
# ITRS Device Review 2016

<table>
<thead>
<tr>
<th>Technology</th>
<th>Speed (min-max)</th>
<th>Dimension (min-max)</th>
<th>Energy per gate-op</th>
<th>Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMOS</td>
<td>30 ps-1 μs</td>
<td>8 nm-5 μm</td>
<td>4 aJ</td>
<td></td>
</tr>
<tr>
<td>RSFQ</td>
<td>1 ps-50 ps</td>
<td>300 nm-1 μm</td>
<td>2 aJ</td>
<td>Larger</td>
</tr>
<tr>
<td>Molecular</td>
<td>10 ns-1 ms</td>
<td>1 nm-5 nm</td>
<td>10 zJ</td>
<td>Slower</td>
</tr>
<tr>
<td>Plastic</td>
<td>100 μs-1 ms</td>
<td>100 μm-1 mm</td>
<td>4 aJ</td>
<td>Larger+Slower</td>
</tr>
<tr>
<td>Optical</td>
<td>100 as-1 ps</td>
<td>200 nm-2 μm</td>
<td>1 pJ</td>
<td>Larger+Hotter</td>
</tr>
<tr>
<td>NEMS</td>
<td>100 ns-1 ms</td>
<td>10-100 nm</td>
<td>1 zJ</td>
<td>Slower+Larger</td>
</tr>
<tr>
<td>Biological</td>
<td>100 fs-100 μs</td>
<td>6-50 μm</td>
<td>.3 yJ</td>
<td>Slower+Larger</td>
</tr>
<tr>
<td>Quantum</td>
<td>100 as-1 fs</td>
<td>10-100 nm</td>
<td>1 zJ</td>
<td>Larger</td>
</tr>
</tbody>
</table>

Data from ITRS ERD Section, quoted from Erik DeBenedictis, Sandia Lab.
Challenge 2015 - 2025: The Coming “Flattening” of Moore’s Law

- There is no active, strong research program anywhere that addresses this challenge
- Alternative technology solutions are feasible, but won’t come by themselves
- After 50 years of exponential growth, how will the industry adjust to a no-growth scenario?
The Three Future Challenges

• Learn how to scale
• Change the ecosystem
• Deal with the “flattening” of Moore’s law
Hans, Congratulations on 20 years of ISC!

Thank you from all of us for providing such an excellent venue for learning and discussions.