The Intel® Omni-Path Architecture (OPA) for Machine Learning

Sponsored by Intel

Srini Chari, Ph.D., MBA and M. R. Pamidi Ph.D.

December 2017

mailto:info@cabotpartners.com

Executive Summary

Machine Learning (ML), a major component of Artificial Intelligence (AI), is rapidly evolving and significantly improving growth, profits and operational efficiencies in virtually every industry. This is being driven – in large part – by continuing improvements in High Performance Computing (HPC) systems and related innovations in software and algorithms to harness these HPC systems.

However, there are several barriers to implement Machine Learning (particularly Deep Learning – DL, a subset of ML) at scale:

- *It is hard for HPC systems to perform and scale to handle the massive growth of the volume, velocity and variety of data that must be processed.*
- *Implementing DL requires deploying several technologies: applications, frameworks, libraries, development tools and reliable HPC processors, fabrics and storage. This is hard, laborious and very time-consuming.*
- *Training followed by Inference are two separate ML steps. Training traditionally took days/weeks, whereas Inference was near real-time. Increasingly, to make more accurate inferences, faster re-Training on new data is required. So, Training must now be done in a few hours. This requires novel parallel computing methods and large-scale high-performance systems/fabrics.*

To help clients overcome these barriers and unleash AI/ML innovation, Intel provides a comprehensive ML solution stack with multiple technology options. Intel's pioneering research in parallel ML algorithms and the Intel® Omni-Path Architecture (OPA) fabric minimize communications overhead and improve ML computational efficiency at scale.

With unique features designed to lower total cost of ownership (TCO) for Machine Learning and HPC, the Intel OPA high-performance fabric delivers 100 gigabits/sec of bandwidth per port and 21% lower latency at scale and 27% higher messaging rates compared with InfiniBand EDR.

Recently, a scale-out cluster system with Intel® Xeon® /Xeon Phi™ processors connected with the Intel OPA fabric broke several records for large image recognition ML workloads. It achieved Deep Learning Training in less than 40 Minutes on ImageNet-1K and the best accuracy and training time on ImageNet-22K and Places-365.

Intel OPA is the top 100G HPC fabric in the Top500 supercomputer list. This lead continues to grow. Globally, many clients from research and academic institutions are already advancing the state-of-the-art of AI/ML applications across many fields using large-scale systems with the Intel OPA fabric.

As clients, across many industries, implement AI/ML for their digital transformation, they should seriously consider investing in systems connected with the Intel Omni-Path Architecture.

This paper was developed with INTEL funding.

Cabot Partners Group, Inc. 100 Woodcrest Lane, Danbury, CT 06810. www.cabotpartners.com

Cabot Partners

Optimizing Business Value

Machine Learning: The Next Wave of Analytics

180 zettabytes of data by 2025

Data from billions of devices is growing exponentially. By 2025, the world is expected to have a total of 180 zettabytes of data (or 180 trillion gigabytes), up from less than 10 zettabytes in 2015.[1] To get actionable insights from this ever-increasing volume of data and stay competitive and profitable, every industry is investing in Analytics and High-Performance Computing (HPC).

As the lines between HPC and Analytics continue to blur, Analytics is evolving from Descriptive to Predictive to Prescriptive and to Machine Learning (ML – Training and Inference) workloads (Figure 1). This requires an IT infrastructure that must deliver higher performance and capabilities to enable rapid and more frequent processing of highly-accurate, data-intensive Training models; leading to better and quicker Inference outcomes.

Leveraging Data and HPC throughout the Analytics and Machine Learning journey

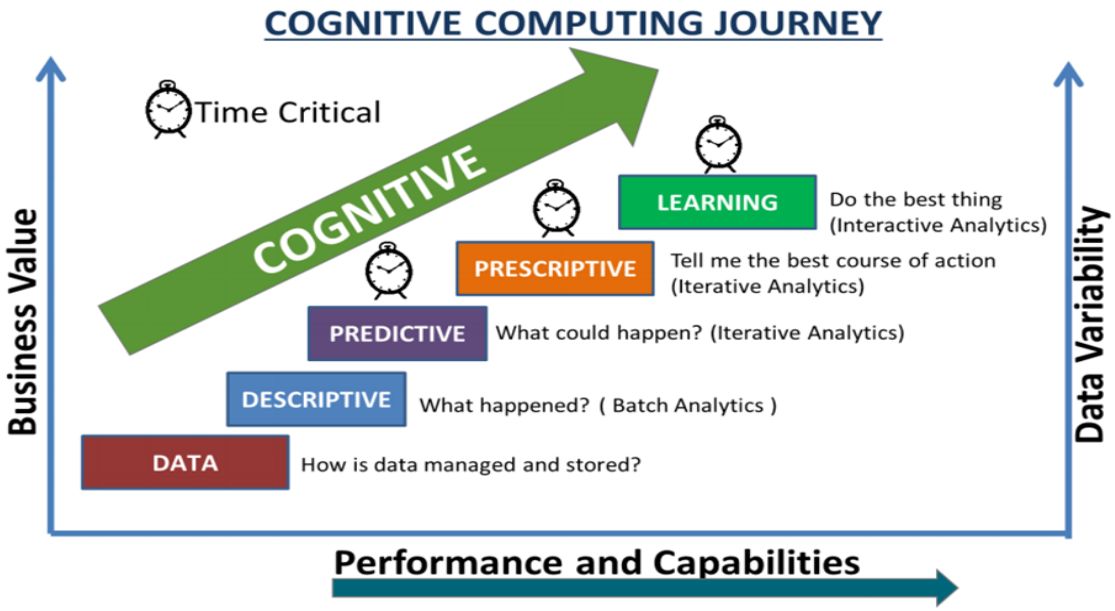

Figure 1: Leveraging Data and HPC for Machine Learning

Deep Learning on HPC-like architectures fueling rapid growth

Deep Learning (DL – a subset of ML) is even more **compute and I/O** intensive because of the enormous amounts of computational tasks (e. g., matrix multiplications) and data involved. In fact, one of Baidu's speech recognition models requires not only four terabytes of training data, but also 20 exaflops of compute – that's 20 billion times billion math operations – across the entire training cycle![2]

