Moving Genomics to the Cloud: Compute and Storage Considerations

Live Webcast
September 9, 2021
10:00 am PT / 1:00 pm ET
Today’s Presenters

Alex McDonald
Independent Consultant
Chair, SNIA Cloud Technologies Initiative

Michael McManus
Director, Precision Medicine & Principal Engineer
Intel

Torben Kling Petersen
Principal Engineer
HPC Storage BU
HPE

Christopher Davidson
Application & Performance Engineering, Manager
HPE
SNIA-at-a-Glance

180
industry leading organizations

2,500
active contributing members

50,000
IT end users & storage pros worldwide

Learn more: snia.org/technical  @SNIA
What We Do

Educate vendors and users on cloud storage, data services and orchestration

Support & promote business models and architectures: OpenStack, Software Defined Storage, Kubernetes, Object Storage

Understand Hyperscaler requirements Incorporate them into standards and programs

Collaborate with other industry associations
SNIA Legal Notice

The material contained in this presentation is copyrighted by the SNIA unless otherwise noted. Member companies and individual members may use this material in presentations and literature under the following conditions:

- Any slide or slides used must be reproduced in their entirety without modification.
- The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.

This presentation is a project of the SNIA.

Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.

The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.
Agenda

- Trends in Genomics & the Need for Data Storage & Management
- Data Management & Storage Considerations
- Public Science in Practice
Trends in Genomics
Michael McManus
Large Scale Sequencing is already underway!

Source: Frost & Sullivan, “Global Precision Market Growth Opportunities, Forecast to 2025,” January 2017 and Intel’s own market research 2017-2021
Cost per Genome is Dropping

Sequencers commonly used for Genomics

- **Illumina**
  - MiniSeq®
  - MiSeq®
  - NextSeq® 100
  - NextSeq® 550
  - NextSeq® 1000/2000
  - NovaSeq 6000®

- **DNBSEQ**
  - G50
  - G400
  - T7

- **ThermoFisher**
  - Ion GeneStudio S5
  - Ion Torrent Genexus

- **Oxford Nanopore**
  - MinION
  - GridION
  - PromethION

- **PacBio**
  - Sequel IIe System
Overview of the Human Genomics Workflow

**Primary Analysis**
- Base Calling → (FASTQ / uBAM)
- Quality Control

**Secondary Analysis**
- Alignment
- Removing duplicates, realignment, recalibration
- Variant Calling

**Tertiary Analysis (Interpretation)**
- Annotate
- Filter & Classify
- Prioritization
- Visualization
- Interpret / Report
More than one million SARS-CoV-2 genome sequences have been shared on the GISAID data-sharing platform since January 2020, and are helping researchers to track the spread of viral variants. Most are from the United States and Europe, but contributions come from every region of the world.

1. **January:** First SARS-CoV-2 genome, from China.
2. **March:** First African sequence, from Nigeria.
3. **April:** Victoria, Australia, has 1,300 cases; 80% are sequenced, identifying clusters from cruise ships and hospitality venues.
4. **May:** UK sequences 6% of cases, more than any other country.
5. **November:** South African surge prompts intensified surveillance. Researchers find a widespread new variant - B.1.351.
6. **December:** 40% of genomes sequenced in Manaus, Brazil, are of the P.1 variant, with mutations linked to increased transmissibility and immune evasion.
7. **March:** US sequencing rate doubles, owing to a government mandate for surveillance and funding from the Centers for Disease Control and Prevention.
Genomic epidemiology of novel coronavirus - Global subsampling
Genomic epidemiology of novel coronavirus - Global subsampling
Genomic epidemiology of novel coronavirus - Global subsampling


Phylogeny:
- Clade A
  - 20H (Beta, V2)
  - 20I (Alpha, V1)
  - 202 (Gamma, V3)
  - 21A (Delta)
  - 21B (Kappa)
  - 21C (Epsilon)
  - 21D (Eta)
  - 21F (Iota)
  - 21G (Lambda)

Geography:

©2021 Storage Networking Industry Association. All Rights Reserved.
Genomic epidemiology of novel coronavirus - Global subsampling

Showing 3534 of 3534 genomes sampled between Dec 2019 and Aug 2021.
Illumina NovaSeq 6000®

360 WGS/month

126TB/month (0.35TB/WGS)

4.2TB/day
389Mbps

Public Cloud

OR

Datacenter (HPC)
Four Data Storage Domains of Big Data in 2025

The World's Population – An Overview (~7.8B)

Source: 2021 World Population Data Sheet [https://interactives.prb.org/2021-wpds/](https://interactives.prb.org/2021-wpds/)

- China, 1,412
- Africa, 1,373
- South America, 433
- Europe, 744
- North America, 371
- Central America & Caribbean, 223
- Oceania (ANZ + others), 43
- Asia (w/o China), 3,239

Source: 2021 World Population Data Sheet [https://interactives.prb.org/2021-wpds/](https://interactives.prb.org/2021-wpds/)
2,743 Exabytes for all ~7.8B people on Earth

Storage for the World's Population in 2021 – 2,743 Exabytes!

