Optimizing High-Performance Computing Systems for Biomedical Workloads

Patricia Kovatch, Lili Gai, Hyung Min Cho, Eugene Fluder, Dansha Jiang

Scientific Computing, Icahn School of Medicine at Mount Sinai, New York, NY

Mount Sinai

HiCOMB Workshop May 18, 2020

Outline

Background

- Growth of Mount Sinai HPC and Data Ecosystem
- Computational Biomedical Research at Mount Sinai

The Mount Sinai HPC System Design

- Compute
- Memory
- Scheduling and Queues
- File System
- Archival Storage
- Cloud
- Sustainability

Future work

Growth of Mount Sinai HPC and Data Ecosystem - Minerva

Highlights:

- Supporting over **$100 million yearly NIH funding** in computational biology research
- Enabling over **900 biomedical publications** since 2012
- Growing our user base nearly **10-fold** since 2012
- Evolving a 70 teraflop machine to a **1.4 petaflop** machine in response to trends, actual usage, and user feedback in seven years
- ~ **18,000** Intel platinum 2.90 GHz compute cores, **28 PB** of parallel storage
- Established a chargeback fee structure for long-term stability and sustainability

Growth of Mount Sinai HPC and Data Ecosystem - Minerva

Minerva compute and online storage over 2012-2019

Compute Partition	Lifetime	Core Type	# of Cores
AMD	2012-2019	AMD Interlagos	7,680
Intel	2014-2019	Intel IvyBridge	2,508
BODE	2014-2019	Intel Haswell	2,484
Chimera	2019-	Intel Platinum	14,304
BODE2	2019-	Intel Platinum	3,840
Total	**Available**	**in 2020**	**18,144**

GPFS Name	Lifetime	Storage Type	Raw PB
SFA10K	2012-2019	DDN SFA10K	1.5
Orga	2014-	IBM ESS BE	3
Orga	2014-2017	IBM Flash	0.16
Orga	2014-	DDN SFA12K	5
Orga	2017-	IBM Flash 840	0.24
Hydra	2017-	IBM ESS LE	4
Arion	2019-	Lenovo DSS	14
Arion	2019-	Lenovo G201 flash	0.15
Total	**Available**	**in 2020**	**28**

Computational Biomedical Research on Minerva

Provide infrastructure, partnership and expertise for > $100 million in yearly NIH funding aiming at better understanding/treatment for diseases

- Including autism, insulin resistance in diabetics, schizophrenia and related neurobehavioral disorders, cardiac care, the origins of drug addiction and depression, and cancer progression

