Extreme Supercomputing

Erik P. DeBenedictis
Sandia National Laboratories

Presentation at University of Notre Dame
September 13, 2004
The Baseline

ETA: January 2005

Performance
- Peak of ~ 40 TF
- Expected MP-Linpack performance > 20 TF
- Aggregate system memory bandwidth ~55 TB/s
- Interconnect
  - Aggregate sustained interconnect bandwidth > 100 TB/s
  - MPI Latency - μs neighbor, 5 μs across machine
  - Bi-Section bandwidth ~2.3 TB/s
  - Link bandwidth ~3.0 GB/s in each direction
- I/O System
  - Sustained 50 GB/s disk I/O bandwidth for each color
  - Sustained 25 GB/s external network bandwidth for each color

Design Parameters
- True MPP, Designed to be a single system
- Fully connected high performance 3-D mesh interconnect
- Topology - 27 X 16 X 24 compute nodes and 2 X 8X 16 service and I/O nodes
- 108 compute node cabinets and 10,368 compute node processors (AMD Sledgehammer @ 2.0 GHz)
- ~10 TB of DDR memory @ 333 MHz (1.0 GB per processor)
- Red/Black switching - ~1/4, ~1/2, ~1/4
- 8 Service and I/O cabinets on each end (256 processors for each color)
- 240 TB of disk storage (120 TB per color)
- Functional hardware partitioning - service and I/O nodes, compute nodes, and RAS nodes
- Functional system software partitioning - LINUX on service and I/O nodes, LWK (Catamount) on compute nodes, stripped down LINUX on RAS nodes
- Separate RAS and system management network (Ethernet)
- Router table based routing in the interconnect
- Less than 2 MW total power and cooling
- Less than 3,000 square feet of floor space

ETA: January 2005
Applications and Computer Technology

System Performance

Applications

No schedule provided by source

Technology

1 Zettaflops
100 Exaflops
10 Exaflops
1 Exaflops
100 Petaflops
10 Petaflops
1 Petaflops
100 Teraflops

Plasma Fusion Simulation

Full Global Climate

[Malone 03]

Compute as fast as the engineer can think

[NASA 99]

↓ 100× ↑1000× [SCaLeS 03]

Geodata Earth Station Range

[NASA 02]

↑ ③ Quantum Dots/Reversible Logic mP (green)

Best-case logic (red)

[DeBenedictis 04]

↑ ② Best-case logic–100k_B T limit

↑ ① μP–125× below 100k_B T limit


Outline

• The Computing of Physics: The Need for Zettaflops
  • Limits of Moore’s Law
    Today’s Technologies
  • An Expert System/Optimizer for Supercomputing
  • The Physics of Computing: Reaching to Zettaflops
  • Roadmap and Future Directions
• Space is divided into cells, each with computer variables representing the physical state of the volume represented by the cell.

• The computer updates the state of a cell for successive time intervals $\Delta T$ based on some physical laws.

• I.e. $S'_{ijk} = f(S_{ijk}, \text{states of nearby cells})$
Fourth Power Scaling Rule

Reference

2× spatial resolution,
2× time steps →
$2^4\times$ FLOPS

Physical Parameter

Time

Physical Parameter

Time
Global Climate

• Objective
  – Collect data about Earth
  – Model climate into the future
  – Provide “decision support” and ability to “mitigate”

• Approaches
  – Climate models exist, but need they more resolution, better physics, and better initial conditions (observations of the Earth)

• Computer Resources Required
  – Increments over current workstation on next slide
# FLOPS Increases for Global Climate

