The Path To Exascale – Challenges and Opportunities

EMIT Workshop, Manchester, 30th July 2015

Gaurav Kaul – Solutions Architect

John Swinburne – HPC Architect
Legal Notices and Disclaimers

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY RELATING TO SALE AND/OR USE OF INTEL PRODUCTS, INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT, OR OTHER INTELLECTUAL PROPERTY RIGHT. Intel products are not intended for use in medical, life-saving, life-sustaining, critical control or safety systems, or in nuclear facility applications.

Intel products may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Intel may make changes to dates, specifications, product descriptions, and plans referenced in this document at any time, without notice.

This document may contain information on products in the design phase of development. The information herein is subject to change without notice. Do not finalize a design with this information.

Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined." Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them.

Intel Corporation or its subsidiaries in the United States and other countries may have patents or pending patent applications, trademarks, copyrights, or other intellectual property rights that relate to the presented subject matter. The furnishing of documents and other materials and information does not provide any license, express or implied, by estoppel or otherwise, to any such patents, trademarks, copyrights, or other intellectual property rights.

Wireless connectivity and some features may require you to purchase additional software, services or external hardware.

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, visit Intel Performance Benchmark Limitations

Intel, the Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Other names and brands may be claimed as the property of others.

Copyright © 2014 Intel Corporation. All rights reserved.
INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS". NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO THIS INFORMATION INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Copyright© 2014, Intel Corporation. All rights reserved. Intel, the Intel logo, Atom, Xeon, Xeon Phi, Core, VTune, and Cilk are trademarks of Intel Corporation in the U.S. and other countries.

Optimization Notice

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice revision #20110804
A bit of History ....

Computer performance and application performance increase ~10³ every decade

- ~100 Kilowatts
- ~5 Megawatts
- 20-30 MW
- ~1 Exaflop/s

1.35 Petaflop/s
Cray XT5
150,000 processors

1.02 Teraflop/s
Cray T3E
1,500 processors

1 Gigaflop/s
Cray YMP
8 processors

100 million or billion processing cores (!)

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1988</td>
<td>First sustained GFlop/s Gordon Bell Prize 1988</td>
</tr>
<tr>
<td>1998</td>
<td>First sustained TFlop/s Gordon Bell Prize 1998</td>
</tr>
<tr>
<td>2008</td>
<td>First sustained PFlop/s Gordon Bell Prize 2008</td>
</tr>
<tr>
<td>2018</td>
<td>Another 1,000x increase in sustained performance</td>
</tr>
</tbody>
</table>
The Top 500 Waterfall

Top 500 <15%
All Other Technical Computing >85%

% of sockets sold

Performance Waterfall*
#1 Top500 System to Single Socket
6-8 years
#1 to #500
~9 years
#500 to Single Socket

*plus.....similar waterfalls for other capabilities in areas like fabrics, storage, software, ...

Source: Top500.org and Intel Estimate of Top500 sockets as % of sum of analysts reports of HPC and branded Workstations sockets. Performance waterfall timelines based on TOP500.org statistics (#1-#500) and Intel estimate (#500 to projected Intel Knights Landing)
Other brands and names are the property of their respective owners.
50 years of Moore’s Law

Enabling new devices with higher functionality and complexity while controlling power, cost, and size.
Moore and Dennard Scaling

Moore’s Law: *The number of transistors per chip doubles every 2+ years*

Dennard Scaling:
- Decrease feature size by a factor of $\lambda$ and decrease voltage by a factor of $\lambda$; then
  - # transistors increase by $\lambda^2$
  - Clock speed increases by $\lambda$
  - **Energy consumption does not change**
    (in reality, voltage decrease was slower; clock speed and energy consumption increased faster)
After ~2004 only the number of transistors continues to increase exponentially.

We have hit limits in:
- Power
- Instruction level parallelism
- Clock speed

Single core scalar performance is now only growing slowly.
Technology Scaling Outlook

- **Frequency**
  - "Almost flat -> Parallelism for performance"

- **Transistor Density**
  - "Lots of transistors!"
  - Pollack’s rule:
    - Single-thread perf. \( \sim \sqrt{\text{Transistor count}} \)
    - Manycore for efficient use of transistors

- **Energy**
  - Transistor density will scale better than power
    - Overprovisioned, energy-limited hardware
    - Dynamic execution, fine-grained power management
  - Signaling power will not scale as well as logic
    - Data movements will dominate energy
The Power & Energy Challenge

TFLOP Machine today

5KW
Disk
Com
Memory
Compute

4550W
100W
100W
150W
200W

TFLOP Machine then With Exa Technology

5W
~3W
~5W
2W
5W

~20W
Promising Technologies

- Near Threshold Voltage (NTV)
- Fine-grain power management
- Stacked memory
- Specialized circuits (SIMD, encryption, ...)

