Minerva HPC and Data Ark Town Hall
Fall 2024

Lili Gai, PhD
Yiyuan Liu, PhD
The Minerva HPC Team
Nov. 22nd, 2024

Icahn
School of
Medicine at
Mount
Sinai

Outline

- Accomplishments & Updates
- 2024-2025 Roadmap
- Appendix I - Minerva HPC & Data Ark Usage

Accomplishments & Updates

Welcome Our New HPC/AI Experts

Senior AI/GPU Computational Scientist Shamimul Hasan Joined May 2024

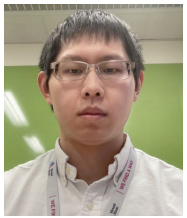
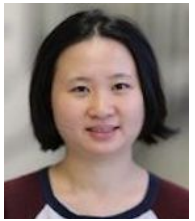
- Research Scientist in Artificial Intelligence for Health, Oak Ridge National Lab
- Graph Analytics, Big Data Management & Analytics, Artificial Intelligence, Deep Learning, Machine Learning, Natural Language Processing, Information Retrieval

HPC Architect Tejas Rao Joined June 2024

- Came from Brookhaven National Lab and UnitedHealth Group
- Over 15 years of experience with High performance computing and extensive experience with operating system internals, continuous integration/automation, enterprise storage and cybersecurity

Staff Summary

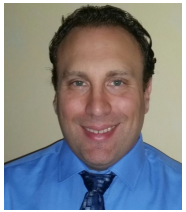
The HPC team consists of **four** computational scientists/bioinformaticians


- Hyung Min Cho, PhD
- Jielin Yu, PhD
- Yiyuan Liu, PhD (part time on Data Ark)
- Shamimul Hasan, PhD, joined this May

...and only "**three**" HPC architects/admins





- Wei Guo, PhD
- Tejas Rao joined this June
- Kali Mclennan left Mount Sinai on Sep. 6th
- Eric Rosenberg (part time on Minerva TSM archival)

Open positions:

- *Sr. HPC admin: starting Dec. 9th*
- Lead HPC Architect - Cybersecurity and Cloud
- *Another Sr. HPC admin*
- *Sr. system admin*
- Associate director

of HPC Tickets in 2024 increased largely

of HPC Tickets opened is increasing largely over years.

Year	2024	2023	2022	2021
# of tickets	5,040	3,754	3,240	2,915

2024 Monthly: # of HPC Tickets opened

Month	Jan	Feb	Mar	Apr	May	Jun	July	Aug	Sep	Oct	Nov(1-20)	Total
# of tickets	564	483	451	413	348	299	536	429	526	514	477	5,040

We will look into the type of tickets open and develop some guideline to reduce the # of basic tickets.

Minerva Refreshment and Expansion

IN PRODUCTION on November 6th 2024

Hardware Purchased:

- **146 high memory** compute nodes with 14,016 cores in total: **~1 PFLOPS**
  - ntel latest 5th Xeon(R) Emerald Rapids 8568Y+ 48C, 2.3GHz
  - 1.5TB DDR5 memory
- **210 GPUs in total**
  - **188 H100**-80GB (SXM5) NVlinked GPUs in 47 nodes
    - **7.9 PFLOPS in GPU (147th place on the Nov 2024 World Top500* list);**
    - **Green500 is 46.2 GFlops/Watt (36th place on the Nov 2024 Green500 list);**
  - 32 L40S GPUs in 4 nodes: 8xL40S per node
- **NDR400 IB networking (400Gb/s)**
  - Integrate all the nodes into one unified network, with quadrupled speed from Enhanced Data Rate (EDR) 100Gbps to Next Data Rate (NDR) 400Gbps
- **Others:**
  - 3.84 TB Local NVME SSD per node for fast local cache
  - 4 login nodes and 6 service node
  - 300 TB memory in total
  - Direct water-cooling solution

The **TOP500 project ranks and details the 500 most powerful computer systems in the world.*

