Virtualizing the Heterogeneous Data Center for Computer Vision Applications: CPUs and GPUs

Michael Gutmann
AvidBeam Technologies, Inc.

The Evolving Data Center

Data Center resources are constantly evolving. The compute, I/O, software and service infrastructures that define data centers evolve as applications demand new capabilities and capacity. Indeed, the most basic goal of *data center virtualization* is to efficiently hide data center complexity and infrastructure in order make accessible its vast resources[1] [2].

Consumer-grade components make-up the back-bone of data center processing. Intel Architecture desktop and laptop CPUs, in various server flavors, have been the processors of the data center for more than 20 years[3]. But this is beginning to change. And the change is not coming from processor designers, but as a result of revolutionary new AI software.

New AI applications are being realized on high-density, multi-core architectures. These are architectures not of 10s of server cores but 1000s of "AI cores" – Cores used for deep neural and convolutional network training[4] [5].

The new machine learning algorithms consume the vast Big Data repositories of the data center. Combining Big Data with the new, massively parallel GPU[6], TPU (Google's Tensor Processing Unit)[7], and similar multi-core platforms is now a goal. The next-generation data center will bring high-performance computer (HPC) capabilities to Big Data, as machine learning is in demand by increasing numbers of developers and applications[8] [9] [10].

The Heterogeneous Data Center

Today's internet runs efficiently on Intel Architecture server cores, whose deployment in data centers represents a large capital investment. The Big Data phenomenon has been achieved by marshalling 1000s of traditional CPUs for parallel search and large file system implementation[11]. For the foreseeable future,

traditional Web applications and transactions will continue to execute using today's data center architecture.

The new HPC machines NVidia, Google and others are beginning to offer, with 1000s of closely interconnected cores, are challenging Intel servers running the new AI. Some public data centers are offering access to the new machines[12]. Yet, for the most part, today's data center consists of a mix of Intel servers, some with attached GPUs, that can be used for computationally intensive algorithms.

The pragmatic challenge, in the video processing space, is how best to distribute computer vision processing across a heterogeneous mix of data center resources.

AvidBeam Technologies has adopted a heterogeneous approach in creating its *Video Big Data* architecture. Video data is distributed on a very fine-grain – up to frame-by-frame distribution – across *ALL* available cores and groups of cores. Those cores may be Intel server CPU cores and non-Intel GPU cores.

In addition, whatever the compute platform of the new computer vision applications, there remains the problem of efficiently distributing source video to allocated computational resources; that is, the efficient aggregation of 1000s of camera feeds and/or video files and their distribution to a computer vision algorithm. Here, the AvidBeam solution plays a vital role, as well.

Heterogenous Server/Core Configurations

The following sketches indicate the variety of CPU and GPU configurations, in general terms, that may be available to computer vision algorithms running in a heterogenous data center.

Figures 1a-1b depict a range of Intel Architecture servers, with a baseline 16-core server (1a)[13] and a higher-end 72-core configuration (1b)[14].

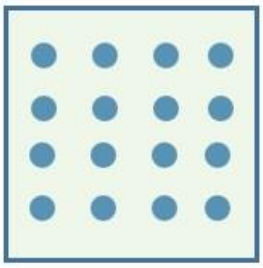

Figure 1a. Intel 16-core server

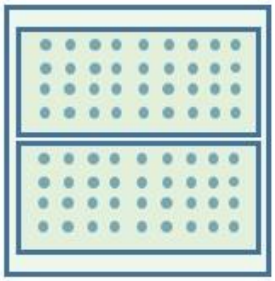

Figure 1b. Intel 2x36 72-core server

The 16-core server with 16GB of RAM is a commonly encountered multi-core server in today's data centers. The 72-core server is an example of a much higher-end server configuration.

Figure 2 depicts an Intel server with attached GPU card[15]. Based on the computer vision algorithm that is deployed, the Intel cores may be used to manage the GPU, as well as provide processor resources for some algorithm functions.

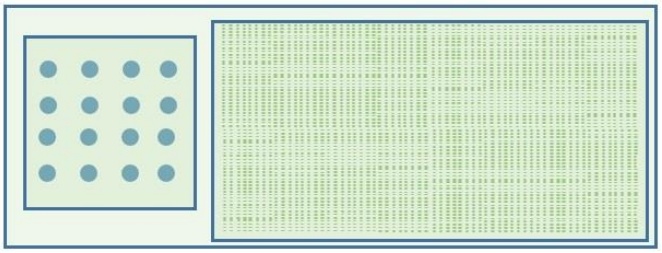

Figure 2. IA 16-core server with attached GPU card

The Figure 2 configuration, above, has been deployed in a number of data centers, where traditional IA servers are upgraded with one or more GPU cards. The GPUs are then available to computer vision and machine learning algorithms. Sharing GPU cards among instances of an algorithm then becomes a challenge when video is distributed across large number of servers without GPUs (Figure 1) and servers with GPUs (Figure 2).

Figure 3 depicts a high-end multi-GPU card complex that is dedicated entirely to machine learning. The front-end, shown on the left-hand side, consists of 72 IA cores, and manages a complex consisting of as many as 28,000 GPU cores[16].

