AI Software: Understanding the Rapidly Expanding Ecosystem

December 20, 2017 by staff
Leave a Comment

This feature continues our series of articles that survey the landscape of HPC and AI. This post focuses on AI software and how to best use and understand this rapidly expanding ecosystem.

The AI software ecosystem is rapidly expanding with research breakthroughs being quickly integrated into popular software packages (TensorFlow, Caffe, etc. and productivity languages (Python, Julia, R, Java, and more) in a scalable and hardware agnostic fashion. In short, AI software must be easy to deploy, should run anywhere, and should leverage human expertise rather than forcing the creation of a “one-off” application.

This whitepaper briefly touched on how Intel is conducting research to help bridge the gap and bring about the much needed HPC-AI convergence. Additional IPCC research insights and scientific publications are available on the IPCC Web resources website.

Even the best research is for naught if HPC and data scientists cannot use the new technology. This is why scalability and performance breakthroughs are being quickly
integrated into performance libraries such as Intel Math Kernel Library – Deep Neural Networks (Intel MKL-DNN) and Intel Nervana Graph.

The performance of productivity languages such as Python, Julia, Java, R and more is increasing by leaps and bounds. These performance increases benefit data scientists and aspects of AI from data preprocessing to training as well as inference and interpretation of the results.

"And important challenge in the convergence of HPC and AI is closing the gap between data scientists and AI programming models."

Julia, for example, recently delivered a peak performance of 1.54 petaflops using 1.3 million threads on 9,300 Intel Xeon Phi processor nodes of the Cori supercomputer at NERSC. The Celeste project utilized a code written entirely in Julia that processed approximately 178 terabytes of celestial image data and produced estimates for 188 million stars and galaxies in 14.6 minutes.

Jeff Regier, a postdoctoral researcher in UC Berkeley’s Department of Electrical Engineering and Computer Sciences explained the Celeste effort: “Both the predictions and the uncertainties are based on Bayesian model, inferred by a technique called Variational Bayes. To date, Celeste has estimated more than 8 billion parameters based on 100 times more data than any previous reported application of Variational Bayes.” Bayesian models are a form of machine learning used by data scientists and the AI community.

**Intel Nervana Graph: a scalable intermediate language**

Intel Nervana Graph is being developed as a common intermediate representation for popular machine learning packages that will be scalable and able run across a wide variety of hardware from CPUs, GPUs, FPGAs, Neural Network Processors and more. Jason Knight (C office, Intel Nervana) wants people to view Intel Nervana Graph as a form of LLVM (Low Level Virtual Machine). Many people use LLVM without knowing it when they compile their software as it supports a wide range of language frontends and hardware backends.

Knight writes, “We see the Intel Nervana Graph project as the beginning of an ecosystem of optimization passes, hardware backends and frontend connectors to popular deep learning networks.”
Intel Nervana Graph also supports Caffe models with command line emulation and Python converter. Support for distributed training is currently being added along with support for multiple hosts so data and HPC scientists can address big, complex training sets even on leadership class supercomputers.

High performance can be achieved across a wide range of hardware devices because optimizations can be performed on the hardware agnostic frontend dataflow graphs when generating runnable code for the hardware specific backend.

Figure 3 shows how memory usage can be reduced by five to six times. Memory performance is arguably the biggest limitation of AI performance so a 5x to 6x reduction in memory use is significant. Nervana cautions that these are preliminary results and "there, more improvements still to come".

Intel Nervana Graph also leverages the highly optimized Intel MKL-DNN library both through direct calls and pattern matching operations that can detect and generate fused calls to Intel Math Kernel Library (Intel MKL) and Intel MKL-DNN even in very complex data graphs. To help even further, Intel has introduced a higher level language called neon that is both powerful in its own right, and can be used as a reference implementation for TensorFlow and other developers of AI frameworks.

**Productivity languages**

An equally important challenge in the convergence of HPC and AI is closing the gap between data scientists and AI programming models. This is why incorporating scalable and efficient AI into productivity languages is a requirement. Most data scientists use Python, Julia, R, Ja...
and others to perform their work.

HPC programmers can be “parallel programming ninjas”, but data scientists mainly use popular frameworks and productivity languages. Dubey observes, “AI Software must address the challenge of delivering scalable, HPC-like performance for AI applications without the need to train data scientists in low-level parallel programming.”

Unlike a traditional HPC programmer who is well-versed in low-level APIs for parallel and distributed programming, such as OpenMP or MPI, a typical data scientist who trains neural networks on a supercomputer is likely only familiar with high-level scripting-language like Caffe or TensorFlow.

The hardware and software for AI devices is rapidly evolving. #HPC

However, the hardware and software for these devices is rapidly evolving, so it is important to procure wisely for the future without incurring technology or vendor lock in. This is why the AI software team at Intel is focusing their efforts on having Intel Nervana Graph called from popular machine learning libraries and packages. Productivity languages and packages that support Intel Nervana Graph will have the ability to support future hardware offerings from Intel ranging from CPUs to FPGAs, custom ASIC offerings, and more.

The insideHPC Special Report on AI-HPC will also cover the following topics over the next few weeks:

- An Overview of AI in the HPC Landscape
- AI and HPC: Inferencing, Platforms & Infrastructure
- AI Technology: The Answer to Diffusion Compartment Imaging Challenges
- AI Systems Designed to Learn in a Limited Information Environment
- Hardware to Support the AI Software Ecosystem

Download the full report: “insideHPC Special Report: AI-HPC is Happening Now” courtesy of Intel.