Minerva Refreshment and Expansion: Timeline

IN PRODUCTION on November 6th 2024

The Refreshment Processes in a Nutshell:

- ✔ RFQ issued for bidding on **Feb. 23**
- ✔ Quotes from all the vendors received on **Mar. 11**
- ✔ PO submitted on **March 28th**
- ✔ Decommission of BODE2 nodes/old racks/private nodes and rerack for space by **July 17th**
- ✔ Construction work in Hess done by **Sep 4th**
- ✔ HPC equipment arrived by **Sep 12th**
- ✔ Installation and acceptance test by **Nov. 3rd**
- ● Network cutover and integration by **Nov. 6th**

Minerva full preventative maintenance (PM) is scheduled for MAJOR Minerva infrastructure upgrades from **8AM Tuesday, Nov. 5, 2024 to 5:00PM Thursday Nov.7, 2024.**

Minerva Operating System OS Upgrade

Thank You!

Thank you for help testing the new Minerva env and provide feedback to us!!!

Procedural Actions:

1.We rebuilt major modules for the new OS Rocky9.4 for better performance/stability, and keep rebuilding as needed ✔

2. We have set up some test nodes with the new OS Rocky9.4 and modules for you to run tests ✔

3. We invited users to run test on their pipeline with the new OS Rocky9.4 in Sep. ✔

4. We completed a rolling upgrade of the current compute nodes by Oct.8th to avoid a cluster-wide Preventive Maintenance (PM), and allow users to do more testing with the new OS and modules. ✔

5. We will formally roll out the upgrade on all compute nodes including newly purchased water-cooled CPU, GPU servers and service nodes on Nov. 5th ✔

Minerva is at Rocky9.4 with new software stack by Nov. 6 2024

Status

We are upgrading the operating system, networking stack and parallel file system software for the current Minerva compute nodes for the later integration with the new nodes. The upgrade includes the base image from Centos7.9 to Rocky9.4, high-speed network software stack (to OpenFabrics Enterprise Distribution (OFED) 24.04) and the Spectrum Scale version (to 5.2.0.1).

Why upgrade?

This is necessary to apply security patches, improve stability and benefit from the better features, provide latest system for new software, and support of the new Spectrum Scale version. CentOS 7 reached its end of life (EOL) and is no longer receiving software updates.

LSF Job Scheduler Upgraded on May 10 2024

Minerva preventative maintenance (PM) for LSF (Load Sharing Facility) job scheduler is scheduled for 6 hours, Friday, **May 10, 2024, from 1:00-7:00 PM.**

What is the plan?

During this PM, we will perform important version upgrades and patching for LSF job scheduler. All the LSF daemon will be restarted on all the compute nodes.

Why upgrade?

1. This upgrade is necessary to better support user application needs and newer versions of NVIDIA driver
2. This is also a preparation for the coming upgrade of the server OS (operating system)

How this PM may impact your work:

During the maintenance:

1. All queues will be inactivated. You cannot submit new jobs or query jobs from anywhere
2. Pending jobs will not be dispatched
3. All the running jobs on compute nodes will **NOT** be affected
4. File access will **NOT** be affected
5. Logins to Minerva and others will still be available

Service Upgraded and Being Migrated to New Hardware

We migrated and upgraded service to new hardware:

1. 2x LSF management nodes
2. 3x Login nodes and data transfer node
3. New NFS server for user home directory
4. LDAP server for user account and group management
5. New subnet manager for network
6. MariaDB database server
7. Posit connect server
8. Xcat server for cluster management and mail, proxy
9. Deploy the next generation high speed (NDR) InfiniBand network to enable faster storage and compute communications
10. Ongoing:
    a. shared web servers for user websites
    b. globus for data transfer

Decommission of Out-of-Warranty Minerva and BODE2 Nodes

[July 17th] We shut down and removed 90 Minerva compute nodes (a total of 4,320 compute cores) permanently for facility construction work to get water piping ready for installing the new direct water cooling (DWC) nodes

This is necessary to get the new equipment installed

One Cluster-Wide PM in last six months

System admins try to minimize the system-wide downtime

- One Cluster-Wide PM (plan for 3 days but complete within 24 hours)
- Some short windows on specific servers and TSM
- Perform changes that doesn't affect operation of cluster as much as possible before the PM
- Well-prepared worksheet by system admin before changes made on system

Unexpected system hang:

- Unexpected outage on Arion file system (7 hours) due to changes on subnet manager

$2M AI Mount Sinai (AIMS) NIH S10 Proposal to Add More GPUs

Thank you for sending us your science story and your publications!!!

