Coverage Initiation: Codeplay aims to open up developer software stack for new AI accelerators

Analysts - John Abbott

Publication date: Tuesday, September 15 2020

Introduction

Codeplay, a compiler specialist that has worked on GPU-powered videogames for the past two decades, has turned its attention to artificial intelligence (AI). Its goal is to help AI accelerator vendors gain design wins by establishing an open standards-based cross-architecture software layer and ecosystem. Current software is typically tied to a single architecture and often supports only a limited number of graph compilers and programming models. Codeplay offers 'silicon enablement' services to customers that include Intel and Renesas. It backs the standard SYCL abstraction layer for the OpenCL framework, designed to execute parallel programming across heterogeneous platforms (CPUs, GPUs, FPGAs, DSPs and AI accelerators). The dominant incumbent in this space is NVIDIA with its CUDA software platform, which has built up a broad and established ecosystem aimed at AI developers. However, CUDA only works on top of NVDIA GPUs.

The 451 Take

Accelerated AI platforms and software development stacks are still relatively immature, but the early dominance of NVIDIA and CUDA as de facto standards has complicated moves toward more open standards and a broader choice of silicon – even though most of the popular frameworks in deployment are already open source. Standards take many years to mature, so companies looking to push the process forward more rapidly are an essential part of the ecosystem. Codeplay may find its efforts supported by powerful partners that would like to gain a slice of NVIDIA's AI business, including AMD, Arm and Intel. The vendor's tools and services could also offer a lifeline to the multiple AI chip startups that are struggling to implement broad enough software stacks of their own.

Context

Edinburgh-based Codeplay was founded in 2002 by game developer Andrew Richards, who remains in place as CEO. Its initial business was to create C compilers for game programmers on platforms including the PlayStation 2 and on x86 Windows PCs. By 2014, the firm had expanded its expertise to implement a vectorizing compiler on the Movidius vision-processing chip (now owned by Intel). By then it was working closely with industry consortium The Khronos Group (which owns the OpenGL and OpenCL specifications) on an implementation of the C++-based SCYL cross-platform abstraction layer for heterogeneous processors.

In 2017, Codeplay worked on tools for running OpenCL programs on Android smartphones and had made the full TensorFlow framework available on the Arm Mali GPU via SYCL. More recently, it has implemented SYCL and OpenCL onto the Renesas R-car system on chip for advanced driver assistance (ADAS) and deployed SYCL as the basis for TensorFlow support on GPUs from Imagination Technology. Williams Advanced Engineering is both a customer and an investor in Codeplay: In April 2018, it contributed £1m (about $1.5m) in funding for ADAS research.

Technology

Standards such as OpenCL, SYCL and Vulkan (cross-platform 3-D graphics) are the basis for Codeplay's Acoran platform and its full set of libraries for AI developers, and there's further support for the multiple open source projects that plug into them, including SYCL-DNN (deep neural networks); the major machine learning frameworks such as TensorFlow, GLOW and TVM AI; Intel's oneAPI unified programming model; and SPIR-V, a compiler intermediate representation for supporting AI graph compilers and new accelerated programming languages.

Acoran runs on – but is not limited to – AMD, Arm, Imagination, Intel and Renesas processors. NVIDIA GPUs are enabled through oneAPI DPC++ SYCL implementations. There are also two more focused packages: ComputeCPP (C++/SCYL for vision and machine learning) and ComputeAorta (compute-focused for OpenCL/SPIR-V, HSA and Vulkan). Typically, processor vendors write their own hand-tuned kernels for specific algorithms to run on the processor of their choice, employing their own tools. Codeplay then analyzes and extracts all of the performance principals from these and maps them to standard programming models. Through its open source ecosystem, a large number of AI operations are supported, including all of the latest applications and programming models for training and inference.