Consequently, HPC (fueled by rapid adoption of DL) is expected to grow by 6.2% annually to $30 billion in 2021.[3] DL growth at traditional enterprise (non-HPC) clients could further increase these healthy projections especially as they begin to deploy HPC-like architectures for DL. However, deploying DL at scale requires a deep computational understanding of the Machine Learning process.

Machine Learning (ML): A Brief Overview

Machine Learning *trains* computers to do what is natural for humans: learn from experience. ML algorithms *learn* directly from data to build the Trained model (typically a Neural Network) whose performance and accuracy improves as the number of data samples available for Training increases. This Trained Model can be used to make Inferences on new data sets (Figure 2).

[1] "IoT Mid-Year Update From IDC And Other Research Firms," Gil Press, *Forbes*, August 5, 2016.
[2] "What's the Difference Between Deep Learning Training and Inference?" Michael Copeland, NVIDIA blog, August 22, 2016.
[3] https://www.hpcwire.com/2017/06/20/hyperion-deep-learning-ai-helping-drive-healthy-hpc-industry-growth/

Machine Learning includes Training and Inference

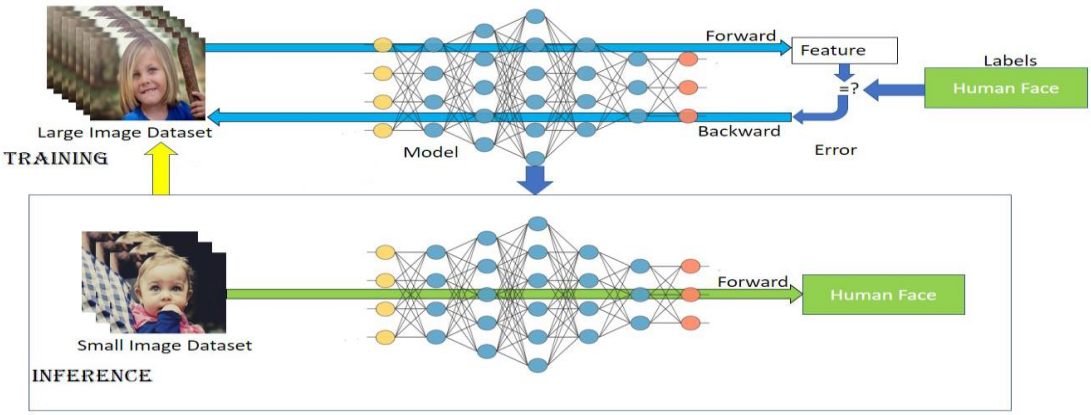

Figure 2: A Typical Human Facial Image Recognition Workflow - Training and Inference

Training a model with a billion parameters (moderately complex network) can take days/weeks unless properly optimized and scaled. Further, this process often needs to be repeated to experiment with different topologies/algorithms/hyper-parameters to reach the desired level of Inferencing accuracy (Figure 3). This typically requires a centralized HPC infrastructure. On the other hand, one Inference instance is less compute intensive. But millions of Inferences may be done in parallel with one Trained Model. So, in aggregate, the computational requirements for Inference could be greater and distributed.

Training is very compute intensive; requires HPC

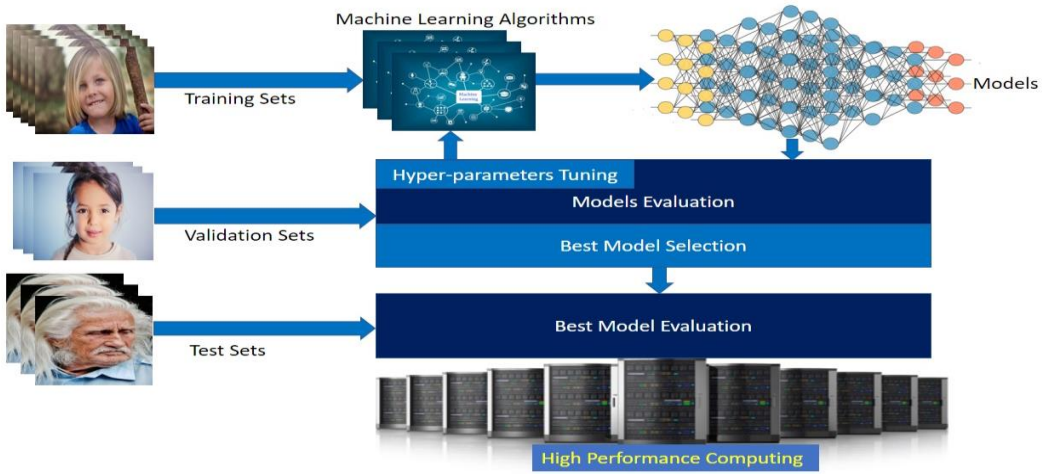

Figure 3: High Level Process and Architecture for Training

Continuously re-training on new data is driving ML computing requirements even higher

Training and Inference are two separate computational steps. Increasingly, to improve predictive accuracy, Training and Inference are being integrated into a highly iterative workflow – as represented with the yellow arrow in Figure 2. This requirement to continuously re-train on new data is driving ML computing requirements to even higher levels – typically found in today's largest HPC environments.