Thank You!

Goal:

Affordable and available modern computational and data resources for AI-driven biomedical research projects are in short supply. To address this gap, we requested a new high-performance instrument with 6x DGX providing state-of-the-art GPU capability and capacity

DGX B200 can deliver 3X the training performance and 15X the inference performance of DGX H100

- B200 offer new capabilities such as floating point 4 bits (FP4)

Status:

Impact Score: 11!!! with almost none weakness noted from the summary statement

- Historical impact score for previous three NIH awards (the lower the better):
  - BODE: 17
  - BODE2: 18
  - CATS: 20

$2M AI Mount Sinai (AIMS) NIH S10 Proposal to Add More GPUs

- **The final AIMS machine will consist of 6x DGX nodes if awarded with a total of**
  - 48 NVIDIA B200 GPUs connected via NVLink and Infiniband NDR400
  - 672 Intel Xeon Platinum 8570
  - 9 TB of memory available on B200, and an additional 12TB available on the server nodes
  - Local high-speed NVME storage and DDR5 RAM enables caching of intermediate results

AIMS total system configuration	
Performance (petaFLOPS) [NVIDIA]	1.92 FP64 3.84 FP32 216 FP16 432 FP8 training 864 FP4 inference
# of GPUs	48 NVIDIA B200
GPU memory size (TB) and type	9 HBM3e
GPU memory bandwidth (TB/s)	384
NVLink bandwidth (TB/s)	87
# of CPU cores	672
System memory (TB)	12
# of nodes & type	6 Lenovo ThinkSystem SR780a V3 8-way GPU

TSM Archival Storage Infrastructure Upgraded

There is a scheduled maintenance on Minerva Tivoli Storage Management (TSM) system on **Thursday, June 27th**, from 1:00 AM to 11:00 PM. The TSM system will be unavailable during this timeframe.

What is the plan?
During this PM we will improve our TSM infrastructure with the following:

1. Install a new TSM server
2. Install two new fibre channel network switches and add more inter-links between the TSM server and the tape libraries
3. Update the TSM server and client software to version 8.1.21

Why upgrade?
After the PM, the Minerva TSM environment will be more stable with higher performance.

- Improved reliability with two isolated fibre channel fabrics for high availability
- **Up to 3x increased speed of data movement between the server and the tape drives due to added fibre channel links**
- Upgraded software with bug fixes as well as improvements to security and performance

New TSM Policy in Effect Starting Aug. 29th

Status: We have updated our Minerva Tivoli Storage Manager (TSM) archival policy via email notice and on our website.

This policy applies to all users that are using Minerva TSM resources. The primary purpose of this policy is to ensure that all users receive consistent performance to archive and/or retrieve files.

We have also updated the TSM archival command dsmc**.** Your sessions may be failed with error messages when you try to archive small files (< 1GB). Please review the Minerva TSM Archival Storage Policy carefully before you perform TSM activities to avoid failures.

The following policy is now in effect:

- 1 GB is the minimum file size to be transferred to TSM. Files <1 GB must be aggregated with tar/zip, resulting in a file size of 1 GB or larger.
- The maximum amount of data archived or retrieved from TSM is limited to 40 terabytes per week.
- TSM is for long-term file storage only, with each file stored once.
- Data archived in TSM has a retention time of six years. We will notify you by email three months in advance before your file will be deleted from TSM.

Reminder on Minerva Data Backup Policy

We are sending emails to reminder users about data backup policy on Minerva

This is a routine reminder on Minerva data backup policy.

