



insideHPC

insideHPC Guide to

QCT Platform-on-Demand Designed for Converged Workloads

by Richard Friedman



Photo courtesy of Quanta Cloud Technology

BROUGHT TO YOU BY



Contents

Introduction.....	2
What converged HPC/AI workloads look like	2
The converged Platform-on-Demand solution from QCT	3
High Performance Computing with QCT POD	4
Deep Learning with QCT POD.....	6
Summary	7
About QCT	7

Introduction

As businesses become more and more data driven, they quickly realize that to stay competitive they need a solution that not only provides advanced capabilities for performing highly complex technical computations, but can support deep data collection and predictive analysis at the same time.

Traditionally, these two domains — High Performance Computing (HPC) and Artificial Intelligence (AI) — existed as separate environments, each with their own unique hardware, software, storage, and networking requirements. HPC usually involves a significant amount of computing power employing state-of-the-art parallel processing techniques. On the other hand, AI (including Machine Learning (ML), and Deep Learning (DL)), employs iterative algorithms to find insights hidden in oceans of data collected over time.

Not too long ago, building such a converged HPC/AI environment would require spending a lot of money on proprietary systems and software with the hope that it would scale as business demands changed.

As we'll see, by relying on open source software and the latest high performance/low cost system architectures, it is possible to build scalable hybrid on-premises solutions that satisfy the needs of converged HPC/AI workloads while being robust and easily manageable.

What converged HPC/AI workloads look like

Lately, we see growth in converged HPC/AI workloads in healthcare, finance, and automotive industries, among others. For example:

- **In healthcare**, pharmaceutical companies search for new medicines for treatment of tumors by applying bioinformatics solutions that utilize HPC molecular dynamics modeling to identify potential compounds, and deep learning to analyze gene expression profiles, and recognize images of human cancers.
- **In finance**, data-rich HPC simulation techniques and predictive modeling with AI are now at the heart of all strategic financial planning and analysis.
- **In the automotive industry**, real-time AI object-recognition training and response techniques for autonomous vehicles, along with HPC simulations of vehicle crash testing, make up most critical workloads.

Building a comprehensive on-premises cluster to handle both a company's HPC and AI workloads

Quanta Cloud Technology has created a converged, on-premises platform to satisfy HPC/AI customer needs while accelerating performance, reducing overall operating cost, and greatly simplifying system management.

from scratch is known to be complicated, costly, and hard to manage. Doing it right demands extensive knowledge of both kinds of workloads and the infrastructures they require.

Quanta Cloud Technology (QCT), a major cloud data center solution provider based in Taiwan, California, Singapore, and Germany, has created a converged, on-premises platform to satisfy HPC/AI customer needs while accelerating performance, reducing overall operating cost, and greatly simplifying system management.

The converged Platform-on-Demand solution from QCT

Building a cluster to satisfy the requirements of both HPC and AI has some significant challenges:

- Introducing an unfamiliar and complex system environment to new users and application developers looking to achieve the highest possible performance.
- Providing system administrators the tools they need to efficiently configure, monitor, and analyze the health and performance of a large cluster.
- Providing additional tools for comprehensive account management, allocation, and control of compute, storage, and networking resources.

Taking these challenges into consideration, QCT has designed a unique Platform-on-Demand, or POD — an on-premises, rack-level system that offers best practice hardware and software integration for both HPC and AI workloads. Leveraging QCT’s own system administration tools, the QCT POD comes pre-configured and pre-validated, ensuring rapid deployment and easy resource management.

QCT POD is constructed out of a set of common building blocks to ensure a high degree of flexibility and scalability — Management Building Block, Compute Building Block, and Storage Building Block — each connected by a network fabric that can be customized to fit user workload demands.

- The Management Building Block, based on RedHat® Enterprise Linux® or CentOS, is a software stack that offers a wide selection of web-based administrative and monitoring tools and dashboards that promote efficient cluster management.
- The Compute Building Block delivers just the right hardware and software combinations to fulfill various workloads, such as HPC, ML, data analytics, cloud service, and edge computing. Tailored for each domain and industry with a comprehensive hardware portfolio, QCT delivers the best configuration to solve their unique challenges and achieve better performance. Kubernetes and Docker form the base for orchestration in the ML building block, and is extendable to other frameworks, such as TensorFlow, Keras, and PyTorch. QCT’s flexible POD design meets many diverse demands and solves most customer challenges.
- The Storage Build Block, designed with the specific HPC & AI requirements for high IOPS and low latency in mind, is now available to customers from all industries with similar demands, such as finance, engineering, life sciences, energy, among others. Today, with AI and Machine Learning (ML) becoming standard practice, these workloads are having a huge impact on storage requirements, especially for large file and block storage.