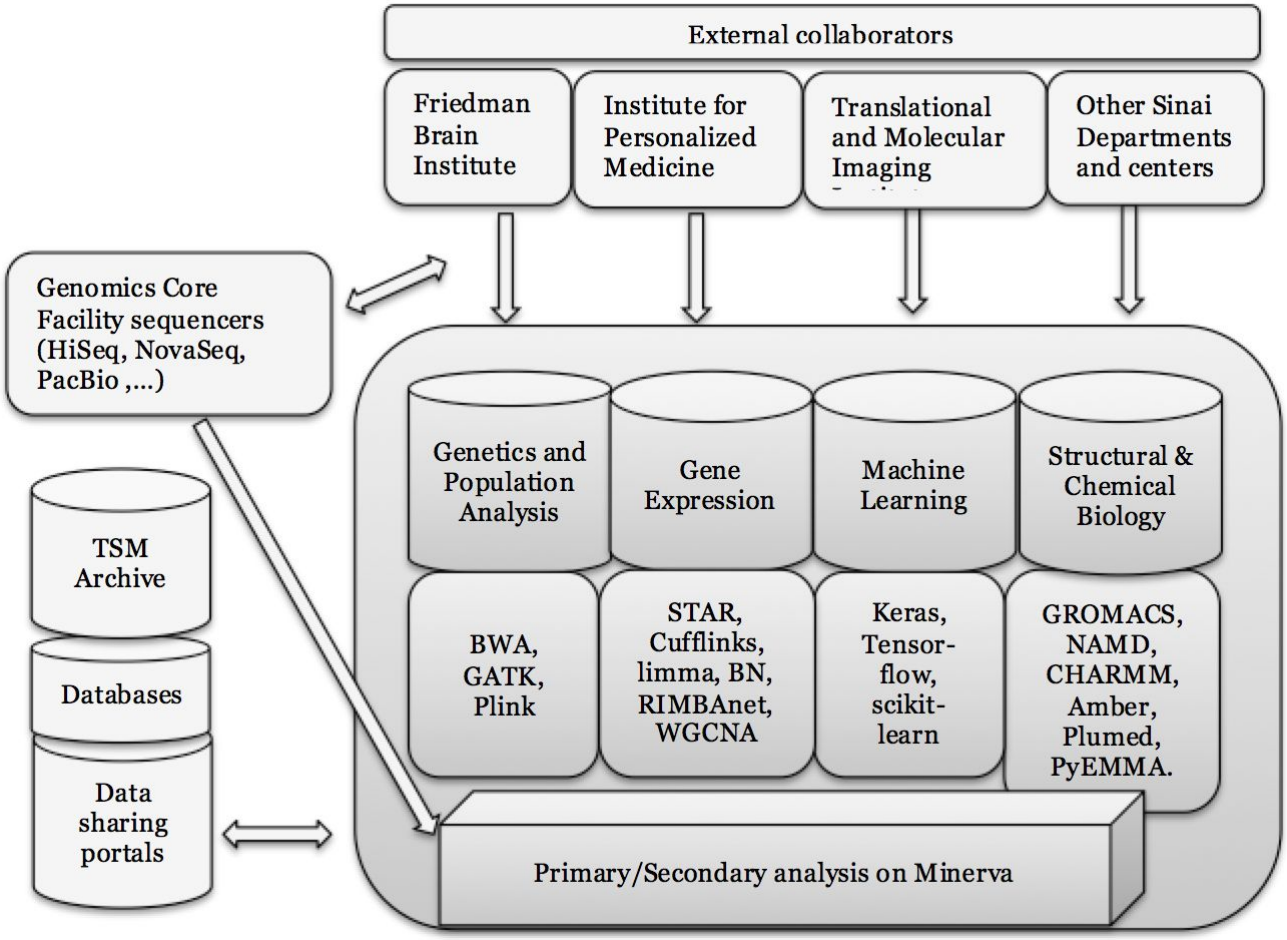
Users and usage on biomedical applications over 2013 - 2019

	2013	2019
Total # of users	339	2,484
# of external users	62	550
# of external institutions	27	75
# of projects	11	312
# of tickets	287	3,454

Biomedical Field	% usage 2013	% usage 2019
Genetics and Genomic Sciences	65%	69%
Structural and Chemical Biology	32%	10%
Machine Learning	0%	10%
Other	3%	11%

