HPC MEETS GPUS: A LOVE STORY

Produced by Tabor Custom Publishing
in conjunction with:

HPC wire

SUSE

Executive Summary

When High Performance Computing (HPC) first encountered the GPU (Graphics Processing Unit) they appeared to have very little in common. HPC was all about parallel computing using powerful supercomputers and Linux cluster management to solve complex computational problems, whereas GPUs were computer hardware that found a niche in speeding up gaming graphics. Fast forward to today, where software-defined HPC is becoming an integral part of computing. Whether it's scale-up capabilities or some variant of scale-out – in the data center, in the cloud, or at the edge – organizations will require massive amounts of compute. "Satisfying this need has resulted in a growing ecosystem of speedy interconnects, new storage technologies, and specialized processors (think GPUs, TensorFlow Processing Unit (TPUs), field programmable gate arrays (FPGAs), and whatever else is coming down the pike," according to panelists at the Advanced Scale Forum 2019 conference.

In today's HPC and artificial intelligence (AI) arena, both Central Processing Units (CPUs) and GPU hardware are being used, along with Linux-based HPC software that supports those two hardware environments. Central processing units running on x86 processors have been used as the brains of the computer and the standard building blocks of data centers for decades. In 2018, 451 Research indicated that the CPU has an extensive user base, market ubiquity, an established ecosystem and ongoing optimization to work in a parallel computing environment. GPUs (also called accelerators) deliver a parallel environment because it is cost-effective to have many of them to provide a highly scalable environment. GPUs have proven efficient in training AI models and have been optimized with frameworks to make them easier to use. However, the role of CPUs versus GPUs being used for HPC and AI is changing. John Abbott, Co-Founder of the 451 Group states, "Without the new accelerators, current infrastructure won't be capable of delivering on the compute performance and power efficiency requirements demanded by tasks such as neural network training and inference[i]."

Expanding use of GPUs

The GPU is widely adopted by data scientists. Both GPUs and HPC have evolved and provide an essential hardware platform for everything from AI and machine learning to four-dimensional simulations to medical imaging to financial modeling. In fact, GPUs have found a home in HPC since many of these workloads run extremely well in a parallel environment (as long as the applications are written and designed to leverage a multi-cluster, multi-processor or multi-GPU environment).

Across industries, new waves of applications are being optimized to take advantage of parallel computing – as the problems needed to solve and their data sets become larger and larger. GPUs are viewed as accelerators while CPUs are ideal for compute intensive workloads. The GPU design lends itself well to massive parallel processing, especially in many math acceleration tasks. The adoption of GPUs as general-purpose accelerators began with data scientists, researchers and developers working on deep learning and neural network projects, some of which are now turning into production deployments across a wide range of businesses running artificial intelligence (AI), machine learning (ML), and deep learning (DL). To further aid in using GPUs for AI, ML and DL processing, the OpenACC ("Open Accelerators") community has a mission of propagating accelerators across GPUs and HPC environments with the help of contributing members like NVIDIA and SUSE.

GPUs speed processing

A GPU is a specialized electronic circuit, also called an accelerator, which was originally developed to quickly manipulate graphic images. GPUs were originally adopted by data scientists due to their highly parallel and vector processing capabilities. But GPUs are seeing major growth in AI and HPC because they are designed to rapidly process tasks such as linear math algorithms which are a major part of DL and ML. 451 Research[ii] indicates that CPUs "are being supplemented or replaced by specialized accelerators capable of boosting the performance of more targeted workloads particularly AI and machine learning by a significant factor." The trend of using GPUs for HPC and AI data intensive processing is also aided by the development of NVIDIA's CUDA language and AI frameworks such as TensorFlow and Caffe which make it easier to get applications running efficiently on GPUs. Here are two examples of how GPUs are used for AI and DL:

- Galaxy formation[iii]: Research teams at the University of California, Santa Cruz, and Princeton University are using AI along with NVIDIA GPUs to study how galaxies are formed.
- Creating better games: Microsoft is using NVIDIA's Volta Tensor Cores[iv] that allow developers to create a game environment, difficulty, or appearance that is tailor-made for the player. It can do this through varying models of neural networks that change and adapt in real-time to the stimuli received from the player.

