AMD EPYC Empowers Server GPU Deep Learning

Whitepaper Sponsored by AMD
June 21, 2017

This paper is a companion to the **AMD EPYC Empowers Single-Socket Servers** white paper[1] and explores AMD's upcoming EPYC server system-on-chip (SoC) and its strong potential as a high-performance host to graphics processing unit (GPU) accelerators in servers. We specifically look at deep learning applications on both single-socket (1S) and dual-socket (2S) server designs – for third-party accelerators and those using the AMD Radeon Instinct GPU accelerators based on AMD's upcoming "Vega 10" architecture.

Deep learning (DL) is the application of large scale, multi-layer neural networks in pattern recognition.[2] DL is a subset of machine learning (ML), which is a subset of artificial intelligence (AI). DL provides the foundational set of algorithms and techniques that are fueling the rapidly evolving capabilities and applications of ML and AI. DL techniques underlie recent AI breakthroughs, such as accurate natural language processing for digital assistants and image recognition for autonomous vehicles. The heart of DL is matrix multiply operations, which can be accelerated by GPUs more efficiently than using the same number of CPU cores.

AMD can address DL with three technologies:

- EPYC server SoC – its processing, throughput, and I/O connectivity
- AMD's Radeon Instinct GPU
- Heterogeneous System Architecture (HSA)[3]

Given that AMD is the only semiconductor company to own both state-of-the-art x86 processor intellectual property (IP) and state-of-the-art GPU IP, this paper will show how an AMD SoC combined with AMD GPUs may provide additional system benefits.

Deep Learning History and Definitions

The recent history of DL is inextricably tied to GPU technology. While neural network algorithms have existed since the 1960s, several breakthroughs were needed to enable the "Cambrian Explosion of AI"[4] that started in earnest early this decade.

In 2001, commodity GPUs combined programmable shaders with floating point capability to support matrix multiplication offload. During the ensuing decade, experimentation on GPU acceleration flourished using brute-force and proprietary software solutions, which resulted in the open source OpenCL programming language launch in the summer of 2009. In the early 2010's

[1] http://www.amd.com/system/files/2017-05/TIRIAS-AMD-Single-Socket-Server.pdf
[2] http://www.theverge.com/2016/2/29/11133682/deep-learning-ai-explained-machine-learning
[3] http://www.hsafoundation.com/
[4] http://fortune.com/ai-artificial-intelligence-deep-machine-learning/

GPUs added faster PCIe Gen 2 and then PCIe Gen 3 system interfaces, fused multiply-add (FMA) instructions, and hardware thread scheduling, all of which made GPUs more appealing for neural net processing.

While GPUs were maturing into capable matrix math offload accelerators, new DL neural network techniques were evolving. Researchers extended the simple neural networks pioneered in the 1960s into more expansive and more interconnected networks of simulated neurons, and then added multiple layers of these new, more complex neural networks.

There are two aspects to the neural network behavior underlying these new DL techniques:

- "Training" feeds massive amounts of representative data through a neural network to train the network to recognize patterns in the data and to optimize the network framework. The training phase often requires floating point math to maintain enough accuracy through the various layers.
- "Inference" (sometimes called "classification" or "recognition") is the production end of neural network, where a service presents data to a trained neural network, and the trained network (using weights developed during the training process) then identifies patterns in the data – it infers (or classifies) the contents of the data. A trained neural network very efficiently classifies patterns in incoming data (the patterns discovered during training) and can do so with much less computational precision than used during training. The classification operation can often be performed with enough accuracy using integer math and can also be more latency sensitive.

Training integrates inference as a feedback mechanism to improve pattern recognition. If an inference neural network incorporates feedback, then it can learn after initial training by adjusting its weights, even while it is in production, recognizing patterns as a service.

The surge in GPU compute power made training new DL networks possible in short, relevant timeframes (such as weeks and then days). As this powerful and scalable combination of software and hardware entered the market, DL-based pattern recognition systems started to show remarkable pattern recognition results.

Because DL is so scalable, both for training and inference, Amazon Web Services (AWS) launched its EC2 Elastic GPU instance in late 2013. In late 2016, Microsoft's Azure service stared offering its N-Series instances. Recently Alibaba announced its PAI 2.0 ML platform and that it will be using Radeon Instinct in its GPU compute cluster. Google Cloud Platform (GCP) "Cloud GPU" instances were still in beta test at the time this paper was written.

