Scaling Tensorflow on up to 512 nodes on CORI Supercomputer
In collaboration with NERSC

Amrita Mathuriya
HPC Application Engineer
DCG/Intel Corporation

November 11, 2017

Intel® HPC Developer Conference 2017, Denver
Presenter: Amrita Mathuriya

- HPC application Engineer in DCG at Intel. Working for code modernization and optimization on Intel® Xeon and Intel® Xeon Phi™

- Working at Intel for past 9 Years.
  - Expert at algorithms and optimizations for IA architectures.
  - Currently working on optimization and scalability of a deep learning framework, Tensorflow.
  - Worked on HPC applications in areas of Computational Geometry, Optical Proximity Correction (OPC), Electromagnetics, Computational Biology, Quantum Monte Carlo.
  - Working on code modernization for Intel® Xeon and Intel® Xeon Phi™ architectures.

- Education:
  - MS in Computer Science with the specialization in Computational Science and Engineering from Georgia Tech
  - Bachelor's degree in Computer Science from Indian Institute of Technology (IIT) Roorkee, India.
Intel/NERSC/Universities BDC Collaboration

How to use Tensorflow on CORI

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.

Speedup of 6-42x for various topologies with MKL!
High level to deep dive – not to scale!

Multi-node Scaling

Performance

Tensorflow

Deep Learning

Artificial Intelligence
Scaling GRPC Tensorflow on up to 512 nodes of Cori Supercomputer

Amrita Mathuriya  
Intel Corporation

Thorsten Kurth  
Lawrence Berkeley National Laboratory

Vivek Rane  
Intel Corporation

Mustafa Mustafa  
Lawrence Berkeley National Laboratory

Lei Shao  
Intel Corporation

Debbie Bard  
Lawrence Berkeley National Laboratory

Prabhat  
Lawrence Berkeley National Laboratory

Victor W Lee  
Intel Corporation

1  Note

This manuscript is submitted to NIPS 2017 "Deep Learning at Supercomputer Scale" workshop.

2  Introduction

We explore scaling of the standard distributed Tensorflow [1] with GRPC primitives on up to 512 Intel® Xeon Phi™ (KNL) nodes of Cori supercomputer [2] with synchronous stochastic gradient descent (SGD), and identify causes of scaling inefficiency at higher node counts. To our knowledge, this is the first demonstration of scaling GRPC TensorFlow to tens of thousands of nodes.
Contributions

- First exploration of distributed GRPC Tensorflow’s scalability on a HPC supercomputer at such large scale.
- Tested Tensorflow on up to 512-1024 nodes.
- Provide detailed analysis of inefficiencies in the current distributed algorithm and its implementation.
Systems

- CORI Haswell
  - Two sockets, each socket is populated with a 16-core Intel® Xeon™ Processor E5-2698 v3 ("Haswell") at 2.3 GHz

- CORI KNL
  - A single-socket Intel® Xeon Phi™ Processor 7250 ("Knights Landing") processor with 68 cores per node @ 1.4 GHz
Tensorflow Resnet50 scaling with dummy data on KNL CORI

83% at 128 workers

Scales well to 128 workers with >80% per worker efficiency!

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.
Tensorflow Resnet50 scaling with dummy data on KNL CORI

Efficiency drops to 23% on 512 workers.

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.
Summary of findings!

Sub-optimal use of interconnect bandwidth at various levels!

1. Distributed algorithm uses centralized servers for gradient averaging. Creates a bottleneck at certain interconnect links.
   - Load imbalance among PS tasks hinders their efficient scaling.

2. Underlying communication primitive GRPC is currently inefficient on Cori's high-speed interconnect. Uses TCP-IP socket communication!
Why Tensorflow?

- Open source software library from Google for dataflow programming

Source: http://deliprao.com/archives/168
Continued... Tensorflow Popularity

Scaling Experiments for Tensorflow

- Synchronous SGD.
- Weak scaling experiments – keep batch size per worker fix.
- A relatively deeper and compute heavy network Resnet50 and a light weight HEP-CNN network.
- Use relatively large batch size of 128 images;
  - Keeps the fraction of time spent on communication low.
- Perform detailed analysis with dummy instead of real data to avoid any potential I/O 26 bottlenecks.

Designed experiments to test scaling limits of Tensorflow at large node counts!
Why ResNet-50?

Details of ResNet-50

- State of the art deep learning network for classification task on natural images such as Imagenet dataset.
- It is a deep network with 54 convolution and 1 fully connected operations.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Convolution layers</td>
<td>53</td>
</tr>
<tr>
<td>Fully connected layer</td>
<td>1</td>
</tr>
<tr>
<td>Total number of trainable varaibles</td>
<td>~ 25.5 million</td>
</tr>
<tr>
<td>Size in MB</td>
<td>~ 100 MB</td>
</tr>
</tbody>
</table>

Resnet50 is a deep network and puts pressure on interconnect bandwidth!
HEP-CNN – Gradient parameters

- Fairly representative of the deep learning networks used in the science community.
- Number of layers = 6
  - 5 convolution layers.
  - 1 fully connected layer.


