Deep Learning Training At Scale
Spring Crest Deep Learning Accelerator
(Intel® Nervana™ NNP-T)

Andrew Yang | 8/16/2019
This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer. No computer system can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit http://www.intel.com/performance.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future, are forward-looking statements that involve a number of risks and uncertainties. A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.

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. For more complete information visit: http://www.intel.com/performance. The products described 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.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

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.

All products, computer systems, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.

Performance results are based on testing as of August 1, 2019 and may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No product or component can be absolutely secure. Check with your system manufacturer or retailer or learn more at http://www.intel.com/.

Intel, the Intel logo, Intel Inside, Nervana, and others are trademarks of Intel Corporation in the U.S. and/or other countries.

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

© 2019 Intel Corporation.
Deep Learning Training

- Real-world performance - time to train and power efficiency
- Compute used for largest models & training sets doubles every 3.5 months**
- GEMMs and Convolutions dominate majority of computation required for Deep Neural Networks.
- Primary factors that drive a DL training accelerator
  - Power
  - Compute
  - Memory & Communication
  - Scale-out

* Based on analysis of ResNet, FaceNet, Open NMT
**https://openai.com/blog/ai-and-compute/
Spring Crest (NNP-T) Architecture Direction

- Train a network as fast as possible within a given power budget, targeting larger models and datasets
- Balance between Compute, Communication, & Memory
- Re-use on-die data as much as possible
- Optimize for batched workloads
- Built-in scale-out support
- Support future workloads
- PCIe Gen 4 x16 EP
- 4x HBM2
- 64 lanes SerDes
- 24 Tensor Processors
- Up to 119 TOPS
- 60 MB on-chip distributed memory
- Management CPU and Interfaces
- 2.5D packaging
SPRING CREST (NNP-T) IMPLEMENTATION

- TSMC CLN16FF+
- 680mm$^2$, 1200mm$^2$ interposer
- 27 Billion Transistors
- 60mm x 60mm/6-2-6 3325 pin BGA package
- 4x8GB HBM2-2400 memory
- Up to 1.1Ghz core frequency
- 64 lanes SerDes HSIO up to 3.58Tbps aggregate BW
- PCIe Gen 4 x16, SPI, I2C, GPIOs
- PCIe & OAM form factors
- Air-cooled, 150-250W typical workload power
NNP-T SOFTWARE STACK

• Full software stack built with open components
• Direct integration with DL frameworks
• **nGraph**: Hardware agnostic deep learning library and compiler for DL platform developers.
  • Provides common set of optimizations for NNP-T across DL frameworks
• **Argon**: NNP-T DNN compute & communication kernel library
• **Low-level programmability**: NNP-T kernel development toolchain w/tensor compiler
Flexible and programmable Tensor-based ISA
- Limited instruction set
- Extensible with uController custom instructions

Same distributed programming model for both intra and inter-chip
- Explicit SW memory management and message passing
- Synchronization primitives
- Compute has affinity to local data

DL workloads are dominated by a limited set of operations
• **Bfloat16 w/ FP32 accumulation**
  • No sacrifice in SOTA accuracy, improved power efficiency & training time*
  • Minimal model changes

---

<table>
<thead>
<tr>
<th>Numeric format</th>
<th>Area/Power efficient</th>
<th>Easy to Converge?</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>FP32</td>
<td>No</td>
<td>Yes</td>
<td>Industry standard</td>
</tr>
<tr>
<td>FP16</td>
<td>Yes</td>
<td>Medium</td>
<td>Good precision for Fprop, Bprop/Update is more challenging.</td>
</tr>
<tr>
<td>16b Integer formats</td>
<td>Yes</td>
<td>No</td>
<td>Better area/power than FP, but hard to use</td>
</tr>
<tr>
<td>Bfloat16</td>
<td>Yes</td>
<td>Yes</td>
<td>Good at Fprop, Bprop, and Update</td>
</tr>
<tr>
<td>8/4/2/1b integer</td>
<td>Extreme</td>
<td>Extremely difficult</td>
<td>Research areas</td>
</tr>
</tbody>
</table>

