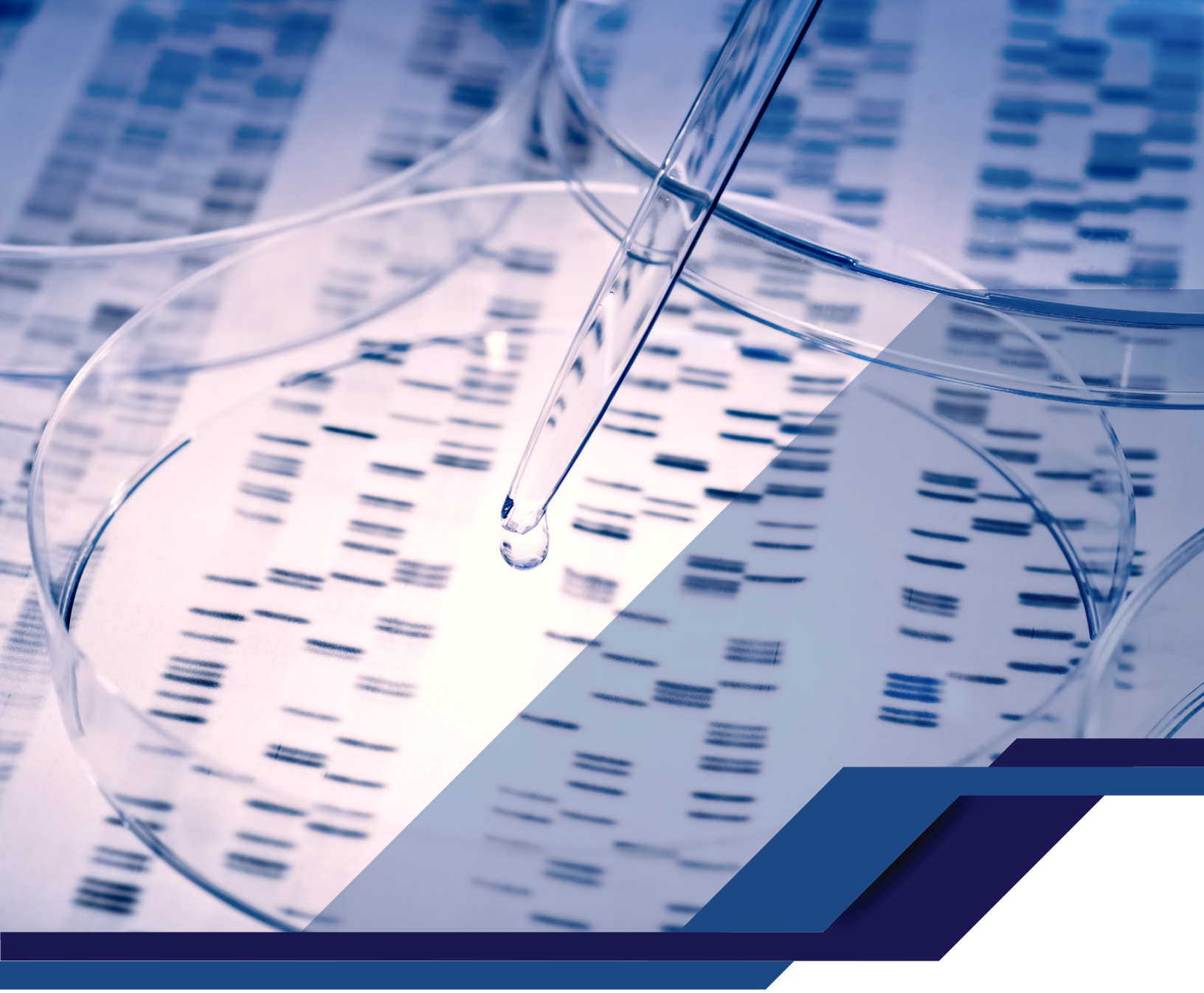

# STREAMLINE PRECISION MEDICINE WORKLOADS WITH QCT POD FOR MEDICAL

Produced by Tabor Custom Publishing
in conjunction with:

HPC wire

QCT | NVIDIA

# Executive Summary

Precision medicine is an approach of treatment and prevention of disease that takes into account the individual variability. Precision medicine life science research often focuses on the areas of next-generation sequencing (NGS), image recognition and molecular dynamics (MD) research used in drug discovery. Precision medicine research results in massive amounts of data and data-intensive workloads which require using High Performance Computing (HPC) systems for computer simulation and running Artificial Intelligence (AI) programs for data analysis. Organizations doing NGS, MD and image recognition research face many challenges due limitations of infrastructure and the difficulty in managing HPC and AI workloads.