The deliverables are a set of customer-selected open source libraries configured for the processor, which can be licensed, integrated and distributed with the processor, kept as open source or kept as closed source branches, and then made available to developers. Through SPIR-V, customers can develop and deploy their own graph compilers if they wish. The same platform runs on all of the elements of a system – CPU, direct memory access, fixed function units, accelerator cores, and RAM and host DRAM, which all need to be run in parallel. System-level, multifunction integration of technologies such as machine learning, computer vision and sensor fusion results in overhead due to the large data movements that will be necessary unless single source code is enabled.

Strategy

Individual accelerator vendors with underdeveloped ecosystems – and that's most of them – have a problem supporting all of the standard graph compilers and programming models. If a customer finds its own individual requirements unsupported by an architecture, then the deal will be off. Codeplay's plan is to build a bridge between AI processors and standards-based AI software. Of course, the easiest option at the moment is to rely on NVIDIA and its CUDA platform, which has the strongest ecosystem.

But as alternatives to NVIDIA GPUs emerge for AI applications, some customers are seeking something more open. Large HPC projects looking to take advantage of emerging technology approaches are a good example, as HPC customers write a lot of their own software. For instance, Codeplay is working with Intel on the Aurora supercomputer, which uses Intel Scalable Processor CPUs and the forthcoming Xe Ponte Vecchio GPUs rather than NVIDIA. The firm evaluated its ComputeCPP SYCL product as a CUDA equivalent and to help developers port their existing CUDA applications across to the new platform.

Codeplay is targeting datacenter (at scale), edge, IoT and automotive devices for its tools, as well as SOC devices that include diverse elements at the system level. Its primary addressable markets are HPC, cloud datacenters, automotive ADAS, smartphones and tablets, medical and industrial equipment running AI, vision processing, machine learning, and big-data compute. Current reference customers include Arm, Broadcom, CEVA, Imagination, Intel, Qualcomm, Renesas and Synopsys, with other unnamed 'major companies' also signed up.

Competition

NVIDIA has built up its own ecosystem based on the CUDA platform, encouraging AI developers to support its platform in much the same way as it encouraged videogame developers a generation before. The introduction of its cuDNN libraries to support deep neural networks back in 2007 was a catalyst for this, leading to the development and support of popular deep learning frameworks such as Caffe, PyTorch and TensorFlow. NVIDIA GPUs are widely available and software developers can start work immediately.

Other established architectures from Intel, AMD and Xilinx, as well as Google for its own TPU, have software development kits (SDKS) available, but they are mostly architecture-specific. Intel's Data Parallel C++ (DPC++) is an interesting example. It's also based on SYCL and is cross-platform, targeting CPUs and accelerators via single source code and enabling custom tuning. However, Intel will focus its efforts on its own CPUs and specialist architectures, such as Movidius, Nervana and Habana.

AMD offers a similar interface as CUDA via its HIP API, a part of its broader ROCm (Radeon Open Compute) software stack, through which it too has gained access to the most widely used frameworks. Arm provides Arm NN, based on OpenCL. But for some new AI chip architectures and more specialized chips such as digital signal processors, SDKs can lack supporting software.

Codeplay's opportunity relies on a more widespread recognition that specialist silicon can deliver greater performance per watt than current GPUs, and that multiple system elements can be unified under a single, standards-based programming model without undue complexity. Challengers to NVIDIA's dominant market position in AI silicon include AMD, Arm, Intel, Qualcomm, Renesas and Xilinx, as well as startups such as Cambricon, Centaur, Cerebras, Esperanto, Graphcore, Groq, Hailo and SambaNova.

SWOT Analysis

Strengths	Weaknesses
Codeplay draws on nearly two decades of experience with compilers and programming APIs for game developers, much of it on GPUs.	AI developers will be reluctant to move away from the hardware and programming tools they already know, unless there are tangible benefits for doing so.
Opportunities	**Threats**
The availability of new architectures will inevitably challenge NVIDIA's dominance. Codeplay's migration	Customer inertia, continued heavy investment and potential acquisitions from NVIDIA could slow down the adoption of

capabilities will help open up the choices.	alternative approaches.

Source: 451 Research, LLC