ICE X

GPUs, MICs and Cooling

Reiner Vogelsang
reiner@sgi.com
Safe Harbor Cautionary Statement
This presentation contains forward-looking statements, including statements regarding management’s expectations about the markets, business, products, operating plans and financial performance of Silicon Graphics International Corp. (“SGI”) as of November 14, 2013. Statements containing words such as "will," "expect," "believe," and "intend," and other statements in the future tense, are forward-looking statements. Any statements contained in this presentation that are not statements of historical fact may be deemed forward-looking statements. Actual outcomes and results may differ materially from the expectations expressed or implied in these statements due to a number of risks and uncertainties, including, but not limited to: SGI operates in a very competitive market which may cause pricing pressure and impair our market penetration; SGI has extensive international business activities which create risks from complex international operations, foreign currency exposure and changing legal, regulatory, political or economic conditions, including SGI’s operations in Japan, which may be negatively affected by earthquakes and other natural disasters, as well as ongoing power supply disruptions, earthquake and tsunami; uncertainty arising from SGI’s increased dependence on business with U.S. Government entities; failure of our customers to accept new products; and economic conditions impacting the purchasing decisions of SGI’s customers. Detailed information about these and other risks and uncertainties that could affect SGI’s business, financial condition and results of operations is set forth in SGI’s Annual Report on Form 10-K under the caption “Risks Factors,” which was filed with the Securities and Exchange Commission on August 7, 2013, as updated by the subsequent filings with the SEC made by SGI, all of which are available at www.sec.gov. Accordingly, you are cautioned not to place undue reliance on forward-looking statements. This presentation is as of November 14, 2013, and the continued posting or availability of this presentation does not imply that forward-looking statements continue to be true as of any later date. We expressly disclaim any obligation to update or alter our forward-looking statements, whether, as a result of new information, future events or otherwise, except where required by law.

The SGI logos and SGI product names used or referenced herein are either registered trademarks or trademarks of Silicon Graphics International Corp. or one of its subsidiaries. All other trademarks, trade names, service marks and logos referenced herein belong to their respective holders. Any and all copyright or other proprietary notices that appear herein, together with this Legal Notice, must be retained on this presentation.
Contents

- SGI HPC Overview
- SGI ICE X
  - SPEED, SCALE AND EFFICIENCY
- GPU and MIC usage example
SGI Servers for Speed, Scale & Efficiency

Technical Computing

- Commercial
- Scientific
- Modeling & Simulation

Big Data

- Hadoop
- In-memory
- Analytics
- Archive

SGI® ICE™ X  Rackable®  SGI® UV™ 2000  Rackable®  SGI® InfiniteData™ Cluster
Total @ Pau, France

- 6912 Nodes in 12 M-Cells
- 24° Celsius direct watercooling
- 2.3PFlop/s
Features per Enclosure Pair:

- 36 blade slots
- Four fabric switch slots
- Four chassis management controller slots
- Unified backplane
- 19” rack mount enclosure*

Separable Power Shelf

One Blade Slot

17.7

21U “Building Block”

Unified Backplane

16.59 (9.5U)

1.75 (1U)
SGI ICE X: IP-113 Compute Blade (Dakota) for “D-Rack” or “M-Rack”

- Two Intel® Xeon® processor E5-2600 v2 series
- Up to eight DDR3 DIMMs per socket @ 1600 MT/s
  - 1866 MT/s support w/ 1 DPC populated
- Up to two 2.5” SATA HDD/ SSD drives
- Traditional heat sinks
- One FDR IB mezzanine card

One dual socket node per one blade slot
SGI ICE X: IP-115 Compute Blade (Gemini Twin) for “M-Rack”

- Four Intel® Xeon® processor E5-2600 v2 series
- Four DDR3 DIMMs/ socket @ up to 1866MT/s
- FDR InfiniBand - single plane
- Up to two 2.5” SATA HDD/ SSD drives
- Liquid cold sinks

**Two** dual socket nodes in **one** blade slot!

One of Two Nodes per Blade Slot
(Bottom node with cold sink shown)
SGI ICE X Compute Node: IP-119 (MedinaQ Twin) for “M-Rack”

- Two Intel® Xeon® processor E5-2600 v2 series
- Two Intel® Xeon Phi™ 5120D Coprocessor
- FDR InfiniBand - single or dual plane
- Four DDR3 DIMMs/ socket @ up to 1866 MT/s
- Up to one 2.5” SATA HDD/ SSD drive
- Liquid cold sinks

**One** dual socket node w/ **two** coprocessors in **one** blade slot!

FDR + 1:1 processor to coprocessor ratio = **Balanced Throughput**
Rack Environmental Options:

Traditional Data Center: D-Rack

Large Scale HPC:

Container Environments: Leverage Cell Solutions
SGI ICE X in M-Rack

**Main Features:**

- 36 blade slots/enclosure pair (72/rack)
- “Single” node blades utilize heat sinks & 10 power supplies/enclosure pair
- “Twin” node blades utilize “cold sinks” & 9 power supplies (8+1)/enclosure
  - Full Socket SKU Stack Support w/ Pdyn 1.2
- Scalable power shelf design
- Deploys in an “M-Cell”
  - Closed loop air flow
  - Warm water support
  - Standard casters and levelers
- 42U - 33”W x 44”D x 78.75
SGI ICE X – M-Rack Cell

Compute and Cooling Racks
Support 480VAC 3-phase electrical feeds
- Reduces cost of the wiring
- Lowers resistive losses

Exterior Trim shown translucent
(to expose internal features)

Cooling-Rack Roof Mounted
Controller not shown

M-Rack Roof Mounted PDUs
not shown
M-Rack ‘Cube’ Cell - Dimensions & Airflow Circulation
(I/O Rack Not Shown)

Cooling Distribution Unit (CDU):
Used when ‘cold sinks’ on blades

CDU
(38W x 34D)
SGI® ICE X – M-Rack Cell CDU Heat Exchanger

CDU Heat Exchanger - Nominal:

115 GPM (26.1 m³/hr) Facilities Flow
12°F (6.7°C) Facilities Temp Rise
115 GPM (26.1 m³/hr) M-Rack Cell Flow
12°F (6.7°C) Facilities Temp Decrease
8°F (4.4°C) “Approach”

57°F (13.9°C) Facilities Supply / Return

45°F (7.2°C)

65°F (18.3°C) M-Rack Supply / Return

53°F (11.7°C)

Facilities
Supply
Return

M-Rack Supply
Return

45°F 7.2°C 57°F (13.9°C)
50°F (10.0°C) 62°F (16.7°C)
55°F (12.8°C) 67°F (19.4°C)
60°F (15.6°C) 72°F (22.2°C)
65°F (18.3°C) 77°F (25.0°C)
70°F (21.1°C) 82°F (27.8°C)
75°F (23.9°C) 87°F (30.6°C)
80°F (26.7°C) 92°F (33.3°C)
85°F (29.4°C) 97°F (36.1°C)
90°F (32.2°C) 102°F (38.9°C)

53°F (11.7°C) 65°F (18.3°C)
58°F (14.4°C) 70°F (21.1°C)
63°F (17.2°C) 75°F (23.9°C)
68°F (20.0°C) 80°F (26.7°C)
73°F (22.8°C) 85°F (29.4°C)
78°F (25.6°C) 90°F (32.2°C)
83°F (28.3°C) 95°F (35.0°C)
88°F (31.1°C) 100°F (37.8°C)
93°F (33.9°C) 105°F (40.6°C)
98°F (36.7°C) 110°F (43.3°C)

M-Rack supply temperature increases linearly with increasing Facilities supply temperature.
PUE 1.067

- **ICE X @ DLR Braunschweig**
  - Processors: 1120 Intel E5-2695 v2 12-core @ 2.4 GHz (13440 cores)
  - Memory: 71.8 TB
  - Interconnect: FDR Infiniband
  - Operating System: Suse Enterprise Linux
  - Rpeak: 258,048 GFLOPS
  - Rmax: 245,876 GFLOPS
  - Nmax: 1,885,596
  - Power Consumption: 238.64 kW
  - TUE: 1.38
  - Perf./Power: 1.038 GFLOPS/W
  - **Warm water cooling:** 30° Celsius
SGI ICE X

TOPOLOGY AND SWITCH BLADE DESIGNS
## Common InfiniBand Topologies

<table>
<thead>
<tr>
<th>All-to-All</th>
<th>Fat Tree (CLOS Networks)</th>
<th>Hypercube</th>
<th>Enhanced Hypercube</th>
</tr>
</thead>
</table>

*All Supported on SGI ICE X*
SGI ICE X

SOFTWARE/ SYSTEM MANAGEMENT DETAIL
SGI® Management Center for SGI® ICE X

- Efficient scalability and manageability of all resources
  - Innovative Hierarchical Management Network
  - Enables administrators to install, configure, provision, and manage a system which scales up to 10,000+ nodes
  - Improved latency providing performance benefits
  - GigE dedicated for system and job management, while InfiniBand dedicated for running applications creating a fast-track for the applications

- Reduce administrative tasks associated with cluster management
  - Boot an entire rack of 144 diskless compute blades in ~5 minutes
  - Image management
  - Powerful scripting through CLI