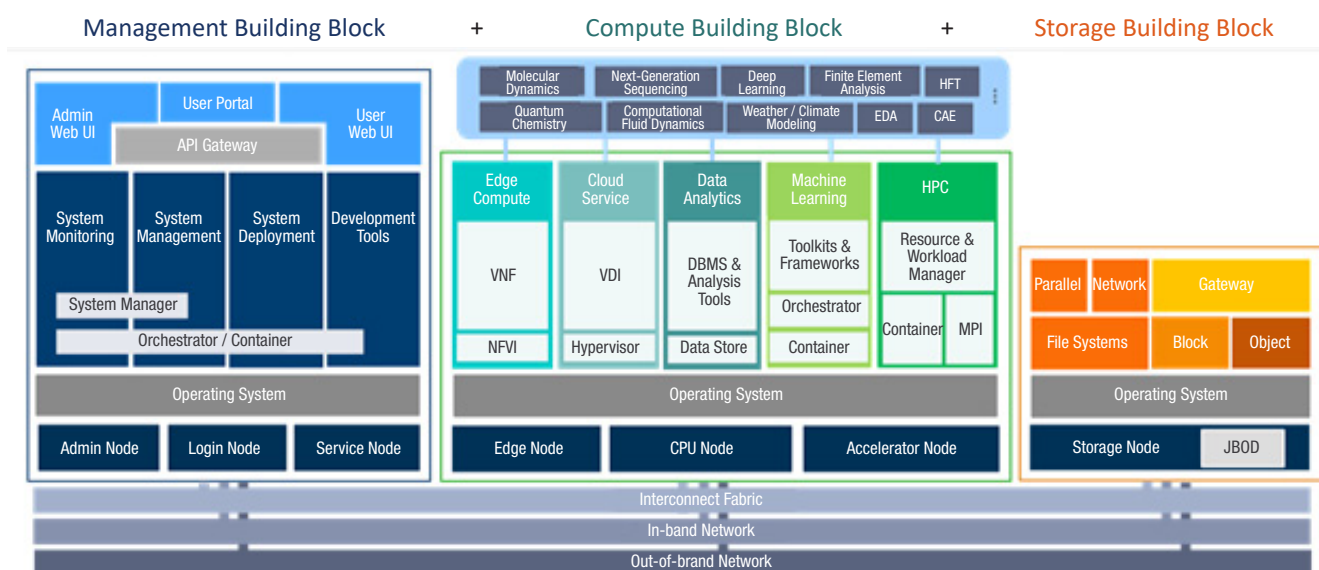


Figure 1: QCT POD Building Blocks

High Performance Computing with QCT POD

Users and administrators migrating from traditional HPC systems will find that QCT POD presents the HPC developer with a familiar environment that they can use easily and efficiently, both on-premises and remotely through web-based tools. They have all the popular compilers, including GCC, Intel® Parallel Studio XE, and the NVIDIA® HPC-SDK™ (for Fortran, C, and C++), and math libraries BLAS, BLACS, and LAPACK, all optimized for high performance, along with Environment Modules for easy runtime setup, and job schedulers such as Slurm for advanced workload management.

The QCT POD also supports Kubernetes and Singularity, permitting users to run HPC jobs the modern way — in containers managed and monitored with browser-based tools.

HPC Example: WRF Model

To see how the QCT POD simplifies and expedites deployment of a popular HPC workload, let's look at what it takes to get the Weather Research and Forecasting model WRF up and running.

WRF is open source and written mostly in Fortran and C. As such, developers can use their preferred high-performance compilers, including the NVIDIA HPC SDK, to build WRF and the libraries it requires.

Now users can employ QCTs Environment Modules to load the specific library versions and the corresponding workload settings they need with just a few quick commands.

Being open source allows developers to adapt and insert other open source codes into the WRF core to enhance performance. But this can cause real difficulties for developers by introducing complicated dependency and compatibility issues with code coming from other environments and using different compilers. To resolve these issues could result in a patchwork of special runtime environments that become ever more difficult to manage.

