Deep Learning Performance with Intel® Caffe – Training, CPU model choice and Scalability

Authors: Alex Filby and Nishanth Dandapanthula.
HPC Engineering, HPC Innovation Lab, March 2018

To get the most out of deep learning technologies requires careful attention to both hardware and software considerations. There are a myriad of choices for compute, storage and networking. The software component does not stop at choosing a framework, there are many parameters for a particular model that can be tuned to alter performance. The Dell EMC Deep Learning Ready Bundle with Intel provides a complete solution with tuned hardware and software. This blog covers some of the benchmarks and results that influenced the design. Specifically we studied the training performance across different generations of servers/CPUs, and the scalability of Intel Caffe to hundreds of servers.

Introduction to Intel® Caffe and Testing Methodology

Intel Caffe is a fork of BVLC (Berkeley Vision and Learning Center) Caffe, maintained by Intel. The goal of the fork is to provide architecture specific optimizations for Intel CPUs (Broadwell, Skylake, Knights Landing, etc). In addition to Caffe optimization, "Intel optimized" models are also included with the code. These take popular models such as Alexnet, Googlenet, Resnet-50 and tweak their hyperparamters to provide increased performance and accuracy on Intel systems for both single node and multi-node runs. These models are frequently updated as the state of the art advances.

For these tests we chose the Resnet-50 model, due to its wide availability across frameworks for easy comparison, and since it is computationally more intensive than other common models. Resnet is short for Residual Network, which strives to make deeper networks easier to train and more accurate by learning the residual function of the underlying data set as opposed to the identity mapping. This is accomplished by adding “skip connections” that pass output from upper layers to lower ones and skipping over some number of intervening layers. The two outputs are then added together in an element wise fashion and passed into a nonlinearity (activation) function.

Table 1. Hardware configuration for Skylake, Knights Landing and Broadwell nodes

	SKL	KNL	BDW
Platform	Single node tests on PowerEdge C6420, R740, R640 Cluster tests on PowerEdge C6420	PowerEdge C6320p	PowerEdge C6320
CPU	Multiple CPU models (see results)	Intel Xeon 7230	Intel Xeon E5-2697v4
RAM	192GB DDR4 @ 2666 MT/s	96GB DDR4 @ 2400 MT/s	128GB DDR4 @ 2400MT/s
Interconnect	Intel® Omni-Path	Intel® Omni-Path	Intel® Omni-Path
Memory Mode (KNL Only)	N/A	Cache	N/A

Table 2. Software details

Software	
OS	RHEL 7.3 x86_64
Linux Kernel	3.10.0-514.el7.x86_64
BIOS	1.2.11
Intel Caffe	1.0.4
Intel MLSL (for multi-node tests)	2017.1.016
Caffe Model	intel_optimized/multinode/resnet_50_256_nodes_8k_batch (with batch size modified)

Performance tests were conducted on three generations of servers supporting different Intel CPU technology. The system configuration of these test beds is shown in Table 1 and the software configuration is listed in Table 2. This space is new and rapidly evolving, with frameworks being continuously updated and optimized. We expect performance to continue to improve, with subsequent releases, as such the results are intended to provide insights and not be taken as absolute.

As shown in Table 2 we used the Intel Caffe optimized multi-node version for all tests. There are differences between Intel's implementation of the single-node and multi-node Caffe models, and using the multi-node model across all configurations allows for an accurate comparison between single and multi-node scaling results. Unless otherwise stated all tests were run using the compressed ILSVRC 2012 (Imagenet) database which contains 1,281,167 images. The dataset is loaded into /dev/shm before the start of the test. For each data point a parameter sweep was performed across three parameters: *batch size, prefetch size, and thread count*. *Batch size* is the number of training examples fed into the model at one time, *prefetch* is the number of batches (of batch size) buffered in memory, and *thread count* is the number of threads used per node. The results shown used the best results from the parameter sweep for each test case. The metric used for comparison is images per second, which is calculated by taking the total number of images the model has seen (batch_size * iterations * nodes) divided by the total training time. Training time does not include Caffe startup time.

Single Node Performance

To determine what processors might be best suited for these workloads we tested a variety of SKUs including Intel Xeon E5-2697 v4 (Broadwell – BDW); Silver, Gold and Platinum Intel Xeon Scalable Processor Family CPUs (Skylake - SKL), as well as an Intel Xeon Phi CPU (KNL). The single node results are plotted in Figure 1 with the line graph showing results relative to the performance of the E5–2697 v4 BDW system.

