QCT POD: AN ADAPTIVE CONVERGED PLATFORM FOR HPC AND AI

Integrated workload-driven system for industrial HPC/AI workloads

Produced by TCI Media Custom Publishing in conjunction with:

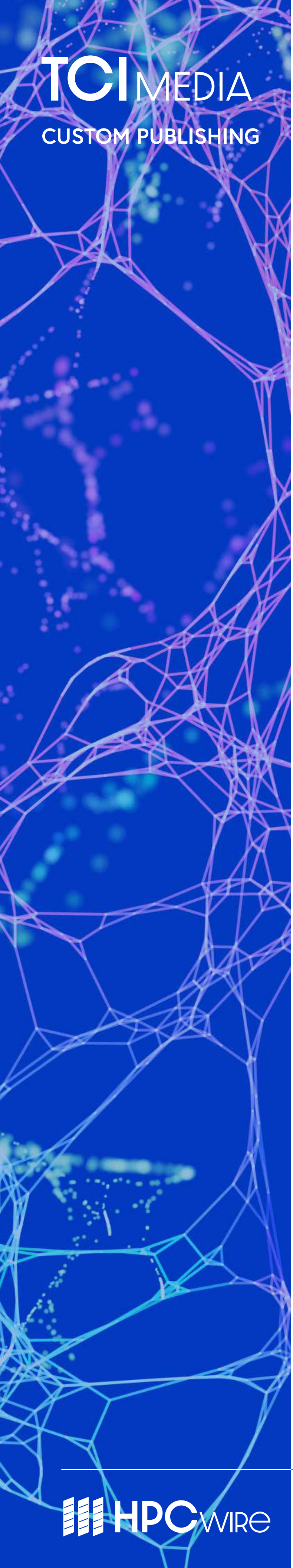
HPCwire QCT intel intel xeon PLATINUM

High performance computing (HPC) workloads are essential for modeling phenomena like weather prediction, computational fluid dynamics, quantum chemistry, and molecular dynamics in scientific research. Life science organizations, researchers and organizations in a variety of industries use artificial intelligence (AI) to help analyze vast amounts of customer and market data to gain insights and make data-driven decisions to optimize business operations, increase efficiency, and drive innovation. Furthermore, the massive growth of data and the need to process, store, and analyze the data requires researchers and many industries to revise their infrastructure and environments.

However, building an integrated workload driven system with optimized performance is time consuming and requires many resources. QCT is a leading cloud data center solution provider with extensive experience in developing HPC and AI solutions for enterprises. For instance, the QCT Platform on Demand (QCT POD) powered by Intel® is an integrated solution that supports both bare-metal and cloud-native environment for HPC and AI workloads. It contains pre-validated and pre-configured system settings designed to meet various workload requirements. This solution brief describes how the QCT POD converged HPC and AI solution aims to simplify the deployment and management of HPC and AI clusters. It also provides an example of how the QCT POD solution is used for medical image diagnosis using the Intel® AI Reference kit. This is one of the accomplishments in various ready-to-use toolkits. QCT worked closely with Intel® to test and validate AI workloads with Intel® software Portfolio, it can aid users to simplify AI development in real-world scenario.

Trends of Convergence of HPC and AI

There is major growth in the use of HPC and AI along with converged HPC/AI workloads. HPC market growth is estimated to reach a compound annual growth rate (CAGR) of 7.1% between 2022 and 2027 (Technavio, March 2023). HPC, AI, machine learning (ML), deep learning (DL) and graphic analysis is forecasted to have a 5 years CAGR expected to exceed 22% growth (Hyperion, November 2022). HPC and AI converged systems will be required to meet this growth and the QCT POD workload-driven system design can fit into a modern AI and HPC workload converged infrastructure.



Challenges in Deploying an HPC/AI System

With the rapidly increasing volume, velocity, and variety of data, HPC and AI workloads are revolutionizing the way organizations optimize business operations, increase efficiency, and drive innovation to stay ahead of the competition. As HPC and AI workloads become increasingly complex to process and analyze, traditional infrastructure and technologies can no longer support these workloads cost-effectively.

