Exascale Computing: Embedded Style

Peter M. Kogge
McCourtney Chair in Computer Science & Engineering
Univ. of Notre Dame
IBM Fellow (retired)
May 5, 2009
• Last 30 years:
  - “Gigascale” computing first in a single vector processor
  - “Terascale” computing first via several thousand microprocessors
  - “Petascale” computing first via several hundred thousand cores
• Commercial technology: to date
  - Always shrunk prior “XXX” scale to smaller form factor
  - Shrink, with speedup, enabled next “XXX” scale
• Space/Embedded computing has lagged far behind
  - Environment forced implementation constraints
  - Power budget limited both clock rate & parallelism
• “Exascale” now on horizon
  - But beginning to suffer similar constraints as space
  - And technologies to tackle exa challenges very relevant

Especially Energy/Power
Topics

• The DARPA Exascale Technology Study
• The 3 Strawmen Designs
• A Deep Dive into Operand Access
Disclaimers

This presentation reflects my interpretation of the final report of the Exascale working group only, and not necessarily of the universities, corporations, or other institutions to which the members are affiliated.

Furthermore, the material in this document does not reflect the official views, ideas, opinions and/or findings of DARPA, the Department of Defense, or of the United States government.

Note: Separate Exa Studies on Resiliency & Software

### The Exascale Study Group

<table>
<thead>
<tr>
<th>NAME</th>
<th>Affiliation</th>
<th>NAME</th>
<th>Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keren Bergman</td>
<td>Columbia</td>
<td>Steve Keckler</td>
<td>UT-Austin</td>
</tr>
<tr>
<td>Shekhar Borkar</td>
<td>Intel</td>
<td>Dean Klein</td>
<td>Micron</td>
</tr>
<tr>
<td>Dan Campbell</td>
<td>GTRI</td>
<td>Peter Kogge</td>
<td>Notre Dame</td>
</tr>
<tr>
<td>Bill Carlson</td>
<td>IDA</td>
<td>Bob Lucas</td>
<td>USC/ISI</td>
</tr>
<tr>
<td>Bill Dally</td>
<td>Stanford</td>
<td>Mark Richards</td>
<td>Georgia Tech</td>
</tr>
<tr>
<td>Monty Denneau</td>
<td>IBM</td>
<td>Al Scarpeli</td>
<td>AFRL</td>
</tr>
<tr>
<td>Paul Franzon</td>
<td>NCSU</td>
<td>Steve Scott</td>
<td>Cray</td>
</tr>
<tr>
<td>Bill Harrod</td>
<td>DARPA</td>
<td>Allan Snavely</td>
<td>SDSC</td>
</tr>
<tr>
<td>Kerry Hill</td>
<td>AFRL</td>
<td>Thomas Sterling</td>
<td>LSU</td>
</tr>
<tr>
<td>Jon Hiller</td>
<td>STA</td>
<td>Stan Williams</td>
<td>HP</td>
</tr>
<tr>
<td>Sherman Karp</td>
<td>STA</td>
<td>Kathy Yelick</td>
<td>UC-Berkeley</td>
</tr>
</tbody>
</table>

**11 Academic**  **6 Non-Academic**  **5 “Government”**  
+ **Special Domain Experts**

**10+ Study Meetings over 2nd half 2007**
The DARPA Exascale Technology Study

- Exascale = 1,000X capability of Petascale
- Exascale != Exaflops but
  - Exascale at the data center size => Exaflops
  - Exascale at the “rack” size => Petaflops for departmental systems
  - Exascale embedded => Teraflops in a cube
- Teraflops to Petaflops took 14+ years
  - 1st Petaflops workshop: 1994
  - Thru NSF studies, HTMT, HPCS …
  - To give us to Peta now
- Study Questions:
  - Can we ride silicon to Exa By 2015?
  - What will such systems look like?
    - Can we get 1 EF in 20 MW & 500 racks?
  - Where are the Challenges?
The Study’s Approach

• Baseline today’s:
  ➢ Commodity Technology
  ➢ Architectures
  ➢ Performance (Linpack)

• Articulate scaling of potential application classes

• Extrapolate roadmaps for
  ➢ “Mainstream” technologies
  ➢ Possible offshoots of mainstream technologies
  ➢ Alternative and emerging technologies

• Use technology roadmaps to extrapolate use in “strawman” designs

• Analyze results & Id “Challenges”
Context:
Focus on Energy
Not Power
CMOS Energy 101

One clock cycle dissipates $C \times V^2$

Dissipate $CV^2/2$
And store $CV^2/2$
From Capacitance
Assume capacitance of a circuit scales as feature size.

90nm picked as breakpoint because that's when Vdd and thus clocks flattened.