1. **We do not backup any user files on any nodes including the private nodes. Please archive/backup your important files by yourselves.**
   a. We have included this in the message of the day (MOTD) after you log into Minerva, User Responsibilities and Acceptable Usage Policy, and Annual HIPAA Policy
   b. We will keep reminding you every quarter via email
   c. Please archive/backup your files following the guide at https://labs.icahn.mssm.edu/minervalab/documentation/access-tsm-with-command-line/

2. **Please don't set the permission of your Minerva files as rwx (read, write and execute) for everyone/others.**
   a. This can result in file deletion by others. Please double check your file permission on Minerva especially for your project directory
   b. Limiting file permissions is the user's responsibility according to the annual HIPAA compliance requirement/training

Expand Open OnDemand with more User-Friendly Features

Goal: Provide easy graphical access to Minerva without Linux command needed

Status: 372 users (~32% of active users) are using it since its launch on Aug 2 2023!!

- We are supporting more apps such as Chimera Desktop, GUI (Matlab, MarketScan, SAS, Stata, Sleap), servers(VS code, Rstudio, Jupyter)

This product offers a fully-compliant job management and desktop portal requiring minimal knowledge of Linux high-performance computing (HPC) environments with no end-user installation requirements other than an up-to-date web browser (Chrome or Firefox recommended)

The service portal is accessed at URL: https://ondemand.hpc.mssm.edu
Documentation is available at https://hpc.mssm.edu Documentation>Open OnDemand or here

All RIF (Research Identifiable Files) CMS Data on Minerva

Goal:

- Centralize the CMS data within a secured and compliant Minerva Ecosystem

Status: ALL RIF CMS Data transferred to Minerva by Sep. 4th 2024

- New encrypted storage server for CMS data in production in Jan. 2024!
  - Cost: $119 per TB per year
  - Current capacity: total storage 100TB with 30TB used
- Minerva is the sole CMS computing environment with an organizational-level plan DMP SAQ for RIF, covering all studies using the approved computing environment
- Coordinated with several groups running CMS RIF data and all transferred to Minerva **by Sep. 4th 2024**

Minerva Training Offered Fall 2024

Eight training sessions in person/Zoom this Spring with more info here

- Four additional training sessions on GPU/AI to lead the AI initiative at Mount Sinai
- We provided training material (including slides & recording) online
- We sent calendar invite to hpcusers email list

Presented Two Minerva sessions and supported BSR2402 graduate course “transcriptomics and epigenomics module” in Sep. (~20 students)

~500 attendees in total

Session	Topic	Date	# of Attendees
1	Minerva Intro	Oct. 2	80
2	Data Ark Introduction to Data Ark	Oct. 4	23
3	LSF Job Scheduler	Oct. 9	26
4	Intro to GPU /AI resources on Minerva	Oct. 16	104
5	Accelerating Biomedical Data Science with GPUs: Practical Approaches And Tools	Oct. 23	60
6	Leveraging Large Language Models in Biomedical Research	Oct. 30	106
7	Access Minerva via web browser Open OnDemand	Nov. 1	41
8	How to Accelerate Genome Analysis Toolkit (GATK) by using Parabricks	Nov. 6	28

Data Ark Data Commons Datasets

There are 18 datasets hosted under Data Ark currently

Access within 24 hours after DUA signed

Public Data Sets

- GTEx
- GWAS Summary Stats
- gnomAD
- eQTLGen
- BLAST
- Reference Genome
- Genebass
- 1,000 Genomes Project
- UKBB-LD
- Partial of the Cancer Genome Atlas (TCGA)
- LDSCORE

Mount Sinai Generated Data

- The CBIPM-BioMe Data Set
- MSDW De-identified OMOP Data set
- MSDW COVID-19 EHR Data Set
- Mount Sinai COVID-19 Biobank
- The Living Brain Project
- STOP COVID NYC Cohort

