



# ADAPTIVE HPC CLUSTER OPTIMIZES NEXT-GENERATION DNA SEQUENCING WORKLOADS

Produced by Tabor Custom Publishing  
in conjunction with:

**HPC** | wire



intel®

## Executive Summary

An important tool in predictive medicine and genomic analysis is the use of Next Generation Sequencing (NGS), also known as massive parallel sequencing. Sequencing provides information about the kind of genetic information that is 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.

The sequencing segment accounted for the dominant revenue share of more than 54% in 2020 as sequencing is a critical step of the workflow. The global next generation sequencing market size was valued at [USD 3.99 billion in 2020](#) and is expected to grow at a compound annual growth rate (CAGR) of 14.4% during the forecast period from 2021 through 2028. The next-generation sequencing market is projected to reach [USD 24.2 billion by 2026](#) from USD 10.3 billion in 2021, at a CAGR of 18.7% during the forecast period. The cost of acquiring NGS data [continues to decline](#), while the NGS data volume grows tremendously.

NGS processing requires the use of high-performance computers and storage solutions due to the massive amounts of data which must be processed, stored, and analyzed. Therefore, efficiency, simplicity, and scalability in processing and storing NGS workloads is critical.

This paper provides an overview of how solutions by QCT, Qumulo and Intel provide an adaptive platform for HPC NGS workloads. The solution is designed to meet the NGS requirements for optimized performance, easy management of data, and the ability to store various data types to obtain fast processing and optimize an organization's total cost of ownership (TCO).

[Quanta Cloud Technology \(QCT\)](#) is a leading cloud datacenter solution provider with extensive experience in developing High Performance Computing (HPC) solutions for companies in a variety of fields. [UniQPOD for Healthcare](#) is a UniQPOD solution powered by Intel to provide pre-validated and pre-configured custom hardware and management tools for use in medical solutions such as NGS. The QCT solution provides a performance boost when processing NGS data along with a Return on Investment (ROI) improvement and process acceleration with easy management of the QCT hardware solution and UniQPOD for Healthcare.

[Qumulo](#) is a leader in enterprise-proven hybrid cloud file storage software running on the Quanta computing platform. Qumulo provides real-time visibility, seamless scale, and real-time control of data across on-prem and cloud. Qumulo provides a complete storage ecosystem for data management. Qumulo's file storage platform efficiently manages billions of files both large and small, and supports a variety of protocols including SMB, NFS, FTP and REST. Qumulo Core dynamically manages file permissions across all protocols simultaneously, which means that all phases of the genomic analysis workflow can use the same Qumulo cluster for real-time collaboration.

## The NGS Workflow

The NGS workflow includes primary, secondary, and tertiary analysis as shown in [Figure 1](#). During primary analysis in laboratories, DNA sequencing instruments generate many small binary base call (BCL) files, containing unordered DNA sequence fragments. A demultiplexing process assembles BCL files into a FASTQ file, a text file that stores the combined output results along with corresponding quality scores. BCL files are typically large files greater than 225 Gigabytes (GB) which are usually stored approximately one month. FASTQ files can range from individual files greater than 120GB each, to a vast number of very small files, all requiring retention of up to three years.

Secondary analysis is the process for genome alignment and variant calling used to identify variants and differences in an individual's genome by comparing it to a DNA reference sequence. This process is the most time-consuming step in NGS processing and can take minutes to days depending on software, computing and storage resource and technology. Secondary analysis is frequently performed by researchers in labs or scientific institutes. During the alignment step, DNA sequence fragments are quality checked, pre-processed, and aligned to a reference genome. A Binary Format (BAM) file containing alignment sequence data is created that stores the alignment data. Variant calling compares differences between the data and the reference genome, with results stored in a Variant Call Format (VCF) file used in bioinformatics for storing gene sequence variations. A GenomeAnalysisToolkit (GATK) with tools for analyzing high-throughput sequencing data is often used in variant discovery. Tertiary analysis begins when scientists and researchers use the NFS, or SMB data files generated by NGS processing in their research.

