Editorial

May 23, 2018

Contact Intel PR

By Naveen Rao

This is an exciting week as we gather the brightest minds working with artificial intelligence (AI) at Intel AI DevCon, our inaugural AI developer conference. We recognize that achieving the full promise of AI isn't something we at Intel can do alone. Rather, we need to address it together as an industry, inclusive of the developer community, academia, the software ecosystem and more.

So as I take the stage today, I am excited to do it with so many others throughout the industry. This includes developers joining us for demonstrations, research and hands-on training. We're also joined by supporters including Google*, AWS*, Microsoft*, Novartis* and C3 IoT*. It is this breadth of collaboration that will help us collectively empower the community to deliver the hardware and software needed to innovate faster and stay nimble on the many paths to AI.

Press Kit: 2018 AI DevCon

Indeed, as I think about what will help us accelerate the transition to the AI-driven future of computing, it is ensuring we deliver solutions that are both comprehensive and enterprise-scale. This means solutions that offer the largest breadth of compute, with multiple architectures supporting milliwatts to kilowatts.

Enterprise-scale AI also means embracing and extending the tools, open frameworks and infrastructure the industry has already invested in to better enable researchers to perform tasks across the variety of AI workloads. For example, AI developers are increasingly interested in programming directly to open-source frameworks versus a specific product software platform, again allowing development to occur more quickly and efficiently.

Today, our announcements will span all of these areas, along with several new partnerships that will help developers and our customers reap the benefits of AI even faster.

Expanding the Intel AI Portfolio to Address the Diversity of AI Workloads

We've learned from a recent Intel survey that over 50 percent of our U.S. enterprise customers are turning to existing cloud-based solutions powered by Intel® Xeon® processors for their initial AI needs. This affirms Intel's approach of offering a broad range of enterprise-scale products – including Intel Xeon processors, Intel® Nervana™ and Intel® Movidius™ technologies, and Intel® FPGAs – to address the unique requirements of AI workloads.

One of the important updates we're discussing today is optimizations to Intel Xeon Scalable processors. These optimizations deliver significant performance improvements on both training and inference as compared to previous generations, which is beneficial to the many companies that want to use existing infrastructure they already own to achieve the related TCO benefits along their first steps toward AI.

We are also providing several updates on our newest family of Intel® Nervana™ Neural Network Processors (NNPs). The

Intel Nervana NNP has an explicit design goal to achieve high compute utilization and support true model parallelism with multichip interconnects. Our industry talks a lot about maximum theoretical performance or TOP/s numbers; however, the reality is that much of that compute is meaningless unless the architecture has a memory subsystem capable of supporting high utilization of those compute elements. Additionally, much of the industry's published performance data uses large square matrices that aren't generally found in real-world neural networks.

At Intel, we have focused on creating a balanced architecture for neural networks that also includes high chip-to-chip bandwidth at low latency. Initial performance benchmarks on our NNP family show strong competitive results in both utilization and interconnect. Specifics include:

> General Matrix to Matrix Multiplication (GEMM) operations using A(1536, 2048) and B(2048, 1536) matrix sizes have achieved more than 96.4 percent compute utilization on a single chip[1]. This represents around 38 TOP/s of actual (not theoretical) performance on a single chip[1]. Multichip distributed GEMM operations that support model parallel training are realizing nearly linear scaling and 96.2 percent scaling efficiency[2] for A(6144, 2048) and B(2048, 1536) matrix sizes – enabling multiple NNPs to be connected together and freeing us from memory constraints of other architectures.
>
> We are measuring 89.4 percent of unidirectional chip-to-chip efficiency[3] of theoretical bandwidth at less than 790ns (nanoseconds) of latency and are excited to apply this to the 2.4Tb/s (terabits per second) of high bandwidth, low-latency interconnects.
>
> All of this is happening within a single chip total power envelope of under 210 watts. And this is just the prototype of our Intel Nervana NNP (Lake Crest) from which we are gathering feedback from our early partners.

We are building toward the first commercial NNP product offering, the Intel Nervana NNP-L1000 (Spring Crest), in 2019. We anticipate the Intel Nervana NNP-L1000 to achieve 3-4 times the training performance of our first-generation Lake Crest product. We also will support bfloat16, a numerical format being adopted industrywide for neural networks, in the Intel Nervana NNP-L1000. Over time, Intel will be extending bfloat16 support across our AI product lines, including Intel Xeon processors and Intel FPGAs. This is part of a cohesive and comprehensive strategy to bring leading AI training capabilities to our silicon portfolio.