Restricted Access

User Group-Acquired Data Sets

- MarketScan®

Data Ark webpage:
https://labs.icahn.mssm.edu/minervalab/resources/data-ark/

Dataset Updates

- September 3, 2024 - **CBIPM-BioMe** de-identified data available through Data Ark
  - To sign the DUA (data use agreement) for this dataset, applicants' direct Sinai affiliation is verified in the background with the HR database, ensuring compliance with the DUA
  - No IRB is required for data access
  - Data access is granted within 24 hours of DUA submission
- October 1, 2024 - **MSDW** (Mount Sinai Data Warehouse) **de-identified OMOP** (Observational Medical Outcomes Partnership) data was re-open to user for access
  - Data is refreshed
- November 5, 2024 - **MarketScan** (proprietary de-identified) data (time scope: 2013-2022) access license was renewed under the new leadership of the user group – Dr. Inga Peter
- December 11, 2024 - UKB genotype data will be offboarded, with PIs and associated members notified. Final offboarding contingent on application PIs' confirmation.

Datasets in Our Pipeline to Be Released Soon

To Be Announced: De-identified Digital Pathology Slides Available Through Data Ark

- Digital pathology slides for over 190,000 individuals treated in the Mount Sinai Health System since 2002
- Nearly every organ system represented including lung, heart, pancreas, kidney, liver, genitourinary, gastrointestinal, hematologic, neuropathologic, etc.
- Slides represent a wide array of pathologic processes including neoplastic, developmental, Inflammatory, toxic, metabolic, genetic, degenerative, traumatic and infectious
- Staining techniques include hematoxylin and eosin (H&E), specialized stains (ex. silver, trichome) and immunohistochemistry

	Count	Anticipated Annual Growth
De-identified Digital Pathology Whole Slide Images (#)	~1.5 million	~1.5 million
Distinct Patients (#) **Female (%)** **Hispanic (%)**	~191,000 63% 19%	
Size of Digital Pathology Images Data Set	~1.3 PB	~1-1.5 PB

To Be Announced: De-identified Digital Pathology Slides Available Through Data Ark

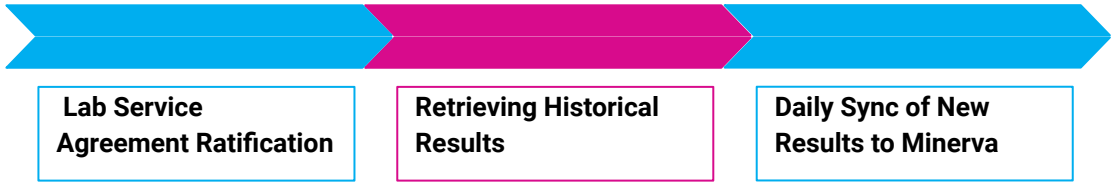
- Previous Accomplishment:
  a. Cohort building functionality had been made available in Leaf
- Recent Accomplishments:
  a. DUA and the webpage have been set up
  b. Images have been linked to patients' EHR data (available at MSDW)
- Access information and data set details available on the Data Ark website

https://labs.icahn.mssm.edu/minervalab/resources/data-ark/digital-pathology-slides/

Genomic Data on Somatic Testing Generated by External Vendors

- **Project objectives**
  - To link the phenotypic and somatic genomic data on Data Ark, facilitating the use of somatic genomic data for analytics, clinical research and clinical operations
  - To make raw and structured somatic genomic results from external vendors available to the Mount Sinai research and clinical community

Vendor	Status
Foundation Medicine	Stage 3: Daily Sync of New Results to Minerva
Caris	Stage 2 &3: Retrieval of Historical Results by Dec 2024; Daily Sync of New Results to Minerva
NeoGenomics	Stage 1: Lab Service Agreement Ratification



Data Transferred from Foundation Medicine to Minerva

File Type	# of Orders Resulted
BAM	8,161
VCF	7,174
PDF	8,161
JSON	3,090
XML	8,161

- About the data from Foundation Medicine:
  - Includes results of orders for a Foundation Medicine test placed at a Mount Sinai facility
  - All historic and current Foundation Medicine results
  - Structured and unstructured results
  - Updated daily with new results

Workflow to Access Somatic Genomic Results

- In Leaf, users can build and save cohorts of patients that have Foundation Medicine test results returned to Data Ark
- Submit a MSDW data request to access the structured results
  - Approved IRB protocol required
- For access to the BAM and/or FASTQ files, the MSDW team will provide a list of the required files to the Data Ark team
  - The Data Ark team will provide the requestor with links to the requested files on Data Ark

What’s Next?

