Architectures for Exascale Computing

Peter Strazdins
Computer Systems Group,
Research School of Computer Science

Computational Science at the Petascale and Beyond Workshop,
Australian National University
13 February 2012

1 Overview

- major components & attributes
- system requirements for exascale computing
- architectural complexity challenges
- resiliency/reliability issues; checkpointing
- energy efficiency challenges: on-chip and system-wide
- examples from recent petascale architectures
  - Jaguar, Tsubame, TianHe (FeiTeng), Blue Gene / Q
- outlook and conclusions
2 Exascale Architecture: Components & Attributes

- processor
  - floating point / integer speed, memory latency and bandwidth
- interconnect: interface chips, switches and wiring
  - bandwidth: point-to-point and bisectional; latency!!
- input / output
  - bandwidth, asynchronicity (output), ||ism
- attributes for all:
  - performance, reliability / resiliency, and energy/power
  - system design must manage these
- must be balanced for their intended use
  - long-running, ultra large-scale simulations
  - highly compute and memory-intensive; moderately I/O intensive
## 3 System Requirements for Exascale Computing

<table>
<thead>
<tr>
<th></th>
<th>2010</th>
<th>2018</th>
<th>Factor Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>System peak</td>
<td>2 Pf/s</td>
<td>1 Ef/s</td>
<td>500</td>
</tr>
<tr>
<td>Power</td>
<td>6 MW</td>
<td>20 MW</td>
<td>3</td>
</tr>
<tr>
<td>System Memory</td>
<td>0.3 PB</td>
<td>10 PB</td>
<td>33</td>
</tr>
<tr>
<td>Node Performance</td>
<td>0.125 Gf/s</td>
<td>10 Tf/s</td>
<td>80</td>
</tr>
<tr>
<td>Node Memory BW</td>
<td>25 GB/s</td>
<td>400 GB/s</td>
<td>16</td>
</tr>
<tr>
<td>Node Concurrency</td>
<td>12 cpus</td>
<td>1,000 cpus</td>
<td>83</td>
</tr>
<tr>
<td>Interconnect BW</td>
<td>1.5 GB/s</td>
<td>50 GB/s</td>
<td>33</td>
</tr>
<tr>
<td>System Size (nodes)</td>
<td>20 K nodes</td>
<td>1 M nodes</td>
<td>50</td>
</tr>
<tr>
<td>Total Concurrency</td>
<td>225 K</td>
<td>1 B</td>
<td>4,444</td>
</tr>
<tr>
<td>Storage</td>
<td>15 PB</td>
<td>300 PB</td>
<td>20</td>
</tr>
<tr>
<td>Input/Output bandwidth</td>
<td>0.2 TB/s</td>
<td>20 TB/s</td>
<td>100</td>
</tr>
</tbody>
</table>

Courtesy of Lucy Nowell & Sonia Sachs, DOE Office of Science
4 System Architectural Complexity Challenges

- trend towards heterogeneous node architectures
  - GPUs, FGPAs, custom accelerators?
- trend towards deeper hierarchies
  - deeper memory hierarchy (now 3 levels cache)
  - I/O: solid-state storage (on node?), I/O nodes
- distributed memory: also within node?
  - also within chip? (100s of cores)
- these, coupled with massive increases in concurrency, present major challenges to the software!
  - debugging, performance analysis, automatic tuning...
5 Challenges for Resiliency as Core Count Grows

(Courtesy of John Daly)
6 Resiliency: Issues with System-wide Checkpointing Approach

- process-level checkpointing: highly I/O intensive
  - increased capacity of I/O system exacerbates MTTI problem
  - also increases power costs!
- note: applications tend to have poor memory scalability
- application-level better but complex
- other approaches
  - hardware redundancy (cores, network links, nodes)
  - algorithmic methods

(Courtesy of Lucy Nowell & Sonia Sachs)
7 Energy / Power Challenges: On-chip

- energy cost to access data off-chip (DRAM) is $100 \times$ more than from registers
  - across system network, only $2 \times$ as much
- within each manycore chip, there will be a non-trivial interconnects too
- maintaining coherency is not scalable
- cache misses and atomic instructions (have $O(p)$ energy cost each ($O(p^2)$ total!)
- cache line size is sub-optimal for messages on on-chip networks
- soon, interconnects are expected to consume $50 \times$ more energy than logic circuits!
8 Energy / Power Challenges: System Wide

