Exascale Systems

Ram Rajamony
Research Staff Member
IBM Research, Austin, TX

Opinions expressed are the author’s personal opinions and should not be construed as being endorsed by IBM
Today: LANL’s RoadRunner*

- A New Programming Model extended from standard, cluster computing
- Hybrid and Heterogeneous HW
- Built around BladeCenter and Industry IB-DDR

<table>
<thead>
<tr>
<th>T/V</th>
<th>N</th>
<th>NB</th>
<th>P</th>
<th>Q</th>
<th>Time</th>
<th>Gflops</th>
</tr>
</thead>
<tbody>
<tr>
<td>WR13C2C8</td>
<td>2236927</td>
<td>128</td>
<td>68</td>
<td>180</td>
<td>7269.80</td>
<td>1.026e+06</td>
</tr>
</tbody>
</table>

* This supercomputer was designed and developed for the DOE and Los Alamos National Laboratory (LANL) under the DOE / LANL project name Roadrunner. The Roadrunner project was named after the state bird of New Mexico.
The Near Future (1): Blue Waters

- **Highly Productive** sustained petaflop system by mid-2011
  - POWER7 multicore chips
  - A large, high-performance memory subsystem, to enable the solution of memory-intensive problems.
  - A low-latency, high-performance interconnect, to facilitate scaling to large numbers of cores.
  - A high-performance I/O subsystem, to enable the solution of data-intensive problems.
  - High reliability, to ensure that the most challenging problems can be addressed.
The Near Future (2): BlueGene/Q

- Mission driven partnership with DOE/NNSA/OoS
  - Push state of the art by >10x
  - Scales > 10PFLOPs
  - Radically new node and system architecture
  - Leverage Blue Gene OS communities
  - Grand challenge science stresses
  - Production system for classified codes
Trends pointing to Exaflop systems
The Quick Answer
What will an Exaflop system look like?

- Peak of an Exaflop

- Reasonable characteristics:
  - Power: 100 to 300 MW → **Must bring down to 30 to 50 MW**
    - Power cost is ~ $1M/MW/year → 1 Exaflop @ 100 MW = $100 million per year!!!
    - Power needs to go down in order to make operational costs affordable
  - “Memory” capacity: 30 to 100 PBytes
  - Productivity: Sustained similar to (or better than) systems in use today
  - Reliability: MTBF of at least multiple days
  - Storage capacity: ~ 1 Exabyte
Getting to an Exaflop: Challenges

- Power
- Cost
- Productivity
- Reliability
- ...

Projected Performance Development

[Graph depicting projected performance development over time, showing various trend lines and markers for different performance metrics.]
## Power: A look at the Green500 list

<table>
<thead>
<tr>
<th>Green500 Rank</th>
<th>Site</th>
<th>Manufacturer</th>
<th>Computer</th>
<th>Mflops Per Watt</th>
<th>Total Power (kW)</th>
<th>TOP500 Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IBM Germany</td>
<td>IBM</td>
<td>BladeCenter QS22 Cluster, PowerXCell 8i 3.2 Ghz, Infiniband</td>
<td>488.14</td>
<td>22.76</td>
<td>324</td>
</tr>
<tr>
<td>1</td>
<td>Fraunhofer ITWM</td>
<td>IBM</td>
<td>BladeCenter QS22 Cluster, PowerXCell 8i 3.2 Ghz, Infiniband</td>
<td>488.14</td>
<td>18.97</td>
<td>464</td>
</tr>
<tr>
<td>3</td>
<td>DOE/NNSA LANL</td>
<td>IBM</td>
<td>BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz , Voltaire Infiniband</td>
<td>437.43</td>
<td>2345.5</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Argonne National Laboratory</td>
<td>IBM</td>
<td>Blue Gene/P Solution</td>
<td>371.75</td>
<td>31.5</td>
<td>304</td>
</tr>
<tr>
<td>4</td>
<td>Dublin Institute for Advanced Studies/ICHEC</td>
<td>IBM</td>
<td>Blue Gene/P Solution</td>
<td>371.75</td>
<td>31.5</td>
<td>305</td>
</tr>
</tbody>
</table>

Source: www.green500.org

Fact: 39 out of the first 40 entries in the list are IBM systems.
76% of the first 100 entries are IBM systems.

To keep total power within 100 MW (!!!!!!), MFLOPS/Watt would have to be 10000.
System power distribution

- Compute and memory generally dominate power today
- But we will be in serious trouble if we let the trends continue

- Need new approaches to compute and memory, must contain interconnect

Assumptions: 0.125 B/F DDR memory, localized interconnect ➔ controlled optics, compute trends continue
System cost distribution

- Compute and memory generally dominate power today
- But we will be in serious trouble if we let the trends continue

First problem is memory

Assumptions: 0.125 B/F DDR memory, localized interconnect ➔ controlled optics, compute trends continue
Reliability

- The LANL Roadrunner* has close to 100K Cell engines (8 to a Cell chip)
- The ½ PFLOP BG/L has over 200K processors

- Compute frequencies are unlikely to increase significantly ➔ we will likely have an atrociously large number of compute elements in exaflop systems.