These public services are derived from the GPU-enabled back-ends for services such as Amazon Alexa and Alexa Voice Service, Microsoft Cortana and Microsoft's Bing Speech application programming interface (API), Google Cloud Speech API, and Alibaba's Ali Xiaomi smart personal shopping assistant on Alibaba's e-commerce sites.

EPYC Can Host Up to Six PCIe Gen 3 x16 Accelerators

What makes AMD's EPYC so unique is that the SoC has 128 external Peripheral Component Interconnect Express (PCIe) Gen 3 lanes. Configuration of those lanes as PCIe Gen 3, Non-Volatile Memory Express (NVMe), or Serial AT Attachment (SATA) links (or second socket interconnect) is determined by the motherboard design (see Table 1). An EPYC SoC can directly connect up to 32 SATA or NVMe devices, or up to eight PCIe Gen 3 x16 devices. PCIe specifies 16 lanes as the maximum link configuration, and leading edge compute offload accelerators such as GPU cards implement all 16 lanes (although GPUs can run with less than 16 lanes for applications that do not require the maximum PCIe bandwidth).

Table 1: AMD EPYC's competitive landscape

Vendor	Intel	Intel	Intel	AMD
Model	Xeon D 1500	Xeon E3-1200 v5	Xeon E5-2600 v4	Epyc
Max Sockets	1	1	2	2
Core Generation	Broadwell	Skylake	Broadwell-EP	Zen
Core Count per Socket	4, 8, 12, 16	4	Up to 22	Up to 32
Thread Count per Socket	8, 16, 12, 32	8	Up to 44	Up to 64
Memory Type	DDR4/DDR3L	DDR4/DDR3L	DDR4/3DS	DDR4/3DS
Speed (MHz)	2400	2133	2400	2667
R/LRDIMM per Socket (GB)	128	N/A	1536	2048
UDIMM per Socket (GB)	64	64	N/A	N/A
DIMM Size - Max (GB)	32	16	128	128
Channels	2	2	4	8
DIMMs per Channel	2	2	3	2
PCIe Gen 3 Lanes	24	16	40	Up to 128
PCIe Gen 2 Lanes	8	No	No	No
Integrated SATA3	6	No	No	Up to 32
Integrated NVMe	No	No	No	Up to 32
Integrated USB3/2	4	No	No	4
Integrated Ethernet	2 x 10G	No	No	No
Requires Chipset	No	C230	C612	No

Source: TIRIAS Research

Each EPYC SoC has the same memory interface, whether in a single socket (1S) or dual socket (2S) configuration. In a 2S configuration, EPYC's system memory capacity doubles, but system

I/O is identical to a 1S solution – half of the I/O lanes on each socket are used for high-speed links between the two SoC sockets. The key attribute is that both 1S and 2S motherboards have the same I/O connectivity.

Each EPYC socket in a 2S configuration can therefore directly connect up to four PCIe Gen 3 x16 devices. However, the practical reality for EPYC to host compute offload accelerators, such as GPU AIBs, is that at least 16 PCIe lanes should be dedicated to networking, another eight to twelve lanes for two or three local storage drives (SATA or NVMe), with a few lanes remaining for miscellaneous other system interfaces, such as management controllers. Typical configurations will most likely be in the five to six GPUs per chassis range, whether single- or dual-socket.

DRAM bandwidth has become a performance limiter for AIB accelerator design, but EPYC overcomes this by matching PCIe bandwidth with memory bandwidth. PCIe Gen 3 supports up to 1 GB/s full-duplex, so 16 GB/s input simultaneously with 16 GB/s output. DDR4-2666 DIMMs have a peak transfer rate of 21.3 GB/s per DIMM. EPYC's eight memory channels combined have an aggregate peak transfer rate of 170 GB/s (half-duplex, or single direction, so transfer rate is for reads or writes but not both simultaneously). Six PCIe Gen 3 x16 accelerator AIBs can therefore generate an aggregate of 96 GB/s of data traffic in each direction, simultaneously. For portions of a workload that stream large amounts of training data into a bank of accelerator cards, 170 GB/s provides 77% performance headroom for a single EPYC SoC to stream data out of memory for each of six PCIe Gen 3 x16 cards.

We expect EPYC-based server end-users to configure their servers with PCIe Gen 3 acceleration cards based on GPUs, FPGAs, and other specialty compute accelerators. One compute SoC socket connecting to six PCIe Gen 3 x16 accelerators can increase 1U compute accelerator density by a large margin.