- Energy Efficiency
- Voltage Range: Low to High
  - Subthreshold
  - NTV (Normal operating range)
  - ~5x Demonstrated
Rethink System Level Architecture

- Emerging memory technologies
- New levels of memory hierarchy
- Minimize data movement across hierarchy
- Innovative packaging and IO solutions
Revise DRAM Architecture

1. Need exponentially increasing BW (GB/sec)

2. Need exponentially decreasing energy (pJ/bit)

Traditional DRAM

- Activates many pages
- Lots of reads and writes (refresh)
- Small amount of read data is used
- Requires small number of pins

New DRAM architecture

- Activates few pages
- Read and write (refresh) what is needed
- All read data is used
- Requires large number of IO's (3D)
3D-Integration of DRAM and Logic

**Logic Buffer Chip**
Technology optimized for:
- High speed signaling
- Energy efficient logic circuits
- Implement intelligence

**DRAM Stack**
Technology optimized for:
- Memory density
- Lower cost

3D Integration provides best of both worlds
DRAM Scaling Using 3D Memory

1Tb/s HMC DRAM Prototype

- 3D integration technology
- 1Gb DRAM Array
- 512 MB total DRAM/cube
- 128GB/s Bandwidth
- <10 pj/bit energy

<table>
<thead>
<tr>
<th>Bandwidth</th>
<th>Energy Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDR-3 (Today)</td>
<td>10.66 GB/Sec</td>
</tr>
<tr>
<td>50-75 pJ/bit</td>
<td></td>
</tr>
<tr>
<td>Hybrid Memory Cube</td>
<td>128 GB/Sec</td>
</tr>
<tr>
<td>8 pJ/bit</td>
<td></td>
</tr>
</tbody>
</table>

10X higher bandwidth, 10X lower energy

Source: Micron
### Needs a Paradigm Shift

**Past and present priorities**—

<table>
<thead>
<tr>
<th>Single thread performance</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Programming productivity</td>
<td>Legacy, compatibility</td>
</tr>
<tr>
<td></td>
<td>Architecture features for productivity</td>
</tr>
<tr>
<td>Constraints</td>
<td>(1) Cost</td>
</tr>
<tr>
<td></td>
<td>(2) Reasonable Power/Energy</td>
</tr>
</tbody>
</table>

**Future priorities**—

<table>
<thead>
<tr>
<th>Throughput performance</th>
<th>Parallelism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power/Energy</td>
<td>Architecture features for energy</td>
</tr>
<tr>
<td></td>
<td>Simplicity</td>
</tr>
<tr>
<td>Constraints</td>
<td>(1) Programming productivity</td>
</tr>
<tr>
<td></td>
<td>(2) Cost</td>
</tr>
</tbody>
</table>

### Evaluate each (old) architecture feature with new priorities
Intel: Investing to Remove 6 Bottlenecks

- Reliability and Resiliency
- Processor Performance
- Interconnect
- Standard Programming Model for Parallelism
- Memory & Storage
- Power Efficiency

Integration
Impact on Applications

How Can I Achieve High Performance?

How to get benefit from Exascale with your code in the future?

Binomial Options DP

Lot of performance is being left on the table

VP = Vectorized & Parallelized
SP = Scalar & Parallelized
VS = Vectorized & Single-Threaded
SS = Scalar & Single-Threaded
DP = Double Precisions

Modernization of your code is the solution

We believe most codes are here
The Many Ways to Parallelism

Serial Code Node Level
Fast Scalar performance, Optimized C/C++, FORTRAN, Threading and Performance Libraries, Debug / Analysis Tools

Parallel Node Level
Multi-core, Multi-Socket, SSE and AVX instructions, OpenMP, Threading Building Blocks, Performance Libraries, Thread Checker, Clk

Multi-Node / Cluster Level
Cluster Tools, MPI Checker
And New Workloads will Emerge
Code Modernization – The 4D Approach

Serial and Scalar
- Choose Proper compiler option
- Use Optimized Library (Intel® Math Kernel Library)
- Choose right precision ……..
- Remove IO bottleneck
- Remove unnecessary computing

Parallelism
- Choose the proper parallelization method
- Load balance
- Synchronization overhead
- Thread Binding

Vectorization
- Auto-vectorization
- Intel® Cilk™ Plus Array Notations
- Elemental functions
- Vector class
- Intrinsics

Memory Access
- Data Alignment
- Prefetch
- Cache Blocking
- Data restructure: aos2soa
- Streaming Store
Intel® Xeon Phi™ Product Family
Based on Intel® Many Integrated Core (MIC) Architecture

“Meet Knight's Landing: Intel's most powerful chip ever is overflowing with cutting-edge technologies”
PCWorld – Jun 23, 2014

