Solution

Clustertech Creates High Performance Computing and AI Integrated Solution Based on the 2nd Generation Intel® Xeon® Scalable Processors

According to a relevant technical director from Clustertech, as technology continues to evolve, AI workloads must be integrated with HPC infrastructure to achieve most efficient and economical results. The 2nd generation Intel® Xeon® Scalable platform is fully optimized for CPU AI training and AI reference functions based on existing high performance. After the integration with the CHESS, it will exert great application value by providing more efficient training and inference acceleration for deep learning applications such as image classification, speech recognition, language translation, and object detection.

AI technology is increasingly being applied to traditional high-performance computing based on modeling and simulation. And HPC itself is becoming the best platform to carry AI massive computing tasks. Clustertech took advantage of the trend to introduce an enhanced CHESS solution on the original classic HPC hybrid cloud platform. The solution integrates HPC and AI applications through the Web interface, and uniformly schedules different tasks to be calculated on a general-purpose CPU cluster, breaking through the limitations of building a GPU heterogeneous platform for AI applications. Through efficient distributed training and reasoning, the performance bottleneck of AI calculation on the CPU is solved without substantially changing the usage habits of HPC users, and unified and efficient HPC/AI fusion is realized.

Overview

Founded in 2000, ClusterTech Limited (hereinafter referred to as "Clustertech") is a leading provider of high-end IT solutions and consulting services in the Asia Pacific region. Clustertech is committed to applying advanced computing technologies to solve various application issues in massive data processing, large-scale computing, deep analysis, AI and uninterrupted services for customers with technologies such as cloud computing, high performance computing (HPC), AI, analysis and big data. To further enhance the performance of HPC and provide computing capability for applications such as AI, Clustertech has launched a high performance and AI integrated solution based on the 2nd generation Intel® Xeon® Scalable processors to achieve a substantial increase in the performance of training and inference.

Challenge: Improve AI and HPC performance

High-performance computing, also known as supercomputing, is an important frontier branch of computer science. It is not only a manifestation of a nation's comprehensive scientific R&D strength, but also vital to its national security, economic and social development. At the same time, as an innovative application that may revolutionize the digital future, AI has changed the looks of all walks of life, from health and precision medicine to transportation, and to automatic vehicles, which are all impacted by its revolutionary power. As AI technology matures and becomes widely applied, we will see more novel applications emerging, and AI will be integrated with existing workloads and technologies.

Of all key factors in AI development, the importance of "computing capability" should not be ignored. Over the past few decades, although researches on AI applications have been carried out extensively, for a long time the data volume and computing capability have not reached the level that supports insight acquisitions and, on such basis, makes critical decisions. As many academic organizations, governments, and enterprises continue to generate massive data, alone with significant improvements in computing capability and cost reduction, AI has finally become a viable option.

HPC is accelerating AI transformation, by applying the power of AI to existing HPC workflows and greatly expanding the scale of AI algorithms to take full advantage of capabilities of HPC systems. These methods have achieved gratifying preliminary results, indicating that the integration of AI and HPC has a very bright future.

Benefiting from exponentially increasing data volume and diversity, HPC cluster has been the engine for evolving HPDA workloads to bring amazing discoveries and insights into understanding business and humans. Machine learning, deep learning and AI combine tremendous computing capabilities with massive data to drive the development of the next generation applications such as automatic driving system and automatic driving vehicle.

At the same time, as user's needs for precision and scale of AI training continue to increase, AI infrastructure is taking on increasingly high loads, which requires more powerful and agile HPC solutions to achieve commercial breakthroughs and provide viable insights so as to lay the foundation for AI applications.

However, an unavoidable reality is that traditional HPC (such as numerical simulation) and AI (such as CNN) have long been distinct in terms of ecology, and there are significant differences in algorithms. For example, traditional HPC requires high data precision and AI generally has low requirements. The traditional HPC computing has a low memory-to-storage ratio and the AI has a high one. The traditional HPC has high parallelism and AI has low parallelism. These differences lead to the traditional HPC not being able to exert its powerful computing power when applied to typical AI scenarios. For example, when the typical Parameter Server architecture in AI is migrated to the HPC scenario, the speedup ratio will sharply decline with the increase of the number of nodes, so the power of multi-node CPU cluster cannot really get played.

Solution: Clustertech launches CHESS*, a high-performance cluster software

Through the high-performance computing system, the scale of AI can be further expanded to greatly benefit practitioners of deep learning. Whether working on PC, local server, or highly parallel HPC infrastructure, deep learning scientists deal with the combination of tasks, languages and environments that looks very similar. It doesn't matter where the neural network is running. What matters is to be fast and accurate. With the help of neural networks running on the HPC system, the performance of deep learning algorithms is enhanced so that data scientists can solve larger and more complex problems.

