The Future of HPC through Driving Challenges and Enabling Opportunities

Thomas Sterling
Professor, Informatics and Computing
Center for Research in Extreme Scale Technologies
Indiana University
June 27, 2016
Sunway TaihuLight

- Peak: 125 Petaflops
- Cores: 10 Million, SW26010
- Linpack: 93 Petaflops, 73% eff.
- Clock: 1.45 GHz
- Memory: 1.3 Petabytes
- Power: 15.4 Megawatts
- Located: National Supercomputing Center in Wuxi
- Vendor: NRCPC
Node Architecture

- 40,960 nodes
  - System Interface – PCIe, 16 GBps
- Node of 4 core groups
  - NoC
  - System Interface (SI) to external devices
  - 32 Gbytes of DDR3 memory
- Each group has
  - a cluster of 64 computing processing elements (CPE)
    - RISC SIMD architecture 8 ops/cycle
    - 64-bit floating point
    - 11.6 Gflops
    - 64KByte scratchpad, 16 Kbyte IC
  - 1 management processing element (MPE)
    - 23.2 Gflops
  - 1 memory controller (MC)
    - Its own memory space
- Designed by the Shanghai High Performance IC Design Center

Source: HPCwire
40 cabinets, 3.1 Pflops each

4 super-nodes per cabinet
256 nodes per super-node
Architecture Constraints

• Memory
  – Really lightweight
  – 125 Pflops with only 1.3 Petabytes for a ratio of 100:1 inverse capacity

• Bandwidth
  – 22.4 flops/byte of transfer

• HPCG 0.3% peak

• Cache-less
  – Small instruction cache (12KBytes)
  – Small scratchpad (16KBytes)

• Bi-section Band Width of only 70 Tbytes/sec.

• Slow clock rate
Three-segment approximation

$R_{max} \text{ [GFLOPS]}$

$\times 10^7$

Nov 2015

(4.42, 6.22e6)

(36.6, 9.52e5)
Three worlds of supercomputing: average $R_{max}$

- **Leaders**
- **"Knee"**
- **Mainstream**

The graph shows the growth of $R_{max}$ in GFLOPS from June 2006 to November 2015, with distinct trends for each category.
Knights Landing
Next Generation Intel® Xeon Phi™ Product Family

Platform
Memory
Up to 384 GB
DDR4

Knights Landing

Compute
- Intel® Xeon® Processor Binary-Compatible
- 3+ TFLOPS¹, 3X ST² (single-thread) perf. vs KNC
- 2D Mesh Architecture
- Out-of-Order Cores

On-Package Memory
- Up to 16 GB at launch
- Over 5x STREAM³ vs. DDR4 at launch

Integrated Fabric
Processor Package

Fabric (optional)

I/O
Up to 36 PCIe 3.0 lanes

1 Over 3 Teraflops of peak theoretical double-precision performance is preliminary and based on current expectations of cores, clock frequency and floating point operations per cycle. FLOPS = cores x clock frequency x floating-point operations per second per cycle.
²Projected peak theoretical single-thread performance relative to 1st Generation Intel® Xeon Phi™ Coprocessor 7120P (formerly codenamed Knights Corner).
³Projected result based on internal Intel analysis of STREAM benchmark using a Knights Landing processor with 16GB of ultra high-bandwidth versus DDR4 memory only with all channels populated.
⁴Intel internal estimate.
Knights Landing Overview

**Chip:** 36 Tiles interconnected by 2D Mesh
**Tile:** 2 Cores + 2 VPU/core + 1 MB L2

**Memory:**
- MCDRAM: 16 GB on-package; High BW
- DDR4: 6 channels @ 2400 up to 384GB

**IO:**
- 36 lanes PCIe Gen3. 4 lanes of DMI for chipset

**Node:**
- 1-Socket only

**Fabric:**
- Omni-Path on-package (not shown)

**Vector Peak Perf:**
- 3+TF DP and 6+TF SP Flops

**Scalar Perf:**
- ~3x over Knights Corner

**Streams Triad (GB/s):**
- MCDRAM: 400+; DDR: 90+

Source Intel: All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. KNL data are preliminary based on current expectations and are subject to change without notice. 1Binary Compatible with Intel Xeon processors using Haswell Instruction Set (version 1.01); *Bandwidth numbers are based on STREAM-like memory access pattern when MCDRAM memory is primary memory. Results have been estimated based on internal Intel analysis and are not a system level measurement. Any difference in system design or software usage may result in actual performance.