<table>
<thead>
<tr>
<th>Issue</th>
<th>Scaling</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Zettaflops</td>
<td></td>
</tr>
<tr>
<td>Ensembles, scenarios</td>
<td>Embarrassingly Parallel</td>
</tr>
<tr>
<td>10×</td>
<td></td>
</tr>
<tr>
<td>Run length</td>
<td>Longer Running Time</td>
</tr>
<tr>
<td>100×</td>
<td></td>
</tr>
<tr>
<td>New parameterizations</td>
<td>More Complex Physics</td>
</tr>
<tr>
<td>100×</td>
<td></td>
</tr>
<tr>
<td>Model Completeness</td>
<td>More Complex Physics</td>
</tr>
<tr>
<td>100×</td>
<td></td>
</tr>
<tr>
<td>Spatial Resolution</td>
<td>Resolution</td>
</tr>
<tr>
<td>$10^4 \times (10^3 \times - 10^5 \times)$</td>
<td></td>
</tr>
<tr>
<td>Clusters Now In Use (100 nodes, 5% efficient)</td>
<td></td>
</tr>
</tbody>
</table>

Based on these inputs, various portions of the Modeling and Data Assimilation System will require anywhere from $10^7$ to $10^{13}$ GFLOPS of computational resources. In other words, the range of computational resources needed is $10^{16}$ to $10^{21}$ Floating Point Operations per Second. For the curious, the range can also be stated as 10 PetaFLOPS to 1 ZettaFLOPS.

4.1.2. Anticipated Computing Technology Capabilities

At first glance, the numbers discussed in the previous section appear so high as to be impossibly ludicrous. However, with the expected growth in computing capabilities, the lower end of this spectrum actually falls within the domain of possibility.

• "…the ultimate goal of making the computing underlying the design process so capable that it no longer acts as a brake on the flow of the creative human thought…"
• Requirement 3 Exaflops
• Note: In the context of this report, this requirement is for one or a few engineers, not a supercomputer center!
Outline

• The Computing of Physics: The Need for Zettaflops
• Limits of Moore’s Law
  Today’s Technologies
• An Expert System/Optimizer for Supercomputing
• The Physics of Computing: Reaching to Zettaflops
• Roadmap and Future Directions
### Best-Case Logic vs. Microprocessor Architecture

<table>
<thead>
<tr>
<th>Expert Opinion</th>
<th>100 Exaflops</th>
<th>800 Petaflops</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>25 Exaflops</td>
<td>200 Petaflops</td>
</tr>
<tr>
<td></td>
<td>4 Exaflops</td>
<td>32 Petaflops</td>
</tr>
<tr>
<td>Assumption: Supercomputer is size &amp; cost of Red Storm: US$100M budget; consumes 2 MW wall power; 750 KW to active components</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Logic**: 2×10²⁴ logic ops/s
- **Microprocessor Architecture**: 800 Petaflops

### Physical Factor

<table>
<thead>
<tr>
<th>Physical Factor</th>
<th>Source of Authority</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability limit 750KW/(80k_BT)</td>
<td>Esteemed physicists (T=60°C junction temperature)</td>
</tr>
<tr>
<td>Derate 20,000 convert logic ops to floating point</td>
<td>Floating point engineering (64 bit precision)</td>
</tr>
<tr>
<td>Derate for manufacturing margin (4×)</td>
<td>Estimate</td>
</tr>
<tr>
<td>Uncertainty (6×)</td>
<td>Gap in chart</td>
</tr>
<tr>
<td>Improved devices (4×)</td>
<td>Estimate</td>
</tr>
<tr>
<td>Projected ITRS improvement to 22 nm (100×)</td>
<td>ITRS committee of experts</td>
</tr>
<tr>
<td>Lower supply voltage (2×)</td>
<td>ITRS committee of experts</td>
</tr>
<tr>
<td>Red Storm</td>
<td>contract</td>
</tr>
</tbody>
</table>
Thermal Noise Limit

This logical irreversibility is associated with physical irreversibility and requires a minimal heat generation, per machine cycle, typically of the order of $kT$ for each irreversible function. – R. Landauer 1961

$kT$ “helper line,” drawn out of the reader’s focus because it wasn’t important at the time of writing – Carver Mead, Scaling of MOS Technology, 1994
Metaphor: FM Radio on Trip to Chicago