330X

15X
The 3 Study Strawmen
Architectures Considered

• Evolutionary Strawmen
  ➢ “Heavyweight” Strawman based on commodity-derived microprocessors
  ➢ “Lightweight” Strawman based on custom microprocessors

• Aggressive Strawman
  ➢ “Clean Sheet of Paper” CMOS Silicon
A Modern HPC System

Silicon Area Distribution
- Processors: 3%
- Routers: 33%
- Memory: 86%
- Random: 8%

Power Distribution
- Random Memory: 9%
- Routers: 33%

Board Area Distribution
- White Space: 50%
- Processors: 56%
- Memory: 10%
- Random: 8%
- Routers: 8%
- Processors: 24%
A “Light Weight” Node Alternative

2 Nodes per “Compute Card.” Each node:
• A low power compute chip
• Some memory chips
• “Nothing Else”

System Architecture:
• Multiple Identical Boards/Rack
• Each board holds multiple Compute Cards
• “Nothing Else”

“Packaging the Blue Gene/L supercomputer,” IBM J. R&D, March/May 2005
“Blue Gene/L compute chip: Synthesis, timing, and physical design,” IBM J. R&D, March/May 2005
**Possible System Power Models: Interconnect Driven**

- **Simplistic:** A highly optimistic model
  - Max power per die grows as per ITRS
  - Power for memory grows *only linearly* with # of chips
    - Power per memory chip remains constant
  - Power for routers and common logic remains constant
    - Regardless of obvious need to increase bandwidth
  - True if energy for bit moved/accessed *decreases* as fast as “flops per second” increase

- **Fully Scaled:** A pessimistic model
  - Same as Simplistic, except memory & router power grow with peak flops per chip
  - True if energy for bit moved/accessed *remains constant*

- **Real world:** somewhere in between
1 EFlop/s “Clean Sheet of Paper” Strawman

Sizing done by “balancing” power budgets with achievable capabilities

- 4 FPUs+RegFiles/Core (=6 GF @1.5GHz)
- 1 Chip = 742 Cores (=4.5TF/s)
  - 213MB of L1I&D; 93MB of L2
- 1 Node = 1 Proc Chip + 16 DRAMs (16GB)
- 1 Group = 12 Nodes + 12 Routers (=54TF/s)
- 1 Rack = 32 Groups (=1.7 PF/s)
  - 384 nodes / rack
- 3.6EB of Disk Storage included
- 1 System = 583 Racks (=1 EF/s)
  - 166 MILLION cores
  - 680 MILLION FPUs
  - 3.6PB = 0.0036 bytes/flops
  - 68 MW w’aggressive assumptions

Largely due to Bill Dally, Stanford
A Single Node (No Router)

Characteristics:
- 742 Cores; 4 FPUs/core
- 16 GB DRAM
- 290 Watts
- 1.08 TF Peak @ 1.5GHz
- ~3000 Flops per cycle

(a) Quilt Packaging

“Stacked” Memory

(b) Thru via chip stack
1 Eflops Aggressive Strawman
Data Center Power Distribution

- 12 nodes per group
- 32 groups per rack
- 583 racks
- 1 EFlops/3.6 PB
- **166 million cores**
- **67 MWatts**
Data Center Performance Projections

But not at 20 MW!
Power Efficiency

Exascale Study:
1 Eflops @ 20MW

Aggressive Strawman

UHPC RFI:
Module: 80GF/W
System: 50GF/W

Historical
Exascale Goal
Top 10
Aggressive Strawman Design
Light Node Simplistic
Heavy Node - Simplistic
Light Node Fully Scaled
Heavy Node - Fully Scaled
Energy Efficiency

Graph showing the energy efficiency over time from 1/1/80 to 1/1/20. The x-axis represents the years, and the y-axis represents energy per flop (pJ/Flop). The graph includes various marks for different categories:

- **X**: Historical
- **▲**: Green 500 Top 10
- **●**: Top 10
- **○**: UHPC Cabinet Energy Efficiency Goal
- **●**: UHPC Module Energy Efficiency Goal
- **■**: Exa Simplistically Scaled Projection
- **□**: Exa Fully Scaled Projection
- **---**: Top System Trend Line
- **-.-.-.-.**: CMOS Technology

The data points show a decreasing trend in energy per flop over time, indicating improved energy efficiency.
Data Center Total Concurrency

- Billion-way concurrency
- Million-way concurrency
- Thousand-way concurrency

Graph showing total concurrency over time with different markers for:
- Top 10
- Historical
- Top System
- Exa Strawman
- Evolutionary Light Node
- Evolutionary Heavy Node