2024-2025 Roadmap

Q4 2024

- ➔ **Continue on module software recompile as needed**
- ➔ **Continue migrating the rest service to new OS and hardware**
- ➔ **Deploy new GPFS storage for expansion and prepare for decommission of old hardware**
  - ◆ We purchased 11PB of raw DSSG storage with 180TB SSD
- ➔ **Set up Checkpoint/Restore In Userspace (CRIU) for job checkpointing**

Q1 2025

- ➔ **Migrate Minerva two factor authentication to Azure MFA**
  - ◆ We worked with DTP and managed to get continuous support on Symantec VIP applications till Jan 2025
- ➔ **Annual Form including HIPAA Form, NIH acknowledgement form, allocation from**
- ➔ **Annual User Survey**
- ➔ **Improve Jupyter notebook and rstudio via ondemand**
- ➔ **Improve audit logging and encryption for security compliance**
- ➔ **Reduce the volume of basic support tickets and improve operational efficiency**

Acknowledgements

- Supported by the Clinical and Translational Science Awards (CTSA) grant UL1TR004419 from the National Center for Advancing Translational Sciences, National Institutes of Health.

CTSA Clinical & Translational Science Awards ®

Thank you!

Appendix I
Minerva & Data Ark Usage
(Apr. 2024 - Sep. 2024)

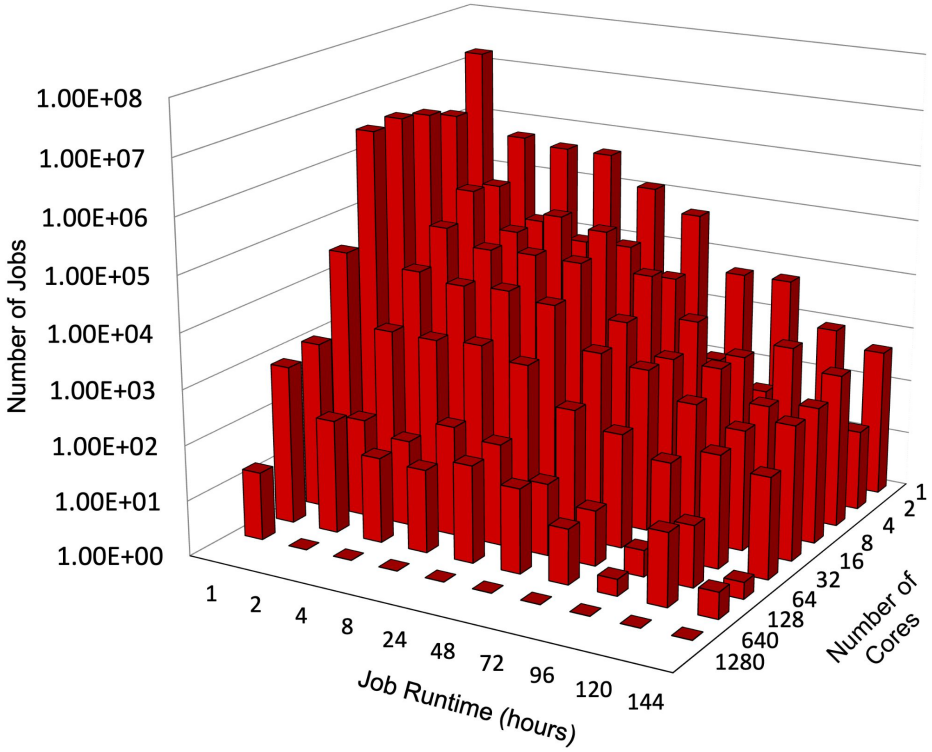
Minerva usage summary (April 2024- Sep. 2024)