Training can now be done in a few hours using scalable data and model parallel algorithms that distribute the ML computational kernels over tens/hundreds of processors. These algorithms can be optimized to reduce communication overheads with high-performance fabrics such as the Intel OPA.

Why High Performance Fabrics to Scale Machine Learning

During Training, the key computational kernels are numerous matrix multiplications throughout the recurring forward and backward propagation steps. Starting with inputs (**I**) and other data, training model weights (**W**) and activations (**O** – outputs) are estimated (Figure 4). Then, stochastic gradient

descent algorithms[4] are used to iteratively adjust the weights/activations until a cost/error function (a measure of the difference between the actual output and predicted the output) is minimized. These final weights determine the Training model that can then be used for Inference.

Numerous matrix multiplications are key computational kernels

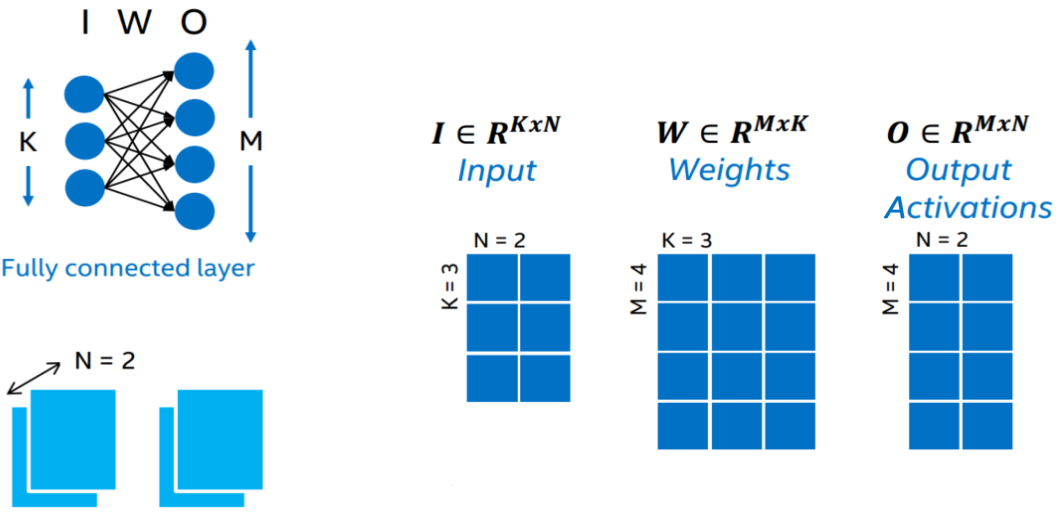

Figure 4: Key Computational Kernels for a Fully-Connected 2-Layer Neural Network[5]

Amount of computation depends on hyper-parameters: # of layers, size of input and # of outputs

The amount of computation depends on the size of the input data (**K**), the number of layers in the network(**N**) and the number of outputs/activations (**M**). The Weights matrix is **M**-rows by **K**-columns.

At each phase, the matrix operations sizes are: Forward propagation: (**M** x **K**) * (**K** x **N**); Backward propagation: (**M** x **K**)^T * (**M** x **N**) and Weight update: (**M** x **N**) * (**K** x **N**)^T. These operations are repeated until the error/cost function is minimized. For larger inputs and deeper networks, these computations grow quickly. Model and data parallel approaches are needed to scale these computations.

Model and data parallel approaches used to scale…but communication overheads must be minimized

Data parallel approaches distribute the data between various cores and each core independently tries to estimate the same weights/activations. Then the cores exchange these estimates to arrive at a consolidated estimate for the step. Whereas in model parallel approaches, the same data is sent to all cores and each core estimates different weights/activations. Then the cores exchange these estimates to arrive at the consolidated estimate for the step.

Generally, for Training on a small number of nodes in a cluster, data parallel approaches are better when the number of activations is greater than the number of weights. While model parallel approaches may work better if the number of weights is greater than the number of activations. For Inference, the data parallel approach works well since each input dataset can be run on a separate node.

As the number of cluster nodes are scaled for Training, data parallelism makes the number of activations per node much smaller than the weights while model parallelism makes the weights per node far less than the number of activations. This reduces computational efficiency and increases communication overheads since skewed (wide and short, or narrow and long) matrices must be split and processed.

Hybrid approaches (Figure 5) that combine data and model parallelism and smart Node-Grouping can reduce communications overhead and improve computational efficiency at scale. Hybrid parallelism partitions activations/weights to minimize skewed matrices. Smart Node-Grouping avoids inefficient global transfers: activations transfer only within a group and weights transfer only across groups.

[4] Ian Goodfellow, Yoshua Bengio and Aaron Courville, "Deep Learning", The MIT Press, 2016
[5] Pradeep Dubey, "Scaling to Meet the Growing Needs of Artificial Intelligence (AI)", Intel Developer Forum, 2016.