Evolution of HPC, AI and Market Drivers

HPC and AI was traditionally limited to use in academia, large research institutions, and government. However, the need to analyze massive amounts of data-intensive workloads is driving the use of HPC into all areas of the business arena as described in the HPC Goes Mainstream paper[v]. This need will only grow in the future as transformative technology such as AI, ML, DL, bioscience analytics, robotics, blockchains, virtual and augmented reality, computational fluid dynamics and simulations require organizations to create, analyze and derive business value from massive amounts of data.

According to Hyperion 2019 research[vi], "Hyperion expects healthy worldwide HPC-based revenue expansion in machine learning, deep learning and AI, growing from just more than $200 million in 2015 to more than $600 million last year to slightly more than $1.2 billion in 2021; for the 2017-2022 period, Hyperion predicts a compound annual growth rate (CAGR) for worldwide HPC-based AI revenues of 26.3 percent; for High Performance Data Analytics (HPDA): 14.9 percent." John Abbott, 451 Research states, "We believe that ML will be embedded into all major business applications within the next five years, automating various processes within them"[vii].

HPC AI Customer Pain Points

Because HPC infrastructure is complex, businesses encounter many obstacles in using HPC systems running data intensive HPC and AI/ML workloads including complexity in the HPC environment relating to cluster management, storage access time and data management. In addition, organizations need to maximize performance as well as scale performance and minimize overhead. "Businesses around the world today are recognizing that a Linux-based HPC infrastructure is vital to supporting the analytics applications of tomorrow, from the edge to the core to the cloud", states Jeff Reser, SUSE HPC Strategist.

HPC Case Studies

HPC and AI have become the engines of human discovery, turbo-charging progress in the fields of medicine, science and engineering. They are integral to curing cancer and uncovering the origins of our world and the universe itself. HPC, AI and use of GPUs are already being used in sectors including financial services, media and entertainment, fraud detection, bioscience, healthcare and pharmaceuticals, oil exploration and weather and climate forecasting.

"As medical science progresses, AI is being used to diagnose patients, offer treatment plans and prescribe medication, all in a matter of seconds. Manufacturers need to quickly build predictive models from historic data that will allow them to react ahead of time to breakages. And computing environments — both private and public — are increasingly under the protection of complex pattern-matching applications that alert administrators to suspicious behaviors that may be malicious. HPC and AI are working together to save lives, protect digital assets and cut corporate costs," states Reser.

Here are some examples of HPC and AI at work:

Pittsburgh Supercomputing Center (PSC) case study:

Researchers at PSC hope to build a diagnostic chip that may identify heart disease in humans. They have been screening more than 100,000 mutant mice to find heart defects, sequencing the genomes and comparing the results to the genome of a healthy mouse. With the PSC-SGI machine running SUSE Linux Enterprise server, processing that had been taking almost two weeks was cut to less than eight hours.

Oil & Gas Research: Total E&P case study:

Total, the world's fifth largest publicly-traded integrated oil and energy companies uses HPC in energy research. Total Exploration and Production (Total E&P) models oil reservoirs – no small task – using GPUs in an HPC environment. All of this requires heavy mathematical processing, which GPUs specialize in. Reserves are found by bouncing sound waves off the Earth's surface and looking for echoes that indicate an oil reserve. Then the reflected wave data is turned into images that geoscientists can use to determine if a reservoir prospect contains hydrocarbons and where the hydrocarbons are located within the image. This determines whether or not it is worth it to drill for oil in that reserve. Using its Pangea supercomputer and SUSE Enterprise Linux Total E&P found, "Using the system, we can run ten times the number of simulations we ran with our previous supercomputer, helping us identify potentials deposits and determine the best extraction methods more easily."