This paper provides information on how the NVIDIA Clara AI solution aids in developing, training, and managing AI medical imaging applications. In addition, it describes how Quanta Cloud Technology (QCT) developed the QCT POD for Medical solution containing common building blocks used to design on-premises, rack-level systems with hardware and software integration capable of running precision medicine HPC and AI workloads. The QCT POD for Medical solution aids researchers and developers because it contains pre-validated and pre-configured custom hardware and management tools designed to meet specific organization workload requirements. The paper provides examples of how the QCT POD for Medical solution makes it easy to launch and manage NVIDIA Clara Train SDK and QOCA® aim AI workflow tasks.

QCT is a leading cloud datacenter solution provider with extensive experience in developing HPC solutions. QCT has expertise on end-to-end HPC and Deep Learning (DL) solutions for companies in a wide variety of fields. The QCT POD for Medical solution provides organizations with best-fit building blocks designed to meet business demands from various workloads. With rack-level integration, QCT enables an organization's ability to reach time and resource efficiency for quick deployment and management of their system.

# Meeting Future Life Science Precision Medicine Workload and Research Needs

According to Precedence Research, the global precision medicine market is poised to streamline at a Compound Annual Growth Rate (CAGR) of 11.5% during the forecast period of 2020 to 2027. PWC research indicates that eighty-two percent of surveyed companies involved in precision medicine focus on molecular testing and diagnostics (IVDs and laboratory testing), followed by bioinformatics software and analytics (74%), and DNA sequencing instrument (70%).

## Next Generation Sequencing

An important tool in predictive medicine and genomic analysis is the use of NGS, also known as massive parallel sequencing. Sequencing provides information about the kind of genetic information carried in a particular DNA segment. NGS is a form of DNA sequencing that analyzes individual strands of DNA simultaneously, which allows it to analyze DNA at a much faster rate than other methods due to the parallel nature of the analysis. NGS analysis begins with DNA data collection followed by gene sequencing. The final step is gene data analysis using BWA, GATK and Deep Variant tools to compare an individual's DNA with reference DNA samples. NGS processing requires the use of HPC compute and storage solutions due to the massive amounts of data which must be processed, stored, and analyzed.

## Medical Imaging

Medical imaging refers to different technologies used to view the human body in order to diagnose, monitor, or treat medical conditions. Sources such as ultrasound, medical X-ray, or Magnetic Resonance Imaging (MRI) provide image data. Each type of technology requires different information about the area of the body being studies or treated. AI is increasingly used for data standardization and labeling used in analyzing medical images. Deep learning frameworks such as TensorFlow, Pytorch, and Keras to help medical imaging models training and inferencing.

## Molecular Dynamics

Molecular Dynamics (MD) uses computer simulation methods for analyzing the physical movements of atoms and molecules. MD-based computational methods are close to becoming routine computational tools for drug discovery. MD simulations are often used in drug research to study the drug-binding process. Increased computer power and use of Graphical Processor Unit (GPU) architectures make these MD simulations possible. MD simulations study protein data collection as well as protein atlas and structure using HPC systems and applications such as GROMACS.

## Workload Attributes in Precision Medicine

Each medical related workload whether NGS, medical imaging or MD requires specific computing resources. For example, MD and NGS typically leverage HPC technology which requires a significant amount of compute power and parallel processing to solve complex computational problems. Image recognition leverages AI deep learning (DL) algorithms and techniques that iteratively learn from data to allow computers to find hidden insights without being explicitly programmed where to look. DL relies on GPU resources and a container-based unified platform would be critical.

## QCT POD for Medical - A Unified Platform Designed for Medical Workloads

Hospital research centers, medical organizations and research institutions face challenges in doing NGS, MD and image recognition research due to the massive amounts of data generated. It is common for organizations to have separate, isolated compute and storage platforms which are not interoperable to support these workload demands. There are often limited technical resources to support complex system deployment, management, and monitoring.