Graph timeline from 1/1/96 to 1/1/20
Data Center Core Parallelism

AND we will need 10-100X more Threading for Latency Management

170 Million Cores

- Historical
- Top 10
- Top System
- Exa Strawman
- Evolutionary Light Node
- Evolutionary Heavy Node
Key Take-Aways

• Developing Exascale systems really tough
  ➢ In any time frame, for any of the 3 classes

• Evolutionary Progression is at best 2020ish
  ➢ With limited memory

• 4 key challenge areas
  ➢ Power:
  ➢ Concurrency:
  ➢ Memory Capacity
  ➢ Resiliency

• Requires coordinated, cross-disciplinary efforts
Embedded Exa: A Deep Dive into Interconnect to Deliver Operands
Tapers

- **Bandwidth Taper**: How effective *bandwidth* of operands being sent to a functional unit varies with location of the operands in memory hierarchy.
  - Units: Gbytes/sec, bytes/clock, *operands per flop time*

- **Energy Taper**: How *energy cost* of transporting operands to a functional unit varies with location of the operands in the memory hierarchy.
  - Units: Gbytes/Joule, *operands per Joule*

- Ideal tapers: “Flat”—doesn’t matter where operands are.

- Real tapers: huge dropoffs
An Exa Single Node for Embedded
The Access Path: Interconnect-Driven

More Routers

MICROPROCESSOR

Memory

Some sort of memory structure
Sample Path – Off Module Access

1. Check local L1 (miss)
2. Go thru TLB to remote L3 (miss)
3. Across chip to correct port (thru routing table RAM)
4. Off-chip to router chip
5. 3 times thru router and out
6. Across microprocessor chip to correct DRAM I/F
7. Off-chip to get to correct DRAM chip
8. Cross DRAM chip to correct array block
9. Access DRAM Array
10. Return data to correct I/R
11. Off-chip to return data to microprocessor
12. Across chip to Routre Table
13. Across microprocessor to correct I/O port
14. Off-chip to correct router chip
15. 3 times thru router and out
16. Across microprocessor to correct core
17. Save in L2, L1 as required
18. Into Register File
# Taper Data from Exascale Report

<table>
<thead>
<tr>
<th>Memory Level</th>
<th>Capacity GB</th>
<th>Bandwidth GB/s</th>
<th>BW Taper Operands/clock</th>
<th>Power mW/flop</th>
<th>Energy Taper Operands/pJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Register File</td>
<td>1KB</td>
<td>35,960</td>
<td>4</td>
<td>8.2</td>
<td>0.56</td>
</tr>
<tr>
<td>L1</td>
<td>64KB</td>
<td>8,992</td>
<td>0.25</td>
<td>5.5</td>
<td>0.068</td>
</tr>
<tr>
<td>L2</td>
<td>256KB</td>
<td>4,496</td>
<td>0.125</td>
<td>3.5</td>
<td>0.054</td>
</tr>
<tr>
<td>L3</td>
<td>97MB</td>
<td>2,288</td>
<td>0.0625</td>
<td>3.8</td>
<td>0.025</td>
</tr>
<tr>
<td>Local DRAM</td>
<td>16GB</td>
<td>712</td>
<td>0.02</td>
<td>10.5</td>
<td>0.0029</td>
</tr>
<tr>
<td>Network DRAM</td>
<td>3.8PB</td>
<td>216</td>
<td>0.006</td>
<td>11.5</td>
<td>0.00008</td>
</tr>
</tbody>
</table>
Bandwidth Tapers

1,000X Decrease Across the System!!
Energy Analysis: Possible Storage Components

Cacti 5.3 Extrapolation Technology for 2017 system

Operands/pJ


- RF: LOP; 4P; 1x128W
- 32KB RAM: HP; 1P; 1x4KW
- 32KB RAM: LOP; 1P; 1x4KW
- 32KB Cache: LOP; 1P, DM; 1W Block
- 32KB Cache: LOP; 1P, DM; 4W Block
- 1M RAM: LP-DRAM; 1P, 1x4KW
- 1M RAM: LP-DRAM, 1P, 8 Bank
- RF from Exa Study
- DRAM from Exa Study