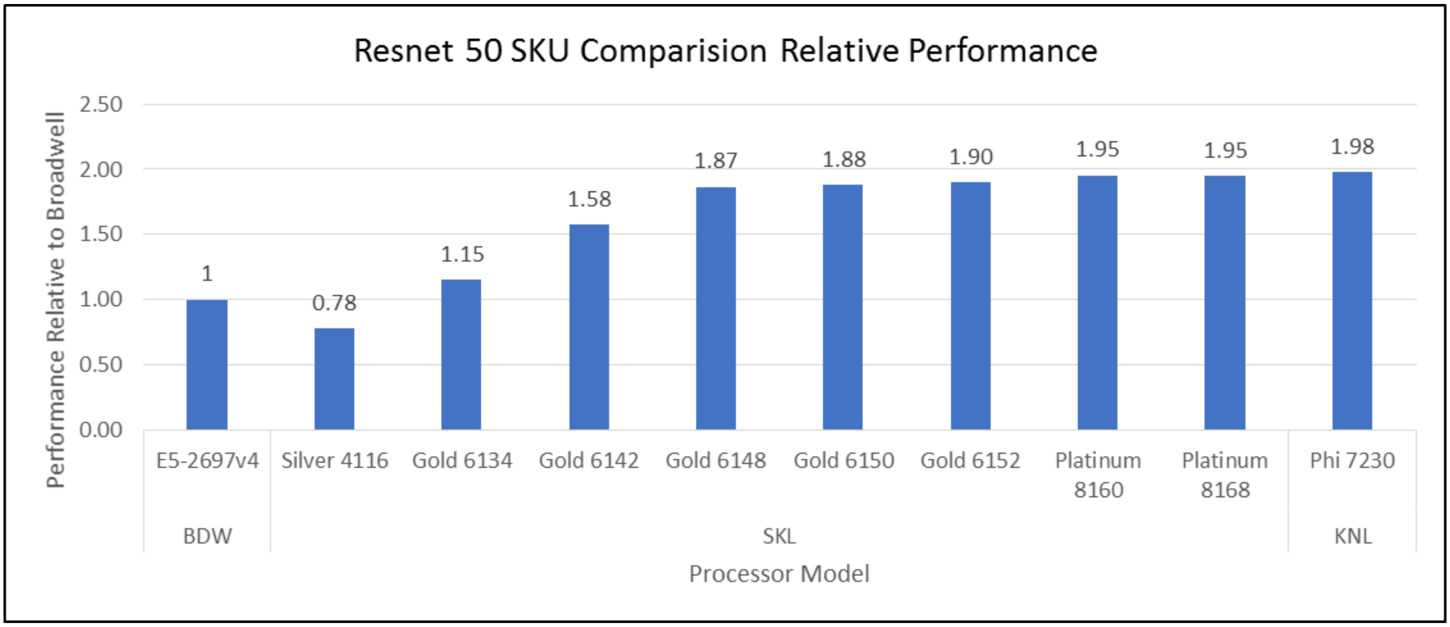

Figure 1. Processor model performance comparison relative to Broadwell

The difference in performance between the Gold 6148 and Platinum 8168 SKUs is around 5%. These results show that for this workload and version of Intel Caffe the higher end Platinum SKUs do not offer much in the way of additional performance over the Gold CPUs. The KNL processor model tested provides very similar results to the Platinum models.

Multi-node Performance and Scaling

The multi-node runs were conducted on the HPC Innovation Lab's Zenith cluster, which is a Top500 ranked cluster (#292 on the Nov 2017 list). Zenith contains over 324 Skylake nodes and 160 KNL nodes configured as listed in Table 1. The system uses Intel's Omni-Path Architecture for its high speed interconnect. The Omni-Path network consists of a single 768 port director switch, with all nodes directly connected, providing a fully non-blocking fabric.

Scaling Caffe beyond a single node requires additional software, we used Intel's Machine Learning Scalability Library (MLSL). MLSL provides an interface for common deep learning communication patterns built on top of Intel MPI. It supports various high speed interconnects and the API can be used by multiple frameworks.