Hybrid methods and smart Node-Grouping reduce communication overheads and improve computational efficiency

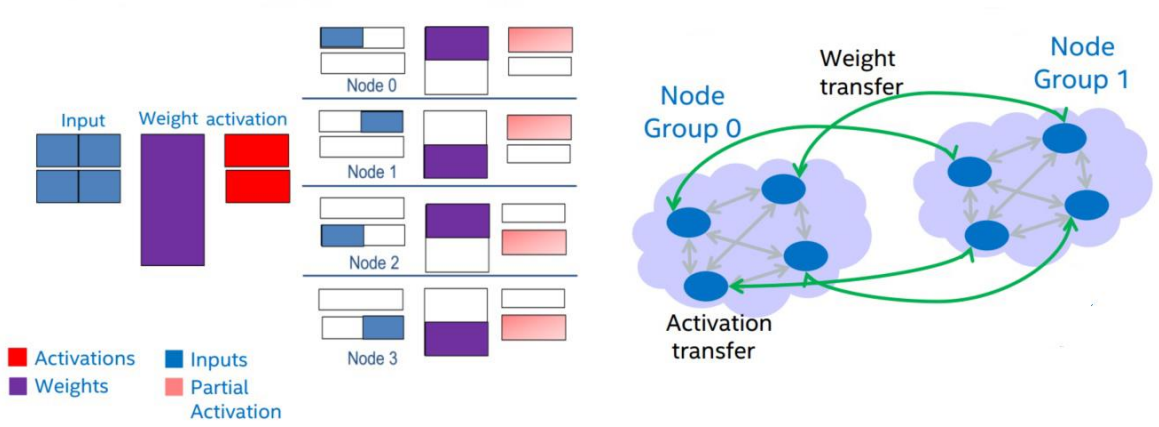

Figure 5: Hybrid Parallelism/Smart Node-Grouping Enhance Computational Efficiency at Scale[5]

Typically, at every layer, Deep Learning communications patterns (Figure 6) involve Reduce and Scatter operations: Reduce the activations from layer N-1 and Scatter at layer N. Common Message Passing Interface (MPI) collectives include: Scatter/Gather, AllGather, AllReduce and AlltoAll.

Several optimized MPI Collectives required to minimize Deep Learning communication overheads

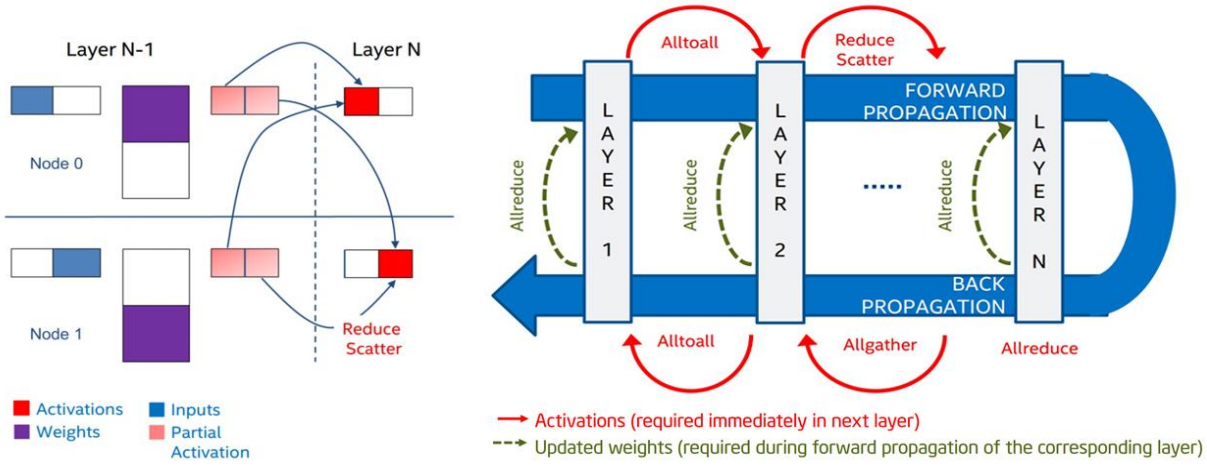

Figure 6: Deep Learning Communications Patterns

Intel is pioneering research in Hybrid parallel approaches and smart Node-Grouping for Machine Learning to improve computational efficiencies. Clients can expect to see these innovations in Math and ML Libraries.[6] Advanced features in the Intel Omni-Path Fabric such as optimized MPI collectives, overlapping compute with communication, smart message and task scheduling, and others enhance computational efficiency and scale even more.

Several advanced OPA features pioneered by Intel reduce communication overheads even more

The Omni-Path Fabric is one of Intel's several high-performance technologies that help clients deploy highly accurate, enterprise-grade ML solutions in record time to accelerate innovation/time-to-value.

How Intel is Unleashing AI/Machine Learning Innovation

To address DL/ML challenges at scale, a cutting-edge HPC architecture is needed. This usually includes high-performance processors, large memory and storage systems and high-performance connectivity between the servers and to high performance storage. Intel provides this end-to-end architecture.

Across many industry-verticals, Intel provides clients multiple HPC technologies and a comprehensive framework (Figure 7) to deploy ML. This portfolio includes hardware, libraries, frameworks, and development tools. Key current hardware technologies include Intel Xeon and Intel Phi processors.