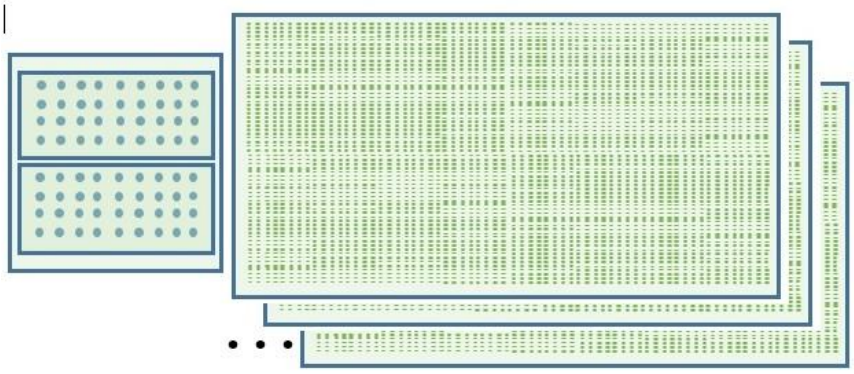

Figure 3. HPC GPU configuration with multiple high-density GPU cards

The illustrations under-score the difficulty of executing computer vision and other machine learning applications in a heterogenous server environment. With the exception of the third example (Figure 3), which will be used as a dedicated appliance for the most demanding machine learning development, the author of a computer vision application needs support in managing the distribution of his or her algorithm across the various platforms, as well as an efficient connection to video inputs, from real-time camera feeds, or stored, off-line sources. The AvidBeam Video Big Data software addresses the deployment of algorithms across the various multi-core platforms.

Executing CV Algorithms in the Data Center/Cloud

Figure 4, below, depicts a high-level view of the AvidBeam infrastructure for hosting computer vision applications in the data center/cloud.

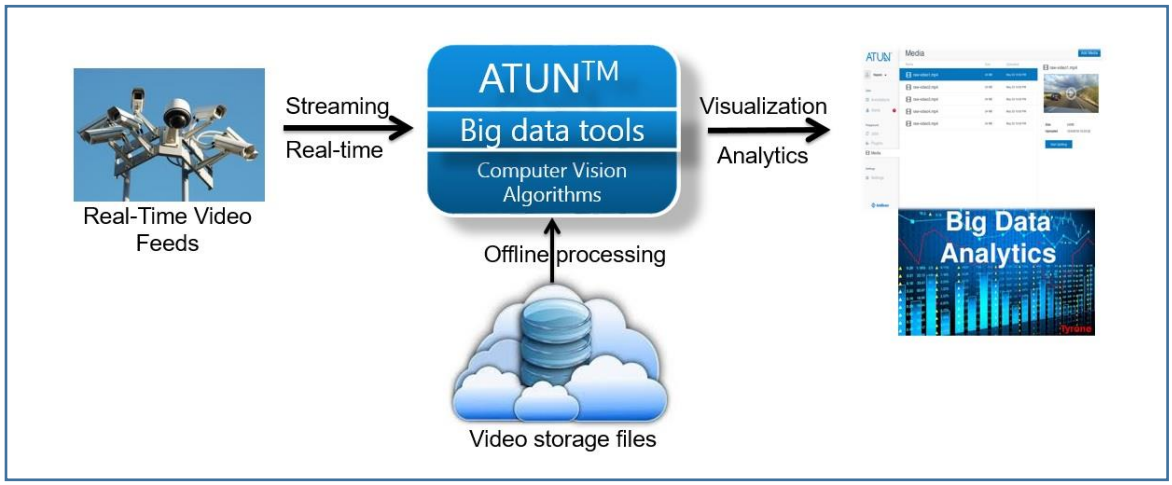

Figure 4. AvidBeam Video Big Data Tools

The ATUN™ product integrates computer vision (CV) algorithms with Big Data Tools.

Three objectives of the software are:

1. Provide a means to easily integrate in-house and third-party CV algorithms[17] [18] within the data center. CV algorithms, wrapped in various industry APIs; e.g., OpenCL, OpenCV, CUDA, etc., are plugged into the ATUN™ framework, which

will then distribute video content to CPU and GPU instances allocated by the application. The CV developer is hidden from details of algorithm distribution and parallel execution within the larger data center infrastructure. Provided the CV algorithm is supplied in both CPU and GPU implementations, it can be mapped to a wide variety of server/core configurations.

2. Distribute video content, from both real-time and off-line sources, to the CV algorithm instances. In the case of state-less image recognition and search functions, video frames may be distributed on a fine-grain, per-frame basis. The system matches video sources to CV algorithm instances virtualized across all available CPU and GPU resources.
3. Provide post-processing interfaces so that classified and tagged output from the CV application can be easily directed to data analytics products and tools (in-house and third-party) for additional machine and human consumption.

Efficiently Managing Server Resources

Figure 5, below, shows a recent relative measure of data center server cost based on the use of heterogenous core resources.