The performance numbers on Zenith were obtained using /dev/shm, the same as we did for the single node tests. KNL multi-node tests used a Dell EMC NFS Storage Solution (NSS), an optimized NFS solution. Batch sizes were constrained as node count increased to keep the total batch size less than or equal to 8k, to keep it within the bounds of this particular model. As node count increases, the total batch size across all the nodes in the test increases as well (assuming you keep the batch size per node constant). Very large batch sizes complicate the gradient descent algorithm used to optimize the model, causing accuracy to suffer. Facebook has done work getting distributed training methods to scale to 8k batch sizes.

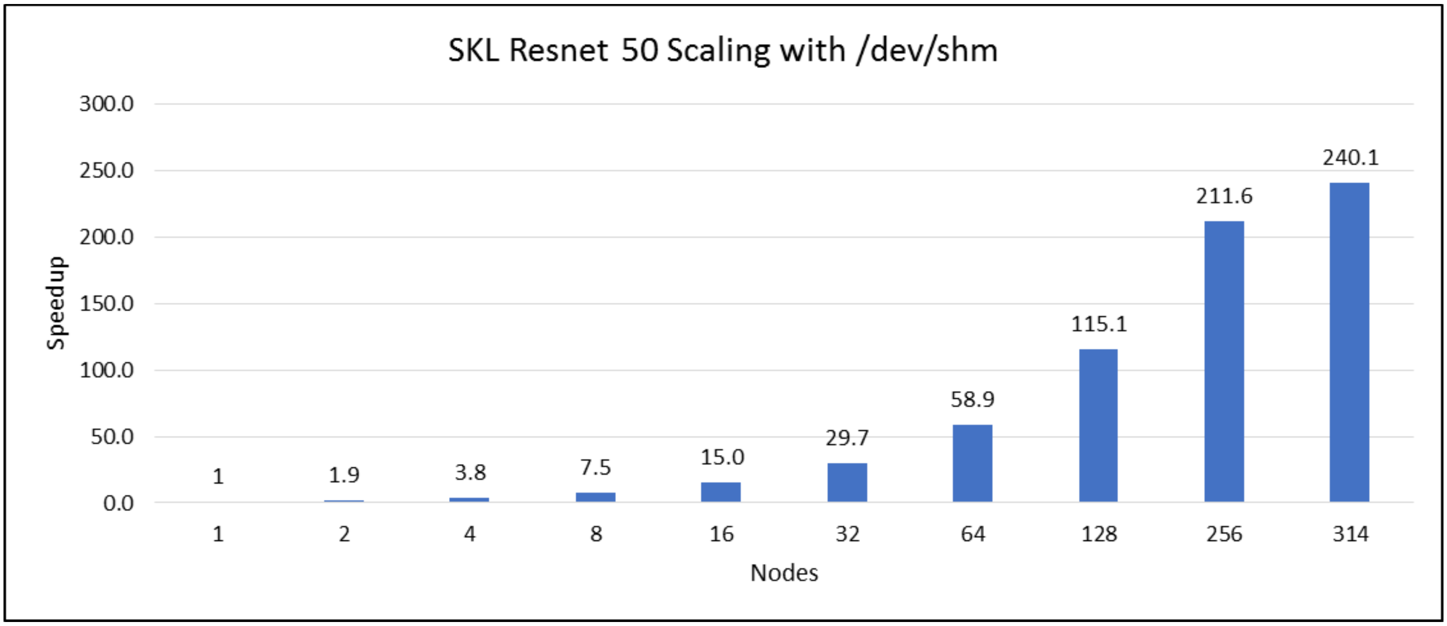

Figure 2. Scaling on Zenith with Gold 6148 processors using /dev/shm as the storage

Figure 2 shows the results of our scalability tests on Skylake. When scaling from 1 node to 128 nodes, speedup is within 90% of perfect scaling. Above that scaling starts to drop off more rapidly, falling to 83% and 76% of perfect for 256 and 314 nodes respectively. This is mostly likely due to a combination of factors the first being decreasing node batch size. Individual nodes tend to offer the best performance with larger batch sizes, but to keep the overall batch below 8k, the node batch size is decreased. Each node is then running a suboptimal batch size. The second is communication overhead; the Intel Caffe default for multi-node weight updates utilizes MPI collectives at the end of each batch to distribute the model weight data to all nodes. This allows each node to ‘see’ the data from all other nodes without having to process all of the other images in the batch. It is why you get a training time improvement when using multiple nodes instead of just training hundreds of individual models. Communication patterns and overhead is an area we plan to investigate in the future.