The massive sample characteristics of deep learning neural networks is ideal for highly parallel HPC environments where extremely high

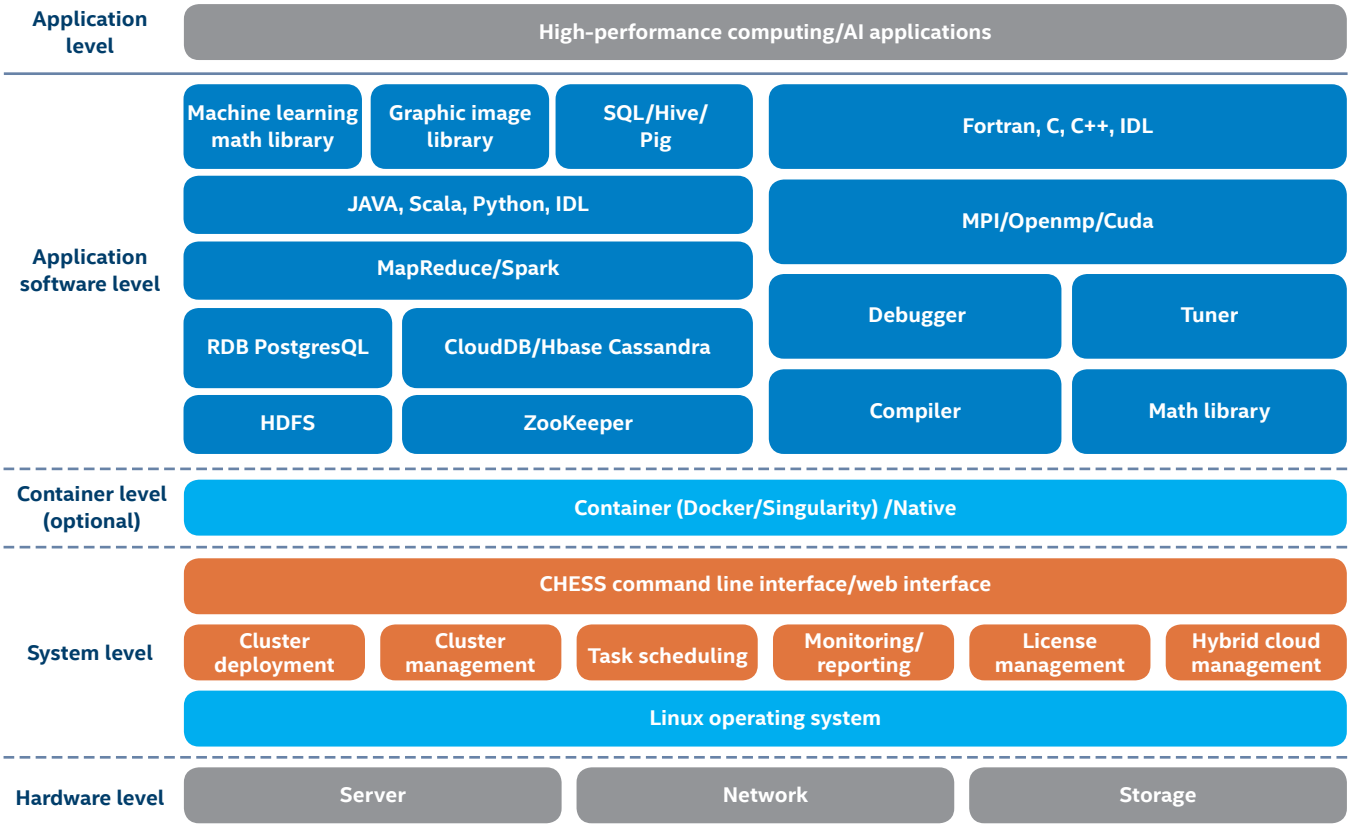

Figure 1. CHESS Software Architecture

computing performance, large memory pools and optimized inter-node communication networks will greatly expand the ability of deep learning networks to identify related structures and patterns.

In terms of hardware, Clustertech solution is powered by the 2nd generation Intel® Xeon® Scalable Processor optimized for AI workloads. Moreover, Intel offers technical support for HPC and AI:

- A series of efficient software tools, optimized libraries, infrastructure modules and flexible frameworks, designed for general-purpose and highly parallel computing, help streamline workflows, assist developers in creating codes, and fully implement Intel® architecture's capabilities in HPC and AI.
- Optimized common deep learning frameworks (including Caffe* and TensorFlow*) for Intel® architecture create greater value and deliver higher performance for data scientists.
- Intel® ParallelStudioXE 2019 contains various performance libraries such as Intel® MKL-DNN for deep neural networks that accelerates deep learning frameworks on Intel® architecture, and Intel® DAAL that accelerates big data analysis.

In terms of software, Clustertech HPC Environment Software Stack (CHESS), a high-performance cluster software independently developed by Clustertech, turns loosely stacked servers into a complete HPC cluster system. CHESS has features such as centralized account management and security control, fine-grained deployment of CPU and GPU, second-level GPU task and resource monitoring, optimized disk IO scheduling and allocation, and improved cluster monitoring and management to achieve unified deployment, management, monitoring, scheduling and reporting of cluster resources, which greatly improves cluster efficiency and simplify cluster management.