Popular 1U "GPU servers" host four full-sized PCIe x16 accelerator cards, but a large portion of the 1U chassis space is devoted to a second processor socket and its memory system. With a little creative design effort, it will be possible to include a fifth and possibly sixth full-sized PCIe x16 card into the chassis using EPYC. As a result, a 1U chassis can be more easily designed to host six short form-factor PCIe x16 accelerators, such AMD's upcoming Radeon Instinct MI8 (see below), especially if local storage is minimized in favor of network-based distributed storage.

AMD Radeon Instinct Architecture Addresses DL

AMD's Vega 10 GPU chips will be packaged in AMD Radeon Instinct graphics add-in board (AIB) form factors that plug into PCIe Gen 3 x16 slots. AMD's Vega 10 GPU chip has 64 graphics cores, each core containing 64 stream processors, for a total of 4,096 stream processors. AMD

Vega 10 and the Radeon Instinct family of AIBs will continue support for AMD's Multiuser GPU (MxGPU) virtualization technology.[5]

Vega 10 architecture is notable for its implementation of packed 16-bit floating point (FP16) instructions. Packing means that AMD implemented a little extra logic to allow two FP16 operations in the same execution pipeline as a single 32-bit floating point (FP32) operation. This can double the performance of algorithms that don't require full FP32 precision.

A single Radeon Instinct MI25 ("MI" stands for "machine intelligence") server GPU AIB can execute a total of 12 FP32 TFLOPS or 25 FP16 TFLOPS.

Table 2: AMD Vega 10 GPU compute capability

	Stream Processor (SP) per Clock	SPs per Core	SPs per Chip	MI25 TFLOPS
FP64	1	64	2,048	0.768
FP32	2	128	4,096	12.3
FP16	4	256	16,384	24.6

Source: TIRIAS Research

Figure 1: AMD Radeon Instinct MI25 AIB

Source: AMD

[5] http://www.amd.com/en-us/solutions/professional/virtualization

AMD is also shipping Radeon Instinct MI6 and MI8 AIBs. The MI6 AIB will use a previous generation “Polaris” GPU with 16 GB of GDDR5 memory that can run either FP16 or FP32 at up to 5.7 TFLOPS in a single-slot full-length 150W AIB. MI8 will use a previous generation “Fiji” GPU with 4 GB of high-bandwidth memory (HBM) in a short form-factor double-slot 175W AIB that can run either FP16 or FP32 at up to 8.2 TFLOPS.

AMD expects the MI25 AIB to primarily be used for DL training and mainstream high-performance computing (HPC) GPU computing, while the MI8 and MI6 AIBs should be attractive to DL inference applications. AMD expects MI6 to also be used as a low-cost Internet of Things (IoT) edge device training alternative.

From a GPU’s point of view, DL typically uses all available memory on a GPU AIB. Also, for each layer in a DL, memory access patterns are very regular and repeat in loops for many iterations, so memory access prefetches are repeatable and they can be optimized and automated. AMD’s use of second generation of highly parallel, high bandwidth memory (HBM2) in Vega 10 multi-chip packaging enables larger memory spaces for each Vega 10 chip while keeping on-chip cache sizes manageable and still maintaining performance.

AMD Open Source DL Developer Tools

AMD open sourced its GPU drivers, GPUOpen developer tools, (including the heterogeneous compute compiler “HCC”), and libraries as the ROCm software platform, available on GitHub[6] and now on revision 1.5. AMD enables DL and HPC programmers to leverage their existing software via a tool called “HIP”, which automatically translates source code from NVIDIA’s proprietary CUDA language to portable C++ source code.

Many developers rely on standard libraries when writing code. DL developers using CUDA also leverage NVIDIA’s cuDNN DL library. AMD recently released their free “MIOpen” 1.0 machine intelligence library, a functional equivalent to cuDNN. MIOpen 1.0 supports the Caffe, TensorFlow, and Torch 7 DL frameworks. Developers can replace cuDNN library calls with MIOpen calls to finish porting their DL code to run on AMD’s Vega GPU architecture. MIOpen includes commonly used DL functions that are optimized for Vega GPU architecture, such as convolution, pooling, activation functions, normalization, and tensors.

[6] https://github.com/RadeonOpenCompute/ROCm and https://rocm.github.io/packages.html

AMD states that HIP's CUDA conversion has up to 96% automated coverage. For example, developers working with a large project of 50,000 lines of code and 75 CUDA kernels ported their project from CUDA to MIOpen and then validated their port in about a week.