• You drive to Chicago listening to FM radio

• Music clear for a while, but noise creeps in and then overtakes music

• Analogy: You live out the next dozen years buying PCs every couple years

• PCs keep getting faster
  – clock rate increases
  – fan gets bigger
  – won’t go on forever

• Why…see next slide

SIA Semiconductor Roadmap

• Generalization of Moore’s Law
  – Projects many parameters
  – Years through 2016
  – Includes justification
  – Panel of experts
    • known to be wrong
  – Size between Albuquerque white and yellow pages

International Technology Roadmap for Semiconductors (ITRS), see
http://public.itrs.net
## Semiconductor Roadmap

<table>
<thead>
<tr>
<th>YEAR OF PRODUCTION</th>
<th>2010</th>
<th>2013</th>
<th>2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM ½ PITCH (nm)</td>
<td>45</td>
<td>32</td>
<td>22</td>
</tr>
<tr>
<td>MPU / ASIC ½ 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</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, $I_{\text{leak}}$ (at 25°C) ($\mu\text{A}/\mu\text{m}$) [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) ($\mu\text{A}/\mu\text{m}$) [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 ($R_{sd}$) (ohms-$\mu\text{m}$)</td>
<td>110</td>
<td>90</td>
<td>80</td>
</tr>
<tr>
<td>Parasitic source/drain resistance ($R_{sd}$) percentage</td>
<td>25%</td>
<td>30%</td>
<td>35%</td>
</tr>
<tr>
<td>Parasitic capacitance percent of ideal gate capacitance</td>
<td>31%</td>
<td>36%</td>
<td>42%</td>
</tr>
<tr>
<td>High-performance NMOS device τ ($C_{\text{gate}} * V_{dd} / I_{dd-NMOS}$) [12]</td>
<td>0.39</td>
<td>0.22</td>
<td>0.15</td>
</tr>
<tr>
<td>Relative device performance [13]</td>
<td>4.4</td>
<td>7.2</td>
<td>10.7</td>
</tr>
<tr>
<td>Energy per ($W/L_{\text{gate}}=3$) device switching transition ($C_{\text{gate}} * (3 * L_{\text{gate}}) * V_{dd}^2$) (fJ/Device) [14]</td>
<td>0.015</td>
<td>0.007</td>
<td>0.002</td>
</tr>
<tr>
<td>Static power dissipation per ($W/L_{\text{gate}}=3$) device (Watts/Device) [15]</td>
<td>9.7E-08</td>
<td>1.4E-07</td>
<td>1.1E-07</td>
</tr>
</tbody>
</table>

---

1,000 $k_B T$/transistor
### Scientific Supercomputer Limits

<table>
<thead>
<tr>
<th>Expert Opinion</th>
<th>25 Exaflops</th>
<th>4 Exaflops</th>
<th>1 Exaflops</th>
<th>100 Exaflops</th>
<th>800 Petaflops</th>
<th>80 Teraflops</th>
<th>40 Teraflops</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Physical Factor</strong></td>
<td>Reliability limit 750KW/(80k_B T)</td>
<td>Derate 20,000 convert logic ops to floating point</td>
<td>Derate for manufacturing margin (4×)</td>
<td>Uncertainty (6×)</td>
<td>Improved devices (4×)</td>
<td>Projected ITRS improvement to 22 nm (100×)</td>
<td>Lower supply voltage (2×)</td>
</tr>
<tr>
<td><strong>Source of Authority</strong></td>
<td>Esteemed physicists (T=60°C junction temperature)</td>
<td>Floating point engineering (64 bit precision)</td>
<td>Estimate</td>
<td>Gap in chart</td>
<td>Estimate</td>
<td>ITRS committee of experts</td>
<td>ITRS committee of experts</td>
</tr>
</tbody>
</table>

Assumption: Supercomputer is size & cost of Red Storm: US$100M budget; consumes 2 MW wall power; 750 KW to active components.