Accounts	
Number of active users	1,002
Number of total users	4,447
Number of project groups	568 (428 active)
Storage	
High-speed storage used (Arion)	19.5 PB (62% utilization)
Archival storage used	19.7 PB
Compute	
Number of jobs run	29,225,632
Core-hours utilized	82,137,717 hrs
System	
Number of maintenance sessions	No preventative maintenance (99.6% uptime)

Jobs and compute core hours by partition

Compute	# Jobs	CPU-hours	Utilization
Chimera	18,811,832	48,346,715	83.7 %
BODE2	2,835,202	9,630,785	59.3 %
Hi-memory nodes	3,118,513	8,078,571	100 %
CATS	4,187,352	14,777,773	89.9 %
GPU nodes	272,733	2,425,516	53.6 %
Total:	**29,225,632**	**82,137,717**	**80.6 %**

Job mix


1.00E+08
1.00E+07
1.00E+06
1.00E+05
1.00E+04
1.00E+03
1.00E+02
1.00E+01
1.00E+00
Number of Jobs
1
2
4
8
24
48
72
96
120
144
Job Runtime (hours)
1
2
4
8
16
32
64
128
640
1280
Number of Cores

Top 10 users compute core hours

PI	Department	# Core-hours	# Jobs
Roussos, Panos	Psychiatry	9,670,087	3,819,119
Raj, Towfique	Genetics and Genomic Sciences	8,661,663	2,375,429
Buxbaum, Joseph	Genetics and Genomic Sciences	8,082,504	1,895,175
Sharp, Andrew	Genetics and Genomic Sciences	7,193,188	2,005,977
Pejaver, Vikas	Institute for Genomic Health	5,963,826	935,353
Zhang, Bin	Genetics and Genomic Sciences	5,930,515	131,874
Kenny, Eimear	Genetics and Genomic Sciences	4,161,654	6,760,431
Bunyavanich, Supinda	Pediatrics	4,091,129	82,180
Charney, Alexander	Genetics and Genomic Sciences	3,638,568	3,441,517
Schlessinger, Avner	Pharmacology	2,491,244	282,588

Top 10 PIs GPFS high speed storage

PI	Department	Storage usage
Zhang, Bin	Genetics and Genomic Sciences	1.5 PiB
Roussos, Panagiotis	Psychiatry	1.3 PiB
Charney, Alexander	Genetics and Genomic Sciences	1.2 PiB
Raj, Towfique	Neuroscience	1.2 PiB
Sebra, Robert	Genetics and Genomic Sciences	772 TiB
Sealfon, Stuart	Neurology	650 TiB
Goate, Alison	Genetics and Genomic Sciences	600 TiB
Nadkarni, Girish|Charney, Alexander	Genetics and Genomic Sciences	599TiB
Buxbaum, Joseph	Psychiatry	584 TiB
Kenny, Eimear	Institute for Genomic Health	451 TiB

Top compute and storage usage department/institute

Department/Institute	Compute Core Hours
Genetics and Genomic Sciences	32,398,961
Psychiatry	18,392,254
Institute for Genomic Health	10,153,580
Neurosciences	9,234,106
Pharmacology	2,511,968
Oncological Sciences	2,272,711
Neurology	1,799,114
Medicine	1,774,569
Structural and Chemical Biology	1,604,732
Mindich Child Health and Development Institute	649,154

Department/Institute	Storage (TiB)
Genetics and Genomic Sciences	7,188
Psychiatry	2,354
Neurosciences	1,447
Oncological Sciences	1,195
Institute for Genomic Health	907
Neurology	740
Medicine	603
Microbiology	324
Structural and Chemical Biology	263
AI and Human Health	177