---

**Package:**
- Omni-path not shown

---

**Nodes:**
- 2 VPU
- Core
- 1MB L2
- Core
Intel® Omni-Path Architecture 100 Series

Many new disclosures released: Intel OPA WEBINAR

High Message Rate
1. 195M messages/s per switch port
   - Up to 73 percent higher switch messaging rate per chip compared to InfiniBand EDR

Low Latency
2. Port-to-port latency as low as 100-110ns.
   - 23% lower port-to-port latency than InfiniBand EDR
   - 60% lower switch fabric latency clusters than InfiniBand

Resiliency without Performance Compromises
- Packet integrity protection
- No additional latency for error detection!

---

1 Based on Prairie River switch silicon maximum MPI messaging rate (48-port chip), compared to Mellanox CS7500 Director Switch and Mellanox SB7700/SB7790 Edge switch product briefs (36-port chip) posted on www.mellanox.com as of July 1, 2015.

2 Latency reductions based on Mellanox CS7500 Director Switch and Mellanox SB7700/SB7790 Edge switch product briefs posted on www.Mellanox.com as of July 1, 2015, compared to Intel measured data that was calculated from difference between back to back osu_latency test and osu_latency test through one switch hop.
ANL Aurora

- Cray Shasta architecture
- Over 50,000 nodes
- Peak performance: 180 PFLOPS
- 3rd generation Xeon Phi cores (Knight’s Hill)
- Over 7PB of DRAM and persistent system memory
- Intel interconnect based on 2nd generation Omni-Path architecture with silicon photonics
- 150+PB data storage using Lustre with >1TB/s throughput
- 13MW peak power consumption
- Software environment includes MPI+OpenMP 4.x, Intel compilers and optimization tools, and Cray compilers and libraries
- Cost: $200 million
- Located at Argonne Leadership Computing Facility (ALCF)
- Delivery in 2018 with anticipated start of production phase Q2 2019
The Negative Impact of Global Barriers in Astrophysics Codes

Computational phase diagram from the MPI based GADGET code (used for N-body and SPH simulations) using 1M particles over four time steps on 128 procs.

Red indicates computation
Blue indicates waiting for communication
Normalized scalability ($\frac{S_{\text{max}}}{2}$) vs. normalized parallelism ($\frac{W}{v_0}$) for different velocity $v$ values:

- $v = \frac{1}{4} \times v_0$
- $v = \frac{1}{2} \times v_0$
- $v = v_0$
- $v = 2 \times v_0$
- $v = 4 \times v_0$

The dotted line represents the $\frac{S_{\text{max}}}{2}$ curve.
Amdahl’s Law with Overhead

\[ T_F = \sum_{i}^{n} t_{Fi} \]

\( v \equiv \text{overhead of accelerated work segment} \)

\( V \equiv \text{total overhead for accelerated work} = \sum_{i}^{n} v_i \)

\[ T_A = \left(1 - f\right) \times T_o + \frac{f}{g} \times T_o + n \times v \]

\[ S = \frac{T_o}{T_A} = \frac{T_o}{\left(1 - f\right) \times T_o + \frac{f}{g} \times T_o + n \times v} \]

\[ S = \frac{1}{\left(1 - f\right) + \frac{f}{g} + \frac{n \times v}{T_o}} \]
Head room, margins, potential innovations
All architectures are von Neumann derivatives
Control is sequential instruction issue, IP

• Costs and burdens
  – Variants: out of order, vector, SIMD, MPPs and clusters
  – Flow control bottlenecks
  – Control state limited to program counters, fork-joins
  – Loss of operational precedence
  – Not effective in asynchronous operation

• Alternatives
  – DAGs
  – Dataflow
  – Systolic arrays
  – unums
Head room, margins, potential innovations
Floating point ALU optimized resource

• Costs and burdens:
  • Cache hierarchy
  • Branch prediction
  • Speculative execution
  • Out of order flow control reservation stations, …
  • Prefetching, many simultaneous in-flight requests

• Alternatives:
  • Emphasis on memory access throughput
  • Response time to incidence of external messages
  • Scratch pad memory
  • Multi-threading
  • Dataflow ISA
  • Asynchronous flow control
Head room, margins, potential innovations

• Separation of CPU and main memory
  – Major bottleneck
  – Worse with multi/many core processor sockets
  – A driver for need for cache
  – Processor in Memory (PIM)
  – On-chip scratch pad memory

• Silicon based semiconductor technology
  – Moore’s Law will flat-line by end of decade, ~ 5 nm feature size
  – Superconducting single flux quantum logic at 100 – 200 GHz, 100X energy advantage
  – Leakage current a challenge
  – Graphene of interest

