Unleash AI Faster

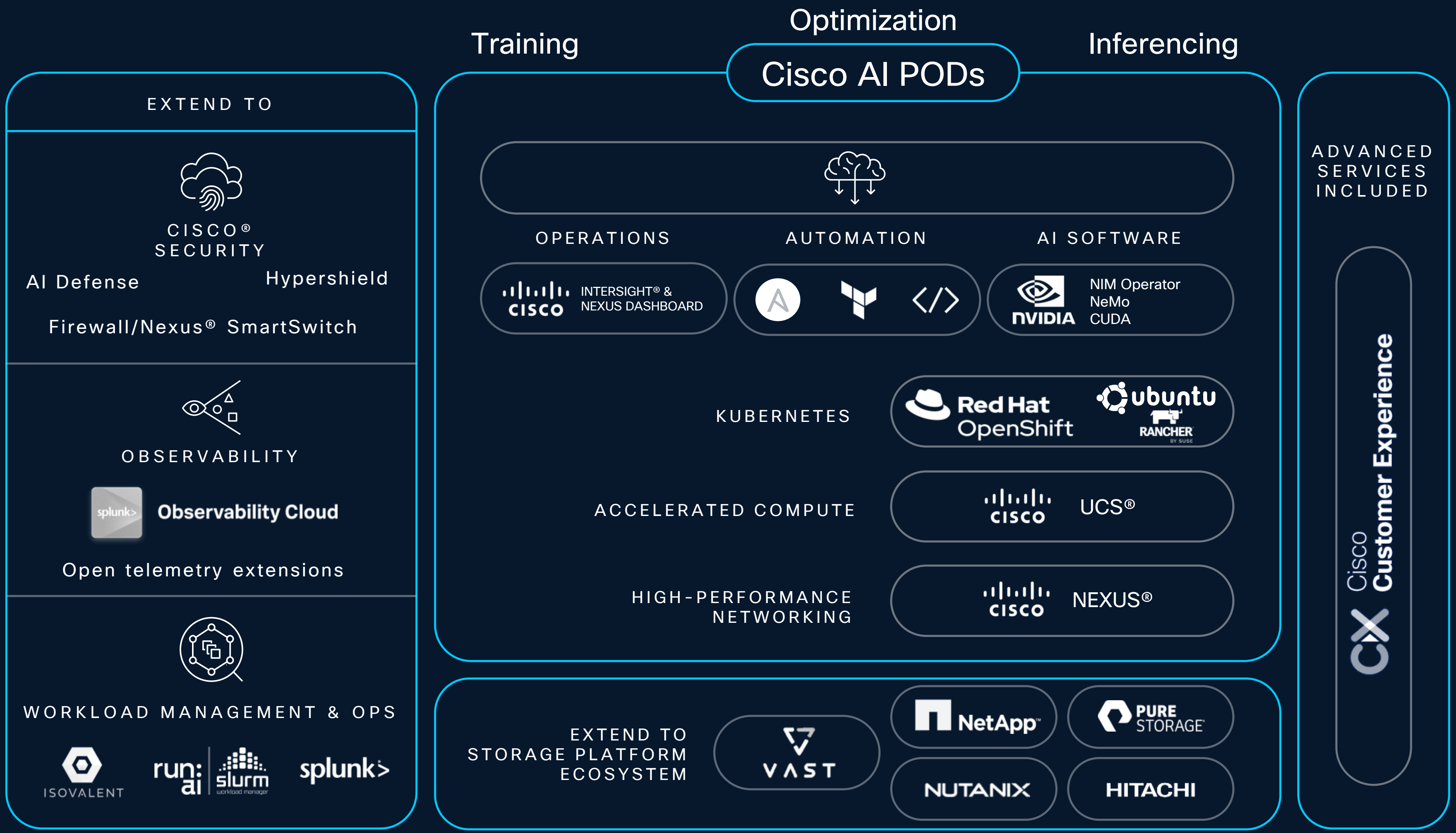
Automated Cisco AI Pod Setup with Intersight