It is not easy to build a unified system designed to meet HPC and AI life science and medical workload requirements. Precision medicine workloads require an optimized infrastructure with CPU, GPU, and storage systems capable of running medical HPC and AI applications and workloads. QCT provides an on-premises rack-level system and developed a platform called QCT POD for Medical specifically designed to meet the needs of the healthcare industry. To reach greater flexibility and scaling, QCT POD for Medical offers common building blocks to meet different medical demands including NGS, MD, and Image recognition. QCT ensures the quality and serviceability of the infrastructure, which could accelerate time-to-value for customers. QCT POD for Medical includes:

- Easy deployment tools and real-time dashboard for easy monitoring, management for administration
- Best-fit workload building blocks that include data processing, analytics, high-performance computing, and deep learning to accelerate medical workloads
- Storage management to simplify data protection and system scalability

# QCT POD for Medical - Meets Medical Needs

The QCT POD for Medical architecture shown in Figure 1 includes Management Building Blocks, Compute Building Blocks, and Storage Building Blocks that are each connected by a network fabric that can be customized to fit specific user workload demands. The solution uses the QCT QuantaGrid D43N-3U GPU server, an NVIDIA-Certified dual-socket accelerated server powered by the 3rd gen AMD EPYC™ series 7003 processors, designed to serve a wide range of workloads. The storage solution provides data lifecycle management for storage system scalability and data protection. The solution also includes building blocks for NGS, MD medical Imaging which makes it easier for users to focus on their work rather than spending time preparing environments, configurations, and connecting to various servers.

QCT provides scalable infrastructure solutions designed to meet the needs of different sized organizations. QCT POD for Medical could be offered in an all-in-one cost effective and compact system with a small footprint that provides necessary infrastructure to run AI workloads. It is an ideal solution for small and local medical institutions or organizations to develop and run AI applications. Larger institutions may select the larger scale of QCT POD for Medical system, which provides an enhanced performance solution that allows users to unleash the power of AI and HPC in diverse use cases.

