Intel’s oneAPI aims to unify AI code efforts across disparate hardware



Enlarge / Intel’s “Mega Trends in HPC” boil down to AI workloads, running on many kinds of hardware, largely in cloud—not on-premise—environments.
Intel Corporation

reader comments

0 with 0 posters participating

Share this story

Saturday afternoon (Nov. 16) at Supercomputing 2019, Intel launched a new programming model called oneAPI. Intel describes the necessity of tightly coupling middleware and frameworks directly to specific hardware as one of the largest pain points of AI/Machine Learning development. The oneAPI model is intended to abstract that tight coupling away, allowing developers to focus on their actual project and re-use the same code when the underlying hardware changes.

This sort of “write once, run anywhere” mantra is reminiscent of Sun’s early pitches for the Java language. However, Bill Savage, general manager of compute performance for Intel, told

Ars that’s not an accurate characterization. Although each approach addresses the same basic problem—tight coupling to machine hardware making developers’ lives more difficult and getting in the way of code re-use—the approaches are very different.

- In this simplified block diagram of AI/ML development, Intel wants to isolate “Languages & Libraries” as the biggest pain point.
  Intel Corporation
- oneAPI aims to uncouple the middleware and frameworks from the gritty details of the underlying hardware they target, making code more re-usable.
  Intel Corporation
- It’s tempting, but inaccurate, to hear oneAPI’s write once run anywhere promises and think “oh, it’s Java for AI.” oneAPI doesn’t produce bytecode, it abstracts optimized, low-level hardware targeting without replacing it.
  Intel Corporation
- In addition to low-level programming in the new Data Parallel C++ language and higher-level use of API calls, Intel is making compatibility, debug, and analysis tools available in the oneAPI layer.
  Intel Corporation

When a developer writes Java code, the source is compiled to bytecode, and a Java Virtual Machine tailored to the local hardware executes that bytecode. Although many optimizations have improved Java’s performance in the 20+ years since it was introduced, it’s still significantly slower than C++ code in most applications—typically, anywhere from half to one-tenth as fast. By contrast, oneAPI is intended to produce direct object code with no or negligible performance penalties.

When we questioned Savage about oneAPI’s design and performance expectations, he distanced it firmly from Java,

pointing out that there is no bytecode involved. Instead, oneAPI is a set of libraries that tie hardware-agnostic API calls directly to heavily optimized, low-level code that drives the actual hardware available in the local environment. So instead of “Java for Artificial Intelligence,” the high-level takeaway is more along the lines of “OpenGL/DirectX for Artificial Intelligence.”

For even higher-performance coding inside tight loops, oneAPI also introduces a new language variant called “Data Parallel C++” allowing even very low-level optimized code to target multiple architectures. Data Parallel C++ leverages and extends SYCL, a “single source” abstraction layer for OpenCL programming.

In its current version, a oneAPI developer still needs to target the basic hardware type he or she is coding for—for example, CPUs, GPUs, or FPGAs. Beyond that basic targeting, oneAPI keeps the code optimized for any supported hardware variant. This would, for example, allow users of a oneAPI-developed project to run the same code on either Nvidia’s Tesla v100 or Intel’s own newly released Ponte Vecchio GPU.

- Intel’s 7nm Xe architecture is intended to cover the entire range of GPU applications, but Ponte Vecchio—the first Xe product—specifically targets high-end deep learning and training in datacenter and supercomputing environments.
  Intel Corporation
- Ponte Vecchio is an exascale GPU aimed at enormous workloads, utilizing a 7nm process, Intel’s Foveros packaging tech, and high-performance interlinks.
  Intel Corporation
- The upcoming Aurora supercomputer at Argonne National Laboratories will include six Ponte Vecchio GPUs and two Xeon Scalable CPUs in each node.
  Intel Corporation

Ponte Vecchio is the first actual product in Intel’s new Xe GPU line and is targeted specifically at HPC supercomputing and data center use. Although neither Savage nor other Intel execs Ars spoke to had timelines or would speak to concrete products, one slide from Intel’s Supercomputing 2019 presentation clearly shows the Xe architecture as encompassing workstation, mobile, and gaming use—so there may be interesting times ahead for rivals in those spaces.

Savage told Ars that although the current version of oneAPI does still require developers to code for a particular architecture family—CPU, GPU, FPGA, etc—Intel plans for a future release to also allow automatic selection of the most optimal hardware type available.

The oneAPI toolkit is available for use and testing now at Intel Devcloud.

https://arstechnica.com/?p=1603199