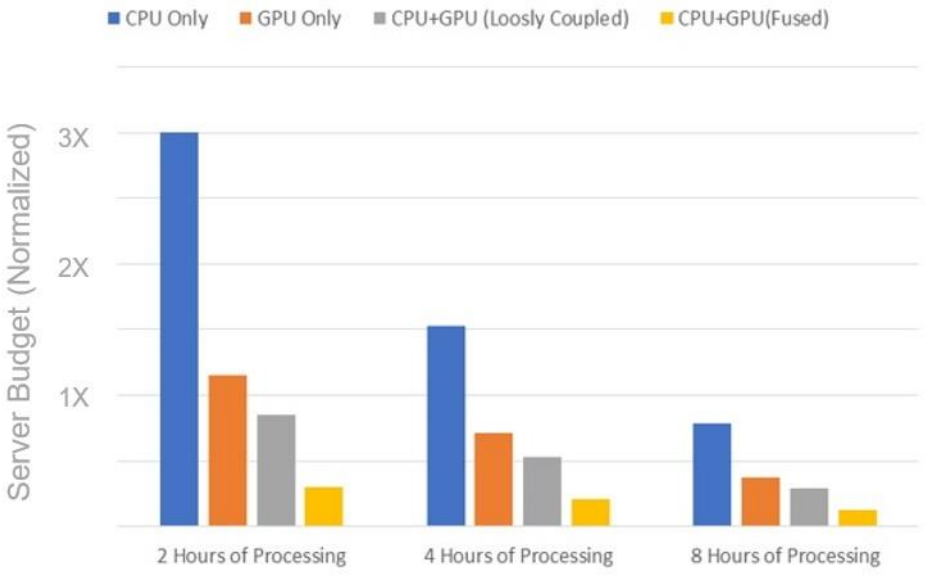

Figure 5. Relative Server Cost as a Function of Core Usage

The benchmark data is taken from an offline video classification task based on eight hours of stored video. The “2 Hours of Processing” data indicate that enough server

resources were allocated to process the video at a 4X rate (eight hours of video content in two hours). The other data show lesser server resources allocated.

Four server/core utilizations were measured: CPU-only; GPU-only; CPU+GPU loosely coupled; and CPU+GPU fused. A dramatic overall savings in server budget is realized when both CPU and GPU cores are allocated and executed simultaneously.

When GPU cards are attached to traditional Intel server CPUs there may occur race conditions among CV algorithm instances and available GPU processors. An unsupervised sharing of limited GPU resources may act to create thrashing among algorithm instances vying for GPUs. This problem has been addressed in the ATUN™ software by using a modified token-based technique, originally used with network contention strategies, for allocating GPUs. It allows for an efficient sharing of GPU resources among competing CV algorithm instances[19].

Conclusion

The goal of the AvidBeam approach is to provide the means to transparently and efficiently virtualizing data center/cloud servers and cores. As traditional Intel Architecture servers are augmented with GPUs, AvidBeam's ATUN™ software relieves the CV algorithm designer from the many details of a data center/cloud deployment across heterogenous server architectures consisting of both CPUs and GPUs.

Realizing a "plug-and-play" CV algorithm environment, while supporting a large number of video feeds, is the goal of AvidBeam Technologies. If successful, a significant reduction in new CV application startup and server costs is possible.

END NOTES

1 Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler, The Hadoop Distributed File System, Yahoo, 2010.

2 Apache Hadoop. http://hadoop.apache.org/

3 Timothy Pricket, X86 Servers Dominate the Datacenter–For Now, www.nextplatform.com, 2015.

4 Jonathan Tompson, Kristofer Schlacter, An Introduction to the OpenCL Programming Model, NYU Media Research Lab, 2013.

5 Alex Krizhevsky, Ilya Sutskever, Geoffery E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, Advances in Neural Information Processing Systems 25, NIPS 2012.

6 https://www.nvidia.com

7 https://cloud.google.com/tpu/

8 https://aws.amazon.com/ec2/Elastic-GPUs/

9 http://www.cirrascale.io/

10 https://azure.microsoft.com/en-us/blog/azure-n-series-general-availability-on-december-1/

11 Cade Metz, Google's Data Center Engineer Shares Secrets of 'Warehouse' Computing, https://www.wired.com/2012/01/google-man/, 2012.

12 Max Smolaks, Google Cloud opens access to GPUs in its data centers, http://www.datacenterdynamics.com, 2017.

13 https://www-ssl.intel.com/content/www/us/en/products/processors/xeon/e5-processors/e5-4660-v4.html

14 Intel® Server Board S7200AP and Intel® Compute Module HNS7200AP TPS

15 PowerEdge C4130 Rack Server with NVidia Tesla GPU. www.dell.com

16 https://www.nvidia.com/en-us/data-center/dgx-1/

17 http://caffe.berkeleyvision.org/

18 https://www.tensorflow.org/

19 Hazem Abdelhafez, Mohamed Rehan, and Hossam A. H. Fahmy, Efficient GPU utilization in heterogeneous big data cluster using token-based scheduler, 30th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), Windsor, ON, Canada, April 2017.