Top 10 PIs - GPU usage hours

PI	Department	GPU hours	# Jobs
Raj, Towfique	Neurosciences	619,066	114,545
Schlessinger, Avner	Pharmacology	250,814	1,356
Nadkarni, Girish	Medicine	201,484	2,885
Roussos, Panos	Psychiatry	125,320	2,097
Fuchs, Thomas	AI and Human Health	123,879	6,195
Sumowski, James	Neurology	119,304	196
Davies, Terry	Medicine	88,258	83
Shi, Yi	Pharmacological Sciences	86,630	85,125
Shen, Li	Neuroscience	83,917	1,482
Osman, Roman	Structural and Chemical Biology	70,946	950

Total TSM archival storage usage (Apr 2024- Sep 2024)

Current archive storage usage	
Archived data	19.7 PB (LTO5: 4.1 PB, LTO9: 15.6 PB)
Number of tapes used	12,904 (10,881 LTO5 + 2,023 LTO9)

Statistics of Apr 2024 - Sep 2024			
Amount of archived data	1,400 TB	Amount of retrieved data	335 TB
# of users who have issued archive commands	57	# of users who have issued retrieve operations	41

Minerva publications > 1,700 since 2012!!

We collect publications twice a year (Jan & June). Thank you!!!

We sent email to PIs and delegate

Year	# pubs
2012	55
2013	59
2014	62
2015	115
2016	149
2017	165
2018	133
2019	178
2020	146
2021	234
2022	174
2023	219
2024	31

Kovatch P, Gai L, Cho H, Fluder E, Jiang D, Optimizing High-Performance Computing Systems for Biomedical Workloads, The 19th International Workshop on High Performance Computational Biology (HiCOMB), IPDPS, IEEE International Parallel and Distributed Processing Symposium, **May 2020**.

Kovatch P, Costa A, Giles Z, Fluder E, Cho H, and Mazurkova S, Big Omic Data Experience, SC'15: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, **November 2015**.

Data Ark Data Usage Summary for Frequently Used Datasets April 2024 - Oct 2024

of unique active users: **48**
# of support tickets: **153**

- *Usage/user metrics of CBIPM-BioMe have been collected since its release on Sep 3rd, 2024.

Dataset	Size (GB)	# of unique users	# of times data accessed
gnomAD	8,628	8	1,799,198
TCGA	154	7	820,936
UK Biobank LD	2,866	10	365,898
GWAS Summary Statistics	6,826	3	145,640
Reference Genome	142	5	37,675
CBIPM-BioMe*	4,608	9	22,148
LD Score Regression	173	2	6,512
UK Biobank	12,695	8	3,301
1000 Genome	143	3	2,170
GTEx	1,888	6	1,886

Data Ark Data Usage Summary for Infrequently Used Datasets April 2024 - Oct 2024

- *User access to MSDW OMOP data on Data Ark had been temporarily closed for compliance with the IRB approval process during Apr-Oct, 2024.

Dataset	Size (GB)	# of unique users	# of times data accessed
Blast	1,116	1	1,130
MSDW Covid	1	1	20
Genebass	903	1	4
MSDW Covid 19 Biobank	378	0	0
STOP Covid NYC	< 1	0	0
eQTLGen	39	0	0
Living Brain Project (LBP)	< 1	0	0
MSDW OMOP*	3,076	0	0

Researcher Engagement Highlights April 2024 - Oct 2024

Events	Date	Attendees	Delivered Service
Joint HPC and Data Ark Town Hall	04/16/24	20+ researchers	Updating on Minerva supercomputing and Data Ark services and roadmap ahead
CTSA Lunch and Learn	05/09/24	5 researchers	Showcasing the benefit of utilizing Data Ark and other Scientific Computing and Data services in research to new scholars
CTSA Translational Science Research Day	06/03/24	72 trainees, junior faculty, and invited speakers	Providing on-demand support and assistance with Data Ark inquiries
Data Ark Training	10/11/24	23 trainees	Training on Data Ark dataset access with a focus on MarketScan data
Digital Concierge	Weekly Wednesday	6 researchers	Providing responsive support for dataset access inquiries