China, 494
Africa, 481
Europe, 260
South America, 152
North America, 130
Central America + Caribbean, 78
Oceania (ANZ + others), 15
Asia (w/o China), 1,134

Source: https://interactives.prb.org/2021-wpds/
Sequencing Everyone on Earth – Storage Fun Fact

• Assume 1 byte of storage = thickness of a 10€ note (0.15mm)

• Imagine a stack of 10€ notes, 1 note for each byte of storage need to sequence the entire Earth’s population

• The required 2,743 Exabytes of storage could be represented as a stack of 10€ notes that would extend all the way from the Earth to the star 58 Eridani, ~43.5 light-years away.*


Ignoring the laws of Physics and General Relativity 😊
The Storage Challenge
Torben Kling Petersen
Convergence of High Performance Storage

Era of convergence of traditional simulation and AI requires NEW HPC storage

- Mainly **WRITING**
- LARGE files
- In mainly **SEQUENTIAL** order.
- Capacity measured in **PETABYTES**

Examples of traditional HPC storage:
- Cray ClusterStor L300
- DDN EXAScaler
- IBM ESS 3000

**CHALLENGE IN NEW ERA:**
An ORDER OF MAGNITUDE less performance for small, random I/O compared to traditional AI storage

- Mainly **READING**
- Files of **ALL SIZES**
- In mainly **RANDOM** order.
- Capacity measured in **TERABYTES**

Examples of traditional AI storage:
- NetApp AFF
- Dell EMC Isilon F-Series
- Pure Storage FlashBlade

**CHALLENGE IN NEW ERA:**
An ORDER OF MAGNITUDE more expensive per terabyte compared to traditional HPC storage

Converged workloads running on one machine in mission- or business-critical workflows
Traditional HPC design – On Prem

Compute system
CPU or CPU/GPU

Parallel File Systems
• Lustre
• Spectrum Scale

Ethernet (1 - 10 Gbps)

IB (FDR/EDR/HDR)
Ethernet (10 - 100 Gbps)

NFS

FC-AL (8 - 160 Gbps)

Tape Archive
Traditional HPC design – Cloud Based

Compute system
CPU or CPU/GPU

Ethernet (1 - 10 Gbps)
IB (FDR/EDR/HDR)
Ethernet (10 - 100 Gbps)

Parallel File Systems
• Lustre
• Spectrum Scale

NFS

Archive
e.g. Blob Storage
The "NEW" world – On prem or cloud based

Compute system
CPU or CPU/GPU

Parallel File Systems
• Lustre
• Spectrum Scale

Data Management

Tape
Cloud Storage
Data Lake (e.g. Ceph)

Zero Watt Storage

Data Placement
Data movement

Hybrid systems
(NVMe and HDD)

RDMA
RoCE
IB/Eth/SlingShot
TCP

RDMA
RoCE
IB/Eth
TCP

Hybrid systems
(NVMe and HDD)

Data Placement
Data movement

Single Virtual NameSpace

Compute

SCM
(e.g 3D Xpoint)
Data Services Requirements

- Data movement NVMe <-> HDDs
  - Policy based data migration based on capacity and age
  - Manual data migration
  - File purging policies
  - WLM directives

- Rapid search facility
  - External to file system -> low impact
  - Query function for advanced searching
  - HSM aware
Data Management

- Data movement - Primary FS to:
  - hot archive
  - object store
  - tape
  - cloud

- Policy based data migration based on
  - Age
  - Size
  - Type
  - Project
  - Classification
  - Usage history etc

- Manage multiple front ends
- Horizontal data movement
- Maintain full namespace mirror
- HSM and Incremental Backups

- Tiers gated by:
  - Cost per PB
  - Capacity growth
  - Retention requirements
  - Access performance
Moving it all “to the cloud”

**Benefits:**
- CAPEX vs OPEX
- Pay as you grow
- Shared resources
- Simple reconfiguration

**Challenges:**
- Moving data to/from the cloud
- Multi-tenancy / Security
- Legal constraints
- Performance
  - Containers kill throughput and IOPS
- Fair cost:
  - By capacity used?
  - By performance?
- Simple reconfiguration
  - Significant data movement
So What’s the Next Step?

Christopher Davidson
Public Science In Practice

The COVID-19 High Performance Computing Consortium

Bringing together the Federal government, industry, and academic leaders to provide access to the world’s most powerful high-performance computing resources in support of COVID-19 research.

100  600

Projects  Petaflops
COVID-19 HPC Consortium - Facts & Timeline

THEN...
- 1 of 13 original members & 1 of 5 original members from Industry
- One of the largest private-public collaborations ever with members from Government, Industry, & Academia

NOW...
- 43 consortium members & global
- Phase 2 – focus on therapeutics and patient impact
- 100 projects
- 600 Petaflops

Dec 2019
First cases of COVID-19

March 22nd
Launch of COVID-19 HPC Consortium

April 2020
Global cases reaches 1M

May 2020
Call for research proposals

May 2020
Europe PRACE & UKRI join Consortium

Aug 2020
Australia NCI & Pawsey join Consortium

Nov 2020
Phase II Focus on Therapeutics & Patient Impact

Nov 2020
HPCwire award for best “HPC Collaboration”

Jan 27, 2021
Global cases reaches >100M
Public Science In Practice

https://www.covid19dataportal.org/

https://covid.molssi.org/
Summary

- Genomic data is growing at an exponential rate
- Work smarter, not harder
- Compute is a small part of the problem; data management & storage are of utmost importance
- Public & Private cloud provide a means to keep pace with the science and democratize the process
- Cloud provides a number of challenges but nothing is impossible
Thanks for Viewing this Webcast

Please rate the webcast and provide us with feedback

This webcast and a copy of the slides will be available at the SNIA Educational Library https://www.snia.org/educational-library

A Q&A from this webcast will be posted to the SNIA Cloud blog: www.sniacloud.com/

Follow us on Twitter @SNIACloud
Thank you

Questions?