• CSP/MPI (well, not unquestioned)
  – MPI + X, where X = OpenMP maybe
  – Fork-joins impose Amdahl bottlenecks
  – X could also be DAGs
  – Asynchronous Multi-Task execution models
<table>
<thead>
<tr>
<th>Old Technology Characteristics</th>
<th>New Technology Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Peak clock frequency</strong> as primary limiter for performance improvement</td>
<td><strong>Power</strong> is primary design constraint for future HPC system design</td>
</tr>
<tr>
<td><strong>Cost - FLOPs are biggest cost for system:</strong> optimize for compute</td>
<td><strong>Cost - Data movement dominates:</strong> optimize to minimize data movement</td>
</tr>
<tr>
<td><strong>Concurrency - Modest growth</strong> of parallelism by adding nodes</td>
<td><strong>Concurrency:</strong> Exponential growth of parallelism within chips</td>
</tr>
<tr>
<td><strong>Memory scaling maintain byte per flop</strong> capacity and bandwidth</td>
<td><strong>Memory Scaling:</strong> Compute growing 2x faster than capacity or bandwidth</td>
</tr>
<tr>
<td><strong>Locality:</strong> MPI+X model (uniform costs within node &amp; between nodes)</td>
<td><strong>Locality:</strong> must reason about data locality and possibly topology</td>
</tr>
<tr>
<td><strong>Uniformity:</strong> Assume uniform system performance</td>
<td><strong>Heterogeneity:</strong> Architectural and performance non-uniformity increase</td>
</tr>
<tr>
<td><strong>Reliability:</strong> It’s the hardware’s problem</td>
<td><strong>Reliability:</strong> Cannot depend on hardware protection alone</td>
</tr>
</tbody>
</table>
Game Changer – Runtime System

• Runtime system
  – is: ephemeral, dedicated to and exists only with an application
  – is not: the OS, persistent and dedicated to the hardware system

• Moves us from static to dynamic operational regime
  – Exploits situational awareness for causality-driven adaptation
  – Guided-missile with continuous course correction rather than a fired projectile with fixed-trajectory

• Based on foundational assumption
  – More computational work will yield reduced time and lower power
  – Untapped system resources to be harvested
  – Opportunities for enhanced efficiencies discovered only in flight
  – New methods of control to deliver superior scalability
Distinguishing Features of ParalleX/HPX

- **LTI**: local thread instantiation
- **RTI**: remote thread instantiation
- **RAMO**: remote atomic memory operation
- **DTA**: depleted thread activation
- **DOT**: dataflow object trigger
- **FVA**: future value access
- **PERC**: percolation

Legend:
- **Locality**: Thread
- **Process**: Suspended Thread
- **Local Memory**: Local Memory Access
- **LCO**: AGAS Address Lookup
- **Accelerator**: Local Action
- **Parcel**
Performance Model, Full Example System

- Example system:
  - 2 nodes,
  - 2 cores per node,
  - 2 memory banks per node

- Accounts for:
  - Functional unit workload
  - Memory workload/latency
  - Network overhead/latency
  - Context switch overhead
  - Lightweight task management (red regions can have one active task at a time)
  - Memory contention (green regions allow only a single memory access at a time)
  - Network contention (blue region represents bandwidth cap)
  - NUMA affinity of cores

- Assumes:
  - Balanced workload
  - Homogenous system
  - Flat network
Gain with Respect to Cores per Node and Overhead;
Latency of 8192 reg-ops, 64 Tasks per Core

Performance Gain of Non-Blocking Programs over Blocking Programs with Varying Core Counts (Memory Contention) and Overheads
Motivation for HPX

• Exploit runtime information through introspection to discover parallelism for scalability and dynamically manage resources to demand for efficiency
• Expose limitations of conventional computer architecture and devise mechanisms for lower overhead and latency
• Based on a crosscutting execution model to determine respective roles, responsibilities, and interoperability
• Serve as a research platform to explore utility, generality, opportunity, and challenges/limitations
• Target and enabler for parallel programming models
• Operation in the presence of uncertainty of asynchrony
• First conceived in support of HTMT project and Cascade
SpMV for parcels and memget
Wavelet Adaptive Multiresolution

DB: mhd.00251.pdb
Cycle: 5020  Time: 0.980469

Pseudocolor
Var: rho

Max: 3.901
Min: 0.0009508

user: manderson
Mon Jun 22 14:10:19 2015

Courtesy of Matt Anderson, IU
Time Required to Check if Memory Address is Local or Remote in HPX5

Chart courtesy of Daniel Kogler, IU
Time Required to Perform a Context Switch Between Lightweight Threads in HPX5

Histogram of context switch timings

Chart courtesy of Daniel Kogler, IU
Time Required to Create a New Lightweight Thread in HPX5

Histogram of thread create timings

Chart courtesy of Daniel Kogler, IU
Laser Interferometric Gravitational-wave Observatory (LIGO)

Hanford, WA

Livingston, LA
LIGO Chirp Filter for Signal Target
Discovery

- 14 September, 2015
- Combined objects of 29 and 36 solar masses
- Produced a black hole of 62 solar masses.
- Missing 3 solar masses converted to gravitational waves
- Travelled 1.3 billion years to Earth
- 50X all the power of all the stars in the universe