[6] https://github.com/01org/MLSL

Intel provides a complete ML portfolio: hardware, libraries, Frameworks, tools and experiences

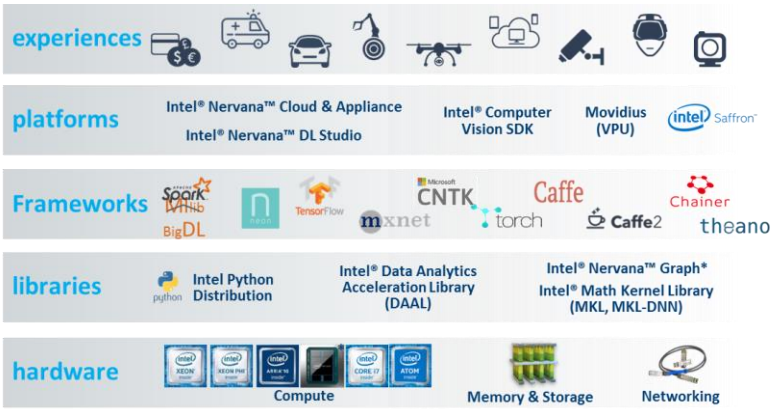

Figure 7. Intel AI/Machine Learning Portfolio

Intel processors optimized for training and inference connected with OPA unleash unmatched innovation

The Intel Nervana Neural Network Processor (formerly codenamed Crest) coupled with the Intel Xeon processor is expected to deliver several times more of raw computing power compared to today's state-of-the-art Graphics Processing Units (GPUs) for Deep Learning Training workloads. In addition, for Inference workloads, the combination of Intel Xeon processors + FPGA (field-programmable gate array) accelerators with the Intel Arria 10 provides a unique, customizable and programmable platform with low latency, flexible precision and high performance-per-watt.

These processors can be connected using the high-performance Intel Omni-Path Fabric. In addition, the Intel OPA supports and performs/scales very well on systems with NVIDIA Graphics Processing Units (GPUs). Intel OPA's support of many heterogenous processor architectures provides clients unmatched flexibility and performance to unleash ML innovation at scale.

The Unique Value of the Intel Omni-Path Architecture (OPA) for ML

Higher bandwidth and lower latency interconnects are needed as clusters scale and communication overheads bottleneck performance. This requires high component density for host fabric adapters, CPU to fabric PCIe adapters, switches and cables; increasing cluster complexity, power usage, rack and floor space requirements and other operational costs. In fact, as clusters scale, interconnect costs are a greater percentage of the Total Cost of Ownership (TCO).[7] For ML, this percentage could be even more.

The Intel OPA 100 Series product line is an end-to-end solution of PCIe adapters, silicon, switches, cables, and management software (Figure 8) designed to lower TCO for Machine Learning and HPC. It delivers 100 gigabits/sec of bandwidth per port and 21% lower latency at scale and 27% higher messaging rates compared with InfiniBand EDR.[6]

Compared with InfiniBand EDR, OPA delivers 21% lower latency, 27% higher messaging rates and reduces power consumption up to 60%

In fact, according to Intel, compared to InfiniBand EDR, OPA lowers power consumption by up to 60% and reduces fabric costs; freeing up cluster budgets to acquire up to 24% more compute capabilities. This greatly improves the economics for Intel Business Partners and their customers. Additionally, software from the OpenFabrics Alliance can be re-used to further enhance deployment economics.

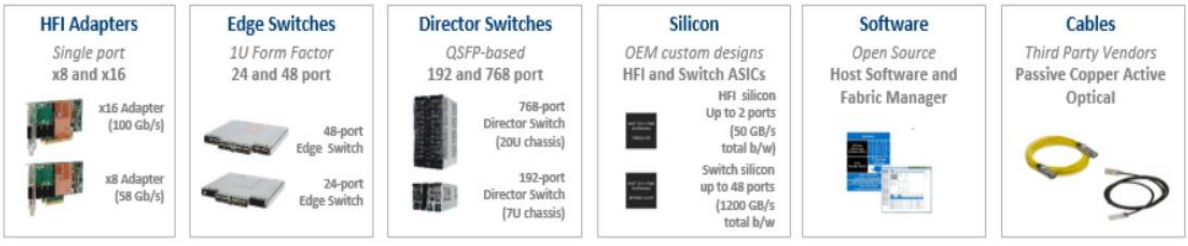

Figure 8. Intel Omni-Path Architecture

[7] Intel internal estimates

Traffic Flow Optimization, Packet Integrity Protection and Dynamic Lane Scaling are innovative features in OPA

Additional innovative features in OPA include improved performance, reliability, and QoS through:

- **Traffic Flow Optimization** to maximize QoS in mixed traffic by allowing higher-priority packets to preempt lower-priority packets, regardless of packet ordering.
- **Packet Integrity Protection** enables error correction fixes at the link level for transparent detection of transmission errors and recovery as they occur rather than at the end-to-end level – as is the case with InfiniBand. This makes network performance predictable and minimizes latency jitter even at scale.
- **Dynamic Lane Scaling**, related to Packet Integrity Protection, maintains link continuity in the event of a lane failure by using the remaining lanes for operations; ensuring the application completes. This improves reliability/resilience. In the case of InfiniBand, the application may terminate prematurely.

