



LLamaMe (LLM) API and RAGged LLaMa

LC User Meeting
October 14, 2025

Otto Venezuela
Livermore Computing

Prepared by LLNL under Contract DE-AC52-07NA27344.



LLamaMe (LLM) API and RAGged LLaMa



LC User Meeting
October 14th, 2025

Otto Venezuela

LC HPC Cloud Enabler Officer

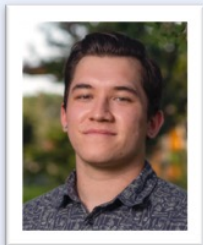


LC Workflow Enablement Group (“WE”-Group)



A deep bench of diverse talent to aid HPC with challenges of disruptive technology

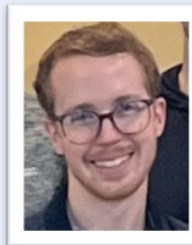
Conner Cook



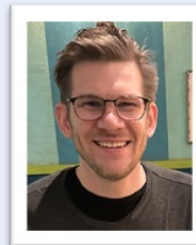
John Consolati
(Team Lead)



James Corbett



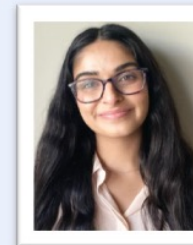
John Westlund



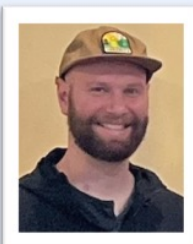
Todd Heer
(Group Lead)



Urwah Mir



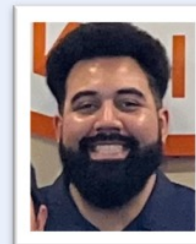
Zeke Morton



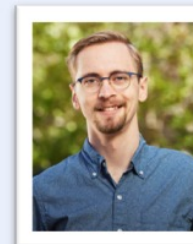
Neil O'Neill



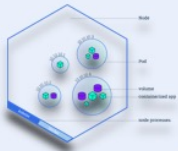
Jordan Dorham



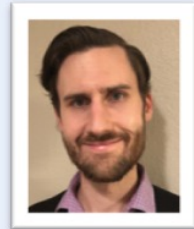
James Taliaferro



Otto Venezuela
(Team Lead)



Paul Bryant
(ORNL)



Thomas Mendoza
(contractor)



David Fox
(Project Lead)



Crucial Cross-Group Collaborators/Contractors

The nexus of cloud technologies, new workflows and traditional LC HPC ecosystems

LLamaMe (LLM) API Service



- LLamaMe API represents an advancement in providing users with access to locally-hosted open-source Large Language Model (LLM) via the REST API endpoint
- Providing this new LLamaMe API access allows users to integrate HPC jobs, integrated development environment (IDE), Retrieval-Augmented Generation (RAG) systems and LLM Agents.
- Currently deployed using Nvidia GPU nodes in CZ, RZ, and SCF Kubernetes clusters

Network	Model	Max Tokens	GPUs
CZ	Meta-Llama-3.3-70B-Instruct	64k	2 NVIDIA A100 80GB
RZ	Meta-Llama-3.1-8B-Instruct	4k	1 NVIDIA A100 40GB
SCF	Meta-Llama-3.3-70B-Instruct	110k	4 NVIDIA H100 80GB



