Optimising Operations of Converged HPC-AI Environments for Faster, More Accurate Research Results and Discoveries

High-performance computing (HPC) and Artificial intelligence (AI) convergence has gained tremendous attention from various industries such as healthcare, High education and research (HER), and manufacturing. Its potential to augment the massive growth of data faster and with high accuracy enables industries to innovate faster with new processes that can solve complex problems at scale.

In Asia, the data used for research is growing exponentially and to make sense of the data, AI workloads will need to converge with HPC infrastructure. The HPC market is set to grow by US$27.15 billion from 2021 to 2026, with 49% of market growth coming from APAC, according to Technavio. There are 24TB of new data created each day and the convergence of HPC and AI can empower HPC users for faster and more accurate results for innovation and new discoveries.

Some key HPC use cases benefiting from advanced AI capabilities include:

- Scientific visualisation and simulation especially in fields such as high-energy physics
- Medical research, life sciences, and genomic sequencing
- Computational fluid dynamics, and computer-aided engineering
- Analytics for financial services (FSI) such as risk and fraud detection
- Industrial product design

QCT POD optimises HPC and AI convergence

As enterprises and organisations move towards adopting infrastructure that is enabled to run HPC and AI workloads, so is the convergence of HPC and AI driven by workload. However, building an integrated workload-driven system from the ground up to support both HPC and AI workloads is complex, time consuming—often taking months—and costly, as HPCs are traditionally created for researchers where ROI is secondary. For users, it involves a steep learning curve for an unfamiliar and complex system and is hard to manage, needing to understand how to allocate and control resources.

> A major challenge that many companies face when creating an HPC-AI platform is that the design and deployment of complex HPC-AI systems usually take months as they often struggle to design the optimal system architecture that suits their business requirements.

Mike Yang
President,
Quanta Cloud Technology (QCT)

With AI and HPC convergence becoming a mainstream trend, more AI workloads will be running on HPC infrastructure. A modern system is needed to support both AI and HPC workloads on a unified platform and this complicates its design. The compute, storage, networking, and software stack need to fundamentally change to combine heterogeneous architecture into a unified platform.

To address this challenge, the QCT Platform-on-Demand (QCT POD) provides an on-premises compute, storage, and networking system with best-configured hardware and software integration for specific workloads designed for HPC, AI, and data analytics. With the common building blocks, it has greater flexibility and scalability to meet business demands from different industries. Furthermore, QCT POD is pre-validated and pre-configured, which saves time and resources for rapid deployment and easy management. This means enterprises and organisations save time and resources while accelerating time to market.

By optimising HPC and AI convergence, QCT POD's integrated storage solutions such as parallel file systems and network-attached storage enable researchers and medical organisations with the necessary infrastructure for HPC, AI computing, and scalability needed for ever-growing data volume.

Streamline System Management and Accelerate Time-to-value with QCT POD Manager and Workload Package

QCT POD addresses one of the main challenges for administrators: to deploy and manage a cluster with converged infrastructure and diverse services. To streamline the process, QCT POD provides a one-key deployment tool to simplify and shorten the system deployment process from days to hours to set up a new cluster with hardware system and software stack. Real-time monitoring support, including web-based dashboards and system alerting with useful administration tools, further helps to simplify management and ease administrators' operational burden.

QCT POD also supports a robust development environment. The workload package includes allowing application developers and AI engineers to accelerate the launch of Jupyter Notebook service and AI development. A repository of pre-installed development tools such compilers, libraries, and AI frameworks are also included in QCT POD empowering application developers and AI engineers to leverage these tools.

With QCT's decades of domain knowledge in HPC workloads, it was able to provide ready-to-use HPC workload applications which have been pre-configured and pre-compiled for optimal performance. Users can quickly leverage these to run their simulations without spending time and resources tuning the application and worrying about software compatibility.

In an example for HER, the QCT POD is composed of cluster management software, built-in application frameworks, compute-optimised hardware, storage technologies, and low-latency networking seamlessly integrated and validated to accelerate the HPC research process. These built-in ready-to-use tools and framework help streamline the deployment and management of the system and reduce time-to-simulation for researchers removing the complexity of managing siloed systems to reduce the total cost of ownership (TCO).

Enabling Deeper Synergy with QCT POD v1.1

The newly released QCT POD v1.1 based on the fundamental building blocks of QCT POD provides additional advanced features of cloud-native scheduler and data tiering tool for enterprises to gain insights from data.