Computational Biomedical Research on Minerva

Research(Data) Workflow and Domain Applications

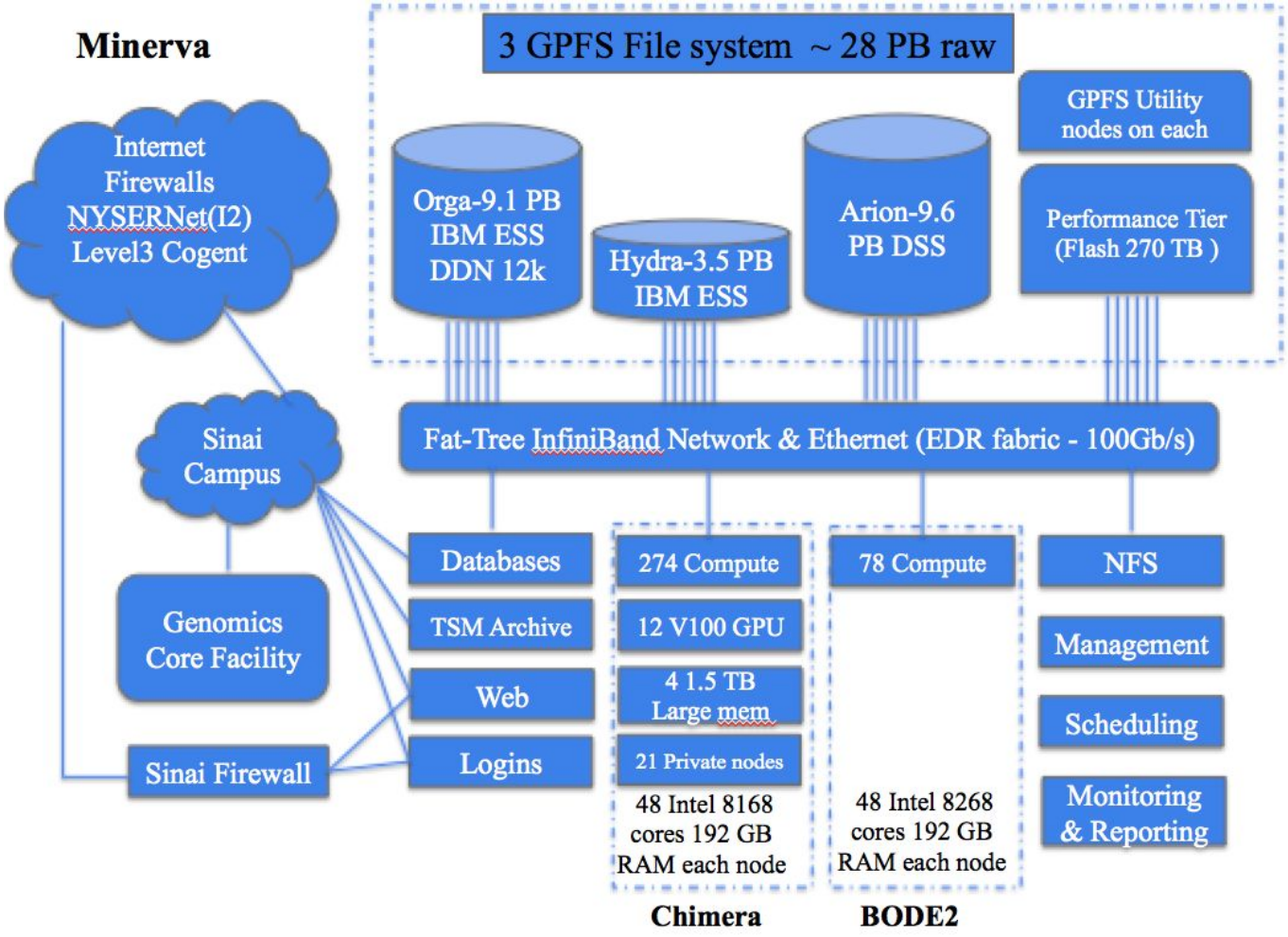


- Major data flow: genomics sequencing systems, archiving on TSM
- Preliminary analysis can be very compute/storage intensive, while final biomedical interpretation can involve simulations/network analyses of high complexity

Mount Sinai Minerva HPC System Design

HPC System Design

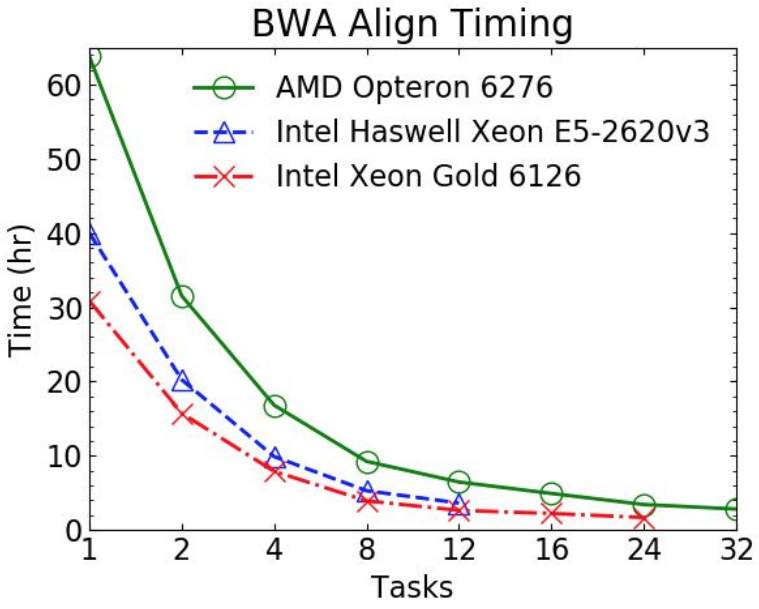
- Maximally improving the throughput of our workload through optimal compute core selection and streamlined queue policies
- Maximally increasing both the performance and size of file system