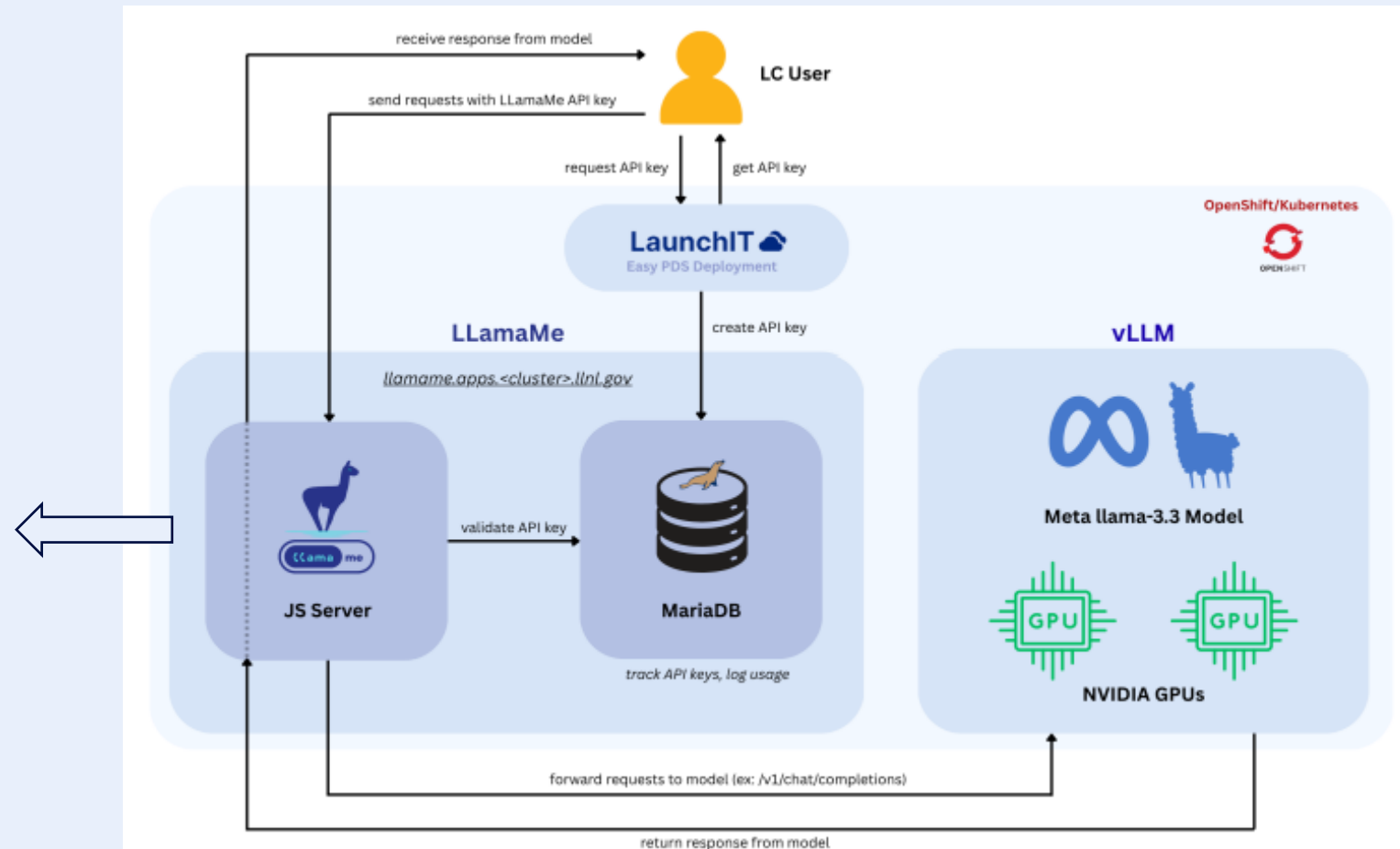
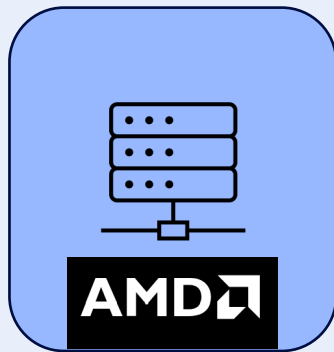
General Availability in all 3 LC network zones (CZ, RZ, and SCF) since April 2025

Increasing LLMs that LC LLamaMe will support; upcoming LaunchIT release v2.3.3

Network	Model	GPUs
CZ	Meta-Llama-3.3-70B-Instruct gpt-oss-120B	4 NVIDIA A100 80GB
RZ	Meta-Llama-3.1-8B-Instruct Codestral-22B gpt-oss-120B Llama-3.3-70B-Instruct Llama-4-Scout-17B-16E-Instruct	1 NVIDIA A100 40GB X AMD MI250 120GB 
SCF	Meta-Llama-3.3-70B-Instruct gpt-oss-120B <Bring your LLMs suggestions>	8 NVIDIA H100 80GB

Models are subject to change at any time
LLM performance and capabilities are improving at an exponential rate!