• Bfloat16 Matrix Multiply Core (32x32)
• FP32 & BF16 support for all other operations
• 2x multiply cores per TPC to amortize SoC resources
• Vector operations for non-GEMM
  • Compound pipeline
  • DL specific optimizations
    • Activation functions, RNG, Reductions & accumulations
    • Programmable FP32 look-up tables
Four stacks of HBM2-2400 8GB devices
- 1.22TBps raw bandwidth and 32GB total device memory. ECC protected
- 2.5MB/TPC of local scratchpad memory
  - 60 MB total distributed memory, ECC protected
  - Native Tensor Transpose
  - Simultaneous read and write on each MRB
- 1.4TBps local read/write bandwidth per-TPC
- Support for direct memory to memory transfer for both HBM and MRB
Bidirectional 2-D mesh architecture to allow any to any communication

Prioritized for throughput and congestion avoidance

Cut-through forwarding and multicast support

2.6TBps total cross-sectional BW, 1.3TBps per-direction

All peripheral devices shared through the mesh (HBM, SerDes)

Separate meshes for different traffic types

Support for direct peer to peer communication between TPCs
• Run the largest models across multiple chips and across chassis
• 16 quads of 112Gbps. 3.58Tbps total bi-directional BW per chip
• Fully programmable router w/multi-cast support enables multiple glue-less topologies
  • Reliable transmission
  • Virtual channels and priorities for traffic management
• Direct low-latency local memory transfer
• Support for up to 1024 nodes
• DL workloads require a variety of GEMM sizes

• Minimize HBM memory boundedness with on-die data re-use => higher utilization of compute resources

• Faster training, less idle resources

• ~2x better than published utilization of competitive products

<table>
<thead>
<tr>
<th>GEMM Size</th>
<th>Spring Crest Utilization</th>
</tr>
</thead>
<tbody>
<tr>
<td>1024 x 700 x 512</td>
<td>31.1%</td>
</tr>
<tr>
<td>1760 x 7133 x 1760</td>
<td>44.5%</td>
</tr>
<tr>
<td>2048 x 7133 x 2048</td>
<td>46.7%</td>
</tr>
<tr>
<td>2560 x 7133 x 2560</td>
<td>57.1%</td>
</tr>
<tr>
<td>4096 x 7133 x 4096</td>
<td>57.4%</td>
</tr>
<tr>
<td>5124 x 9124 x 2048</td>
<td>55.5%</td>
</tr>
</tbody>
</table>

* All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.
Performance measured on July 10, 2019 on pre-production NNP-T Spring Crest silicon, using 22 TPCs at 900MHz core clock and 2GHz HBM clock. Host is an Intel® Xeon® Gold 6130T CPU @ 2.10GHz with 64 GB of system memory. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

** Other names and brands may be claimed as the property of others. Deepbench data from: https://github.com/baidu-research/DeepBench/blob/master/results/train/DeepBench_NV_V100.xlsx
Based on NVIDIA DGX-1, NVIDIA V100 GPU, Linux Kernel 4.4.0-124-generic, CUDA 10.0.130, CuDNN 7.3.1.20, NVIDIA Driver 410.48, Intel® Xeon® CPU E5-2698v4@2.2GHz
SPRING CREST CONVOLUTIONS

- Various convolution hyperparameters are required by DL workloads
- Support multiple tensor layouts for maximum on-die data reuse resulting higher compute efficiency

```
<table>
<thead>
<tr>
<th>Description</th>
<th>Spring Crest Utilization</th>
</tr>
</thead>
<tbody>
<tr>
<td>c64xh56xw56_k64xr3xs3_st1_n128</td>
<td>86%</td>
</tr>
<tr>
<td>c128xh28xw28_k128xr3xs3_st1_n128</td>
<td>71%</td>
</tr>
<tr>
<td>c512xh28xw28_k128xr1xs1_st1_n128</td>
<td>65%</td>
</tr>
<tr>
<td>c128xh28xw28_k512xr1xs1_st1_n128</td>
<td>59%</td>
</tr>
<tr>
<td>c256xh14xw14_k1024xr1xs1_st1_n128</td>
<td>62%</td>
</tr>
<tr>
<td>c256xh28xw28_k512xr1xs1_st2_n128</td>
<td>71%</td>
</tr>
<tr>
<td>c32xh120xw120_k64xr5xs5_st1_n128</td>
<td>87%</td>
</tr>
</tbody>
</table>
```