2×10²⁴ logic ops/s
Personal Observational Evidence

• Have radios become better able to receive distant stations over the last few decades with a rate of improvement similar to Moore’s Law?

• You judge from your experience, but the answer should be that they have not.

• Therefore, electrical noise does not scale with Moore’s Law.
Outline

• The Computing of Physics: The Need for Zettaflops
• Limits of Moore’s Law
  Today’s Technologies
  • An Expert System/Optimizer for Supercomputing
• The Physics of Computing: Reaching to Zettaflops
• Roadmap and Future Directions
Supercomputer Expert System

Application/Algorithm run time model as in applications modeling

Logic & Memory Technology design rules and performance parameters for various technologies (CMOS, Quantum Dots, C Nano-tubes ...)

Interconnect Speed, power, pin count, etc.

Physical Cooling, packaging, etc.

Expert System & Optimizer (looks for best 3D mesh of generalized MPI connected nodes, µP and other)

Results
1. Block diagram picture of optimal system (model)
2. Report of FLOPS count as a function of years into the future

Time Trend Lithography as a function of years into the future
Sample Analytical Runtime Model

- Simple case: finite difference equation
- Each node holds $n \times n \times n$ grid points
- Volume-area rule
  - Computing $\propto n^3$
  - Communications $\propto n^2$

$$T_{\text{step}} = 6 n^2 C_{\text{bytes}} T_{\text{byte}} + n^3 F_{\text{grind/floprate}}$$
Expert System for Future Supercomputers

- Applications Modeling
  - Runtime
    \[ T_{\text{run}} = f_1(n, \text{design}) \]
- Technology Roadmap
  - Gate speed = \( f_2(\text{year}) \),
  - chip density = \( f_3(\text{year}) \),
  - cost = \$(n, \text{design}), …
- Scaling Objective Function
  - I have \( C_1 \) & can wait \( T_{\text{run}} = C_2 \) seconds. What is the biggest \( n \) I can solve in year \( Y \)?

- Use “Expert System” To Calculate:
  \[ \text{Max} \quad n: \$<C_1, T_{\text{run}}<C_2 \]
  All designs

- Report:
  \[ T_{\text{run}}(n, \text{design}) \]
  and illustrate “design”
Outline

• The Computing of Physics: The Need for Zettaflops
• Limits of Moore’s Law
  Today’s Technologies
• An Expert System/Optimizer for Supercomputing
• The Physics of Computing: Reaching to Zettaflops
• Roadmap and Future Directions
Candidate Technologies for Zettaflops

- CMOS per Moore’s Law
  - Cluster/µP solution exceeds limits by 10,000×
    - Trillion US$ cost
    - 10 × Hoover Dam for power supply
  - Custom logic solution exceeds limits by 100×
    - US$10 billion cost
    - 100 MW power
  - ∴ worth our while to consider alternatives

- Limiting search for Alternatives to CMOS
  - Digital (not Analog)
  - Plausible to lots of Floating point
  - Controllable by something recognizable as “programming”
  - Mature enough for above issues to be addressed in published papers
  - Rules out coherent quantum, neural nets, DNA computing, optical interference, …
Alternatives to CMOS for Zettaflops

• New Devices
  – Superconducting: RSFQ (a. k. a. nSQUID, parametric quantrons)
  – Quantum Dots/QCA
  – Rod Logic
  – Helical Logic
  – Single Electron Transistors
  – Carbon Nanotube Y Junctions
  – ...

• Logic and Architecture
  – “Reversible logic” will be unfamiliar to today’s engineers but has been shown to be sufficient
  – Arithmetic elements and microprocessors have been demonstrated
  – Leading architecture:
    • Reversible ALU/CPU
    • Irreversible memory
1 Zettaflops Scientific Supercomputer

- How could we increase “Red Storm” from 40 Teraflops to 1 Zettaflops?