Marc Abu El Ghait
Customer Success Specialist – Compute & AI

DEVNET-2488

Cisco AI PODs

Introducing AI POD “Integrated Offerings”



Training

Optimization

Inferencing

Cisco AI PODs

EXTEND TO

CISCO® SECURITY

AI Defense

Hypershield

Firewall/Nexus® SmartSwitch

OBSERVABILITY

splunk> Observability Cloud

Open telemetry extensions

WORKLOAD MANAGEMENT & OPS

ISOVALENT

run:ai

slurm workload manager

splunk>

OPERATIONS

cisco INTERSIGHT® & NEXUS DASHBOARD

AUTOMATION

AI SOFTWARE

NVIDIA

NIM Operator

NeMo

CUDA

KUBERNETES

Red Hat OpenShift

ubuntu

RANCHER BY SUSE

ACCELERATED COMPUTE

cisco UCS®

HIGH-PERFORMANCE NETWORKING

cisco NEXUS®

EXTEND TO STORAGE PLATFORM ECOSYSTEM

VAST

NetApp

PURE STORAGE

NUTANIX

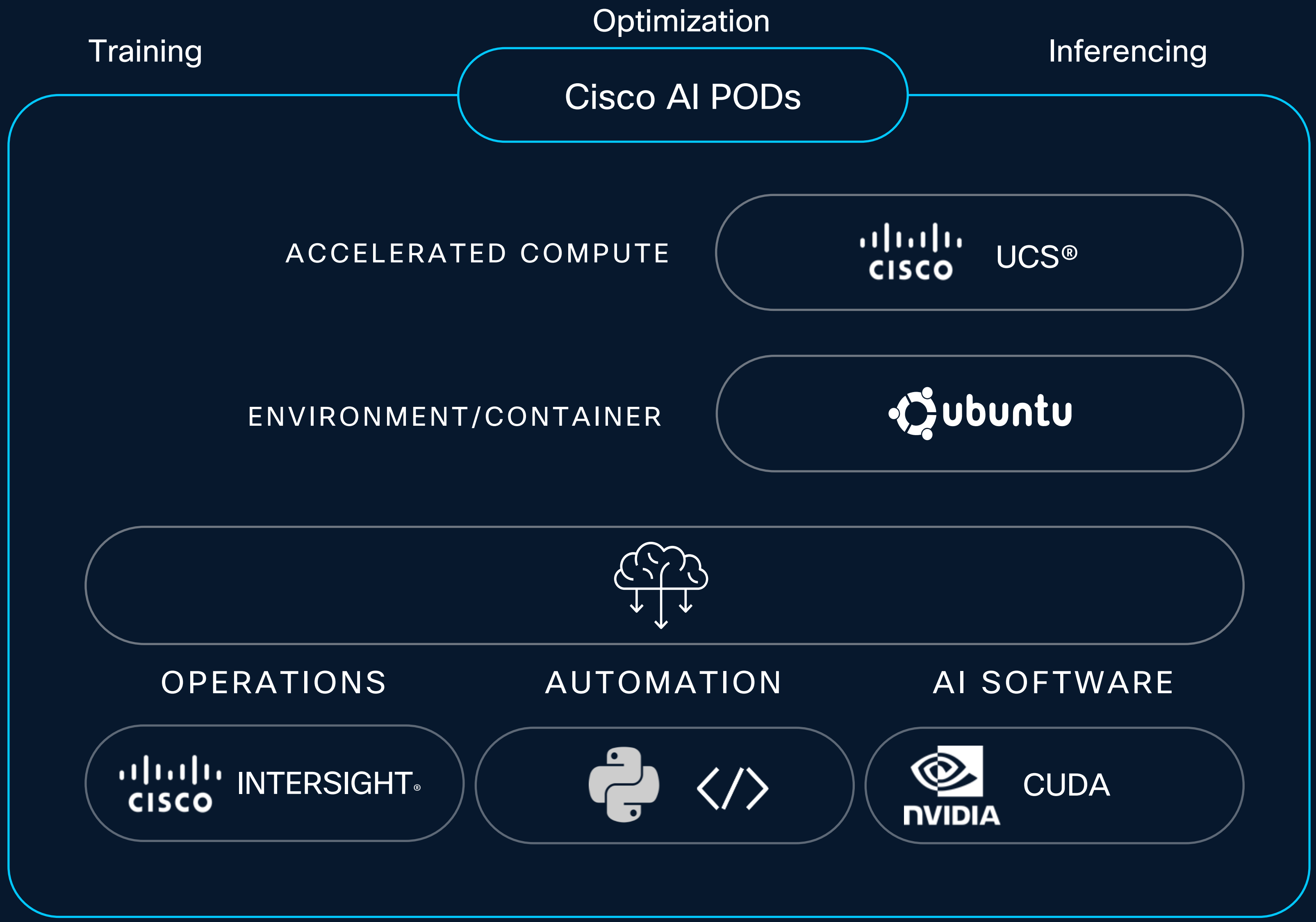
HITACHI

ADVANCED SERVICES INCLUDED

CX Cisco Customer Experience

Cisco AI PODs

Compute AI POD



Training

Optimization

Inferencing

Cisco AI PODs

ACCELERATED COMPUTE

cisco UCS®

ENVIRONMENT/CONTAINER

ubuntu

OPERATIONS

AUTOMATION

AI SOFTWARE

cisco INTERSIGHT®

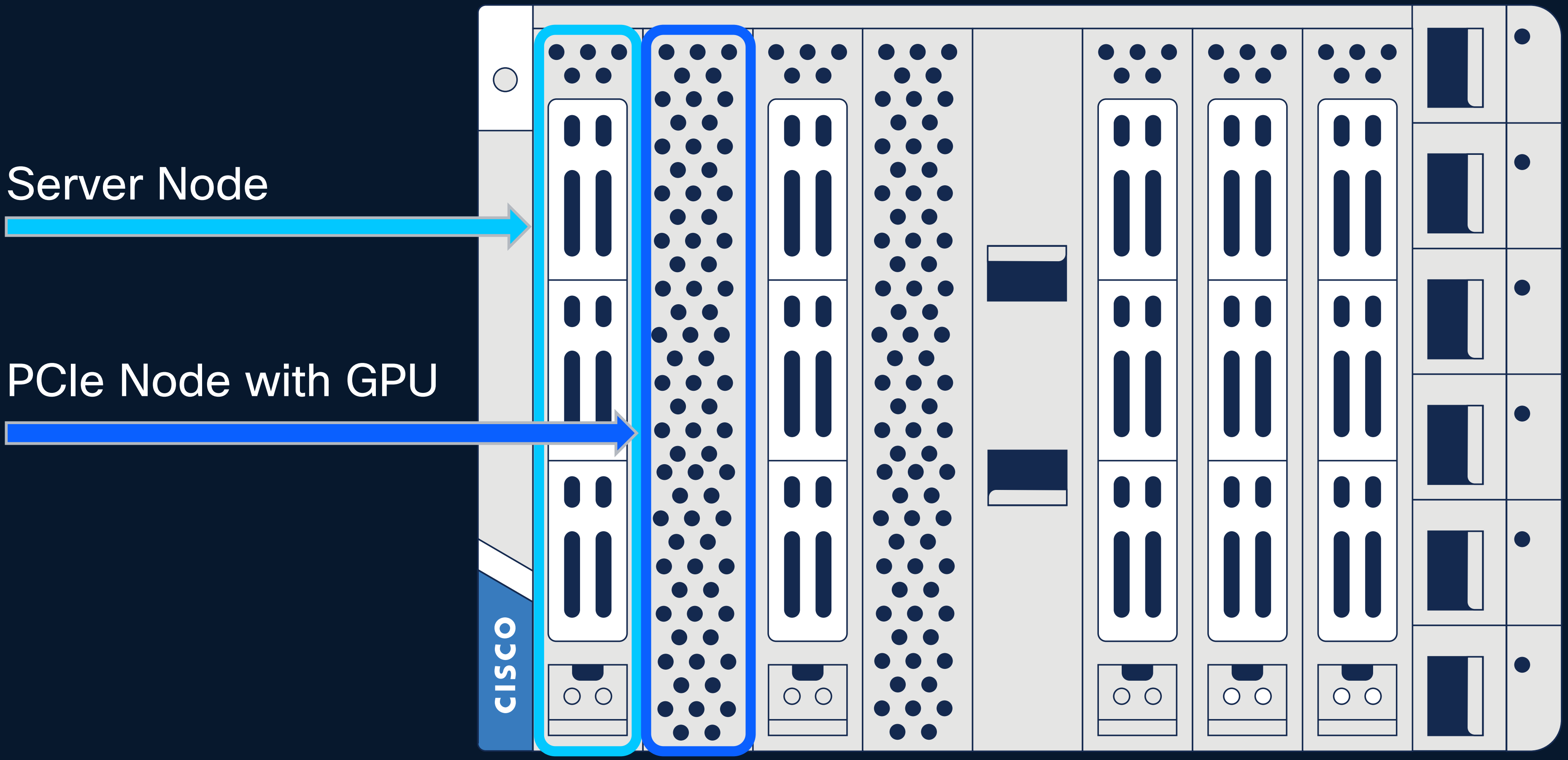
</>

NVIDIA CUDA

Cisco AI POD Compute Infrastructure

AI PODs can come in multiple types of UCS products :

- UCS C845A
- UCS C885A
- UCS C880A
- **UCS X-Series**



Server Node

PCIe Node with GPU

What's needed to setup AI?

From a compute point of view

Hardware:

- UCS Servers
- GPUs

Software Solutions:

- **Intersight** for physical configuration
- Operating System
- Containers to host workloads
- GPU software stack

Challenges:

- Respect Cisco Validated Designs
- Learn new technological dimensions
- Find proper AI use cases