- essential not only the system’s *peak energy* be bounded, that *current energy* is minimized
- need to be able to *throttle down* components when not in (full use)
  - network and I/O as well as compute processors
  - e.g. on the Intel Single-chip Cloud Computer, has frequency and voltage domains \((P = fV^2)\)
- exascale system will need a hierarchical power and thermal management system
9 Current Systems: TSUBAME2: Commodity + Accelerators

- system: 2.4 PFLOPS (peak), 1.2 (LINPACK); 1.6 MW (peak)
- 1408 Intel Westemere nodes (6 cores 2.93 GHz)
  - 3 NVIDIA Tesla M2050 GPUs (8GB/s PCI Express) each!
  - memory bandwidth is now 32 GB/s per CPU
- 2 PB of local fast SSD-backed local storage + Lustre 9 PB (30 nodes)
- unified system and storage interconnect:
  - fat-tree, Infinband (12×324 + 179×38 port switches) 8GB/s per node
- racks: controlled heat-exchangers for high density (32 cooling towers)
- only system in top ten of Top 500, Green 500 and Graph 500
10 Current Systems: Jaguar: Proprietary + Accelerators (↑)

- 1.75 PFLOPs (LINPACK), 2.35 (peak); 3 MW (peak), air cooled
- 18,688 Cray XT5 compute nodes
dual hex-core AMD Istanbul
- Seastar2+ interconnect: 3D torus,
  - based on HyperTransport and proprietary protocols
- 1 ASIC per node, 6 ports, 9.6 GB/sec each
- external 10 PB Lustre file system, 240 GB/s read/write
- reliability & management system has own 100 Mb/s management fabric between blades & cabinet-level controllers
- expected to be upgraded in 2012 with GPU accelerators (20 PF peak)
Current Systems: TianHe-1A: Proprietary + Accelerators (↓)

- 4.7 PFLOPS (peak), 2.3 (LINPACK); 4 MW
- $700 m^2$ area (112 compute, 8 service, 6 comm. & 14 I/O racks)
- 7168 compute nodes: 2 Xeon X5670 hex-core CPUs + NVIDIA M2050 GPU
- service nodes: 2 FT-1000 CPUs (similar architecture to UltraSPARC T3); 2PB storage
- high-radix routing chip, $8 \times 1.2$ GB/s links;
- hierarchical fat tree ($11 \times 384$ port switches, optical fibres) to connect racks ($8 \times 8$ torus)
- closed air cooling (2 aircond units per rack)
- 3-layer power management
- predecessor system, based on VLIW ‘stream processors’ (SPARC ISA)
- successor system (FeiTeng, v3): SIMD multicore chip (SPARC ISA)
Current Systems: Blue Gene/Q: Proprietary + Multicore

- nodes: 1.6GHz PowerPC A2: 18 cores (1 OS, 1 spare), 4-way HMT, quad d.p. SIMD (205 GFLOPS @ 55 watts)
- on-chip network: crossbar to 0.8GHz 32MB L2$
- integrates logic for chip-to-chip communications in a 5D torus
- supports Transactional Memory and speculative execution (!)
- 4-rack system achieved 0.67 PFLOPs (0.84 peak) in Nov 11
- top on Green 500: 165% (77%) more efficient Tianhe-1A (Tsubame 2.0) @ 2 GFLOPS/watt
- Sequoia system at LLNL expected to reach 20 PFLOPs (peak) in 2012 (96 racks, 6 MW)

(courtesy A Register)
13 Outlook and Conclusions

• 2 main swim-lanes: accelerators vs customized multicore
  • accelerators will be integrated on-chip (e.g. Fusion, Knights Corner)
  • in either case, some heterogeneity (cores to support OS, comm.)
• NVIDIA’s vision: the Echelon architecture
  • heterogeneous chips: 28 in-order stream-processors (VLIW with register hierarchy) + 8 post-RISC cores
  • self aware RTS + OS, locality aware compiler + autotuner
• custom (proprietary) components needed for power efficiency
• 100s of cores per chip required: issues for programming models (coherency, core hierarchies) and scalability of communication
• data locality is crucial: deeper hierarchies over several scales expected
• interconnects: fat-tree or high-D torus or combination?
• ‘self-aware’ systems to manage power and reliability (also redundancy)
• latency (esp. for globals) and power are likely the limiting issues