DL software ported to C++ using HIP will still run well on NVIDIA GPUs prior to replacing cuDNN library calls. HIP-generated C++ code can be compiled with either NVIDIA's compiler toolset or (after replacing cuDNN calls with MIOpen calls) with AMD's compiler toolset, so that developers can evaluate GPU accelerator performance, efficiency, and cost options.

AMD also continues their support for the HSA Foundation and the heterogeneous compute compiler (HCC), as well as language runtimes in ROCm, such as HIP. ROCm tools run on Ubuntu Linux versions 16.04 LTS (Canonical's "long term supported" enterprise releases), which are popular with and supported by major public cloud providers.

AMD SoC and GPU Better Together

There is a time-honored tradition in the high-tech industry of trying to provide proprietary "better together" solutions, but these single-vendor match-ups nearly always fail because they make unacceptable compromises when operating in a multi-vendor deployment.

AMD took a different path when looking at EPYC and Radeon Instinct together. Both products must operate in multi-vendor deployments – EPYC with a range of PCIe accelerator AIBs and Radeon Instinct with other vendors' processors. Because of this, using proprietary hardware or software lock-ins was and is unacceptable to AMD. AMD's better together story is about adding a bit of extra logic to the memory controllers on both products without making any compromises.[7]

AMD started with the fact that both products are HSA-compliant.[8] The original plan was to surface memory page faults and report precisely what happens. As a result, AMD implemented IOMMUv2 Address Translation Cache (ATC), which enables a GPU to directly access a processor's memory page tables.[9] AMD's implementation of ATC will be supported in hardware and is intended to work with existing Linux distributions. ATC will not be enabled at the launch of either EPYC or Radeon Instinct; AMD intends to enable the ATC feature at a later date.

ATC enables AMD's Vega 10 GPU to keep pace with AMD's EPYC SoC page table operations in real-time. The SoC does not have to pin virtual memory to a physical location before copying

[7] http://pdfpiw.uspto.gov/.piw?Docid=08578129

[8] http://www.hsafoundation.com/conformance-2/

[9] See pages 9-10 at http://www.hotchips.org/wp-content/uploads/hc_archives/hc26/HC26-11-day1-epub/HC26.11-2-Mobile-Processors-epub/HC26.11.220-Bouvier-Kaveri-AMD-Final.pdf

the contents of that memory to the GPU, and then unpin memory after the copy to free it back into the virtual memory pool. Pinning and unpinning takes time, which affects performance. The net effect of IOMMUv2 ATC will be more efficient data copies between SoC and GPU in many cases.

With virtual memory pointers freely shared between SoC and GPU, the GPU's high bandwidth memory cache sees one virtual memory pool spanning SoC and GPU. The underlying memory system will migrate data to the GPU's HBM as it is needed. Software developers writing applications on Radeon Instinct will be able to reference memory allocated by processes running under Linux on EPYC, and processes running under Linux on EPYC will be able to see Radeon Instinct stack variables and memory allocations. This bi-directional memory access should make it easier to develop new compute acceleration applications for or to port existing applications to Radeon Instinct.

IOMMUv2 ATC is invisible to applications code because the required changes are within AMD's ROCm software stack. AMD is collecting data to evaluate the performance impact before ATC is enabled in the market, and expects that many applications will run a little faster.

Conclusion

There are two inexorable trends in cloud computing today – the move to open source and the use of deep learning as an underpinning for intelligent services. The high-tech industry is still in the early days of the Cambrian Explosion of AI, and there is experimentation happening across the industry. Mastery of both open source and DL are strategic short-term imperatives.

AMD has been a force in open source infrastructure for over a decade. Now AMD's EPYC SoC, Radeon Instinct AIB, and ROCm software stack bring AMD's AI, machine learning, and deep learning strategy into focus. It is a solid strategy for AMD.

AMD has also reinforced its commitment to driving unified memory architectures, starting with AMD's founding role in HSA. It is getting more difficult to innovate in basic architecture, after decades of evolution and maturity, but AMD used the combination of its EPYC SoC and Radeon Instinct AIB to provide addition innovation without compromising interoperability with other vendors' hardware or software.

The combination of AMD's EPYC SoC and Radeon Instinct AIB with ROCm software platform will deliver a highly capable DL acceleration platform for AI based services.