ML Benchmark Results: For large image recognition datasets, the Intel Scalable Processor 8160/Intel Phi and Intel OPA deliver very high performance/scale to train DL approaches based on Convolutional Neural Networks (CNN) on models optimized with Stochastic Gradient Descent. Figures 9-10 summarize the results for Intel Caffe Resnet-50 on Imagenet-1K on: Stampede2 at the Texas Advanced Computing Center (TACC) and MareNostrum 4 at the Barcelona Supercomputer Center.[8]

Over 90% scaling efficiency from 1 to 256 nodes

Figure 9: Over 90% Scaling Efficiency with Intel Omni-Path Architecture from 1 to 256 Nodes

Time to Train reduces beyond 256 nodes

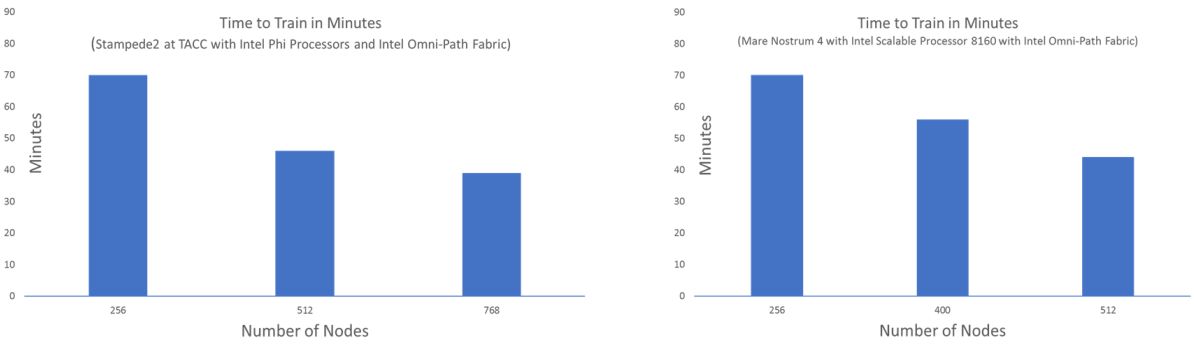

Figure 10. Time to Train Continues to Reduce Beyond 256 Nodes with Intel OPA

The Intel OPA fabric broke several records for large image recognition workloads by achieving Deep Learning Training in less than 40 Minutes on ImageNet-1K. Clients are leveraging these impressive performance and scaling results to unleash AI/ML innovation with the Intel Omni-Path Fabric.

[8] https://blog.surf.nl/en/imagenet-1k-training-on-intel-xeon-phi-in-less-than-40-minutes/

Good scaling with OPA on TensorFlow

TensorFlow Results: TensorFlow from Google is the leading ML Framework. It grew from about 14,000 GitHub Stars in November 2015 to 44,508 in February 2017 and to 68,684 in September 2017. [9] OPA based systems support TensorFlow through several communications interfaces. Preliminary multi-node measurements[10] performed by Intel with Resnet-50 show good scaling (Figure 11).

OPA systems scale well on GoogLeNet with Intel Caffe and MLSL

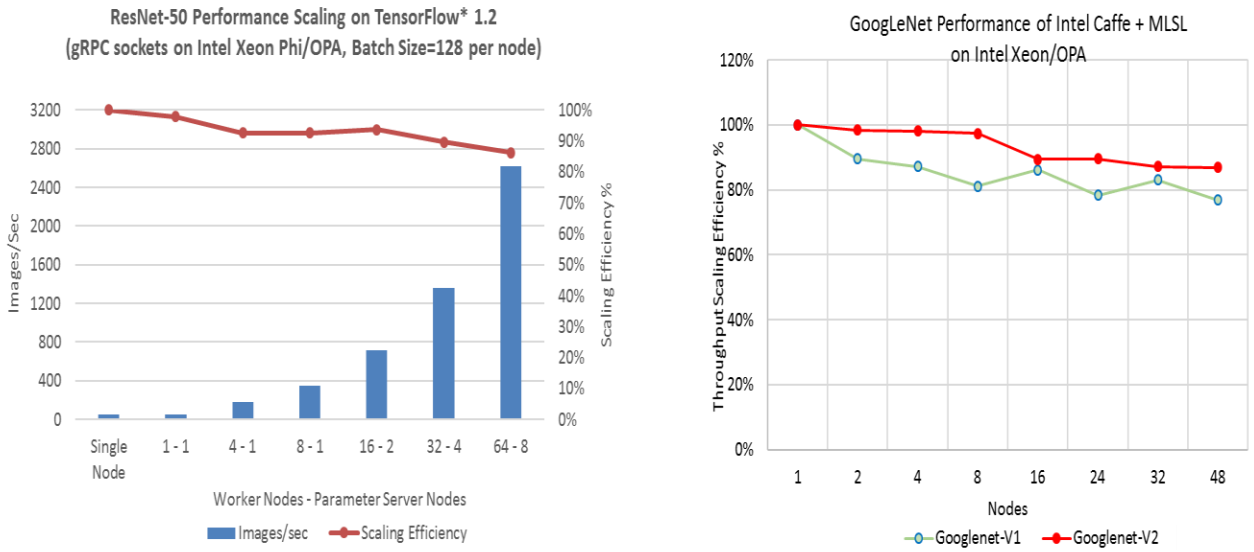

Figure 11: Good Scaling on TensorFlow (Leading ML Framework) and GoogLeNet (Computer Vision)

GoogLeNet with Intel Caffe Results: With 22 layers deep, Google's GoogLeNet is used to measure improvements in computer vision technology. Parallelization and implementation was done by Intel with Machine Learning Scaling Library (MLSL) and Intel Caffe (both optimized for Intel Xeon and OPA).