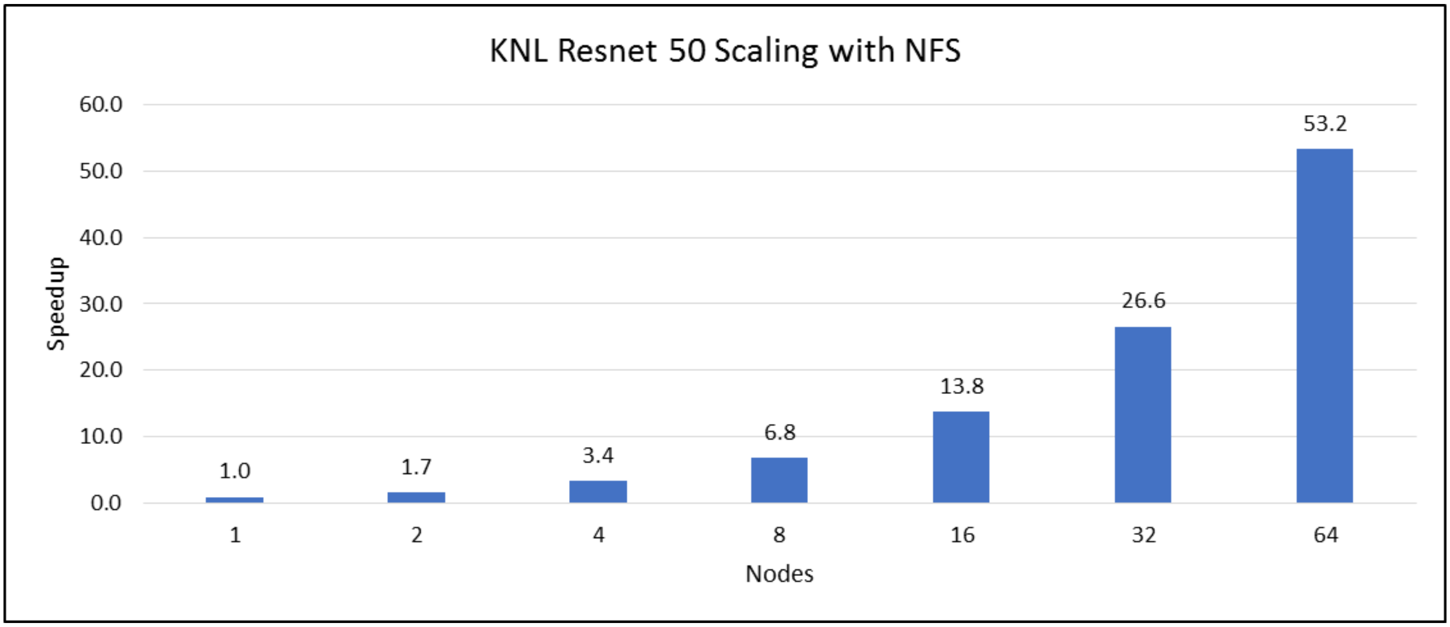

Figure 3. Scaling Xeon 7230 KNL using Dell NFS Storage Solution

The scalability results on the KNL cluster are shown in Figure 3. The results are similar to SKL results in Figure 2. For this test, batch size was able to remain constant due to the smaller number of nodes and the fact that a smaller batch size was optimal on KNL systems. With multi-node runs some performance is lost due to threads being needed for communication, and not pure computation as with single node tests.

Conclusions and Future Work

For this blog we have focused on single node deep learning training performance comparing a range of different Intel CPU models and generations, and conducted initial scaling studies for both SKL and KNL clusters. Our key takeaways are summarized as follows:

- Intel Caffe with Intel MLSL scales to hundreds of nodes.
- Skylake Gold 6148, 6150, and 6152 processors offer similar performance to Platinum SKUs.
- KNL performance is also similar to Platinum SKUs.

Our future work will focus on other aspects of deep learning solutions including performance of other frameworks, inference performance, and I/O considerations. TensorFlow is a very popular framework which we did not discuss here but will do so in a future part of this blog series. Inferencing is a very important part of the workflow, as a model must be deployed for it to be of use! Finally we'll also compare the various storage options and tradeoffs as well as discuss the I/O patterns (network and storage) of TensorFlow and Intel Caffe.