By design, QCT provisions the POD with pre-compiled optimized runtime libraries for core workloads, and gives developers recipes for rebuilding those libraries as needed. Now users can employ QCTs Environment Modules to load the specific library versions and the corresponding workload settings they need with just a few quick commands. This lets them build their HPC applications quickly and start running benchmark tests as soon as they get access to the QCT HPC POD.

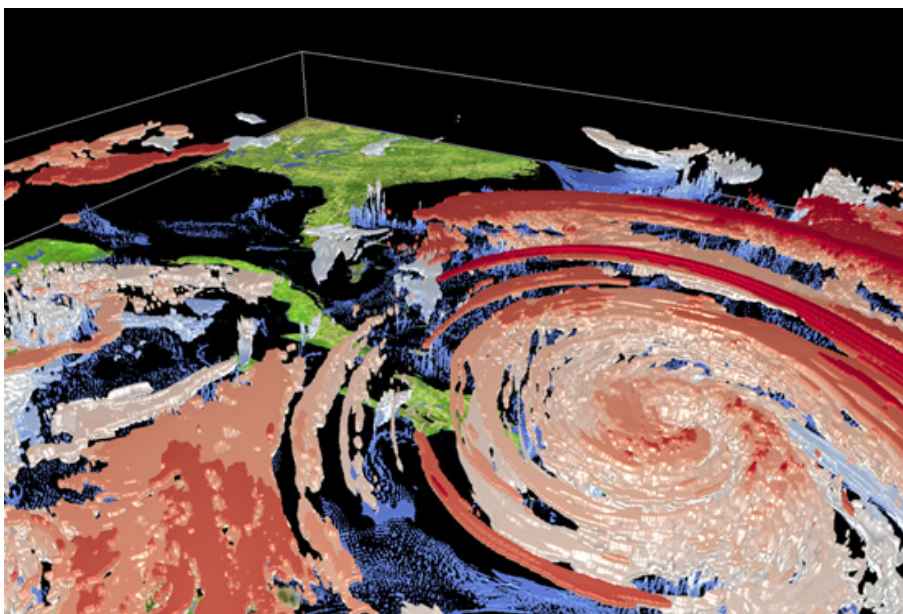


Figure 2: Hurricane Maria WRF simulation

This was very much the case with the WRF model. In particular, WRF simulations use pre- and post-processing tools to prepare the input for the model and to analyze and visualize the simulation output. Figure 3 shows the entire WRF workflow ecosystem.

Each of the pre- and post-processing tools have their own unique runtime library dependencies that need to be built, tested, and deployed. Building the entire WRF ecosystem typically takes weeks to months to migrate and get running on a new HPC system. By using the pre-configured software provided by QCT, developers can get the WRF ecosystem ready for input and generating output for analysis quicker than on traditional HPC systems.

By using the pre-configured software provided by QCT, developers can get the WRF ecosystem ready for input and generating output for analysis quicker than on traditional HPC systems.

By providing optimized and pre-compiled runtime libraries with POD, QCT eliminates most of the pain developers encounter when preparing a runtime environment for HPC applications on a new HPC system. Users are able to get up and running with their first test runs with minimal effort and wasted time.