LLamaMe Architecture (RZ)



Expanding our LLMs support to use NVIDIA and **AMD** GPUs

Retrieval Augmented Generation (RAG)

LC RAGged LLaMa



Image generated via LivChat (DALL.E 3)



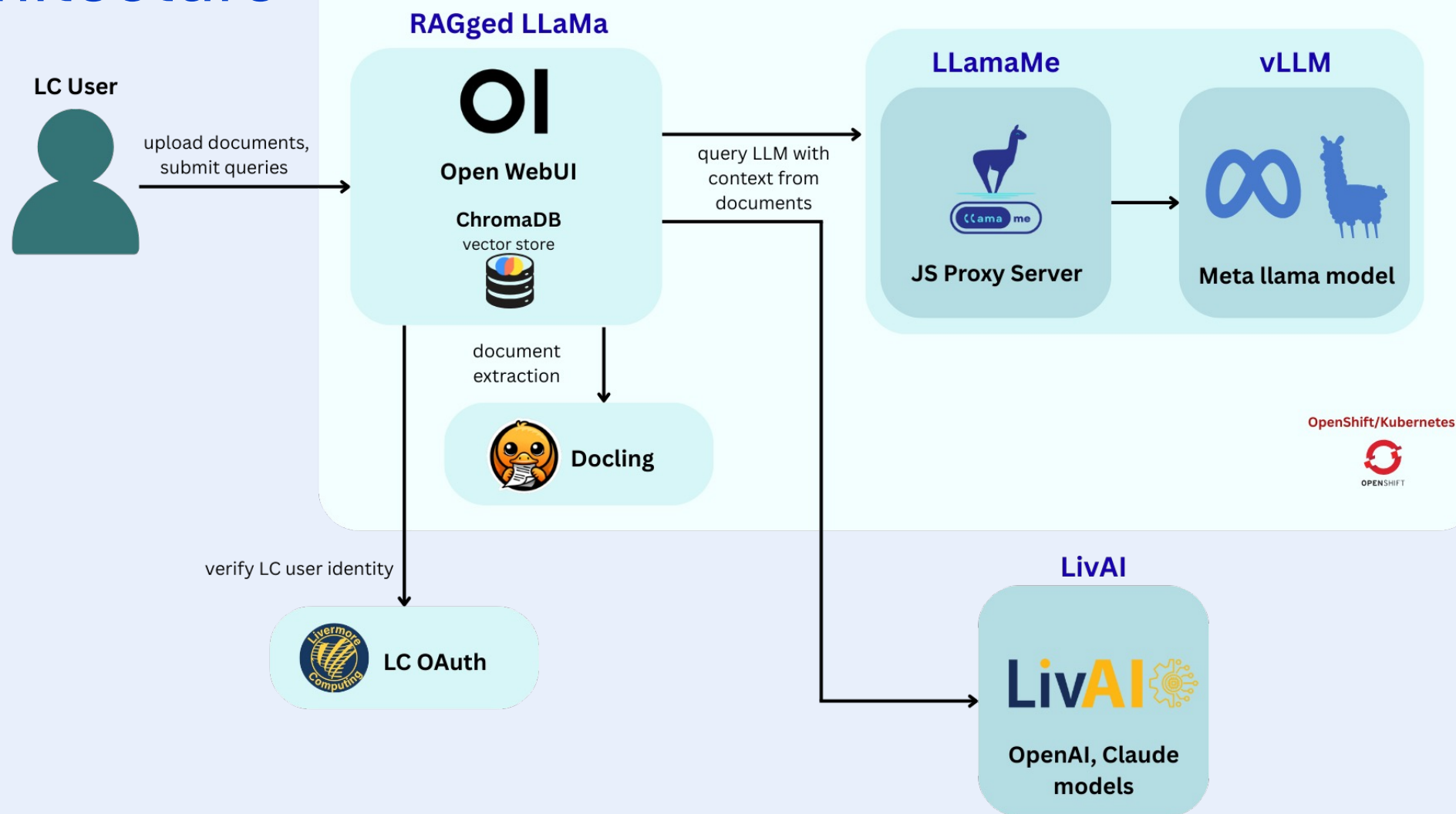
LC RAGged LLaMa = Open WebUI



Open-source AI platform with a built-in RAG engine: web interface + API server to interact with LLMs.

