PacBio + VAST

Success Story

# Accelerating Genomic Discovery with VAST Data



## The Challenge: Growing Data and Compute Requirements

Genomics leader PacBio is a pioneer in long-read DNA sequencing, empowering scientists to better understand genomes for research ranging from healthcare to agriculture. However, generating extreme amounts of data from developing and manufacturing PacBio's highly accurate sequencers like the Revio and Sequel IIe Systems presents massive data challenges.

Adam Knight, Director of IT Infrastructure at PacBio, explained, "For us, it's the time it takes to sequence genomes, the cost of sequencing, and the error rates and accuracy that are important. We're trying to increase accuracy, speed, and throughput, while reducing error rates and overall cost of sequencing runs." Knight added, "We had to do something different. It had to be more than just storage; it had to be a seamless fit into our existing operations." Knight continues, "Our sequencers produce gigabytes of data per sequencing run for our customers, so you can imagine how much data we need to generate in order to design, develop, and manufacture an instrument platform like Revio. We also have a geographically distributed team working around the world, so we were facing delays due to the logistics of getting data to the right people at the right time."

With the totality of datasets growing into the petabyte range, performance and scalability were top concerns. Like many growing companies, PacBio's infrastructure sprawled across locations and legacy infrastructure. Knight noted, "We faced challenges with diverse file access protocols and bottlenecks when moving data between our instruments and our network file shares." Consolidating on a high-performance and easily scalable data platform became a priority.

### About Us

VAST is the data platform for simplicity at scale. VAST engineered the first new scale-out architecture in 20 years to take you beyond the conventional limits of scale, resilience and cost.

**Looking to simplify your data center and unlock insights from all of your data? Contact us at hello@vastdata.com**

*"It's a robust, high-performance platform that's easily scalable. We recently added another 2PB to our cluster with no issues. VAST has been reliable as we continue to expand it."*

**- Adam Knight,**
**Director of IT Infrastructure, PacBio**

PacBio

## Seeking a Unified High-Performance Platform

As PacBio was looking to upgrade its existing HPC storage infrastructure, they investigated what else they could use. Knight says "We don't just hit the easy button and go with the status quo. We see what's out there and spend time and effort to architect the best solution for our business needs. It would have been easy to extend support on what we had but we're nerds; we wanted to see what was new in the market."

According to Karun Kallakuri, Staff HPC Engineer at PacBio, the company sought three key attributes in a solution: "Performance, ease of management, and a single namespace across protocols like SMB, NFS, and S3."

Kallakuri explained, "Our top criteria are performance and ease of management. VAST is a no-brainer for that. And then a single namespace across various protocols like NFS and S3 changed how we manage our data architecture. It saved us so much time." Leveraging the VAST multiprotocol namespace allowed PacBio to transfer data runs off of their instruments eight times faster than previous technologies.

After evaluating options, PacBio selected VAST Data's DataStore which addressed PacBio's requirements of extreme performance, limitless scalability, and multiprotocol support. The VAST Data Platform is composed of several components including a DataStore, DataBase, DataSpace and DataEngine.

As Knight explained, "VAST initially caught our eye as a more affordable high-performance tier for HPC data management. It was positioned as highly scalable with a solid roadmap and good support. It beat out our existing solution." Beyond performance for HPC workloads, VAST Data offered ease of management and the ability to consolidate file and object data under a global namespace.

> *"We had to do something different. It had to be more than just storage; it had to be a seamless fit into our existing operations."*
>
> **- Adam Knight, Director of IT Infrastructure, PacBio**
>
> PacBio

## Accelerating Genomic Workflows

With VAST Data's scale-out architecture, PacBio has significantly accelerated how quickly they can process and analyze data after sequencing. Kallakuri explained, "We used to use rsync to transfer manufacturing data between sites, which sometimes took days. VAST enabled using parallel S3 commands to achieve line speed and reduce transfer times from days to hours for multi-terabyte quantities of data."

He continued, "The same data is immediately available via NFS for our HPC workloads to kick off. This was a game changer, improving time to results."

In addition, VAST enabled a major consolidation of legacy NAS and object storage systems. Knight said, "VAST is now the majority of our data platform spend. It's used for HPC data management, primary storage, instrument data, software builds - we've expanded it into more areas."

Kallakuri added, "We leverage the global namespace for simplicity and time savings in managing the file system." Avoiding data movement across tiers improves performance and productivity.

*"The increased throughput is a huge help for our deep learning workloads where high I/O is critical. Being able to mount and operate seamlessly across platforms without file permission or flag issues makes our lives much easier. There's zero IT overhead regarding management, since VAST handles that for us."*

- **Adam Knight, Director of IT Infrastructure, PacBio**

PacBio




## Powering Ongoing Innovation

With VAST Data, PacBio can keep pace with its sequencing innovations and increasing data demands. Knight explained, "It's a robust, high-performance platform that's easily scalable."

He elaborated, "We recently added another 2PB to our cluster with no issues. VAST has been reliable as we continue to expand it. And we're excited to see how easy it is to scale as we grow our data platform infrastructure."

The software-defined architecture simplifies adding performance and capacity. Knight noted, "As we grow our cluster, VAST has terrific dashboards and tools showing bottlenecks and data flows."

The company also benefits from the robust roadmap of the VAST Data platform, like the new data catalog. "The ability to manage metadata on-box, with no external queries needed, will be massively useful" remarked Knight. "It will immediately accelerate workflows that currently require heavy offline processing."

Kallakuri added, "We are in the process of exploring it and are excited to provide API access to the catalog for our developers instead of crawling the file system every 10 minutes."

He concluded, "Our R&D teams love the increased speed achieved through S3 transfers and so we expanded its use across our internal network." By delivering breakthrough performance today, while rapidly enhancing intelligence, automation, and multi-cloud capabilities, VAST Data is powering PacBio's innovation engine now and into the future.



vastdata.com/demo hello@vastdata.com



©2024 VAST Data, Inc. All rights reserved. All trademarks belong to their respective owners.
03/2024