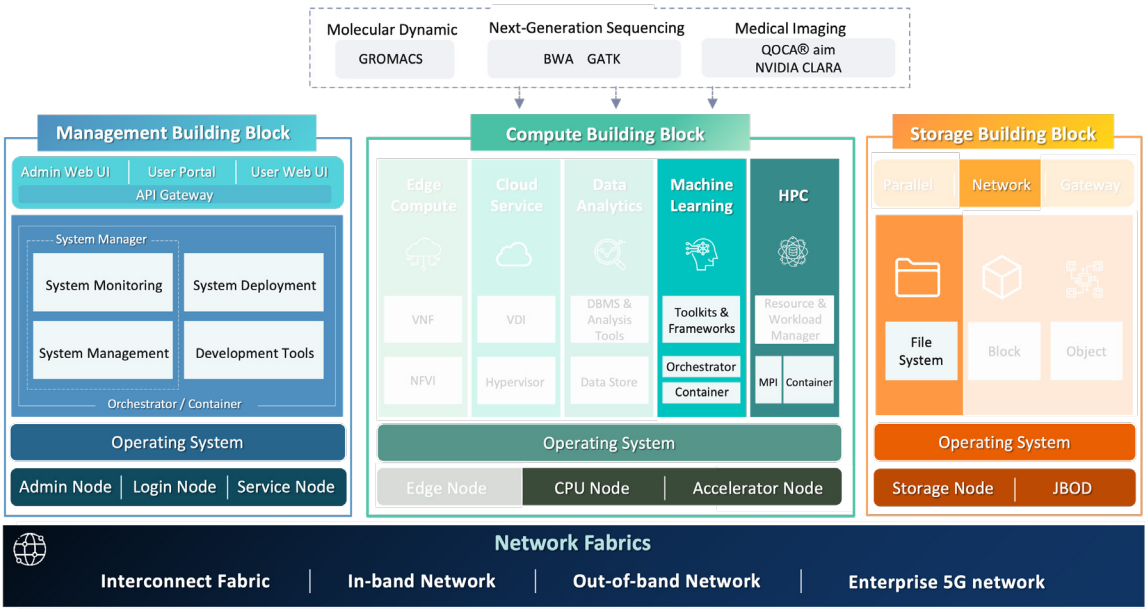


*Figure 1. QCT POD for Medical Architecture*

## Building Blocks for Administrators

It is a challenge for administrators to deploy and manage a cluster with converged infrastructure and diverse services. The QCT POD for Medical solution contains a one-key deployment tool, which administrators can use to simplify their deployment process and help alleviate IT staff stress. Using QCT POD for Medical may shorten the offline system deployment process from days to hours to set up a new cluster with hardware, container, and storage environments. The solution allows administrators to provision their software environment automatically and also provides the latest BIOS and BMC firmware, NVIDIA GPU accelerators, core firmware and drivers.

QCT POD for Medical provides real-time monitoring support including web-based dashboards and system alerting with useful administration tools. Administrators can check the status about central processing units (CPUs), memory, disk, and GPUs, etc. A web-based dashboard enables administrators to monitor and interpret the system status easily for both container and bare metal environments. Administrators can check the resource utilization for Central Processing Units (CPUs), memory, disk, and Graphic Processing Units (GPUs). Management tools also simplify cluster system management for log, resource, containerized applications, event alerts, and account management. An integrated job scheduling system helps resolve job conflict issues because the system performs automatic job queuing. The QCT POD for Medical solution is highly reliable, high available, and scalable.

## Robust Development Environment for Developers and Users

The QCT POD for Medical system includes a robust development environment that makes it easier for developers and users to launch their tools and help optimize AI workflow and HPC workload performance. It is composed of pre-validated workload packages as well as pre-installed diverse development tools composed of compilers, libraries, an application framework, and data visualization tools.

The system allows users to automatically provision a Jupyter notebook environment and multiple developers can access the web-based computational platform to run code and do data analysis interactively. Using the QCT POD for Medical platform makes work easier for developers because they can focus on their area of expertise and reduce their efforts for environment preparation.

# NVIDIA Clara Train SDK Use Case: Accelerate the AI imaging workflow

The NVIDIA Clara Imaging solution is a healthcare application framework for AI-powered imaging, genomics, and the development and deployment of smart sensors. It includes full-stack GPU-accelerated libraries, Software Development Kits (SDKs), and reference applications for developers, data scientists, and researchers to create real-time, secure, and scalable solutions. The Clara imaging solution provides easy-to-use, domain-optimized tools to create high-quality, labeled datasets used to train robust AI models based on their unique medical data in clinical health organizations. The NVIDIA Clara helps developers build robust AI models with techniques such as AutoML, privacy-preserving Federate and Transfer Learning. Once the AI model is developed, it can be deployed into an application to interface with hospital environments.

NVIDIA GPU Cloud (NGC) is a GPU-accelerated cloud platform optimized for deep learning and scientific computing, which accelerates the workflow for developers. The Clara Train SDK is a critical SDK for medical imaging scenarios. The user typically needs to perform a few steps to use the Clara Train SDKs. On the server side, the user needs to prepare the Clara Train SDK container and start the AI-Assisted Annotation (AIAA) server manually. On the client side, the user needs to install visualization tools such as MITK and configure it to connect to AIAA server. Some other operations for preparing data are required and must be manually set up between the client and server during the whole process.

With the QCT POD for Medical solution, QCT has done extensive work to help automate and streamline the NVIDIA Clara setup process to save time and effort for users to allow them to just focus on leveraging NVIDIA Clara to build their AI applications.

The foundation of QCT POD for Medical is Kubernetes orchestration of the deployment and execution of pods on mixed CPU-only and GPU-accelerated compute resources. As shown in Figure 2, QCT POD for Medical gives users access to a web-browser-based desktop environment that has the necessary visualization software pre-installed. The Clara Train SDK container is already pre-loaded and running on a GPU-backed server when users enter their personalized GUI environment. On the server side, users download the Clara Train SDK Docker image and then run the SDK with the proper configuration. On the client side, users only need to download the desired training model, open the virtualization client, upload patient or medical data-sets to the Clara container, then initiate the segmentation and start the analysis. Multiple users can access the system with enhanced resource allocation.