Reenergized HPC Sector with GPUs and Applications

The introduction of new accelerators for AI workloads has helped re-energize the HPC sector and provides a way for HPC vendors to broaden out their total addressable markets. At the same time the hyperscalers, such as Amazon and Google, are upgrading their infrastructure with HPC-like capabilities to meet the growing demand for computationally complex applications such as deep learning and large-scale analytics. Similarly, more HPC capacity is likely to be delivered from the cloud.

Reser states, "When integrating revolutionary applications into existing infrastructure, HPC is indispensable. Therefore, careful thought must be given to the HPC architecture that brings together hardware and software to optimize resource consumption for potent scalability and performance delivery. When one thinks about supercomputing, you get visions of astronomical expense. But today, such deployments are much more common in the mid-range commercial space. HPC lends itself naturally to AI and ML applications, which typically require churning through a lot of data, very quickly. Efficiency is of paramount importance, and that's what HPC provides.

As a result, multi-cloud and edge computing will open up whole new areas of business for both HPC and hyperscalers. It's all about extending standard Linux clusters with more specialized requirements – such as SMP (Symmetric Multiprocessing), shared memory, different flavors of input/output (I/O), and the addition of vector-like capabilities (through GPUs) to the primarily scalar resources of CPUs. This hybrid model will enable classic HPC workloads (numerically intensive, data-driven), to be run side-by-side with more complex workloads where lots of tasks are being run concurrently."

Introducing SUSE HPC Products: Meet Growing AI Needs

SUSE supports the AI/ML market in a way that brings value to their partners and customers, starting with an essential HPC platform that incorporates the technologies and tools necessary to manage the parallel cluster environment. SUSE delivers an HPC infrastructure that is ML-ready and provides the tools needed to gain valuable business insights faster. Coupled with scalability and interoperability – HPC and GPUs form a match made for parallel computing success.

Harness HPC power: SUSE Linux Enterprise High Performance Computing (SLE HPC) manages data-intensive workloads and runs on multi-core processors (both x86- and ARM-based) and GPUs. It includes performance monitoring, high performance message passing, advanced multipathing and I/O capabilities. It delivers and supports tools like Slurm for cluster management, Ganglia for workload monitoring, OpenMPI for networking and much more as part of an HPC module that comes with the OS platform. In addition, there are AI/ML development frameworks (i.e., TensorFlow), file systems (e.g., Ceph-based SUSE Enterprise Storage, Lustre, NSF), and scientific and mathematical libraries that are available separately through SUSE Package Hub or SUSE partners that are tested and proven to run on the SUSE HPC platform.

Increase Storage for AI and HPC data: Businesses use HPC systems to analyze large amounts of data which leads to the need for additional methods to store all the data. For example, organizations with large, complex problems to solve (e.g., weather forecasting) need an HPC environment that supports parallel clusters and powers advanced analytics or modeling workloads. The more data made available to these applications, the more useful and accurate the results. Therefore, data is collected from more places and big data environments are the norm in HPC-based infrastructure. SUSE Enterprise Storage is a software-based solution powered by Ceph technology. SES provides low-cost, easy to manage software-defined storage for two use cases:

1. A smaller HPC environment with two through eight clusters (where this is less of a need for extremely low-latency storage).

2. Tier 2 (backup/archival) storage for HPC data.

Utilize OpenStack Cloud flexibility: SUSE OpenStack Cloud delivers enterprise-ready technology for building Infrastructure-as-a-Service (IaaS) private clouds.

Consolidate using Container virtualization: According to 451 Research[viii], "Container technology, specifically system containers, is essentially operating system virtualization – workloads share operating system resources such as libraries and code. Containers have the same consolidation benefits as any virtualization technology, but with one major benefit – there is less need to reproduce operating system code." So containers effectively provide the same capability with fewer resources. The SUSE Container as a Service Platform (CaaSP) is an enterprise class container management solution that enables IT and DevOps professionals to more easily deploy, manage, and scale container-based applications and services. It includes Kubernetes to automate lifecycle management of modern applications, and surrounding technologies that enrich Kubernetes and make the platform itself easy to operate. In addition, SUSE CaaSP provides trusted prebuilt container images, the ability to meet compliance requirements, perform audits and inspect the images of containers, including the running containers and management of containers at scale.