C=# input dimensions, H=height, W=width, K=# filters, R=filter X, S=filter Y, ST=stride N=minibatch size

*All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice. Performance measured on July 10, 2019 on pre-production NNP-T Spring Crest silicon, using 22 TPCs at 900MHz core clock and 2GHz HBM clock. Host is an Intel® Xeon® Gold 6130T CPU @ 2.10GHz with 64 GB of system memory. For more complete information about performance and benchmark results, visit [www.intel.com/benchmarks](http://www.intel.com/benchmarks).
• Benchmarked on ring topology (intra and inter-chassis)
• Support for different All-reduce algorithms with different communication patterns

<table>
<thead>
<tr>
<th>Communication Kernels</th>
<th>Within-chassis (8 cards)</th>
<th>Cross-chassis (16, 32 cards)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spring Crest 16x ICL</td>
<td>Spring Crest 16x ICL</td>
</tr>
<tr>
<td>2-card Send/Recv BW (GB/s)</td>
<td>161</td>
<td>161</td>
</tr>
<tr>
<td>Allreduce BW (GB/s)</td>
<td>151</td>
<td>151</td>
</tr>
<tr>
<td>Broadcast BW (GB/s)</td>
<td>147</td>
<td>147</td>
</tr>
</tbody>
</table>

* All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.
Performance measured on July 10, 2019 on pre-production NNP-T Spring Crest silicon, using 22 TPCs at 900MHz core clock and 2GHz HBM clock. Host is an Intel® Xeon® Gold 6130T CPU @ 2.10GHz with 64 GB of system memory. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.
**COMMUNICATION PERFORMANCE**

- Low overhead and direct memory transfer result in low latency
- High efficiency even at moderate transfer sizes
- Cross-chassis scale-out with the same network/connectivity

<table>
<thead>
<tr>
<th>Allreduce Latency, 2KB (µs)</th>
<th>Spring Crest 16x ICL</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 cards (in-chassis)</td>
<td>3</td>
</tr>
<tr>
<td>4 cards (in-chassis)</td>
<td>8</td>
</tr>
<tr>
<td>8 cards (in-chassis)</td>
<td>9</td>
</tr>
<tr>
<td>16 cards (cross-chassis)</td>
<td>30</td>
</tr>
<tr>
<td>32 cards (cross-chassis)</td>
<td>36</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Allreduce Data Rate (GB/s)</th>
<th>Message Size</th>
<th>Spring Crest (8 chips)</th>
<th>Spring Crest (32 chips)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 MB</td>
<td>68.7</td>
<td>39.9</td>
</tr>
<tr>
<td></td>
<td>8 MB</td>
<td>115.8</td>
<td>92.2</td>
</tr>
<tr>
<td></td>
<td>32 MB</td>
<td>137.5</td>
<td>130.2</td>
</tr>
<tr>
<td></td>
<td>128 MB</td>
<td>147.1</td>
<td>147.4</td>
</tr>
</tbody>
</table>

* All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice. Performance measured on July 10, 2019 on pre-production NNP-T Spring Crest silicon, using 22 TPCs at 900MHz core clock and 2GHz HBM clock. Host is an Intel® Xeon® Gold 6130T CPU @ 2.10GHz with 64 GB of system memory. For more complete information about performance and benchmark results, visit [www.intel.com/benchmarks](http://www.intel.com/benchmarks).
CONCLUDING REMARKS

- Domain specific acceleration has a place in DL training
  - Training time and model size continue to be bottlenecks
- Numerics and compute tailored for DL
- No legacy workloads to support
- Architect from ground up to reduce data movement and keep compute units fed
- Higher utilization and efficiency on micro-benchmarks translate into better overall WL performance => Faster, more power efficient training