MLSL abstracts communication patterns and supports data/model/hybrid parallelism and leverages Intel MPI optimizations for OPA for communication. MLSL can also use other runtimes/message layers.

The MLSL API is designed to support a variety of popular Frameworks (e.g. Caffe, Torch, Theano, etc.). It provides statistical data collection to monitor time spent on different operations; including computation and communication. Figure 11 demonstrates good scaling on GoogLeNet with OPA.[11]

Worldwide, many clients benefit from the unique value provided by systems with the Intel OPA fabric.

Growing lead over InfiniBand EDR – 30% more in Top100 list

Machine Learning Client Case Studies on the Intel Omni-Path Fabric

Although volume shipments just started in February 2016, Intel OPA quickly established 100G HPC fabric leadership in the Top500[12] supercomputer list. This lead over InfiniBand (IB) EDR is growing. In fact, in the June 2017 Top500 list, Intel OPA has 22% more total entries than IB EDR and 30% more entries in the Top100. Globally, Intel OPA has been deployed successfully in numerous systems at leading institutions – many of whom are doing pioneering work in AI and Machine Learning.

[9] "Top Deep Learning Projects," GitHub, September 2017.
[10] 1. Pre-released Intel Xeon Phi Processor codenamed Knights Landing-Fabric (KNL-F) w/integrated Omni-Path Architecture. B0 Stepping. QS upto 68 cores. 1.40 GHz. 98GB (6x16B) DDR4-2400 RDIMMS. OmniPath (OPA) Si 100 series. 48 port OPA switch with dual leaf switches per rack. 48 nodes per rack, 24 spine switches. Oracle Linux Server release 7.3. Kernel:3.10.0-514.6.2.01.el7_x86_64.knl1.
[11] Intel® Xeon® Processor E5-2697A v4 dual-socket servers with 64GB 2133 MHz DDR4 memory per node. 2 cores for MLSL and 30 MPI ranks per node. Intel® Turbo Boost and Hyper-Threading technology enabled. Red Hat Enterprise Linux* Server release 7.2 (Maipo). Intel® Parallel Studio XE 2017.4.056, Intel MPI 2017.3.196 MLSL 2017 Update 1 https://github.com/intel/MLSL/releases/tag/v2017.1-Preview Intel Caffe : https://github.com/intel/caffe
[12] www.top500.org

6X computing capacity enabling interactive supercomputing for ML innovations

MIT Lincoln Laboratory Supercomputer Center (LLSC)

Empowers Users to Unleash Innovations in Autonomous Systems/Machine Learning

Description/ Challenges	• Needed code modernization of legacy applications to exploit new multicore, multi-chip heterogeneous infrastructure architectures. • Bring both internal and external users up to speed without disrupting their normal work routine. • Exploring new areas, such as autonomous systems, device physics, and machine learning, demanded additional computational capabilities.
Solution/ Results	• Deployed a 1 petaflop supercomputer system with Intel Xeon Phi processors connected by the Intel OPA Fabric for all parallel processing projects. • Delivered a 6X computing capacity boost and an environment-friendly, "extremely green" computing center running 93% carbon free. • Quickly and efficiently processed Big Data from Autonomous systems and developed algorithms, allowing these systems to make intelligent choices. • Fast turnaround for rapid prototyping with in-device physics and interactive supercomputing on new Intel-based HPC systems.
Benefits	• Enabled users to focus on their core HPC and data analysis jobs and not worry about computing resources. The system now determines the best hardware/software for the job and handles heterogeneous workloads well. • Pioneering work on machine learning (ML) and neural networks enabled with interactive supercomputing. Also investigating speech/video processing and additional ML topics: theoretical foundations, algorithms, and applications.

Advancing petascale Deep Learning and AI across many fields

Pittsburgh Supercomputing Center

Advances the State of the Art in Petascale Computing and Artificial Intelligence

Description/ Challenges	• Traditional HPC systems limited since users needed to solve problems with varying paradigms using components across Big Data, AI, and other domains. • Needed a much more flexible system design without sacrificing performance.
Solution/ Results	• Built and installed system in two Phases. Phase 1 had 822 servers with Intel OPA. Phase 2 added 86 nodes and 126 TB of system memory. • Intel OPA fabric converged computing with very high bandwidth I/O and increased long-term storage capacity to 10 PB; enhancing community data collections, advanced data management, and project-specific data storage. • Improved flexibility by partitioning into groups of nodes for different workloads from traditional HPC, high-performance analytics, ML and visualization.
Benefits	• Early access to an advanced supercomputer improved science and research. • Expanded capacity for applying Deep Learning and AI across a wide range of fields in the physical and social sciences and the humanities.

Compared to InfiniBand, OPA consumed less power and was thermally more stable and reliable

Tokyo Institute of Technology (Tokyo Tech)

Extends AI/Machine Learning to New Areas: Energy, Healthcare, and Transportation

Description/ Challenges	• The existing TSUBAME 2.5 supercomputer outgrew its capabilities and was unable to meet computing needs in emerging areas and machine learning. • Issues with power consumption, thermal stability and reliability at scale.
Solution/ Results	• The upgrade, TSUBAME 3.0, is a hybrid CPU-GPU system with Intel Broadwell processors, NVIDIA Pascal Tesla P100 GPUs, and Intel OPA. • TSUBAME 3.0 uses water cooling to reduce power consumption, allowing the system to achieve the world's best performance efficiencies of 4.5 gigaflops/watt, achieving #1 on the Green500 List and a Power Usage Efficiency (PUE) of 1.033, one of the lowest in the industry. • Intel OPA consumed less power compared to InfiniBand, was thermally more stable, and adaptive routing enabled higher reliability at scale. • With 12.15 petaflops of peak double precision performance (24.3 petaflops single precision and 47.2 petaflops half precision), the system is flexible and ideal for neural networks common in deep learning and machine learning.
Benefits	• The merging of GPU with AI and HPC accelerates computation, enabling unprecedented advances in new areas: energy, healthcare, and transportation.