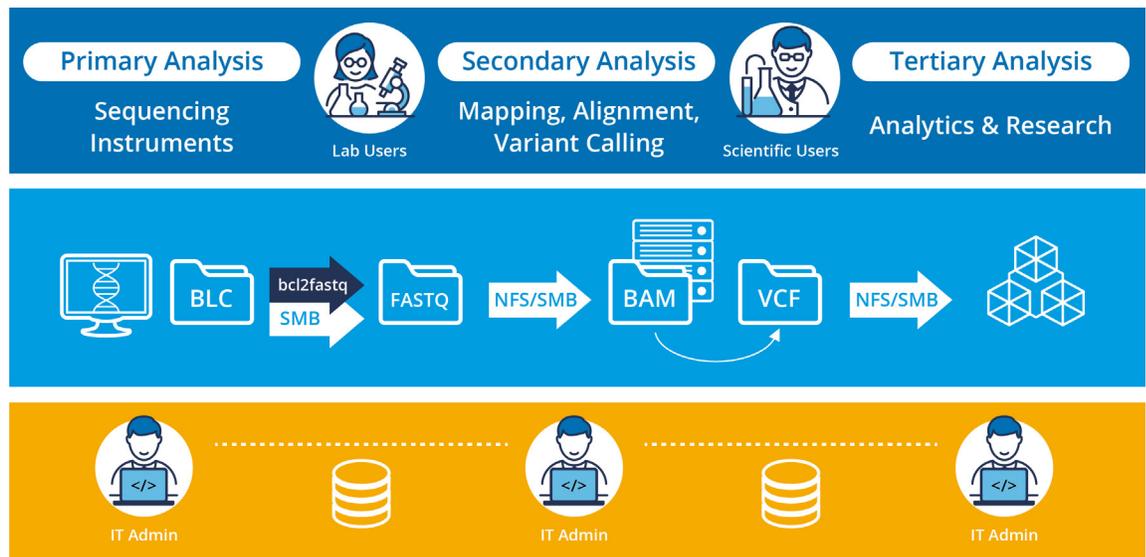


Figure 1. NGS Workflow

Processing and storage of NGS data are major pain points for organizations. It can take over 30 hours using a standard GATK best practice tool to perform 30X whole genome sequencing. Organizations need a way to speed processing during the NGS secondary analysis phase.

Organizations also face challenges with processing and storing NGS files because legacy compute, storage and file systems are often based on old designs which cannot meet the demands of modern NGS workflows.

## How QCT and Qumulo Meet NGS Workloads Needs

Organizations need an adaptive HPC infrastructure that can handle NGS data sets and provide optimized performance, easy management for data processing and storage as well as meet data storage retention requirements for regulatory compliance. QCT hardware and the UniQPOD for Healthcare solution meets NGS workload needs. UniQPOD for Healthcare is a Platform on Demand (POD) solution, which offers an innovative technology system with building blocks designed to meet different medical demands, like Next-Generation Sequencing (NGS), Molecular Dynamics (MD) and Image Recognition. UniQPOD for Healthcare innovations aid organizations with processing medical workloads, streamlined deployment, and simplified system management to reduce complexity and provide flexibility for specialized NGS workloads.

Storage requirements for genomic data has increased dramatically with the advancement of next-generation sequencing. Organizations need a solution that provides data management and storage tiering to reduce the overall TCO. NGS processing and analysis generates massive amounts of data, often in an unstructured data format containing both small and large files. Qumulo’s file data storage software is integrated with QCT solutions to allow organizations to easily store and manage file data generated by NGS workloads at any scale. [Figure 2](#) shows the infrastructure of the combined QCT and Qumulo solution.

The QCT-Qumulo solution provides dedicated management services along with a high-performance compute infrastructure capable of meeting diverse workload demands as well reducing the analysis time during the NGS secondary analysis phase. The Qumulo high-performance data storage solution works with many different file types while providing easy management, scalability, and real-time data visibility and system control.

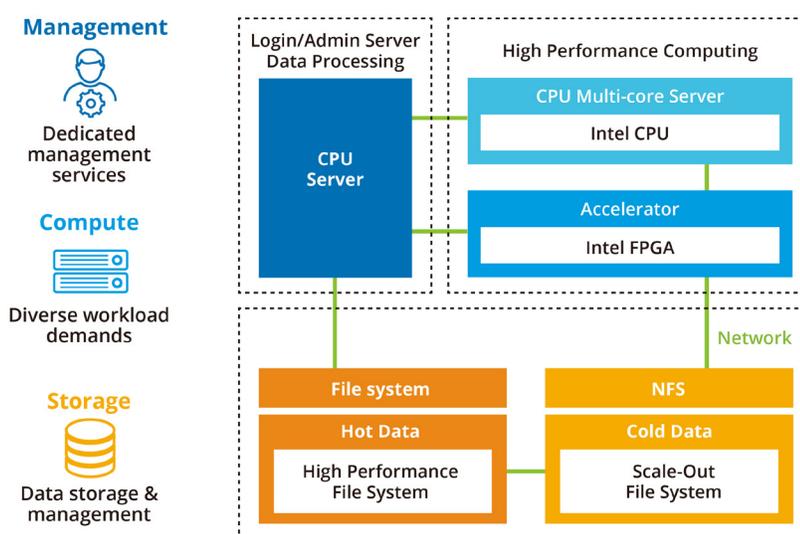


Figure 2. QCT-Qumulo Infrastructure and Benefits

## QCT Hardware Solutions use Latest Intel Components

The [QCT D53XQ-2U](#) server is the compute hardware used in the NGS solution. The D53XQ-2U provides scalability up to 10x PCIe 4.0 expansions. The server is optimized for AI acceleration and can support up to 2x dual-width accelerators in a 2U system. It supports all 24x NVMe flash drives as hot tier storage, targeting HPC and enterprise workloads. As part of the solution, Qumulo leverages a QCT 1U high density storage server as an appliance, which supports a 12 large form factor to build a large storage capacity while only consuming a small space in a modern enterprise data center to achieve data storage demands.