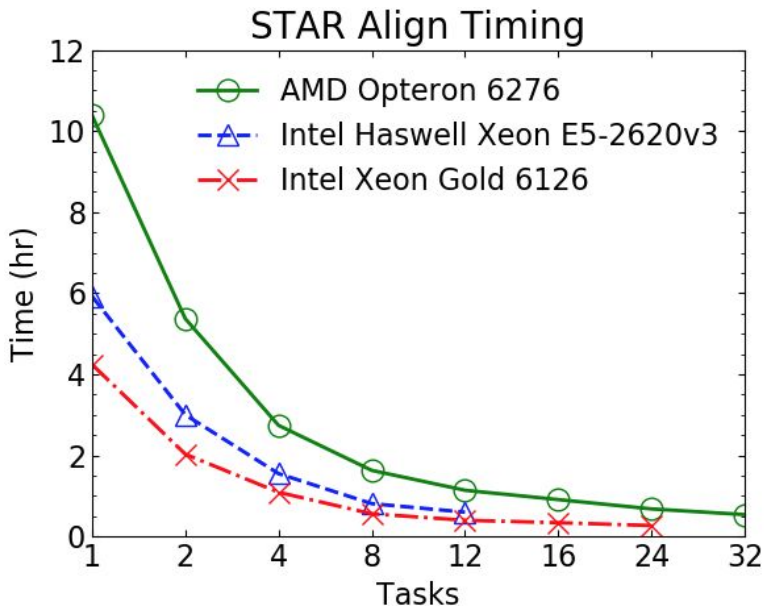


HPC System Design - Compute /CPU Architecture

Benchmark (at 2018) with major aligners shows moderately larger intel CPU cores per node gives better performance and throughput

- Aligning is the most computationally intense portion of the genomics workflow
- Both aligners exhibit good linear scaling with the number of threads; only starts to lose some performance for larger processors
    - Gold node offers (32 ~ 48 cores) > 2x speedup over whole Haswell node





12 nodes with 4X V100 GPUs added to accelerate some of the structural and chemical biology and machine learning studies

HPC System Design - Memory

Larger memory per node needed for higher throughput and larger job capability

- Most genomic applications are memory intensive and are growing
- The average per job and per core memory usage ***more than doubled***
- Large shared memory node needed to extend the job capability

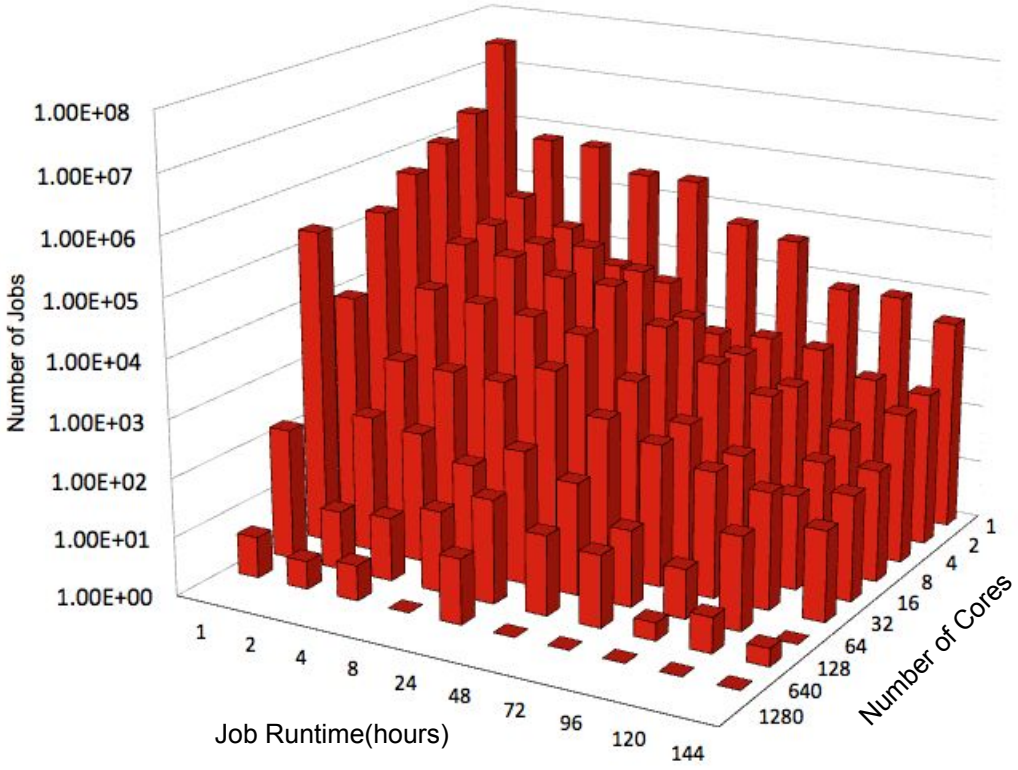
	Minerva Usage 2013	**Minerva Usage 2019**
Max memory used per core	787 GB	960 GB
Min memory used per core	0.005 MB	0.004 MB
Average memory used per core	885 MB	2.95 GB
Median memory used per core	320 MB	470 MB
Max memory used per job	5.6 TB	2.1 TB
Min memory used per job	0.01 MB	1.0 MB
Average memory used per job	1.8 GB	3.8 GB
Median memory used per job	651 MB	675 MB