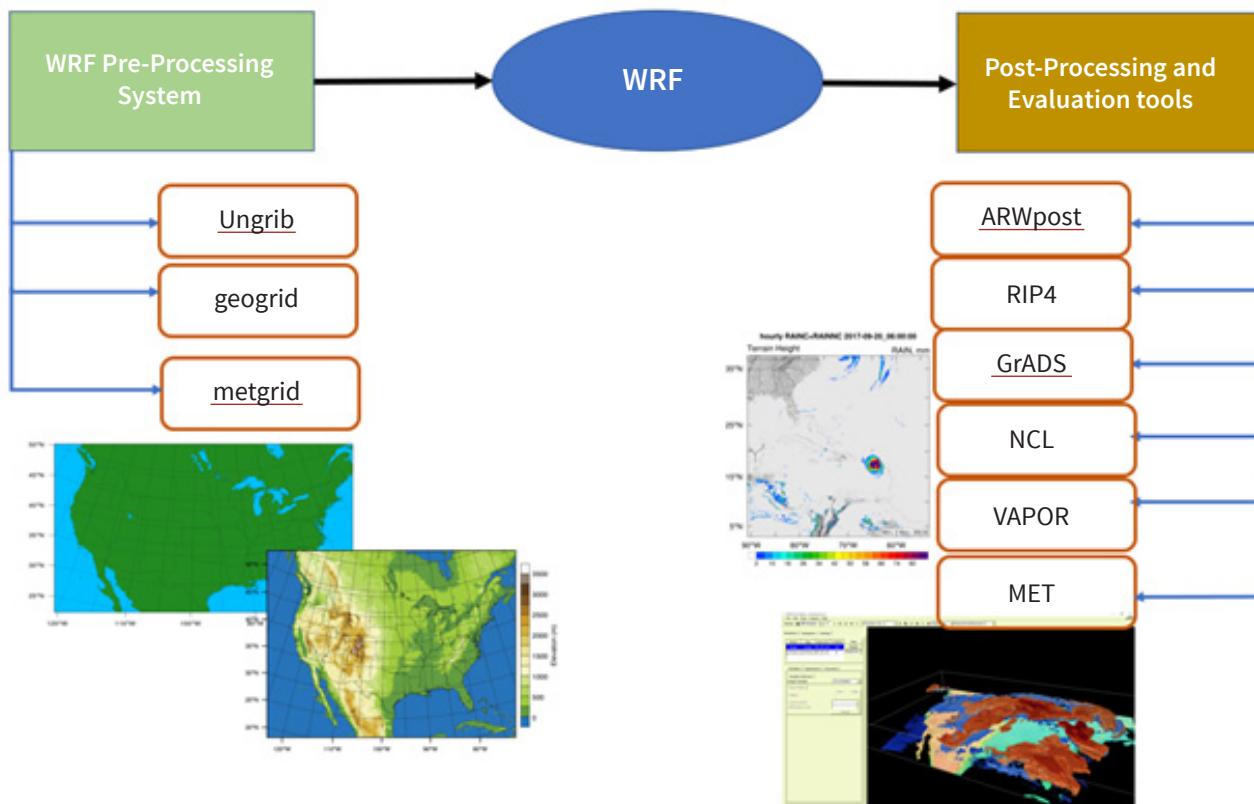


Figure 3: Typical WRF workload ecosystem, showing pre- and post-processing tools

Deep Learning with QCT POD

Containerization and prepared frameworks also enable rapid deployment of AI/ML/DL workloads on QCT POD.

QCT POD management tools efficiently allocate Docker and Kubernetes containerized resources, giving users access to an NVIDIA GPU enabled JupyterHub container for their Deep Learning development.

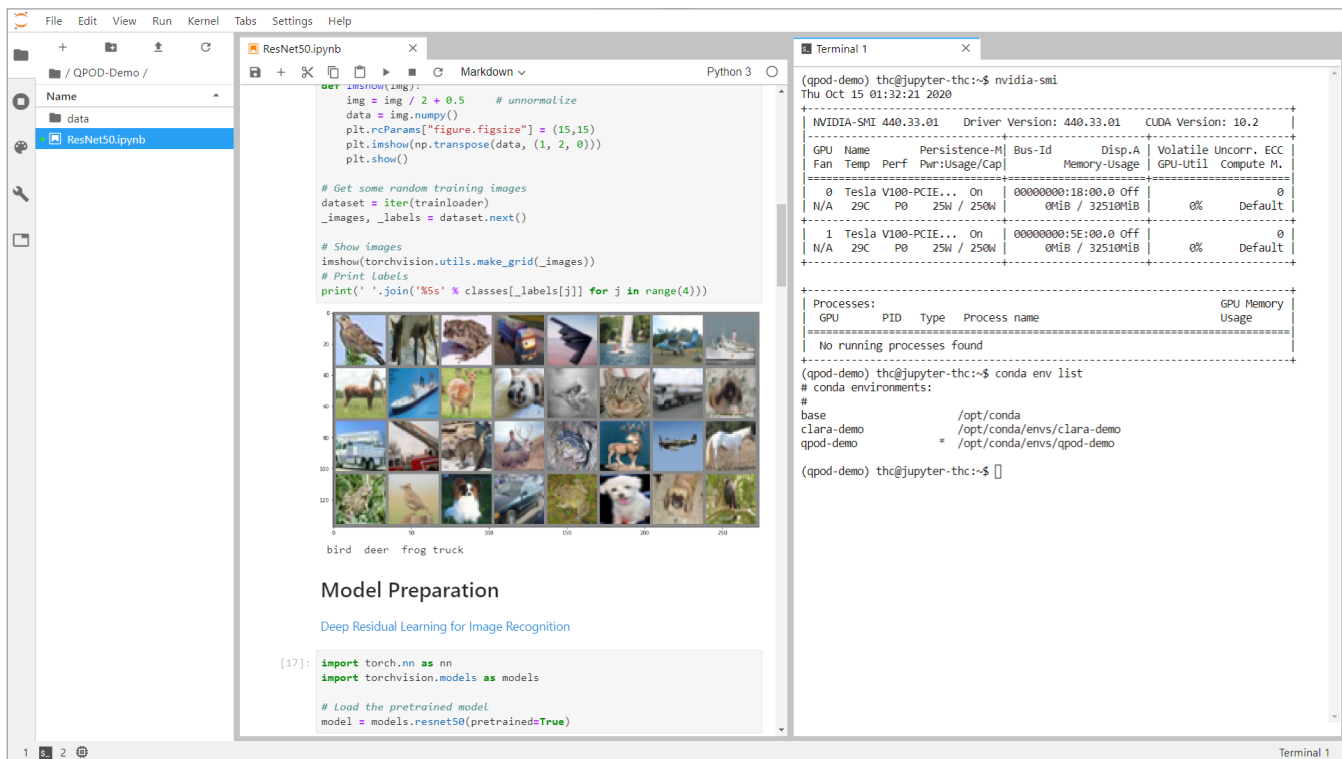
QCT POD fully integrates JupyterLab, providing users with a default web-based environment workspace where they can directly upload their datasets. Included are:

- Jupyter notebook, code editor, and command-line tools, made available through a single webpage.
- NVIDIA GPU-compatible Deep Learning frameworks such as Tensorflow, Pytorch, and others, as well as Anaconda for data science Python developers. All these frameworks are pre-installed and immediately available for users to manage their diverse projects.

QCT POD provides a comprehensive Deep Learning development environment that is ready for use over a widely diverse set of workloads.

In this example, we've prepared ResNet50 with Pytorch and performed the training using two NVIDIA® Tesla® V100 GPUs. In a future roadmap, QCT will also provide NVIDIA® Tesla® A100 GPUs to accelerate performance on deep learning training and inference. Users can pull pretrained models from NVIDIA® NGC™ (NGC), and the whole process — allocating resources, preparing the data model, data preprocessing, training — can be performed in a single workspace on the QCT POD.

QCT POD provides a comprehensive Deep Learning development environment that is ready for use over a widely diverse set of workloads.



The screenshot displays a JupyterLab environment. The main notebook, titled 'ResNet50.ipynb', contains Python code for image classification using ResNet50. The code includes data loading, preprocessing, and training steps. Below the code, a grid of 25 small images is shown, representing the dataset used for training. The images are labeled with their corresponding classes: 'bird', 'deer', 'frog', and 'truck'. The notebook also includes a section for 'Model Preparation' with code to load the pretrained ResNet50 model.

On the right side of the interface, a terminal window is open, showing the output of the 'nvidia-smi' command. The output displays the status of the NVIDIA GPUs, including the Tesla V100-PCIe... cards, their memory usage, and the driver version. The terminal also shows the output of the 'conda env list' command, indicating the active conda environment is 'qpod-demo'.

Figure 4: Image classification example

Summary

QCT and their partners are dedicated to providing best in class solutions using modular designs and integration of NVIDIA HPC/DL SDK to meet the most demanding HPC, AI, and data science workloads.

The adoption of NVIDIA GPUs allows users to accelerate performance for many varied workloads. With QCT POD management, storage, and compute building blocks, NVIDIA GPUs provide a powerful environment for diverse HPC, Machine Learning, Cloud Services, and Edge Computing workloads. Designed to meet the demands from

each domain and industry, QCT POD delivers the unique hardware and software combinations configured to solve these challenges and achieve best performance on customer applications and workloads.

The QCT POD, with its pre-installed frameworks and pre-compiled libraries, support for containers and advanced job/resource management tools, and embracing the latest open source technologies, offers the best in class solution for a company's converging HPC and AI workflows.

About QCT

Quanta Cloud Technology (QCT) is a global data center solution provider. We combine the efficiency of hyperscale hardware with infrastructure software from a diversity of industry leaders to solve next-generation data center design and operation challenges. QCT serves cloud service providers, telecoms and enterprises running public, hybrid and private clouds.

Product lines include hyper-converged and software-defined data center solutions as well as servers, storages, switches, integrated racks with a diverse ecosystem of hardware components and software partners. QCT designs, manufactures, integrates and services cutting-edge offerings via its global network. The parent of QCT is Quanta Computer, Inc., a Fortune Global 500 corporation.

To learn more about the QCT, please visit go.qct.io.