2013
Knights Corner
Intel® Xeon Phi™
x100 Product Family
22 nm process
Coprocessor
Over 1 TF DP Peak
Up to 61 Cores
Up to 16GB GDDR5

2016
Knights Landing
Intel® Xeon Phi™
x200 Product Family
14 nm process
Processor &
Coprocessor
Over 3 TF DP Peak
Up to 72 Cores
On Package High-
Bandwidth Memory
3X Single-Thread
Out-of-order core

Future
Knights Hill
Next generation of
the Intel® MIC
Architecture Product
Line
In planning

*Per Intel's announced products or planning process for future products
No Positioning Change
Knights Landing Targeted for Highly-Vectorizable, Parallel Apps

Most Commonly Used Parallel Processor*
Parallel, Fast Serial
Multicore + Vector
Leadership Today and Tomorrow

Optimized for Highly-Vectorizable Parallel Apps
Many Core
Support for 512 bit vectors
Higher memory bandwidth
Common SW programming

*Based on highest volume CPU in the IDC HPC Qview Q1'13
Is Xeon Phi compelling vs Xeon?

"Rifle shot" approach targeted with customers in FSI, Oil & Gas, and Life Sciences based on affinity/readiness

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Source: Intel measured as of Q1 2014  Configuration Details: Please slide speaker notes. For more information go to http://www.intel.com/performance

Intel® Xeon® processor
Xeon Phi = Intel® Xeon Phi™ coprocessor

Intel® Xeon Phi™ coprocessor vs. 2S Intel® Xeon® Processor

Higher is Better

DCC Energy
Financial Services
Life Sciences
Manufacturing
PhysicsWeather

Relative performance vs. 2S Intel® Xeon® processor

Intel Measured Results: Different hardware architectures may require different source code. Results are based on Intel best efforts to use code optimized to run best on all architectures and perform the same work. Future code optimizations may result in different results.
Three Knights Landing Products

Knights Landing Processor
"Self-boot" Intel® Xeon Phi™ processor platform

- Knights Landing is the host processor
- Boots standard off-the-shelf OS’s

Benefits:
- Higher performance density for highly parallel applications
- Reduced system power consumption
- Higher perf/Watt & perf/$$

Knights Landing Coprocessor:
- Solution for general purpose servers and workstations

Benefits:
- Targeted for applications with larger sections of serial work
- Upgrade path from Knights Corner as PCIe card

1. Projections based on early product definition and as compared to prior generation Intel® Xeon Phi™ Coprocessors
2. Based on Intel internal analysis. Lower power based on power consumption estimates between (2) HCA’s compared to 15W additional power for KNL-F. Higher density based on removal of PCIe slots and associated HCA’s populated in those slots.
3. Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance. Results based on internal Intel analysis using estimated theoretical Flops for KNL processors, along with estimated system power consumption and component pricing in the 2015 timeframe. See backup for complete system configurations.
A Paradigm Shift for Highly-Parallel Server Processor and Integration are Keys to Future

- **Memory Bandwidth**: ~500 GB/s STREAM
- **Memory Capacity**: Over 25x* KNC
- **Resiliency**: Systems scalable to >100 PF
- **Power Efficiency**: Over 25% better than card¹
- **I/O**: Up to 50 GB/s BW with int fabric
- **Cost**: Less costly than discrete parts¹
- **Flexibility**: Limitless configurations
- **Density**: 3+ KNL with fabric in 1U³

*Comparison to 1st Generation Intel® Xeon Phi™ 7120P Coprocessor (formerly codenamed Knights Corner)

¹Results based on internal Intel analysis using estimated power consumption and projected component pricing in the 2015 timeframe. This analysis is provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

²Comparison to a discrete Knights Landing processor and discrete fabric component.

³Theoretical density for air-cooled system; other cooling solutions and configurations will enable lower or higher density.
Knights Landing Architectural Diagram

- **2x 512b VPU per core** (Vector Processing Units)
- **Based on Intel® Atom Silvermont processor with many HPC enhancements**
  - Deep out-of-order buffers
  - **Gather/scatter** in hardware
  - Improved branch prediction
  - **4 threads/core**
  - High cache bandwidth & more

**Up to 72 cores** 
2D mesh architecture

- **6 channels DDR4**
- **Up to 384GB**
- **Common with Grantley PCH**

- **2 ports Storm Lake Integrated Fabric**
- **On-package 50 GB/s bi-directional**

- **PCIe Gen3 x36**

- **DMI**

**Up to 16GB high-bandwidth on-package memory (MCDRAM)**
- Exposed as NUMA node
- ~500 GB/s sustained BW

- **Wellsburg PCH**
- **Integrated Fabric On-package**

- **~3 TF DP peak**
- Full Xeon ISA compatibility through AVX-512
- ~3x single-thread vs. compared to Knights Corner