- connections to LLMs with OpenAI-compatible APIs
 - ex: LLamaMe, LivAI, Venado*
- user-friendly chat interface
- supports uploads for a variety of file types
- multi-user access control
 - private profiles with flexibility to share documents for collaboration
- designed to run offline

Architecture





RAGged LLaMa Connects to OpenAI APIs: LLamaMe and LivAI

OI

RAGged LLaMa (Open WebUI)

New Chat

Search

Notes

Workspace

Chats

Previous 7 days

LC CZ VNC Setup

Previous 30 days

Configuring LC CZ VNC

August

Joke Time

Roman Empire Urine Tax

Users

Evaluations

Functions

Settings

General

Connections

Models

Evaluations

Tools

Documents

Web Search

Code Execution

Interface

Audio

Images

Pipelines

Database

General

OpenAI API

Manage OpenAI API Connections

https://livai-api.llnl.gov/v1

https://llamame.apps.czapps.llnl.gov/v1

Ollama API

Direct Connections

Direct Connections allow users to connect to their own OpenAI compatible API endpoints.

Cache Base Model List

Base Model List Cache speeds up access by fetching base models only at startup or on settings save—faster, but may not show recent base model changes.

Models

The screenshot displays the Open Inference UI (OI) interface for managing models. The main panel shows a list of 19 models, each with a status toggle switch. The models listed are:

- anthropic.claude-3-5-sonnet-20240620-v1:0
- claude-haiku-3
- claude-sonnet-3.7
- gpt-35-turbo
- gpt-4.1
- gpt-4.1-mini
- gpt-4.1-nano
- gpt-4o
- gpt-4o-commercial
- gpt-4o-mini
- gpt-5
- gpt-5-mini
- gpt-5-nano

The left sidebar contains navigation options: New Chat, Search, Notes, Workspace, Chats, Previous 7 days, LC CZ VNC Setup, Previous 30 days, Configuring LC CZ VNC, August, Joke Time, and Roman Empire Urine Tax. The top navigation bar includes Users, Evaluations, Functions, and Settings. The bottom left corner shows the user profile for Otto Venezuela.

 RAGged LLaMa (Open WebUI)

 New Chat

 Search

 Notes

 Workspace

 Chats

Previous 7 days

LC CZ VNC Setup 

Previous 30 days

 Configuring LC CZ VNC

August

Joke Time 🎉

Roman Empire Urine Tax

 Otto Venezuela

 Lawrence Livermore National Laboratory

ModelsKnowledgePromptsTools

Knowledge3

 Search Knowledge

COLLECTION

sync-test

Sync test

By Otto VenezuelaUpdated 13 days ago

COLLECTION

test-1

Testing

By Otto VenezuelaUpdated 21 days ago

COLLECTION

RealVNC

documentation and tickets

By LC DOERSUpdated 23 days ago

ⓘ Use '#' in the prompt input to load and include your knowledge.

oi

RAGged LLaMa (Open WebUI)

📝

New Chat

🔍

Search

📖

Notes

🛠️

Workspace

📄

Chats

Previous 7 days

LC CZ VNC Setup 🖥️

Previous 30 days

🖥️ Configuring LC CZ VNC ...

August

Joke Time 🎭

Roman Empire Urine Tax

gpt-4o ▾ +



📄

RealVNC

Collection

How can I configure LC CZ VNC?

< 2/2 >

oi **gpt-4o** 09/08/2025 at 3:20 PM
To configure LC CZ VNC:

- Machines and Access:** RealVNC is installed on dedicated VNC clusters, including `czvnc.llnl.gov` for the CZ center. You can connect to these nodes by navigating to port 5999.
- Settings:** No special server-side settings are required for using RealVNC on LC's infrastructure.
- File Systems:** The VNC nodes mount the typical LC file systems, such as home directories, parallel file systems, and workspaces.
- Programming Environment:** Be aware that the VNC nodes provide a reduced programming environment. Many software tools, including compilers and MPI builds available on other nodes, will not be available on VNC nodes. If additional packages are needed, send a request along with a justification to the LC Hotline.
- Support and Resources:** If you need assistance or encounter issues, you can contact the LC Hotline at `lc-hotline@llnl.gov` or call `(925) 422-4531`. Additional tips and tricks for using VNC at LC can be found in the CZ Confluence documentation at `lc.llnl.gov/confluence/display/LC/VNC+How-to+for+LC` `Help.txt` `Environment.txt`.