AI for the Real World

The breadth of our portfolio has made it easy for organizations of all sizes to start their AI journey with Intel. For example, Intel is collaborating with Novartis on the use of deep neural networks to accelerate high content screening – a key element of early drug discovery. The collaboration team cut time to train image analysis models from 11 hours to 31 minutes – an improvement of greater than 20 times[4].

To accelerate customer success with AI and IoT application development, Intel and C3 IoT announced a collaboration featuring an optimized AI software and hardware solution: a C3 IoT AI Appliance powered by Intel AI.

Additionally, we are working to integrate deep learning frameworks including TensorFlow*, MXNet*, Paddle Paddle*, CNTK* and ONNX* onto nGraph, a framework-neutral deep neural network (DNN) model compiler. And we've announced that our Intel AI Lab is open-sourcing the Natural Language Processing Library for Python* that helps researchers begin their own work on NLP algorithms.

The future of computing hinges on our collective ability to deliver the solutions – the enterprise-scale solutions – that organizations can use to harness the full power of AI. We're eager to engage with the community and our customers alike to develop and deploy this transformational technology, and we look forward to an incredible experience here at AI DevCon.

Naveen Rao is vice president and general manager of the Artificial Intelligence Products Group at Intel Corporation.

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 www.intel.com/benchmarks.

Source: Intel measurements on limited release Software Development Vehicle (SDV)

[1] General Matrix-Matrix Multiplication (GEMM) operations; A (1536, 2048), B(2038, 1536) matrix sizes

[2] Two chip vs. single chip GEMM operation performance; A (6144, 2048), B(2038, 1536) matrix sizes

[3] Full chip MRB-CHIP MRB data movement using send/recv, Tensor size = (1, 32), average across 50K iterations

[4] 20X claim based on 21.7X speed up achieved by scaling from single node system to 8-socket cluster.

8-socket cluster node configuration: CPU: Intel® Xeon® 6148 Processor @ 2.4GHz ; Cores: 40 ; Sockets: 2 ; Hyper-threading: Enabled; Memory/node: 192GB, 2666MHz ; NIC: Intel® Omni-Path Host Fabric Interface (Intel® OP HFI); TensorFlow: v1.7.0 ; Horovod: 0.12.1 ; OpenMPI: 3.0.0 ; Cluster: ToR Switch: Intel® Omni-Path Switch

Single node configuration: CPU: Intel® Xeon® Phi Processor 7290F; 192GB DDR4 RAM; 1x 1.6TB Intel® SSD DC S3610 Series SC2BX016T4; 1x 480GB Intel® SSD DC S3520 Series SC2BB480G7; Intel® MKL 2017/DAAL/Intel Caffe

Tags: Naveen Rao

Other News

April 8, 2021
SD Supercomputer Center Selects Habana, Intel for Efficient AI

April 6, 2021
Intel Launches Its Most Advanced Performance Data Center Platform

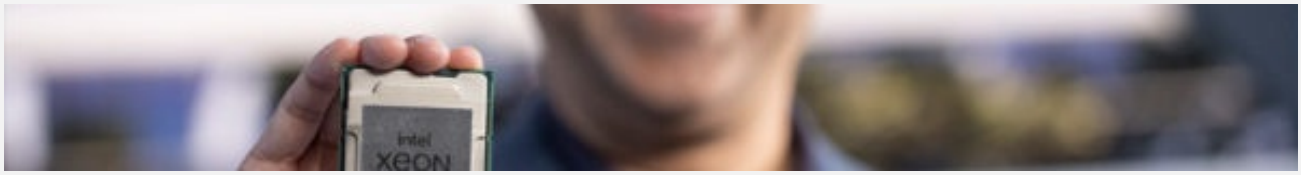
April 1, 2021
At John Deere, 'Hard Iron Meets Artificial Intelligence'

About Intel

Intel (Nasdaq: INTC) is an industry leader, creating world-changing technology that enables global progress and enriches lives. Inspired by Moore's Law, we continuously work to advance the design and manufacturing of semiconductors to help address our customers' greatest challenges. By embedding intelligence in the cloud, network, edge and every kind of computing device, we unleash the potential of data to transform business and society for the better. To learn more about Intel's innovations, go to newsroom.intel.com and intel.com.

Latest News: Artificial Intelligence

April 8, 2021
SD Supercomputer Center Selects Habana, Intel for Efficient AI



April 6, 2021
Intel Launches Its Most Advanced Performance Data Center Platform

April 1, 2021

At John Deere, ‘Hard Iron Meets Artificial Intelligence’

Read More