Answer

- $>2.5 \times 10^7$ power reduction per operation
- Faster devices $\times$ more parallelism $>2.5 \times 10^7$
- Smaller devices to fit existing packaging

Ref. “Maxwell’s demon and quantum-dot cellular automata,” John Timler and Craig S. Lent
JOURNAL OF APPLIED PHYSICS 15 JULY 2003
An Exemplary Device: Quantum Dots

- Pairs of molecules create a memory cell or a logic gate

Ref. “Clocked Molecular Quantum-Dot Cellular Automata,” Craig S. Lent and Beth Isakson, IEEE TRANSACTIONS ON ELECTRON DEVICES, VOL. 50, NO. 9, SEPTEMBER 2003
Not Specifically Advocating Quantum Dots

- A number of post-transistor devices have been proposed.
- The shape of the performance curves have been validated by a consensus of reputable physicists.
- However, validity of any data point can be questioned.
- Cross-checking appropriate; see →

Reversible Multiplier Status

- 8×8 Multiplier Designed, Fabricated, and Tested by IBM & University of Michigan
- Power savings was up to 4:1
QCA Microprocessor Status

- M. Niemier Ph. D. Thesis, University of Notre Dame
- 12 Bit µP
- CAD design tool principles
  - 10× circuit density of CMOS at same λ
- Applies to various devices
  - Metal dot 4.2 nm²
  - Molecular 1.1 nm²

Figure 4.6. A 2-bit QCA Simple 12 ALU with registers
Reversible Microprocessor Status

• Status
  – Subject of Ph. D. thesis
  – Chip laid out (no floating point)
  – RISC instruction set
  – C-like language
  – Compiler
  – Demonstrated on a PDE
  – However: really weird and not general to program with +=, -=, etc. rather than =
CPU Design

• Leading Thoughts
  – Implement CPU logic using reversible logic
    • High efficiency for the component doing the most logic
  – Implement state and memory using conventional logic
    • Low efficiency, but not many operations
  – Permits programming much like today
Supercomputer is 211K chips, each with 70.7K nodes of 5.77K cells of 240 bytes; solves 86T=44.1Kx44.1Kx44.1K cell problem. System dissipates 332KW from the faces of a cube 1.53m on a side, for a power density of 47.3KW/m². Power: 332KW active components; 1.33MW refrigeration; 3.32MW wall power; 6.65MW from power company. System has been inflated by 2.57 over minimum size to provide enough surface area to avoid overheating. Chips are at 99.22% full, comprised of 7.07G logic, 101M memory decoder, and 6.44T memory transistors. Gate cell edge is 34.4nm (logic) 34.4nm (decoder); memory cell edge is 4.5nm (memory). Compute power is 768 EFLOPS, completing an iteration in 224µs and a run in 9.88s.
Outline

- The Computing of Physics: The Need for Zettaflops
- Limits of Moore’s Law
  Today’s Technologies
- An Expert System/Optimizer for Supercomputing
- The Physics of Computing: Reaching to Zettaflops
- Roadmap and Future Directions
Where to Go Next I: Workshop

• Don’t believe me? Believe the Experts
• Workshop Agenda October 12
  – Applications session – Climate expert Phil Jones
  – Advanced Architectures – PIM expert Peter Kogge
  – Limits of Current Architectures – Me
  – Panel I: Important Applications
  – New Logic – Reversible Logic Expert Michael Frank
  – New Devices – Quantum Dot Developer Craig Lent
  – Panel II: Do Opportunities Justify the Effort?
Where To Go Next II: Roadmap

Notes:
* Not necessarily one machine; different applications may require different machines
* Specifics are just my ideas
Will Supercomputers Grow Forever?

• Will supercomputer simulations scale up forever, or will there be a maximum?
  – Zettaflops simulates the Earth, and the Earth is the largest thing that we care about in detail

• Will progress in science always come through “simulating physics on a computer”?
  – Perhaps future problems could be formulated as a combination of symbolic reasoning (artificial intelligence) and floating point