Empire AI Town Hall

Patricia Kovatch, Dean for Scientific Computing and Data
Lili Gai, PhD, Director for High Performance Computing and Data
The Minerva HPC Team

Oct 10, 2025

Icahn School of Medicine at Mount Sinai

Outline

1. Introduction of Empire AI
2. Empire AI Hardware Resources
3. How to Access Empire AI
4. Next Steps
5. Discussion & Questions

About Empire AI

- Launched in April 2024 by Governor Kathy Hochul, Empire AI is a bold partnership of New York's leading public and private universities, establishing a **state-of-the-art artificial intelligence computing center, housed at SUNY's University at Buffalo**

- Provides members access to high-performance computing power for responsible AI research and development

- Positions New York as the national model in responsible AI innovation, with its leading research institutions pioneering safe, equitable, and accessible AI research and development that benefits the whole state

Good News! Mount Sinai Joined Empire AI

- **Thanks to Eric Nestler, Lisa Stump, Girish N. Nadkarni & Patricia Kovatch, we joined Empire AI**
  - Enable access to new large-scale GPU resources and collaborations within New York State!
  - Mount Sinai will leverage shared resources to accelerate AI research focused on clinical translation
  - Achieve economies of scale beyond a single university

- Empire AI is a consortium of **ten** New York State institutions with support from New York State and private philanthropy, that oversees a shared computing facility to promote responsible research and development

CONSORTIUM MEMBERS & PARTNERS

Empire AI Hardware Resources

Empire AI deploys large-scale GPU clusters in phases over 10 years funded by over $400 million. Empire AI GPU Hardware Information – Fall 2025

Alpha

- **144 H100** in 18 HGX nodes each with
  - 8 H100 80GB GPUs per node
  - 400Gb/s IB connection
  - 30TB NVMe caching space
  - 2TB of system memory

Beta (in production in Dec. 2025)

- NVIDIA DGX GB200 NVL72 SuperPOD with **288 B200 GPUs** for AI applications
- 60 nodes of NVIDIA Grace-Grace ARM Superchip for HPC and data processing
- Beta will be up to ~7x faster than Alpha on training and up to ~20x on inference

Storage

- 20+PB flash VAST as primary storage initially served via NFS
- 10+PB flash from DDN for high-performance data served via Lustre

How to Onboard Projects to Empire AI?

- **!!!Mount Sinai requirement: all PIs and users sign an annual DUA first!!!**
  - ➔ *I agree not to store and process any sensitive, HIPAA, controlled genomic, other controlled data or any data from any patient care process from MSHS (Mount Sinai Health System) such as MSHS clinical data including de-identified clinical data, EHR, genetics and transcriptomics etc on the Empire AI cluster*
  - ➔ There is no cybersecurity, encryption or regulatory framework in place currently
  - ➔ Efforts are underway to develop a HIPAA-compliant environment

- **Detailed Procedures**
  1. Mount Sinai PIs must initiate the process by submitting the Empire AI Data Use Agreement Form
  2. After DUA received, PIs and users will be contacted with instructions for onboarding via email
     a. PIs submit “Empire AI project onboarding Form” with projects details and users listed
     b. Listed users submit Empire AI Data Use Agreement Form & Empire AI user onboarding form
     c. The link to the forms will be provided via email. Will not be posted publicly.

How to Use Empire AI?

- **As this is Alpha phase, only GPUs/CPU and storage are available**
  - No any other extended Service (such as extensive software stack, Open Ondemand, globus, data archival, database etc)

- **Empire AI uses SLURM job scheduler**
  - You will need to learn SLURM Syntax for job submission

- **Quick Empire AI guide** at Minerva website: https://labs.icahn.mssm.edu/minervalab/documentation-new-york-states-empire-ai/

- **Empire AI general info page**: https://empireai.freshdesk.com/support/solutions

Minerva GPUs or Empire AI GPUs?

- ❑ **Empire AI GPUs will complement existing Minerva infrastructure with extra GPU resources for general data as needed**

- ❑ **Minerva remains the only cybersecurity- and HIPAA-approved GPU platform for Mount Sinai Health System (MSHS) data**

- ❑ **You can always use Minerva GPUs with all the extensive service as needed**
  - *Currently Minerva have larger # of H100 and GPUs than Empire AI* (installed in Nov. 2024)
    - See details here
  - Minerva will also install **48x B200** as announced which will be in production in **Dec. 2025**
    - *8x NVLinked B200 GPUs per node, 192 GB of memory per GPU, for a total of 48 xB200 GPUs and 9 TB of memory available on B200*
    - *112 Intel Xeon Platinum 8570 2.1GHz Cores, 2 TB memory, 25 TB high-speed NVME local storage per node for a total of 672 cores and 12 terabytes of memory on servers.*

Charging Model on Empire AI

- No charge for the current System till Dec. 2025.

- Cost recovery will be starting in Dec. 2025 (for both Alpha + Beta resources)
  - Empire AI suggested a uniform charge based on **Service Units (SU). The rate is TBD**

- Allocations will be made in *service units (SUs)* that can be expended upon any Empire AI computer resource. Each member allocation corresponds to 700K SUs per year. For example,
  - 1K SUs corresponds roughly to 1 month of H100 GPU time on Alpha, and might be an appropriate request for startup projects seeking to become familiar with use of AI/HPC or the Empire AI environment.
  - 20K SUs corresponds to over 2 H100 or 1 B200 GPU dedicated for an entire year (or 12 B200 GPUs for a month, or an entire NVL72 for nearly 6 days) and might be viewed as an "average-sized" allocation.

Questions and Other Resources

Questions?

1. If you have general questions about this Empire AI resources, please open a ticket at hpchelp@hpc.mssm.edu for now
2. All future ticket and communication will go via Empire AI ticketing system at support@empireai.edu

The Empire AI cluster is administrated and managed by Empire AI, not Minerva Team. Please note that the Empire AI team is operating with limited support capacity during this phase.

Additional Resources

- Empire AI – Overview
- Governor Hochul's Press Release on Empire AI Expansion
- Empire AI GPU Hardware Information – Fall 2025

Next Steps

Empire AI is collecting proposals to use Beta system that will be in operation by Dec. 2025

- The application form will be sent to PIs after PI signed DUA. The form will not be posted publicly.
- Project proposal submissions are open now to Oct 22, 2025 (midnight ET)
- Allocations made as a result of this solicitation will commence in Dec. 2025 and expire at the end of Nov. 2026

Minerva HPC will improve the data user agreement form with authentication

Define a charging model for cost recovery

Empire AI support is limited (all part time)

- Empire AI is hiring system admins

Icahn School of Medicine at Mount Sinai

Thank You

Discussion & Questions?