Organizations face the complexity of designing, deploying and managing a converged HPC/AI cluster. Building a cluster from the ground up is usually complex, time consuming and often takes months. It is critical for users to have an efficient cadence to allocate and control resources. Pain points for AI/HPC workloads include:

- Limited human resources to manage multiple systems
- Inefficient resource allocation
- Increased cost without synergy

A modern heterogeneous architecture includes bare-metal, virtual machines, and containers that work with converged HPC/AI workloads. The solution provides easy management to effectively handle evolving workloads. In addition, it provides cost-effectiveness and reduced total cost of ownership (TCO) for the organization.

Introducing QCT Platform on Demand (QCT POD)

A converged HPC and AI system architecture provides a scalable and efficient platform for running compute-intensive HPC/AI applications. QCT POD simplifies the journey to design, deploy and manage a converged HPC/AI, saving time and resources and accelerating time-to-market. QCT POD provides a variety of compute nodes, storage solutions, networking and management building blocks. QCT POD is an on-premises AI and HPC converged system containing pre-configured and pre-validated tools including best-practice software and hardware integrated for specific workloads.

QCT POD workload packages include compilers, libraries, frameworks and development tools and development environment. Workload packages are provided for numerical weather prediction, computational fluid dynamics, quantum physics, and AI imaging workloads, etc. The comprehensive development environment can simplify the development journey for users.

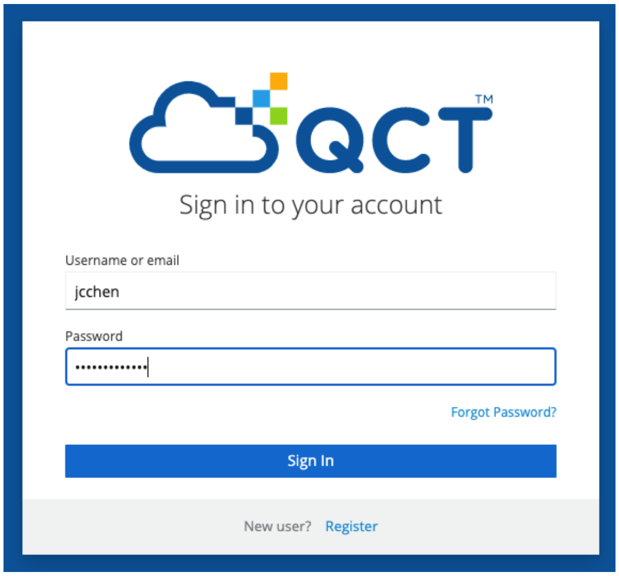
The solution also provides an easy-to-use management system through QCT POD Manager, which includes rapid deployment and provisioning, simplified system management and real-time monitoring. QCT POD Manager provides:

- **Rapid system deployment**: QCT POD enables users to shorten the system deployment process from days to hours to set up a new cluster through a text-based UI configurator.
- **Real-time monitoring:** It supports a real-time web-based dashboard and event alerting with useful administration tools.
- **Simplified system management:** It comprises user authentication, resource management, container orchestration tools, and data tiering management that allow administrators to easily control access of users and resources.

QCT POD Enhances Security with User Authentication

User authentication is the process of verifying the identity of a user. It is an essential security measure for protecting systems and data from unauthorized access. The security of user authentication is a critical issue. QCT POD enhances security of user authentication using login with One-Time Password (OTP) and two-factor authentication (2FA). Users log into QCT POD with OTP verification using the QCT login screen shown below left.