HPC System Design - Scheduling and Queues

LSF job scheduler handles large number of pending jobs O(million)

Job mix shows the majority are short single or low-core count jobs

- Jobs are scheduled on a per core basis rather than by node
- Queues with different priorities, time limits and dedicated node pools



	Minerva Usage 2013	Minerva Usage 2019
Total core hours used	18,075,749	106,871,360
Number of jobs run	2,560,896	52,190,640
Max core hours used per job	22,161	59,511
Average core hours used per job	4.2	2.05
Max job run time	205 h	336 h
Avg job run time	1.6 h	1.42 h
Median job run time	5 min	5 min

HPC System Design - File System GPFS

Metadata access: flash employed to hold the metadata and small files due to its low latency and high input/output Operations Per second

- The sheer number of files with tiny files as the majority from genomic analyses
  - Job runtime and interactive command line directory listings are slow
  - Storage space is wasted
  - High latency responses

Storage usage has grown at the rate of > 1PB per year

	Minerva Usage 2013	Minerva Usage 2019
Number of files	54,026,071	2,351,357,662
Total storage used	0.7 PB	8.1 PB
Average file size	15 MB	3.7 MB
Median file size	29 B	774 B
Max file size	1.3 TB	14.3 TB
Min file size	0	0
80% of files smaller than	10 KB	0.3 MB
Number of zero-length files	1,731,148	52,394,913
Amount of archival storage used	1 PB	19 PB

Interconnect between nodes/storage

Infiniband HDR(100 Gb/s)

Data migration between file systems during upgrade

Active File Management (AFM) migrated the files from the DDN and ESS BE pools to the DSS pool

HPC System Design - Archival Storage

TSM encrypts and saves copies of data on tape to two geographically disparate locations for disaster recovery (6 years)

Current archive storage usage	
Archived data	9.83 PB
Total data with offsite copy	19.66 PB
Number of tapes used	13,710

Statistics of 2019			
Amount of archived data	1.7 PB	Amount of retrieved data	510 TB
# of archive operations	19,709	# of retrieve operations	1,294
# of archive users	86	# of retrieve users	54

HPC System Design - Cloud Technology

Singularity container is supported since 2019

- Enabled users to provision the infrastructure and pull desired containers without any requirements for administrative and privileged access
- About 2% of our users are now regularly computing with Singularity

Apache Spark-based technologies via GPFS's Hadoop is also supported

.

HPC System Design - Cloud Cost

Cloud (such as AWS or Azure)

More expensive and less efficient for the scale of our science

- Cloud compute costs are for hyperthreads or virtual cores, not physical cores so compute performance is variable
- Cloud still needs computational scientists and HPC admins to help build the environment and provide support
- Costs for Amazon data transfer and read/write costs are additional and not included. Azure and other cloud services have similar pricing

.

HPC System Design - Sustainability

Minerva HPC have facilitated over 900 publications since 2012

- Twice a year we collect publications utilizing Minerva from the research community

Year	2012	2013	2014	2015	2016	2017	2018	2019	**Total**
# of pubs	54	59	64	121	151	171	152	176	**948**

Chargeback on storage

- Rate is re-calculated yearly, *i.e.*, $109 per TB per year in 2019, including access to compute cycles, GPUs, large shared memory nodes, and archival storage
- Bill on actual usage
- Free < 1TB usage for new faculty and unfunded experimentation

Minerva HPC Futures

New Services/Resources on Minerva in 2020

HIPAA Compliant Minerva

- Will facilitate new applications such as feature extraction from identified radiological images and machine learning on raw Electronic Health Record (EHR) data
- We are in the final stages of getting approval

Rstudio-Connect

- RStudio Connect is a publishing platform for your work created in R and Python
- Share ***Shiny applications, R Markdown reports, dashboards, plots, Jupyter Notebooks***, etc in one convenient place
- Publish with push-button from the RStudio IDE, with flexible security policies
- Will also set up open source Shiny Server for hosting only shiny applications

Visualization Portal

- Set up **Open OnDemand** as better visualization portal to access Minerva through modern web browsers

More compute nodes with higher memory

Thank you!