1 Help.txt 2 Environment.txt 3 CONFIGURATION_17.txt



tell me more

Send a Message

+ | 📄 Code Interpreter



New PDS – Vector Database

PGVector



UNDER DEVELOPMENT

FY26 – Q1 We will provide a new Vector Database PDS called PGVector via LaunchIT

- Open-source extension for PostgreSQL that enables efficient storage, indexing, and querying of vector embeddings directly within your PostgreSQL database
- enables vector similarity search, allowing applications like semantic search, and other AI-driven features

 LaunchIT

Service



pgvector

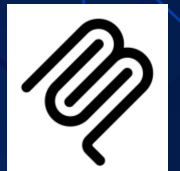
LC PGVector PG-16 TLS

Provided by LC IaaS (WEG)

PostgreSQL database service with persistent storage and the postgres vector extension.

What is next?

- Evaluating MLflow (Early Access)
- Provide PGVector via LaunchIT
- Integrate LC RAGged LLaMa into LaunchIT via LC Wormhole
- Evaluating Model Context Protocol (MCP) Server – securely
- Providing robust support for DataHub
- Integrating Fireworks workflow tool via LC Wormhole
- Trying to keep up with new Large Language Models



Stay connected ... call me (llámame)



Microsoft Team: [LC LLamaMe](#) (LLM) | General | Microsoft Teams



Docs: <https://hpc.llnl.gov/cloud>

Gitlab: [LLamaMe](#) Example Repo



Feedback: lit-help@llnl.gov

Bugs: lc-hotline@llnl.gov

Me: otto@llnl.gov



Thank you!



The future is already here – it's
just not very evenly distributed.

William Gibson



EXTRA slides

RAGged LLaMa - LC Customizations

- Deployment: **OpenShift**
 - container running in LC OpenShift Kubernetes clusters
- LLM Connection: **LLamaMe API**
 - keeps data within LC network (when using LC LLamaMe)
- Authentication: **LC OAuth**
 - accounts linked to LC user on login
- Document extraction: **Docling**
 - handles multiple document formats
 - advanced PDF understanding
 - OCR support for scanned PDFs & images
 - converts content to Markdown, HTML, JSON formats



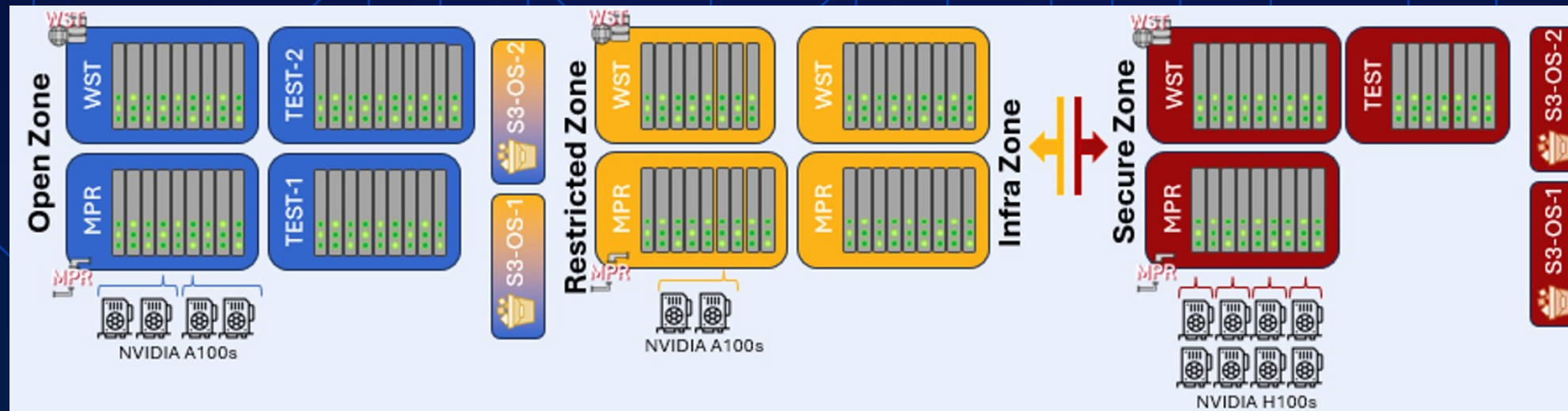
AMD MI250x Nodes

- We are using dedicated MI250 GPU nodes to host LLMs
- Kudos to Dave Fox and John Westlund for cluster configuration.
- Thanks to Jon Belof and CJ's vision and support
- Special thanks to Kailey Wong for expanding LLamaMe and LaunchIT integration.



Tenaya, RZVernal, Tioga

On-Prem cloud infrastructure across LC Zones



LC Kubernetes (K8s) Fleet