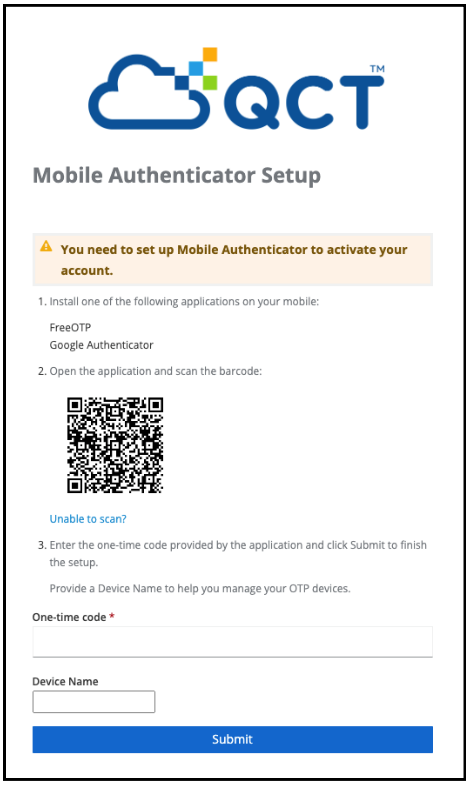
Users can use their mobile phone with either FreeOTP or Google Authenticator applications to authenticate the user identity for two-factor authentication (2FA). The process is shown to the right.



Sign in to your account

Username or email
jcchen

Password
••••••••••••••



Mobile Authenticator Setup

⚠ You need to set up Mobile Authenticator to activate your account.

1. Install one of the following applications on your mobile:

FreeOTP
Google Authenticator

2. Open the application and scan the barcode:

Unable to scan?

3. Enter the one-time code provided by the application and click Submit to finish the setup.

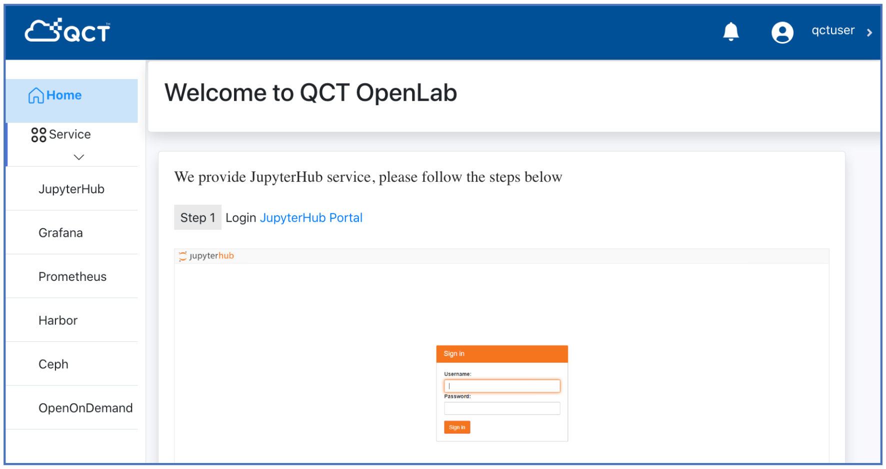
Provide a Device Name to help you manage your OTP devices.

One-time code *

Device Name

QCT POD Enables Single Sign-On

QCT POD also provides Single Sign-On (SSO) for multiple services, which allows users to securely authenticate with multiple services using one set of login credentials. This solution also simplifies the management role and permissions to save time users spend logging in to provide a better user experience.