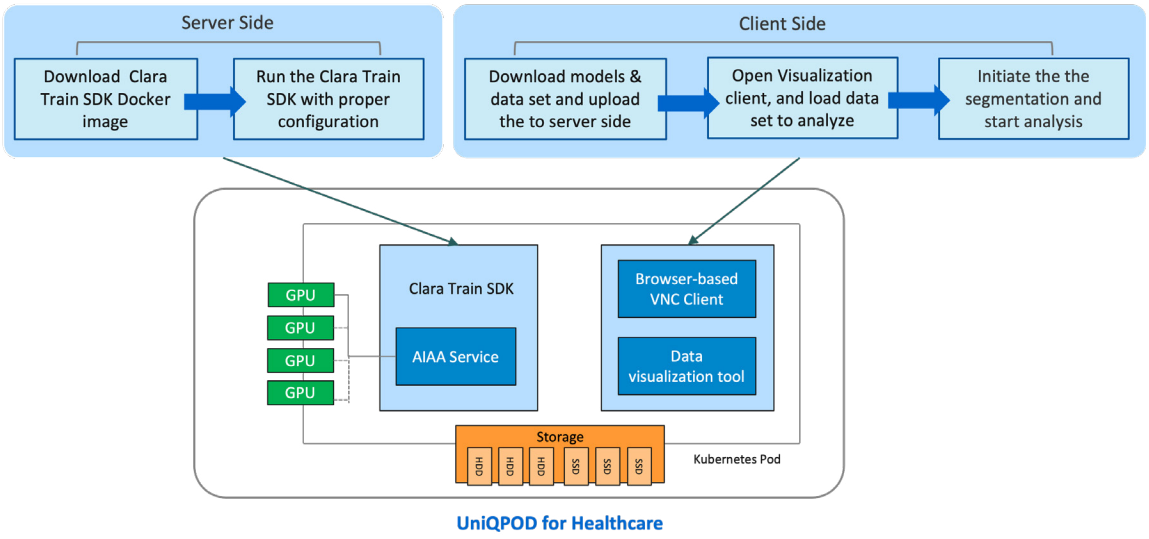


*Figure 2. QCT POD for Medical integrates Clara Train SDK and streamlines the workflow*

# QOCA® aim Use Case: Simplifying the AI Pipeline for Medical Researchers and Scientists

Quanta Computer, a Global Fortune 500 company located in Taiwan, created QOCA® aim as an integrated AI platform for medical research that provides data management, modeling, and inferencing. QOCA® aim is a Quanta AI aided medical data and automatic AI model building and inference platform. Researchers, physicians, hospitals, and developers use the tool for data processing of medical information, AI data model training and to generate clinical meaningful AI results for diagnosis.

The QCT POD for Medical solution supports QOCA® aim service and environments that make it easy to launch and run the application. In addition, QCT POD for Medical provides administrators with simplified deployment, management, and enhanced resource allocation. An integrated, robust development environment provides an integrated AI framework and fine-tuned AI pipeline and workflows.

## Summary

The growing field of precision medicine provides great promise in areas such as Next-generation DNA sequencing, medical imaging, and molecular dynamics. However, running precision medicine workloads typically requires HPC infrastructure and systems capable of running AI workloads. Many organizations have separate compute and storage systems that are not capable of meeting precision medicine demands.

The QCT infrastructure solves the issue of siloed compute and storage systems that cannot manage precision medicine needs. QCT can tailor infrastructure with server-level systems containing powerful computing resources designed to meet HPC and AI workload needs.

QCT's integrated QCT POD for Medical solution includes a web-based GUI management systems that provides administrators and developers with system monitoring and job scheduling tools. In addition, the QCT POD for Medical solution best practice system contains pre-configured and pre-validated files, frameworks, environments, and libraries making it easy for users to launch and run tools such as NVIDIA Clara and QOCA® aim. The QCT POD for Medical solution provides medical workload building blocks, a streamlined deployment process, and a simplified solution which enables organizations to meet their HPC and AI precision medicine workload needs.

---

For more information on QCT and how QCT POD for Medical can help your organization, see:

QCT POD for Medical:
https://go.qct.io/qct-pod/qctpod-for-medical/