Intel® Xeon Phi™ Processor Family (formerly codenamed "Knights Landing")

Performance Fact Sheet

Unlock deeper insights for business innovation and scientific research with the Intel® Xeon Phi™ processor

June 20, 2016 - The new platforms based on the Intel® Xeon Phi™ product family x200 enables customers to gain deeper insights to pursue new discoveries, drive business innovations or shape the future using advanced analytics. One key to unlocking these deeper insights is the new Intel Xeon Phi processor – the first offering to deliver the performance of an accelerator with all the benefits of a server-class processor.

By contrast, special-purpose offload accelerators, like GPUs, often remain underutilized since applications are not always suitable or optimized for them. Moreover, standardizing on Intel® architecture means you can use a single programming model for all your code, reducing operational and programming expenses through a shared developer base and code reuse.

Based on Intel internal analysis, using the new Intel Xeon Phi processor delivers better performance, performance per watt and performance per dollar and can run full simulations as compared to the NVIDIA Tesla* GPUs. For example, using the life sciences modeling found in the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS)* classical molecular dynamics workload, a single server using the Intel Xeon Phi Processor 7250 (16GB, 1.40 GHz, 68 core) compared to a hosted NVIDIA Tesla* K80 GPU[1] delivers up to:

- 5 times more steps per second to simulate complex models faster
- 8 time higher performance per watt for lower energy costs per solution for compelling value
- 9 times better performance per dollar to maximize future potential needs

FUJITSU

Fujitsu* has achieved a world-record, single-processor result on the technical computing applications benchmark, SPECfp*_rate2006, demonstrating the Intel Xeon Phi processor 7250 can solve the most complex data challenges faster and with greater efficiency using the compatible binaries used on all Intel Xeon processors to support the broadest set of workloads for maximum asset utilization across the data center.[7]

The Intel Xeon Phi processor is a true evolution in design and architecture that delivers the performance of an accelerator with the benefits of a server-class processor for your most demanding tasks. With proven scaling proven at up to 128-nodes delivering 50x faster training versus a single-node[2] on Intel Xeon Phi processors, leadership is demonstrated on many other benchmark and application workloads with better performance over NVIDIA Tesla* that fuel breakthroughs in science and industry:

- Deep learning image classification training scalability – up to 2.3 times faster training per system[3] with up to 38% better scaling efficiency at 32-nodes[4]
- High performance ray tracing visualization – up to 5.1 times faster renderings at 7.6 times better performance per dollar[5]
- Financial risk modeling – up to 2.7 times more options evaluated per second, 2.8 times better performance per watt and 5.1 times superior performance per dollar value[6]

See more performance, performance per watt and performance per dollar advantages for Intel Xeon Phi processors at http://www.intel.com/performance/datacenter.

Configurations and Disclaimers

[1] Up to 5x more timesteps per second, 8x higher performance per watt and 9x better performance per dollar claims based on LAMMPS* Course-Grain Water Simulation using Stillinger-Weber* potential comparison of:

- BASELINE CONFIGURATION: Dual Socket Intel® Xeon® processor E5-2697 v4 (45 M Cache, 2.3 GHz, 18 Cores) with Intel® Hyper-Threading and Turbo Boost Technologies enabled, 128 GB DDR4-2400 MHz memory, Red Hat Enterprise Linux* 6.7 (Santiago), Intel® Omni-Path Host Fabric Interface Adapter 100 Series 1 Port PCIe* x16, Intel® Server Board S2600WT2R, BMC 1.33.9832, FRU/SDR Package 1.09, 1.0 TB SATA drive WD1003FZEX-00MK2A0 System Disk + one NVIDIA Tesla* K80 GPUs, NVIDIA CUDA* 7.5.17 (Driver:

352.39), ECC enabled, persistence mode enabled. Number of MPI tasks on host varied to give best performance. CUDA MPS* used where possible. Mean Benchmark System Power Consumption: 683W. Estimated list price including host: $13,750 source http://www.colfax-intl.com/ND/Servers/CX1350s-XK6.aspx.

- NEW CONFIGURATION: One node Intel Xeon Phi processor 7250 (16 GB MCDRAM, 1.4 GHz, 68 Cores) in Intel® Server System LADMP2312KXXX41, 96GB DDR4-2400 MHz, quad cluster mode, MCDRAM flat memory mode, Red Hat Enterprise Linux* 6.7 (Santiago) running Intel® Compiler 16.0.2, Intel® MPI 5.1.2.150, Optimization Flags: "-O2 -fp-model fast=2 -no-prec-div -qoverride-limits", Intel® Omni-Path Host Fabric Interface Adapter 100 Series 1 Port PCIe x16, 1.0 TB SATA drive WD1003FZEX-00MK2A0 System Disk. Mean Benchmark System Power Consumption: 378W. Estimated list price: $7300 source Intel Recommended Customer Pricing (RCP).