Welcome to QCT OpenLab

We provide JupyterHub service, please follow the steps below

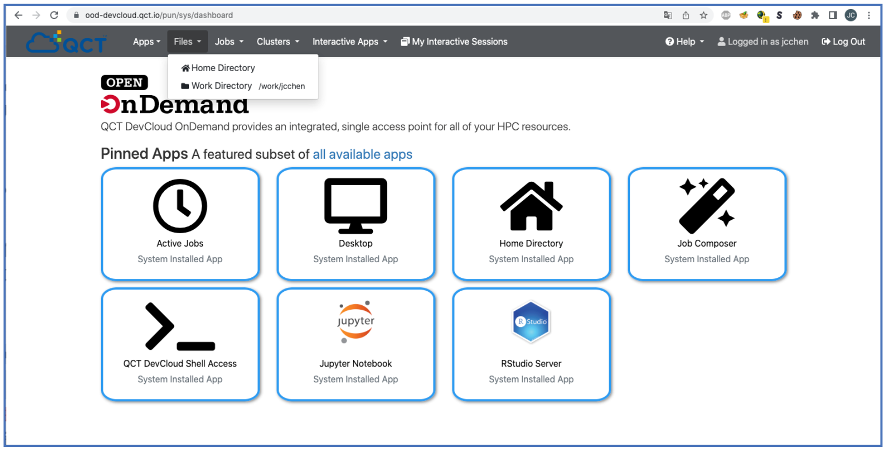
Step 1 Login JupyterHub Portal

Username:

Password:

HPC Web-based User Portal

After users login into QCT POD, some users may expect an intuitive way to do user operation and submit jobs. QCT integrates Open OnDemand (OOD) into QCT POD, a web portal that provides web access to compute or storage resources on a bare-metal environment. This eliminates the steep learning curve of command-line interface and users of all skill levels can login through Open OnDemand for HPC workloads as shown in the screen below. The Open OnDemand web portal allows access to Active Jobs, Desktop (NoVNC) Service, Home Directory, Job Composer, Active Job Status, Shell Access (Terminal Service), Jupyter Notebook Service, and RStudio Service.



QCT DevCloud OnDemand provides an integrated, single access point for all of your HPC resources.

Pinned Apps A featured subset of all available apps

How QCT POD Aids Resource Utilization in a Cloud-Native Environment

Users may face challenges related to cluster resource utilization in a containerized environment. QCT POD provides advanced cloud-native job scheduling tools on Kubernetes to support batch-mode job scheduling such as fair share and gang-scheduling for compute-intensive workloads.

Users also have challenges for job submission on Kubernetes using YAML files. QCT developed a command line interface (CLI) to let users submit jobs just like in a bare-metal environment they are familiar with using. QCT CLI interface functions include:

- Data source mounting
- Reduced CLI input (auto job-naming, simplified resource specifications, GPU flags)
- Logs streaming/output
- Support with Job ID ledger for job monitoring

QCT Hardware Portfolio in the QCT POD Solution

QCT provides a wide range of state-of-the art hardware and infrastructure solutions. The QCT POD solution uses various QCT hardware for the compute node, storage node, and GPU nodes. The following QCT hardware is used:

Compute Node-QuantaGrid D54X-1U:

This server includes a 4th Gen Intel® Xeon® Scalable processor, 32x DDR5 RDIMMs, up to 5 PCIe x 16 slot (Include 2*OCP3.0), up to 12x 2.5" and NVMe/ SATA SSD. Liquid cooling is also supported to help reduce equipment overheating and decrease energy usage. Upgraded data protection with Intel SGX, Intel® Crypto Acceleration, and PFR (optional)

Compute/Storage Node-QuantaGrid D54Q-2U:

This server includes a 4th Gen Intel® Xeon® Scalable processor and 32x DDR5 RDIMMs, providing the ultimate resilience and scalability with up to 10x PCIe 5.0 expansions. The server is optimized for AI acceleration and can support up to 2x dual-width accelerators in 2U system with support for 24x NVMe/ SATA SSD. Liquid cooling is supported.

GPU Node-QuantaGrid D54U-3U:

The GPU server is powered by 4th Gen Intel® Xeon® Scalable processors with up to 350W TDP and is PCIe 5.0 and DDR5 platform ready. It can use up to 4x double-width accelerators or 8x single-width accelerators. Support is for active type and passive type.

GPU Node-QuantaGrid D74H-7U:

The GPU server is powered by 4th Gen Intel® Xeon® Scalable processors with up to 350W TDP. It is also PCIe 5.0 and DDR5 platform ready. It uses an HGX H100 8-GPU with PCIe 5.0 OCP expansion slots for GPUDirect RDMA, and all-NVMe drive bays for GPUDirect Storage.