- **Full 2 ports Storm Lake**
- **HFI**
- **Connect**

- **50 GB/s bi-directional**
- **~500 GB/s**

- **Up to 72 cores**

**Tile**

- **2 VPU HUB 2 VPU**
- **1MB L2 Core**
- **2x 512b VPU per core** (Vector Processing Units)

**Diagram is for conceptual purposes only and only illustrates a CPU and memory - it is not to scale and does not include all functional areas of the CPU, nor does it represent actual component layout.**
Today’s Parallel Investment Carries Forward

Sustained threading, vectorization, cache-blocking and more

**MOST** optimizations carry forward with a recompile

**Incremental** tuning gains

Recompile

MKL
MPI
TBB
OpenMP
Cilk Plus™
OpenCL

KNL Enabled Performance Libraries & Runtimes

AVX-512
Cache Mode For High Bandwidth Memory

KNL Enabled Compilers

Recompile

KNL Enhancements (memory, architecture, bandwidth, etc.)

Native or Symmetric or Offload

1st Generation Intel® Xeon Phi™ Coprocessor

Knights Landing

Intel® Xeon Phi™
Parallel is the Path Forward
Intel® Xeon® and Intel® Xeon Phi™ Product Families are both going parallel

<table>
<thead>
<tr>
<th>Core(s)</th>
<th>Threads</th>
<th>SIMD Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>128</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>128</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>128</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>128</td>
</tr>
<tr>
<td>8</td>
<td>16</td>
<td>256</td>
</tr>
<tr>
<td>12</td>
<td>24</td>
<td>256</td>
</tr>
<tr>
<td>tbd</td>
<td>tbd</td>
<td>512</td>
</tr>
</tbody>
</table>

Future Intel® Xeon® processor
- Knights Corner
- Knights Landing

<table>
<thead>
<tr>
<th>Core(s)</th>
<th>Threads</th>
<th>SIMD Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>57-61</td>
<td>Up to 72</td>
<td>512</td>
</tr>
<tr>
<td>228-244</td>
<td>Up to 288</td>
<td>512</td>
</tr>
</tbody>
</table>

More cores → More Threads → Wider vectors

*Product specification for launched and shipped products available on ark.intel.com. 1. Not launched or in planning.
Intel® Omni-Path Fabric - CPU Integration

**Integration Value Vectors:**
- Greater Density
- Better & Balanced Performance
- Reduced Cost
- Reduced Power
- Improved Reliability

**Time**

<table>
<thead>
<tr>
<th>Time</th>
<th>Technology</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HSW</td>
<td>PCIe HFI</td>
</tr>
<tr>
<td></td>
<td>BDW</td>
<td>MCP Integration</td>
</tr>
<tr>
<td></td>
<td>KNL</td>
<td>MCP Integration (Preserves PCIe Lanes)</td>
</tr>
<tr>
<td></td>
<td>Next Xeon</td>
<td>Tight Integration</td>
</tr>
<tr>
<td></td>
<td>Next Xeon-Phi</td>
<td>Additional integration, improvements and features</td>
</tr>
</tbody>
</table>

**MCP = Multi Chip Package**
Intel’s Technical Computing Portfolio
Technologies & Products
HW-SW Co-design

Applications and SW stack provide guidance for efficient system design

Applications
System SW
Programming Sys
Architecture
Circuits & Design

Limitations, issues and opportunities to exploit
## What will matter in 10 years

<table>
<thead>
<tr>
<th></th>
<th>Now</th>
<th>2025</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perf/$</td>
<td>Linpack, Real Applications</td>
<td>Real Applications</td>
</tr>
<tr>
<td>Perf/Watt</td>
<td>Limited by worst case application</td>
<td>All applications will be able to run at chosen power level. Dynamic, optimal energy management.</td>
</tr>
<tr>
<td>Reliability</td>
<td>Use of file system checkpoint restart (spinning disks)</td>
<td>Transparent hardware and system software recovery. Checkpoints in non-mechanical media.</td>
</tr>
<tr>
<td>Big Data</td>
<td>Parallel IO</td>
<td>New storage paradigm</td>
</tr>
</tbody>
</table>
SW Challenges

1. Extreme parallelism (1000X due to Exa, additional 4X due to NTV)
2. Data locality—reduce data movement
3. Intelligent scheduling—move thread to data if necessary
4. Fine grain resource management (introspective)
5. Applications and algorithms incorporate paradigm change
Summary

- Exascale will be there by 2022 or so
- "Business as usual" (riding on Moore's Law and commodity technology) is becoming increasingly harder
- Supercomputers are becoming more "special purpose"
  - Expect most/all supercomputers to use floating point accelerators in a few years; more specialized accelerators to follow
- Can continue to push performance to zetascale
  - Will need to think of supercomputers as unique facilities, such as particle accelerators – not clusters of PCs
- *Supercomputing will become much more interesting*