SUSE Joins OpenACC supporting HPC compilers to aid in parallel computing

OpenACC ("Open Accelerators") is a directives-based programming approach to parallel computing designed for performance and portability on CPUs and accelerators for High Performance Computing. SUSE joined OpenACC and is proud to be working on the GNU Compiler Collection (GCC) to aid in making parallel computing tools to run cross-platform. "SUSE is a full-fledged member of the OpenACC community and continues to work closely with GPU vendors like NVIDIA – joint efforts that are necessary and foundational for progressing AI/ML workloads," states Reser.

Conclusion

The HPC market is changing, with opportunities in almost every sector driven by the need to support new data-intensive workloads like AI/ML and analytics. From the edge to the core to the cloud, these applications are benefiting from a software-defined HPC infrastructure. HPC was once the domain of academia, researchers and governments using supercomputers to solve complex computational problems. GPUs were originally created to aid in graphics processing and speed gaming graphics. GPUs can provide a cost-effective highly scalable environment that accelerates processing on a massive scale. In addition, GPUs excel at massive parallel processing, especially in many math acceleration tasks that are commonly used in AI, ML and DL processing. AI is now moving into mainstream processing as more organizations use HPC applications and GPUs and to process, categorize and analyze massive amounts of data.

"Today, both an HPC software infrastructure and GPU acceleration have evolved and become essential elements of the platform for managing data-intensive workloads –for everything from artificial intelligence and machine learning to four-dimensional simulations to medical imaging to financial modeling. In fact, GPUs have found a home in HPC since many of these workloads run extremely well in a parallel environment. Across industries, new waves of applications are being optimized to take advantage of parallel computing – as the problems needed to solve and their data sets become larger and larger," states Reser.

SUSE addresses HPC infrastructure needs with an OS and popular tools and libraries designed to help manage workloads running in parallel cluster environments. SUSE software-defined storage provides Ceph-based storage for HPC archival and backup. SUSE also delivers Kubernetes for container management and OpenStack for cloud deployment – both becoming even more important in high performance data analytics, AI and machine learning.

About SUSE

SUSE® provides and supports enterprise-grade Linux and open source solutions with exceptional service, value and flexibility. With partners and communities, we innovate, adapt and deliver secure Linux, cloud infrastructure and storage software to create solutions for mixed enterprise IT environments – all backed by always-on and experienced SUSE Support.

---

i John Abbott, 451 Research, The New Accelerators: How the Performance and Efficiency of the Datacenter is Transforming for New Workloads. 2019.

ii John Abbott, 451 Research, The New Accelerators: How the Performance and Efficiency of the Datacenter is Transforming for New Workloads. 2019

iii Tony Kontzer, NVIDIA, Scaling the Universe: How Large GPU Clusters Aid Understanding of Galaxy Formation. December 07, 2018. https://blogs.nvidia.com/blog/2018/12/07/gpu-clusters-aid-understanding-of-galaxy-formation/

iv Jacob Ridley, PCGames, Microsoft want to use your GPU's AI skills to make your games prettier and more deadly. 2018. https://www.pcgamesn.com/microsoft-ai-powered-gaming

v HPCwire, Tabor Networks, HPC Goes Mainstream, 2018. https://www.hpcwire.com/hpc-goes-mainstream/

vi Doug Black, HPCwire, AI and Enterprise Datacenters Boost HPC Server Revenues Past Expectations – Hyperion, April 09, 2019. https://www.hpcwire.com/2019/04/09/ai-and-enterprise-datacenters-boost-hpc-server-revenues-past-expectations-hyperion/

vii John Abbott, 451 Research, The New Accelerators: How the Performance and Efficiency of the Datacenter is Transforming for New Workloads. 2019

viii Owen Rogers, Jay Lyman, 451 Research, Containers: economically, they appear to be a better option than hardware virtualization. Sept. 23, 2016