The QCT servers include 3rd Gen Intel® Xeon® Ice Lake Scalable processors which use [Intel® Turbo Boost Technology](#) and [Intel® Hyper-Threading Technology](#) to efficiently use processor cores which can increase processors speeds. In addition, [Intel® Stratix 10](#) and [Intel® Arria 10](#) accelerators are part of the solution which can drastically decrease secondary NGS analysis time. In addition, QCT servers use [WASAI-Lightning™](#) acceleration technology running on Intel FPGAs with UniQPOD for Healthcare to reduce DNA analysis time and provide better management.

## Qumulo Solution Solves these Typical NGS Storage Issues

The Qumulo C-Series Cached Performance Class & K-Series Active Archive Class appliances used in the QCT solution are both based on the QCT 1U 12LFF Intel platform which supports up to four flash disks as cache. QCT with Qumulo storage appliances as part of the solution is used because of [Qumulo's ability to effectively store and manage files](#) generated as part of the NGS DNA sequencing processing. Many storage systems cannot efficiently store NGS data due to the fact that genomics data sets are large with a mix of small and large files, use various file protocols with both a sequential and random-access pattern, and require low I/O latencies.

The Qumulo Core software pre-installed as part of the QCT 1U 12LFF Intel platform provides:

- An [easy-to-use management dashboard](#) that shows what is happening in the storage system in a real-time event at billion-file scale
- The ability to create petabyte data lakes in minutes
- ML intelligent predictive caching that enables all-flash levels of performance even on hybrid clusters
- Unique Qumulo Scalable Block Store (SBS) stores very large and very small files in a highly space-efficient manner
- The ability to perform analysis in real-time for cost efficiencies
- Scalability – clusters scale up linearly
- Cost effectiveness because 100% of storage capacity is available for file storage
- Simultaneous mixed protocol support (SMB, NFS, FTP and REST) so all phases in genomic pipeline can use the Qumulo cluster collaboratively

## QCT Use Case: Speeding NGS Processing

QCT cooperates with WASAI Technology Inc, a company that specializes in Big Data acceleration, to process the massive amounts of data in NGS. QCT uses the WASAI-Lightning™ with an Intel accelerator card running on the UniQPOD for Healthcare Platform to get a significant performance boost during DNA sequencing.

The WASAI Lightning Acceleration solution running on UniQPOD for Healthcare can efficiently lower time required by the GATK Best Practices pipeline during the secondary phase of NGS sequencing. While processing a WGS (whole genome sequencing) workflow by GATK Best Practices, it usually takes more than 32 hours with CPU-only computation. With a Genome-in-a-Bottle (GIAB) genome sample (NA12878 30x), the Intel® Stratix 10 with WASAI Lightning can process a WGS in 2 hours and 28 minutes. The Intel® Arria 10 with WASAI Lightning can process a WGS in 4 hours and 58 minutes. This processing time is approximately 11 times faster than a CPU-only solution. For more information on the testing results, see the [Next Generation Sequencing application with WASAI on UniQPOD for Healthcare white paper](#).

### PrecisionFDA NA12878 30X Execution Time on QuantaGrid Acceleration Server



Figure 3. WASAI Lightning running BWA & GATK with Intel® FPGA on QuantaGrid Acceleration Server

## Qumulo Genetic Sequencing Use Case: Helping Progenity Plan for the Future

[Progenity, Inc.](#) is a privately-held biotechnology company that provides clinicians with complex molecular and specialized diagnostic tests for women's health, reproductive medicine, and oncology. Progenity's work in genetic sequencing has generated more than a billion files. Progenity selected Qumulo's hybrid cloud file storage system to replace its legacy vendor in 2016. Qumulo was able to offer an affordable solution for Progenity, without sacrificing performance or scalability. For more information on the solution, see this [Progenity case study](#).

### Summary

Next Generation Sequencing (NGS) is a DNA sequencing technology that requires HPC parallel file processing to sequence and analyze individual strands of DNA molecules and compare them to a DNA reference sequence. NGS requires the use of HPC computers and storage systems capable of processing and storing massive amounts of data.

Processing and storage of NGS data cause problems for organizations because their compute systems cannot meet the processing needs of NGS. Storing NGS data is a problem because genomics data sets contain massive amounts of data with a mix of large and small files, use various file protocols and require low I/O latencies that most storage solutions cannot handle.

Quanta Cloud Technology (QCT), a leading cloud datacenter solution HPC provider, using the QCT UniQPOD for Healthcare solution, provides pre-validated and pre-configured custom hardware and management tools for use in medical solutions including NGS. QCT with Qumulo storage appliances are part of the NGS storage solution. The QCT-Qumulo solution powered by Intel helps organizations increase their ROI by providing pre-configured hardware optimized for performance in processing NGS data, easy management of data, and the ability to easily store and manage various types of data generated across the NGS cycle.

*- QCT brings agility and innovation with a purpose-built solution to your end-to-end workflow in NGS-*

For more information on QCT and how UniQPOD for Healthcare can help your organization, see: UniQPOD for Healthcare: <https://go.qct.io/qct-pod/uniqpod-for-healthcare/>