<table>
<thead>
<tr>
<th></th>
<th>Kernel Sizes</th>
<th>Total params</th>
</tr>
</thead>
<tbody>
<tr>
<td>'conv1_w'</td>
<td>3 3 3</td>
<td>128</td>
</tr>
<tr>
<td>'conv2_w'</td>
<td>3 3 128</td>
<td>128</td>
</tr>
<tr>
<td>'conv3_w'</td>
<td>3 3 128</td>
<td>128</td>
</tr>
<tr>
<td>'conv4_w'</td>
<td>3 3 128</td>
<td>128</td>
</tr>
<tr>
<td>'conv5_w'</td>
<td>3 3 128</td>
<td>128</td>
</tr>
<tr>
<td>'fc1_b'</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>'fc1_w'</td>
<td>128 2</td>
<td></td>
</tr>
<tr>
<td>Total Parameters</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total MB</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Light weight CNN with ~593K trainable parameters with 2.26 MB size!
Tensorflow – Synchronous SGD

All reduce operation among workers with centralized servers.
HEP-CNN Network – Scaling on KNL CORI

For the light weight HEP-CNN network, 1 PS suffices for 256 workers.
ResNet-50 – Efficiency vs. workers
Tensorflow – Synchronous SGD

The communication time increases as $O(\text{number of workers})$.

All reduce operation among workers with centralized servers.

Network Bandwidth Bottleneck?
Tensorflow – Synchronous SGD

All reduce operation among workers with centralized servers.
Resnet50 – Scaling Parameter Servers on KNL CORI

Per worker efficiency does not significantly increase beyond 32 PS tasks!
Load Imbalance Among PS Tasks – Resnet50

- Resnet50 – Total number of trainable parameters 161.
  - 53 conv kernel tensors
  - 53*2 1D vectors for batch-norm
  - 1 fc-weight 2D tensor + 1 fc-bias 1D vector.
- 99% of total size is located in 54 variables – kernel for conv and weight for fc layer!
- NO chunking of variables.
- One variable can only be assigned to a single PS task.

Total number of large trainable variables determines the PS scalability!
Resnet50 scaling with dummy data on KNL CORI

Scales well to 128 workers with >80% per worker efficiency, but efficiency drops to 23% on 512 workers.

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.
GRPC an efficient communication protocol on high-speed interconnects?

- Uses TCP-IP
- Doesn't go through the software layer to use high-speed Cray Aries interconnect.
- Crude estimates show 5-6x gap in communication time with 1 PS and 16 workers w.r.t. the peak achievable!
Summary of findings!

Sub-optimal use of interconnect bandwidth at various levels!

1. Distributed algorithm uses centralized servers for gradient averaging. Creates a bottleneck at certain interconnect links.
   ▪ Load imbalance among PS tasks hinders their efficient scaling.

2. Underlying communication primitive GRPC is currently inefficient on Cori's high-speed interconnect. Uses TCP-IP socket communication!
Outlook and Future Work

- Potentially, more efficient all-reduce algorithms such as ring or tree-reduction can better utilize the network bandwidth.
  - No need to allocate extra nodes as parameter servers.
  - No load imbalance issue.
- MPI communication layer is capable of utilizing the HPC high-speed interconnects.
  - Replacing GPRC with MPI layer should lower down the time spent in communication.
LEGAL DISCLAIMERS

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. 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.

A "Mission Critical Application" is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL'S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS' FEES ARISING OUT OF, DIRECTLY OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENCE IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.

Intel may make changes to specifications and product descriptions at any time, without notice. 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. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document 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. Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product or order.

Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or go to: http://www.intel.com/design/literature.htm

Knights Landing and other code names featured are used internally within Intel to identify products that are in development and not yet publicly announced for release. Customers, licensees and other third parties are not authorized by Intel to use code names in advertising, promotion or marketing of any product or services and any such use of Intel's internal code names is at the sole risk of the user

Intel, Look Inside, Xeon, Intel Xeon Phi, Pentium, Cilk, VTune and the Intel logo are trademarks of Intel Corporation in the U.S. and other countries.

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

Copyright © 2017 Intel Corporation
Legal Disclaimers

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 SSE3 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
Legal Disclaimers

- 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 information go to [http://www.intel.com/performance](http://www.intel.com/performance).

- Estimated Results Benchmark Disclaimer:
  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.

- Software Source Code Disclaimer:
  Any software source code reprinted in this document is furnished under a software license and may only be used or copied in accordance with the terms of that license.

  Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

  THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.
Thank you for attention!
Amrita Mathuriya
amrita.mathuriya@intel.com