- Flexible OS Provisioning
  - Install RHEL or SLES on System Admin and Rack Leader Nodes
  - Compute nodes can run different releases of RHEL, SLES or CentOS
# Dynamic, Fine Grained Energy Management

| Monitoring | Real-time monitoring of actual power and inlet temp data\* aggregated to rack, system, and user-defined logical groups  
Provides delta inlet temperature in rack for thermal monitoring  
Aggregated Energy/Thermal events: Pre-defined and user defined |
|-----------|---------------------------------------------------------------------------------------------------------------|
| Trending  | Log power and thermal data  
Save history data for capacity planning  
Query for key metric trends e.g. Max/Min/Avg rack power, energy cost, etc. |
| Control   | Supports multiple power policy types,  
✓ Schedule by time of day  
✓ Custom – Maintains group power, dynamically manage power to changing server workloads. Minimizes group performance impact  
✓ Minimum – Puts group to minimum energy  
Co-existing policies at multiple hierarchy levels e.g. rack, group  
Maintain group energy limit during server boot |
| Scalability | Manage across datacenter hierarchy (rack, system, group)  
Enterprise scaling and support (supports up to 1000s nodes) |
SGI ICE X
SUMMARY
SGI ICE X Summary

- Twin-Node Compute Blade Options
  - 36 compute nodes per Enclosure
  - ‘On Node’ accelerator support
- Independently Scalable Power Architecture
- Separable Cooling Solutions
  - Air-cooled for lower power blades
  - Options for open-loop chilled-water or closed-loop warm water cooling
  - Options for integrated liquid cooling (cold sinks) at the node level
- Unified Backplane Fabric
  - Single enclosure pair solution
  - Fabric switches enable topology choices
- Multiple Data Center Environment Support
  - Traditional hot/cold aisle
  - High density closed-loop cooling
  - Large scale container solution
Kepler GPU Architecture – Evolving for HPC

- 15 SMX - Streaming Multiprocessors
- 192 Processing cores
- Local fast memory

GPU Chip (device)

L2 Cache

Memory ~6GB

GPU Board

~8GB/s PCIe Gen2 x16 Bus 250 GB/s
Optimization for Xeon Phi

- Performance analysis showed
  - Inadequate vectorization from misalignment and loop peeling
  - Large memory footprint not fitting in L1 and L2 caches
  - Memory latency bound and not saturating memory bandwidth

Optimizations

1. Break outer loop over \( j \) into small chunks. More thread parallelism; smaller footprint per thread.
2. Compute using thread-private statically sized arrays. Improved vectorization.
3. Combine/eliminate temporaries to reduce footprint from 100KB \( \rightarrow \) 60KB thread. More threads/core hide memory latency.

- Host CPU and Intel Phi run identical source code
- Effort optimizing for Phi benefits host
## Whole code performance

<table>
<thead>
<tr>
<th>CONUS 12km*</th>
<th>w/o wsm5 optim.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 node Xeon (2x SNB-EP, 16 cores):</td>
<td>109.2s</td>
</tr>
<tr>
<td>1 node Intel Phi (KNC, 61 cores):</td>
<td>109.6s</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CONUS 2.5km*</th>
<th>with and w/o Phi</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 nodes 2x SNB per node</td>
<td>328.4s</td>
</tr>
<tr>
<td>8 nodes 2x SNB + KNC per node</td>
<td>223.5s</td>
</tr>
</tbody>
</table>

- **Standard 12km and 2.5km benchmark cases**
  - Times in seconds for 3 hour simulations, compute-only
  - [http://www.mmm.ucar.edu/wrf/WG2/bench](http://www.mmm.ucar.edu/wrf/WG2/bench)

- **CONUS 12km** (*Indraneil Gokhale, Intel Corp.*)
  - Phi performance equivalent to two Xeon SNB-Eps
  - Phi-optimized WSM5 code yielded 10-13% improvement on both systems

- **CONUS 2.5** (*Alexander Knyazev, Intel Corp.*)
  - Comparison of run times with and without using Phi on nodes
  - Here, the Phi card adds the equivalent of another Xeon SNB-EP processor
Intel® Xeon Phi™ Coprocessor Performance Scaling
Berlin Quantum ChromoDynamics (BQCD)* with multiple nodes

SE10 Performance

Higher Is Better
- Great Node Scaling
- Great Symmetric Performance (E5-2670 + SE10 working together)

BQCD Performance (CF/s)

1 Node 2 Nodes 4 Nodes 8 Nodes
21.8 46.2 88.5 172.5
31.4 54.5 99.0 176.5
44.1 90.9 168.5 273.9

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 Jan 2013. Configuration Details: Please see backup slide. For more information go to http://www.intel.com/performance

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