- How to handle compute failures?
- What about memory failures?
- What about optics failures?
Getting to an Exaflop: Challenges

- Power
- Cost
- Productivity
- Reliability
- ...
Enabling Technologies
Phase Change Memory

Based on Chalcogenide Glass material property

Crystalline

Δ

Amorphous

Low res

High res

Heat applied via Joule Heating – an accompanying transistor or diode drives a current through the material

- Overview
  - Has been used in CD-RW and DVD-RW technologies
  - Can be switched between Crystalline and Amorphous states with the addition of heat.
  - Heat applied via Joule Heating – an accompanying transistor or diode drives a current through the material
  - Non-Volatile, true Random Access capability
  - Density limited by the size of the drive transistor or diodes and not the memory cell size

- Details of Switching Process
  - Drastic difference in resistance (ratios > 1000):
    - SET – amorphous to crystalline - requires heating to ~ 320C (above crystalline temp but below melt temp) and holding until recrystallization can occur \( \leq \) Speed limiting step - published data indicates 50 ns (IEDM 2003)
    - RESET – crystalline to amorphous – Requires heating to ~ 650C (above the melt temp) and doing a quick quench to lock in the amorphous state \( \leq \) Power limiting step - published data indicates 30 ns pulse length to get to \( 10^{12} \) cycles

Source: Stefanie Chiras, IBM Research
Storage Class Memory: Game changer for cost?

- NAND assumes ITRS Road Map

<table>
<thead>
<tr>
<th>Year</th>
<th>½Pitch (nm)</th>
<th># of bits</th>
<th># of 3D layers</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>45</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2009</td>
<td>40</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>2010</td>
<td>36</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>2011</td>
<td>32</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>2012</td>
<td>28</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>2013</td>
<td>25</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>2014</td>
<td>22</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>2015</td>
<td>20</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

Source: Chung Lam, IBM Research
Interconnect: Communication Locality

Before threshold filtering

After threshold filtering

Over a large range of investigated *scientific* workloads, most communication is:

- Localized
- Slow-varying in terms of partner assignments

Optical Circuit Switches

Source: http://electronicdesign.com/Files/29/5942/Figure_02.gif
Interconnect Structure

- Optics-based, but what if we exploit locality?
- Possibility: Leverage optical *circuit switching* technologies

Cost and Power → 3D Integration

What new systems can 3D enable?
Compute Processors

- Frequency likely to remain around 4 GHz
  - Probably 18nm technology

- Huge number of transistors possible: 30B to 50B
  - Depends on cost and commercial viability
  - Large number of cores/compute engines/accelerators

- High degree of 3D integration
  - Onboard eDRAM
  - PCM integration needs work

- Latencies and Bandwidth
  - Latencies depend on memory technology characteristics
  - High-speed elastic interfaces
  - ???
Strawman Exaflop System(s)

- Multitude of 3-4 GHz compute cores per chip
  - Sea of engines with control processors, smaller caches, interconnects (optics) integrated on chip, contained cache coherence

- Memory
  - Combination of eDRAM and other technologies – 3D stacking
  - Exact combination depends on PCM bandwidths and latencies

- Interconnect: Need to make a very hard choice
  - Localized workloads: High degree toroidal interconnect with non-local “warps”
  - Non-local workloads: Interconnect with broader reach
Productivity

Development and Execution time
The architectural landscape

Multicore processors, many with accelerators
- e.g. Sun Niagara
- e.g. Intel multicore, IXP
- e.g. IBM Cell
- e.g. GPGPU

Road Runner: Cell-accelerated Opteron

Programmers must understand concurrent structure of their applications

Applications seeking to leverage these architectures will need to go beyond data-parallel, globally synchronizing MPI model.

Source: Vijay Saraswat, IBM Research
The Partitioned Global Address Space Model

- Computation is performed in multiple places.
- A place contains data that can be operated on remotely.
- Data lives in the place it was created, for its lifetime.
- A datum in one place may reference a datum in another place.
- Data-structures (e.g. arrays) may be distributed across many places.
- Places may have different computational properties (e.g. PPE, SPE, ...).

A place expresses locality.

Source: Vijay Saraswat, IBM Research
X10: An Asynchronous PGAS language

- Asynchrony
  - Simple explicitly concurrent model for the user: `async (p) S` runs statement S “in parallel” at place p
  - Controlled through `finish`, and local (conditional) `atomic`

- Used for active messaging (remote asyncs), DMAs, fine-grained concurrency, fork/join concurrency, do-all/do-across parallelism
  - SPMD is a special case

Concurrency is made explicit and programmable.

Source: Vijay Saraswat, IBM Research
Concluding Thoughts

- **Getting to an exaflop system is going to be hard**
  - We need significant innovation in all system areas

- **Cost and power are going to be tremendous challenges**
  - How do we make it affordable?
  - How do we feed it?

- **Software and Hardware will need to work together to make system affordable**

- **Nothing like a good challenge to get things going!**