Experience the QCT POD Solution Using the QCT DevCloud Program

QCT co-works with Intel® to provide the DevCloud Program which includes a comprehensive HPC/AI/DA environment for users to experience the QCT POD solution and infrastructure expertise. The DevCloud Program includes a heterogenous computing platform with HPC, AI, and data analysis (DA) tool kits along with a cloud-native and bare-metal environment with resource and job management tools. The program can also provide software defined storage with data tiering management and QCT precompiled workloads. Development tool kits include support across a range of QCT infrastructure hardware platforms with infrastructure expertise. The DevCloud program allows users to remotely access and test their applications in a cluster environment. Click here to register for the DevCloud Program.

The DevCloud Program includes a comprehensive HPC/AI/DA environment for users to experience the QCT POD solution and infrastructure expertise.

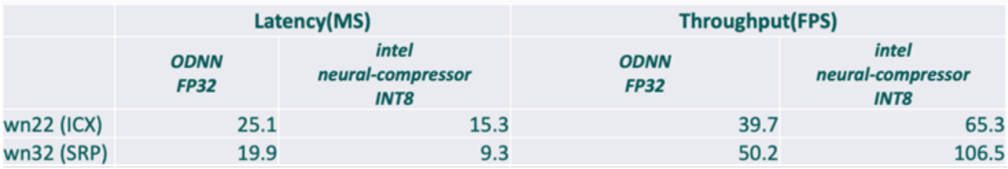
Medical Image Diagnosis using Intel® AI Reference kit (Medical Imaging) on DevCloud

In this example, the Intel® AI Reference kit was used to perform a diagnosis of chest x-rays in the DevCloud environment to distinguish between two classes (normal versus pneumonia). The result of running the Intel® AI Reference kit was image output results which classified images as either pneumonia or normal. With the use of machine learning (ML) and deep learning, the AI algorithm identifies images within a study that warrants further attention by the radiologist/reader to classify the diseases. This aids in reducing the read time as it draws the reader's attention to the specific image and identifies abnormalities. The AI medical imaging diagnostics test was run using the QCT DevCloud environment.

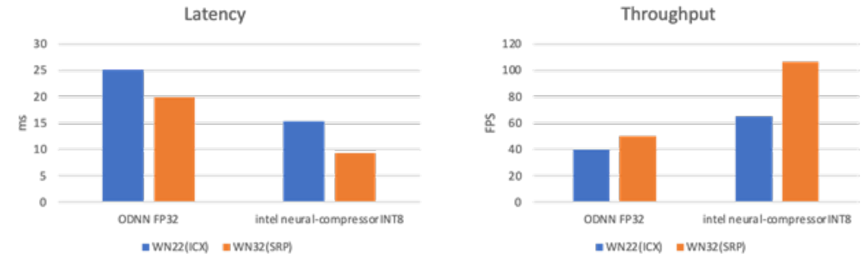
Case Summary: Medical Imaging Test

The medical diagnostic test ran on the Intel® CPU Platform QuantaGrid D54X-1U server with dual Intel® Xeon® Platinum 8458P CPUs of 4th Generation Intel® Xeon® Scalable Processors, and another QuantaGrid D53X-1U server with dual Intel® Xeon® Gold 6330 CPUs. The user first accesses the web-based JupyterLab notebook service used with integrated account management, then submits AI inference jobs using batch-mode job scheduling support in the cloud-native environment. After jobs are processed, users can monitor a Job ID ledger through the QCT developed command line Interface. When the jobs are done, the user can examine the medical diagnosis result directly.