[2] Up to 50x faster training on 128-node as compared to single-node based on AlexNet* topology workload (batch size = 1024) training time using a large image database running one node Intel Xeon Phi processor 7250 (16 GB MCDRAM, 1.4 GHz, 68 Cores) in Intel® Server System LADMP2312KXXX41, 96GB DDR4-2400 MHz, quad cluster mode, MCDRAM flat memory mode, Red Hat Enterprise Linux* 6.7 (Santiago), 1.0 TB SATA drive WD1003FZEX-00MK2A0 System Disk, running Intel® Optimized DNN Framework, training in 39.17 hours compared to 128-node identically configured with Intel® Omni-Path Host Fabric Interface Adapter 100 Series 1 Port PCIe x16 connectors training in 0.75 hours. Contact your Intel representative for more information on how to obtain the binary. For information on workload, see https://papers.nips.cc/paper/4824-Large image database-classification-with-deep-convolutional-neural-networks.pdf.

[3] Up to 2.3x faster training per system claim based on AlexNet* topology workload (batch size = 1024) using a large image database running 4-nodes Intel Xeon Phi processor 7250 (16 GB MCDRAM, 1.4 GHz, 68 Cores) in Intel® Server System LADMP2312KXXX41, 96GB DDR4-2400 MHz, quad cluster mode, MCDRAM flat memory mode, Red Hat Enterprise Linux* 6.7 (Santiago), 1.0 TB SATA drive WD1003FZEX-00MK2A0 System Disk, running Intel® Optimized DNN Framework (internal development version) training in 10.5 hours compared to 1-node host with four NVIDIA "Maxwell" GPUs training in 25 hours (source: http://www.slideshare.net/NVIDIA/gtc-2016-opening-keynote slide 32).

[4] Up to 38% better scaling efficiency at 32-nodes claim based on GoogLeNet deep learning image classification training topology using a large image database comparing one node Intel Xeon Phi processor 7250 (16 GB MCDRAM, 1.4 GHz, 68 Cores) in Intel® Server System LADMP2312KXXX41, DDR4 96GB DDR4-2400 MHz, quad cluster mode, MCDRAM flat memory mode, Red Hat* Enterprise Linux 6.7, Intel® Optimized DNN Framework with 87% efficiency to unknown hosts running 32 each NVIDIA Tesla* K20 GPUs with a 62% efficiency (Source: http://arxiv.org/pdf/1511.00175v2.pdf showing FireCaffe* with 32 NVIDIA Tesla* K20s (Titan Supercomputer*) running GoogLeNet* at 20x speedup over Caffe* with 1 K20).

[5] Up to 5.1x faster renderings at 7.6 times better performance per dollar claim based on frames per second (FPS) results with a 1024x1024 image workloads with Intel Embree 2.10.0 using one node Intel Xeon Phi processor 7250 (16 GB MCDRAM, 1.4 GHz, 68 Cores) in Intel® Server System LADMP2312KXXX41, 96GB DDR4-2400 MHz, quad cluster mode, MCDRAM flat memory mode, Red Hat* Enterprise Linux 6.7 scoring 32.5 FPS with an estimated list price of $7,300 compared to hosted NVIDIA Titan X* GPU scoring 6.28 FPS with an estimated list price of $13,750.

[6] Up to 2.7x more options evaluated per second, 2.8 times better performance per watt and 5.1 times superior performance per dollar value claims based on Monte Carlo DP workload results comparing one node Intel Xeon Phi processor 7250 (16 GB MCDRAM, 1.4 GHz, 68 Cores) in Intel® Server System LADMP2312KXXX41, 96GB DDR4-2400 MHz, quad cluster mode, MCDRAM flat memory mode, CentOS* 7.2, quadrant cluster mode, MCDRAM flat memory mode scoring 4.43M options/second, 355 average watts and estimated list price of $7,300 compared to Supermicro* SYS-1028GR-TR server using Intel® Xeon® processor E5-2699 v4 (55 MB Cache, 2.2 GHz, 22 Cores) with Intel Hyper-Threading and Turbo Boost Technologies enabled, 256 GB DDR4-2133 MHz, Red Hat Enterprise Linux* 7.1 (Maipo) plus NVIDIA Tesla K80* scoring 1.62M options/second, 358 average watts (host system was essentially idle) and estimated list price of $13,750.

[7] World record claim based on a SPECfp*_rate2006 base score of 842 and peak score of 870 submitted to SPEC.org (considered estimated until published) compared to all other 1-chip results published at https://www.spec.org/cpu2006/results/rfp2006.html as of 14 June 2016. Configuration: Fujitsu PRIMERGY* CX1640 M1 using Intel® Xeon Phi™ processor 7250 (16 GB MCDRAM, 1.4 GHz, 68 Cores) with 192 GB memory, Red Hat Enterprise Linux* 7.2 (3.10.0-327.13.1.el7.mpsp_1.3.2.100.x86_64) running Intel Intel® C++ and Fortran Compilers 16.0.2.181.