Cloud Storage & Data Management Strategy
for NovaSeq X+ Data

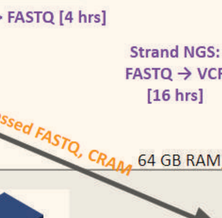
Srikant Sridharan, Aman Saxena, Priyanshu Agarwal, Radhakrishna (RK) Bettadapura

Strand Life Sciences, Bangalore, India

VISIT OUR **WEBSITE**

strand
WE SOLVE PROBLEMS IN GENOMICS.

Conference Name
Bio-IT World Conference & Expo 2024
Dates - April 15th to 17th
Place - Boston, MA, USA



Contact

Radhakrishna (RK) Bettadapura
Vice President, Business Development
+1 (415) 917-9605
rk@strandls.com

Introduction

The NovaSeq X Plus (NSX+) generates 1.5 TB of data on each run of the 10B flow cell. Our data operations team has developed a streamlined pipeline, incurring AWS costs of $300 per run to process 100 samples. The total AWS cost includes compute and transfer at $220 and $80 per run respectively.

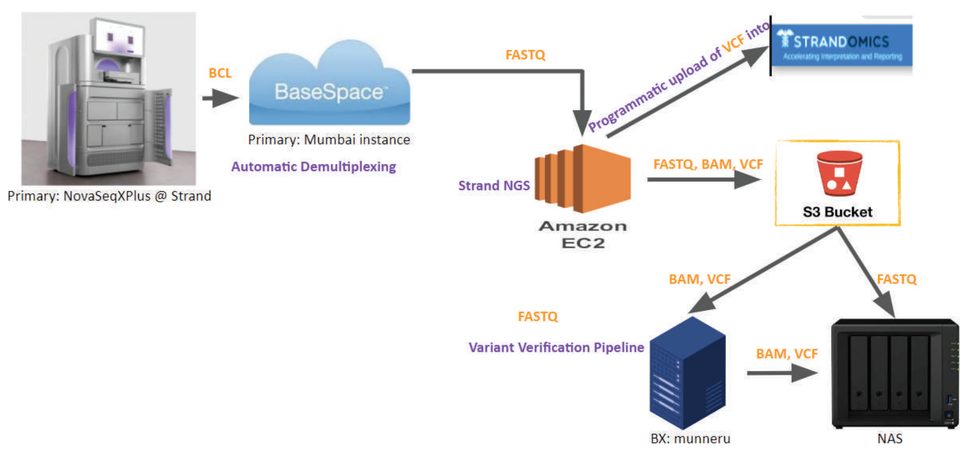
Further, assuming we generate 5 TB of data/month for the next 120 months and interpolate between the minimum and maximum NSX+ capacity, we anticipate storing 3.5 PB of data.

Considering the growth in genome sequencing rates, we have designed a long-term storage strategy to accommodate 16 PB of data as we scale our operations over the next decade.

Preliminary Data Flow Architecture

1. Transitioned to streaming data from NSX+ instead of NovaSeq 6000 for improved cost and time efficiency.
2. Initially, the team implemented a basic data flow architecture.
3. Raw BCL files from the sequencer were uploaded to BaseSpace, Illumina's cloud storage solution, where they underwent demultiplexing, and FASTQ files were generated.
4. Secondary bioinformatics pipelines were executed on an Amazon EC2 instance to generate VCFs from FASTQs.
5. VCF files were subsequently uploaded to StrandOmics, our tertiary interpretation and reporting software.
6. Simultaneously, data from the EC2 instance was transferred to an Amazon S3 bucket.
7. From S3, the BAM, VCF and FASTQ files were transferred to our on-premise infrastructure consisting of network attached storage (NAS) solution and a server attached to NAS (BX server) for triggering a variant verification (VV) pipeline.

Preliminary Data Flow



Cost-Optimized Data Flow & Operations

We improved our data flow for optimal cost-efficiency based on insights from processing over 1000 samples in the last two months:

Cost Savings via Improvements in Data Operations:

1. Tightened the run cycle from one month to 28-36 hours by optimizing start and end points.
2. Previously, data storage on S3 led to the pipeline being replicated in different instances, and this approach proved inefficient due to prolonged EC2 instance activity. This process is now streamlined to a single-run.
3. VV shifted to local servers, reducing costs and ensuring effective usage of compute resources.
4. Improved data access by downloading StrandNGS data locally within 48 hours.