OPA with GPUs enabling ML use at scale in energy, healthcare and transportation

Texas Advanced Computing Center

Manages Epidemics, Forecasts Hurricanes/Hail and Creates More Viable Energy Sources

Description/ Challenges	• Increasing user demands to enable researchers to answer complex questions in epidemiology, weather forecasting, and efficient energy resources utilization. • Existing infrastructure unable to handle growing ML and analytics workloads.
Solution/ Results	• Deployed an 18 Petaflops Stampede 2 system with Intel Xeon and Intel Xeon Phi processors connected by Intel OPA; lowering power consumption. • Last system phase will integrate upcoming 3D XPoint non-volatile memory. • End-to-end solution for advanced modeling, simulation, and analysis needs of thousands of researchers across the country and scales to meet new increasing workloads such as ML at scale.
Benefits	• Massive performance increase will support data assimilation, design, control, uncertainty quantification, training/inference and decision-making for large-scale complex models in natural/social sciences, engineering and medicine. • Algorithms, software and libraries being developed for visualization, machine learning, analytics and bio and health informatics will be used to manage epidemics, forecast hurricanes and create more viable energy sources. • ML provides the most accurate forecasts of hail size, severity and locations; providing actionable alerts to mitigate damage to life and property.

ML predictions more accurate for catastrophic weather events and epidemics

High value ML use cases across many industries with Intel OPA

Intel OPA Poised to Accelerate Value Across Multiple Industries

The Intel Omni-Path Fabric has already unleashed unmatched ML innovation at research laboratories and academic institutions. Next, as many industries embrace AI and ML for digital transformation, the Intel OPA is well poised to profoundly accelerate adoption of many high value HPC-like use cases:

Automotive/Manufacturing: Predictive maintenance, Advanced Driver Assistance Systems (ADAS), dynamic pricing, Manufacturing-as-a-Service with online contracting, and optimizing supply chains.

Banking and Financial Services: Anti-money laundering, fraud detection, risk analysis, GRC (governance, risk, and compliance), security, privacy, customer segmentation, high-frequency trading, credit-worthiness evaluation, cross-selling and upselling services.

Consumer/Retail: Predictive inventory planning, real-time store monitoring analytics, cross-channel and upsell marketing, converting window shoppers to real shoppers, visual analytics, facial recognition, sentiment analysis, customer ROI and lifetime value, and recommendation engine.

Healthcare/Life Sciences: Clinical decision support and predictive analytics, electronic health records (EHR), image analytics (CT scans, MRI, and ultrasound), personalized medicine, and pathology.

Accelerates value across Manufacturing, Financial Services, Retail, Healthcare, Energy and Telco.

Oil & Gas and Utilities: Seismic data processing, preventive maintenance and equipment monitoring, power usage analytics with demand/supply optimization, smart building/cities/grid/meter management, customer-/time-of-day-specific pricing; essentially converting a *power grid* into an *information grid*.

Telecommunications: Reducing customer churn with targeted marketing, network performance/ capacity optimization, fraud detection and prevention, network security, and deep-packet inspection.

Conclusions

The ever-increasing volume, velocity and variety of data is creating both challenges and opportunities in every industry as they increasingly deploy AI /Machine Learning for their digital transformation. Training a model with larger datasets typically produces more Inferencing accuracy but this could take weeks/days. Parallel HPC-like systems coupled with productivity-enhancing software tools and libraries are increasingly being used to scale ML efficiently, particularly as faster Training and re-Training on new data become key requirements to improve Inference accuracy.

Intel provides a comprehensive AI/ML solutions portfolio with optimized frameworks, development tools, libraries, multiple high-performance processors and the Intel Omni-Path Architecture (OPA).

Intel's pioneering parallel computing research in AI/ML and unique Intel OPA features reduce communications overheads and enhance computational efficiency at scale. In fact, recently, a large-scale cluster system with Intel processors connected with the Intel OPA fabric broke several records for image recognition and other ML workloads.

Prospective clients interested in deploying and cost-effectively scaling AI/ML should seriously consider investing in systems connected with the Intel OPA for the following additional reasons/benefits:

Lower TCO and several performance and scaling records for Machine Learning

- Lower TCO for Machine Learning and HPC compared with InfiniBand EDR
- Many global research/academic institutions have already realized significant value for ML
- Growing lead in the Top500 supercomputer list compared with InfiniBand EDR
- Lastly, Intel is making significant investments to unleash AI/Machine Learning innovation.

Cabot Partners is a collaborative consultancy and an independent IT analyst firm. We specialize in advising technology companies and their clients on how to build and grow a customer base, how to achieve desired revenue and profitability results, and how to make effective use of emerging technologies including HPC, Cloud Computing, Analytics and Artificial Intelligence/Machine Learning. To find out more, please go to www.cabotpartners.com.