QCT Cloud-native Scheduler Streamlines Job Scheduling for Compute-intensive Workloads on Cloud-native Environments

Traditionally, HPC and AI run on different environments: HPC on bare-metal and AI on containerised environments with each having a different workload manager. However, QCT POD employs an approach of having a containerised environment to support HPC and AI workloads to create an AI/HPC converged system. Having a containerised environment can be a challenge to traditional HPC users using Kubernetes as a workload manager for their HPC workloads; after all, successfully using Kubernetes to schedule jobs comes with a steep learning curve.

To address the AI & HPC convergence challenge, QCT developed Cloud-native Scheduler a job scheduling feature to remove the pain of job scheduling in a containerised environment for HPC users. Users can submit jobs on Cloud-native Scheduler using the CLI command tool just like using typical HPC job schedulers such as Slurm or Open PBS. Cloud-native Scheduler works with Volcano and Kubernetes behind the scenes to provision and manage the compute resources to complete the HPC jobs in a containerised environment. Volcano inherits the design of Kubernetes APIs and allows users to run applications requiring intensive and high-performance computation on Kubernetes.

Users can achieve the following with Cloud-native Scheduler:

- Shorten learning curve from bare metal to containerised job scheduling
- Support traceable logs output
- Enable easy monitoring of jobs ID status in accounting data for administration

Ultimately, Cloud-native Scheduler enables HPC users to save time and effort in learning Volcano and Kubernetes to submit jobs in a containerised HPC environment.

Data tiering tool made easier with data management

In the exascale era, the exponential data growth has created a need for effective data management. But many research institutions and enterprises often land up with data silos, making it difficult to optimise data, especially in the face of an increasing demand for data management platforms to break data silos and effectively manage their data.

Integrated Rule-Oriented Data System (iRODS) is a grid data management system that organises and connects data with metadata. As a rule engine, it allows flexible definition of data storage, data access, and data processing. iRODS creates a data virtualisation layer on top of the different types of storage, so users can access and manage their data on both on-premises and cloud environments, and from different locations. It integrates storage systems into one single namespace enabling data tiering management. This is critical for enterprises and organisations with huge data growth or with multiple different storage systems.

The new QCT POD v1.1 supports integration with iRODS and facilitates the deployment in a containerised environment. With automated deployment tools, users can install iRODS with lesser manual process and lesser time required for installation. Organisations can enjoy the following benefits when leveraging QCT value-add features for iRODS:

- Simplifying the process of iRODS and storage integration process
- Leveraging policy templates to quickly set up data management policies by administrators
- Deploying iRODS in a containerised environment to enjoy the benefits of a cloud-native environment

What this means for users is the ability to drive an efficient end-to-end workflow while ensuring data security and integrity of metadata information, which in turn can accelerate time-to-value of data.

Advanced Features of QCT POD1.1 Enhance Productivity, and Efficiency

Apart from time- and cost-savings, QCT POD 1.1 also increases productivity for both administrators and users with faster results.

- The pre-defined and pre-configured attributes help enterprises and research institutions reduce the complexity of system design and deployment process to accelerate time-to-value/time-to-insight.
- The monitoring and management features (QCT POD Manager) release the burden of data centre admin, helping enterprises improve on overall TCO.
- Qbatch in QCT POD v1.1 helps HPC users remove the painful learning process of cloud-native job scheduling technology to accelerate time-to-insight.
- The support for iRODS in QCT POD v1.1 helps enterprise and research institutions facilitate data management to achieve higher storage efficiency, improving overall TCO for storage.

Intel technology enabling QCT POD

At the heart of QCT POD is the QCT hardware platforms, providing the compute power to empower AI and HPC. QCT POD hardware portfolio includes QuantaGrid D53X, QuantaGrid D53XQ, and QuantaGrid T43Z which are powered by 3rd Gen Intel® Xeon® Ice Lake Scalable processors to deliver optimal performance.

3rd Gen Intel® Xeon® Ice Lake Scalable processors feature Intel® Advanced Vector Extensions 512 (Intel® AVX-512) and built-in Intel DL Boost instructions to incomparable performance acceleration on deep learning workloads to empower use cases such as image classification, speech recognition, and object detection.

The exponential growth of data has led to many enterprises struggling to meet the increased demand for workload processing, storage, and analytics efficiency. Enterprises need to have the right platform with an integrated system capable of running both HPC and AI with increased workloads. Using the QCT POD automates and simplifies the process from system preparation to deployment and offloads management process with its pre-installed system management and monitoring.

"In the next two years, as more AI workloads are utilising HPC infrastructure, and more AI techniques are combined and applied to HPC simulation to improve the result, enterprises and research institutions need a unified platform to support both workloads. Having a modern AI/HPC converged platform becomes more important," said Yang.

Learn how to optimise operations of converged HPC-AI environments with QCT at https://go.qct.io/qct-pod/