- RF: HP; 4P; 1x128W
- 32KB Cache: LOP; 1P, DM; 1W Block
- 1M RAM: LOP, 1P, 1 Bank
- 1M RAM: Comm-DRAM, 1P, 1 Bank
- 32KB SRAM Exa Study
## Summary Transport Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Value</th>
<th>units</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core to L1</td>
<td>2.1</td>
<td>mm</td>
<td>Scaled to 2015 dimensions</td>
</tr>
<tr>
<td>Cross chip</td>
<td>21.3</td>
<td>mm</td>
<td>Unchanged from Exa study</td>
</tr>
<tr>
<td>High swing wire</td>
<td>0.107</td>
<td>pJ per bit per mm</td>
<td></td>
</tr>
<tr>
<td>Low swing wire</td>
<td>0.018</td>
<td>pJ per bit per mm</td>
<td></td>
</tr>
<tr>
<td>One word chip-chip - pad</td>
<td>144</td>
<td>pJ per word</td>
<td>Traditional pad</td>
</tr>
<tr>
<td>One word chip-chip - TSV</td>
<td>0.8</td>
<td>pJ per word</td>
<td>11 fJ/bit</td>
</tr>
<tr>
<td>One word chip-chip - capacitive</td>
<td>144</td>
<td>pJ per word</td>
<td></td>
</tr>
<tr>
<td>One word chip-chip - inductive</td>
<td>10.1</td>
<td>pJ per word</td>
<td></td>
</tr>
</tbody>
</table>

### On-chip Optical

<table>
<thead>
<tr>
<th>Option</th>
<th>Value</th>
<th>units</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>E/O modulator</td>
<td>7.2</td>
<td>pJ per word</td>
<td>Modulate 1 word onto laser beam</td>
</tr>
<tr>
<td>O/E Receiver</td>
<td>7.2</td>
<td>pJ per word</td>
<td>Receive one word back to digital</td>
</tr>
<tr>
<td>On-chip Broadband Router</td>
<td>0.5</td>
<td>mW</td>
<td>per router</td>
</tr>
<tr>
<td>Power per laser</td>
<td>10.0</td>
<td>mW</td>
<td>per wavelength</td>
</tr>
<tr>
<td>Power for 250 lasers</td>
<td>2.5</td>
<td>W</td>
<td></td>
</tr>
</tbody>
</table>

### Off chip Optical

<table>
<thead>
<tr>
<th>Option</th>
<th>Value</th>
<th>units</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laser, per word</td>
<td>96</td>
<td>pJ per word</td>
<td>Using 0.3mW per channel @10% activity and 2 G</td>
</tr>
<tr>
<td>Modulator, per word</td>
<td>0.72</td>
<td>pJ per word</td>
<td></td>
</tr>
<tr>
<td>RX + TIA, per word</td>
<td>360</td>
<td>pJ per word</td>
<td>Again at 10% activity</td>
</tr>
<tr>
<td>Temperature Control</td>
<td>?</td>
<td>pJ per word</td>
<td></td>
</tr>
</tbody>
</table>

1 Operand/word = 72 bits
## Energy Tapers vs Goals

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Register File</td>
<td>9.77E-04</td>
<td>0.7</td>
<td>0.0</td>
<td>0.7</td>
<td>0.0%</td>
<td>1.436</td>
</tr>
<tr>
<td>L1 Hit</td>
<td>3.13E-02</td>
<td>3.9</td>
<td>0.0</td>
<td>3.9</td>
<td>0.0%</td>
<td>0.260</td>
</tr>
<tr>
<td>L1 Miss, Local L2 Hit</td>
<td>2.50E-01</td>
<td>30.6</td>
<td>13.3</td>
<td>17.3</td>
<td>43.4%</td>
<td>0.033</td>
</tr>
<tr>
<td>L1 Miss, L3 Hit</td>
<td>3.04E+02</td>
<td>158.3</td>
<td>141.0</td>
<td>17.3</td>
<td>89.1%</td>
<td>0.006</td>
</tr>
<tr>
<td>L1,L2/L3 Miss On-Module Access</td>
<td>6.55E+04</td>
<td>292.1</td>
<td>255.0</td>
<td>37.1</td>
<td>87.3%</td>
<td>0.003</td>
</tr>
<tr>
<td>L1,L2/L3 Miss Off-Module Access</td>
<td>2.52E+07</td>
<td>6620.8</td>
<td>6607.2</td>
<td>13.7</td>
<td>99.8%</td>
<td>0.000</td>
</tr>
</tbody>
</table>

![Graph showing the comparison of different access and taper goals](image-url)
Reachable Memory vs Energy

Energy per Operand (pJ) vs. Reachable memory (MB)
What Does This Tell Us?

- There’s a lot more energy sinks than you think
  - And we have to take all of them into consideration

- Cost of Interconnect Dominates

- Must design for on-board or stacked DRAM, with DRAM blocks CLOSE
  - take into account physical placement

- Reach vs energy per access looks “linear”

- For 80GF/W, cannot afford ANY memory references

- We NEED to consider the entire access path:
  - Alternative memory technologies – reduce access cost
  - Alternative packaging costs – reduce bit movement cost
  - Alternative transport protocols – reduce # bits moved
  - Alternative execution models – reduce # of movements