Cost Savings via Storage:

1. Detached Amazon EBS volumes from EC2 instances and attached them to more cost-effective instances for on-premise downloads eliminating S3 dependency. The current compute infrastructure utilizes existing storage without the need for additional third-party solutions, resulting in savings of $80/run.
2. Routed ORA compressed FASTQ and CRAM files from the sequencer to NAS and then to Azure Blob storage, bypassing the previously expensive S3-based solution.

Infrastructure Decisions:

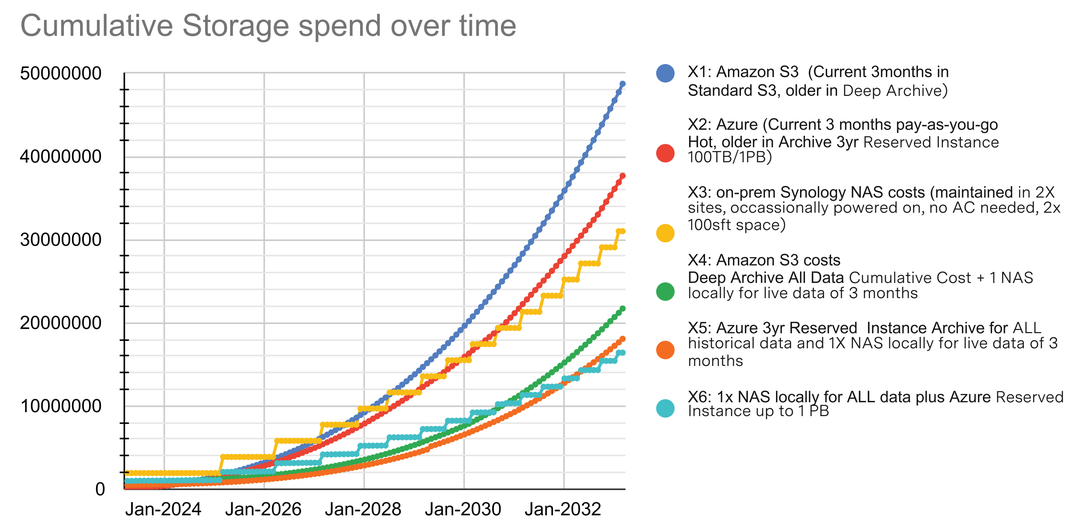
1. Chose EC2 for computing and temporary storage due to its parallelization capabilities, eliminating the need for multiple in-house servers.
2. Opted for Azure for long-term storage due to 15% cost reduction obtained from using reserved instances.

Overall, the solution lowers costs, and the new data flow architecture achieves a **$650 reduction per run.**

Cost-Optimized Data Flow

Long Term Storage Strategy

1. We started Month 1 of operations generating 5 TB of data and projected that over 120 months, it will grow to 53 TB of data. This was based on linear interpolation from the minimum NovaSeq X+ capacity (1x 10B per week) to the maximum capacity (3x per week of 2x 25B) in ten years. So, at the end of 10 years, we projected to generate around 3.5 PB of data.
2. With some aggressive assumptions regarding the growth rates of genome sequencing, if it takes off, we estimate that around **16 PB of data** will need to be stored over the next decade.
3. We needed to arrive at a storage strategy around this vast data, which would allow us to keep this data safe at the lowest cost possible for the next 7-8 years as we scale our operations and bring in more revenue via an increased volume of samples.
4. The access pattern was of archival mode, where retrieval will be rare, and there can be a gap between data availability and retrieval.



5. We compared corresponding costs incurred over the next ten years for the above strategies.
6. After looking at the comparison in cost, we chose Azure 3-year Reserved Instance Archive for all historical data (X5).
7. We also decided to store hot data (last three months) and sample data in on-prem Synology NAS to allow faster retrieval of recent data. By choosing this strategy, we projected to spend around **$250K USD** to store all the data generated over ten years.
An interesting observation we arrived at regarding hot data storage is that bypassing the hot tier of S3 saves costs (X1). By skipping to store three months' worth of hot data in the cloud, we reduced the costs from **$600K to $250K** USD, saving **60% of cloud costs.**