At the cluster software level, CHESS provides function modules such as cluster deployment, cluster management, cluster monitoring, task scheduling, task scheduling management, cluster reporting, license management, visualization, and public cloud management. It also provides WebPortal interface interaction and supports HA functions to avoid any single point of failure that affects cluster system operation. These functions centrally manage and monitor the resources of all nodes in the cluster system to enable a single system image of the entire cluster system, making users feel that they are using a high-performance computer.

At the application environment level, CHESS supports the HPC toolset that includes parallel commands, debugging and tuning tools, messaging library, math library and compiler. They support traditional HPC applications in different industries, including CAE simulation, basic disciplines, life science, oil exploration, meteorology and AI.

Effect: Unprecedented AI infrastructure performance

HPC is no longer limited to large research institutions. Enterprises consume an increasing large number of HPC computing cycles. In some of the largest HPC clusters in the world, there are private oil and gas companies. Individualized medical research applies HPC to highly focused treatment programs. The new HPC facility is bringing innovative and integrated architectures to non-traditional usages that integrate simulation, AI, visualization and analysis into one supercomputer. Benefiting from the 2nd generation Intel® Xeon® Scalable platform, the Clustertech CHESS delivers unprecedented performance.

Take a department of Beijing Jiaotong University as an example. The main research field of the department is the one on machine learning-

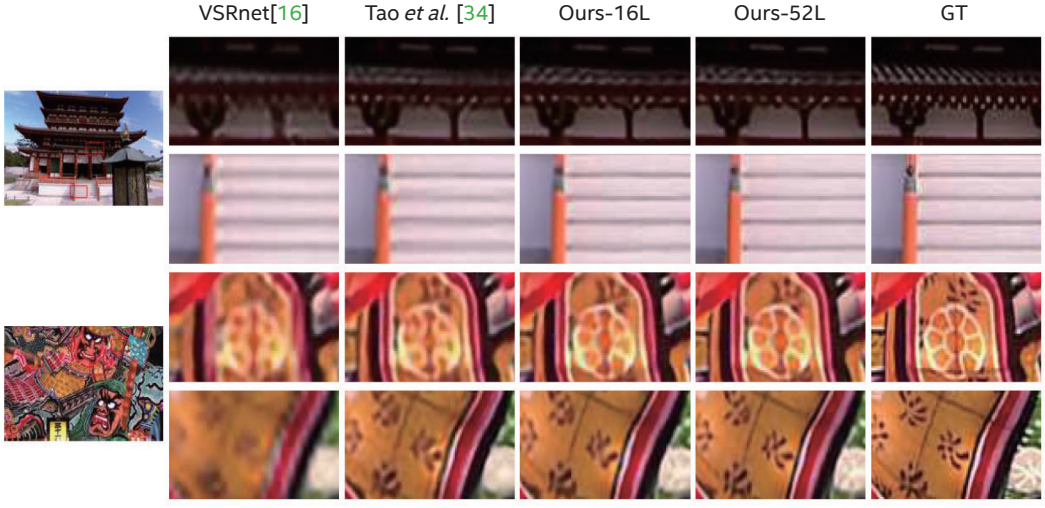

Figure 2. Qualitative comparison on videos from [34]

Note: VSRnet is an early video ultra resolution effect. Tao et al is a video ultra resolution effect in 2017. ours-16L, and ours-52L are the VSR-DUF video ultra resolution effect. GT is the expected real ultra resolution video frame.

based AI video processing and image recognition. By deploying the Clustertech CHESS, it has achieved video ultra quality enhancement and recognition to meet user's computing needs for AI research, to achieve efficient management and refined scheduling of clusters, and to fully meet user's computing needs for AI:

This user has achieved improved video resolution using machine learning method on the Clustertech CHESS. For example, input a low-resolution video stream (e.g. 480P) and output a high-resolution video stream (e.g. 1080P). High performance and AI integrated solutions based on the 2nd generation Intel® Xeon® Scalable processors provide users with high performance and agile services and unprecedented capabilities.

In addition, when CHESS uses the Slurm to call the TensorFlow framework for distributed ResNet-50 image classification training, the measured results show that the speed of a single node based on the second generation Intel® Xeon® Scalable Processor is N pictures per second, the speed of 2 nodes is 2N pictures per second, and the processing speed of 4 nodes is very close to 4N pictures. From the processing speed point of view, as the number of nodes increases, the speed of processing pictures increases linearly with almost no performance loss. According to the test data released by foreign large-scale super-computing centers, the linearity is still very good even if it expands to hundreds of Xeon nodes.

Conclusion: Provide more efficient training and inference acceleration for AI applications

Breaking through the border to control AI, and empowering the new mission of traditional HPC. Clustertech Enhanced CHESS conforms to the trend and integrates the AI mainstream framework and the Horovod distributed system optimized for Intel® architecture. With the unified cluster resource scheduler, distributed training and inference based on the mainstream framework (Tensorflow/Caffe) can be easily implemented, and excellent speedup ratio on multi-node distributed training examples is achieved. This greatly improves the usability of the CPU cluster to deal with AI massive computing tasks, and reduces the threshold for traditional HPC users to use AI. The HPC/AI integration is unstoppable.