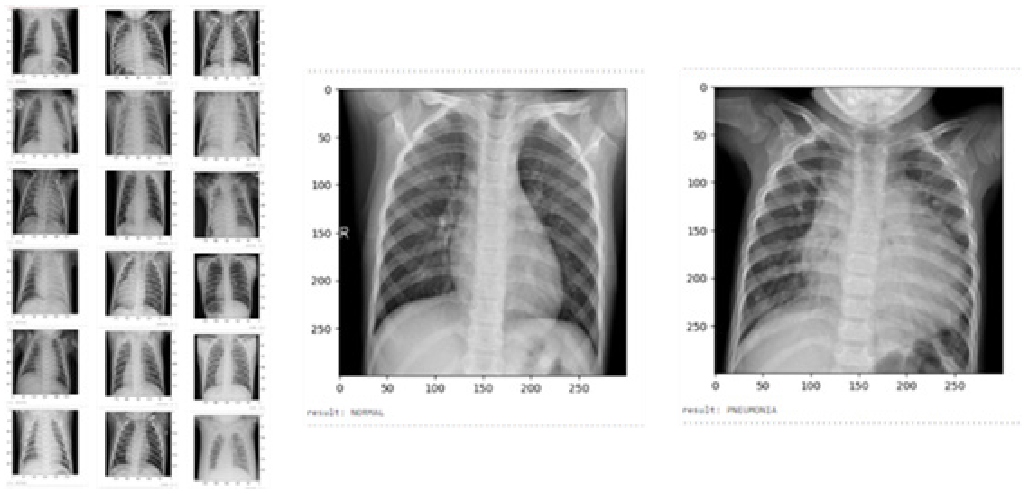
In this case, the user ran all tests on the server using the QCT and Intel® joint solution. The Intel® software Portfolio includes Intel® Advanced Matrix Extensions (Intel® AMX), OpenVINO™ Toolkit and Intel® oneAPI optimized Tensorflow running on a QuantaGrid D54X-1U server. The Intel® AMX instruction set could boost AI inference efficiency significantly by using 2D register files (tiles) and TMUL supporting two data types: INT8 and BF16. Another QuantaGrid D53X-1U server ran without the Intel® AMX instruction set to observe the performance difference. Tests show that even if we convert the model from FP32 to INT8, the accuracy of the model does not change (98%->98%), but the latency is reduced while the throughput is almost 2x higher at the same time. We can also see that 4th Generation Intel® Xeon® Scalable Processors CPUs have more improvements in both latency and throughput than the 3rd Generation CPUs.



	Latency(MS)		Throughput(FPS)	
	ODNN FP32	*intel neural-compressor INT8*	*ODNN FP32*	*intel neural-compressor INT8*
wn22 (ICX)	25.1	15.3	39.7	65.3
wn32 (SRP)	19.9	9.3	50.2	106.5



The sample input and output is shown below. The left pane of the figure depicts some images of the input data, the sample results shown in the right pane indicate whether the images were classified as "normal" or "pneumonia" by inference based on the trained and fine-tuned model.



The result of running the Intel® AI Reference kit was image output results which classified images as either pneumonia or normal. With the use of machine learning (ML) and deep learning, the AI algorithm identifies images within a study that warrants further attention by the radiologist/reader to classify the diseases.

Summary

HPC and AI are increasingly being used in life sciences research and across many industries to gain insight from massive amounts of data. Many organizations have an infrastructure that is not capable of HPC processing or meeting their modern AI workload needs. Modern infrastructure requires a comprehensive hardware portfolio, with high quality software and hardware system integration.

The QCT POD solution powered by Intel® provides the building blocks of compute and storage for performance optimization with best-practice software and hardware system integration. QCT POD also supports HPC and AI converged workloads in a HPC and AI environment with pre-configured and pre-compiled toolkits to ease system management and development. QCT cooperates with ecosystem partners to deliver industrial HPC/AI solutions to meet specific industrial needs.

QCT encourages you to register for the QCT DevCloud Program to enjoy a QCT POD development environment with remote access. Once permitted, you can experience the QCT POD hardware and software environment to develop or test application, including Intel® oneAPI toolkits, QCT pre-compiled workloads, and other development toolkits.

> ***The QCT POD solution powered by Intel® provides the building blocks of compute and storage for performance optimization with best-practice software and hardware system integration.***

For more information on QCT POD, see: https://go.qct.io/qct-pod/

Intel, the Intel logo, Xeon, and Xeon Inside are trademarks or registered trademarks of Intel Corporation in the US and/or other countries.