All those considerations will **delay the value realization**

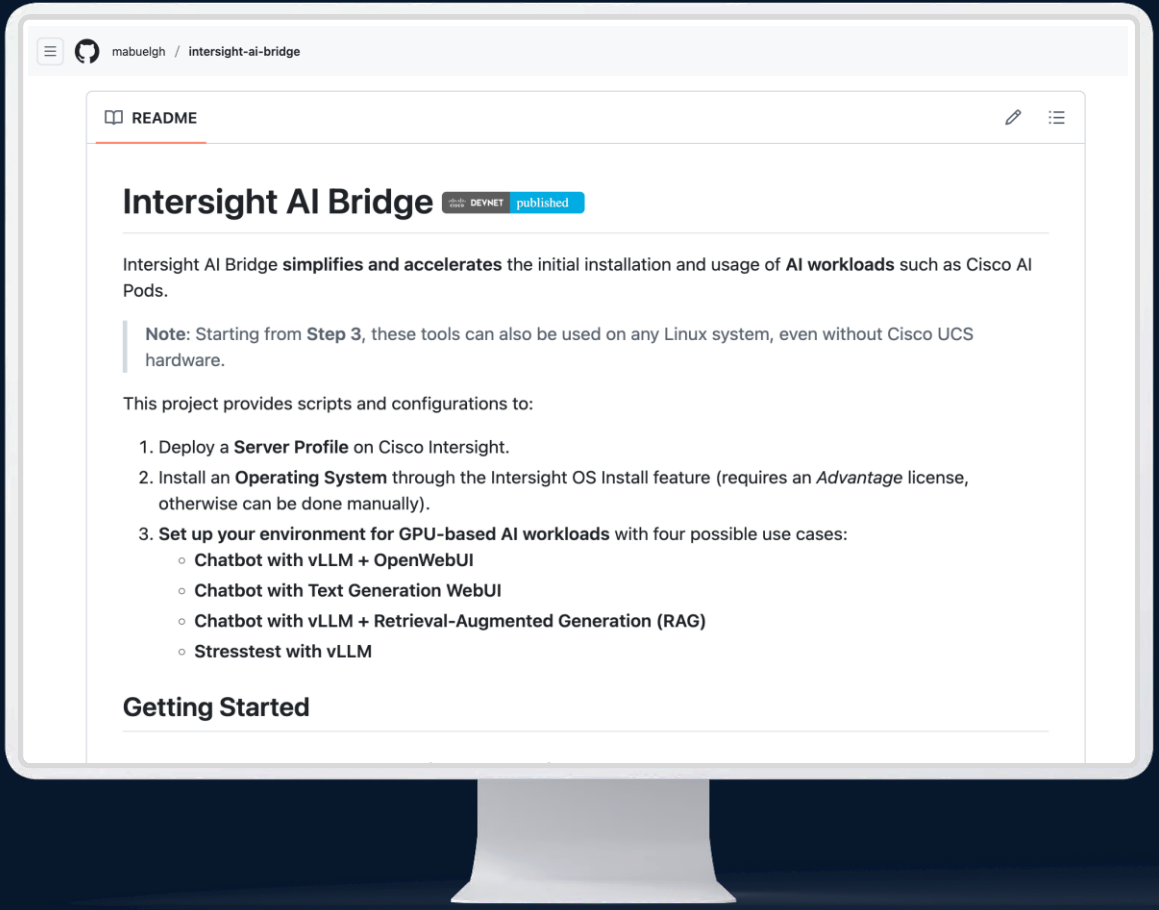
Intersight AI Bridge

A free community project

Complete **automated** build-up:

From **brand new AI Pod servers** to configured AI platform with **embedded use-cases**

- **Deploy** Intersight server configuration
- **Install** Operating System
- **Setup** OS and GPU configurations
- **Validate** with AI scenarios



Intersight AI Bridge

Intersight AI Bridge **simplifies and accelerates** the initial installation and usage of **AI workloads** such as Cisco AI Pods.

> **Note**: Starting from **Step 3**, these tools can also be used on any Linux system, even without Cisco UCS hardware.

This project provides scripts and configurations to:

1. Deploy a **Server Profile** on Cisco Intersight.
2. Install an **Operating System** through the Intersight OS Install feature (requires an *Advantage* license, otherwise can be done manually).
3. **Set up your environment for GPU-based AI workloads** with four possible use cases:
   - **Chatbot with vLLM + OpenWebUI**
   - **Chatbot with Text Generation WebUI**
   - **Chatbot with vLLM + Retrieval-Augmented Generation (RAG)**
   - **Stresstest with vLLM**

Getting Started

Intersight AI Bridge

Goals:

- Accelerate **time-to-value** for AI Pods
- Ensure consistent and **compliant deployment**
- Follow the **guidelines from Cisco Validated Designs**, NVIDIA and Ubuntu
- Show the automation capabilities of the products and solutions

Available:

- On GitHub, for free : https://github.com/mabuelgh/intersight-ai-bridge

Intersight AI Bridge

Operational Workflow

Deploy Server Config

Create Intersight server configuration based on the **CVD** and deploy automatically using Python.

Install OS

Deploy and install the OS with **recommended releases.** Use Intersight's OS Install feature to deploy **Ubuntu.**

Setup system & NVIDIA software

Use scripts to **install all the required system packages** to use NVIDIA hardware and toolkit on the OS.

Validate AI scenario

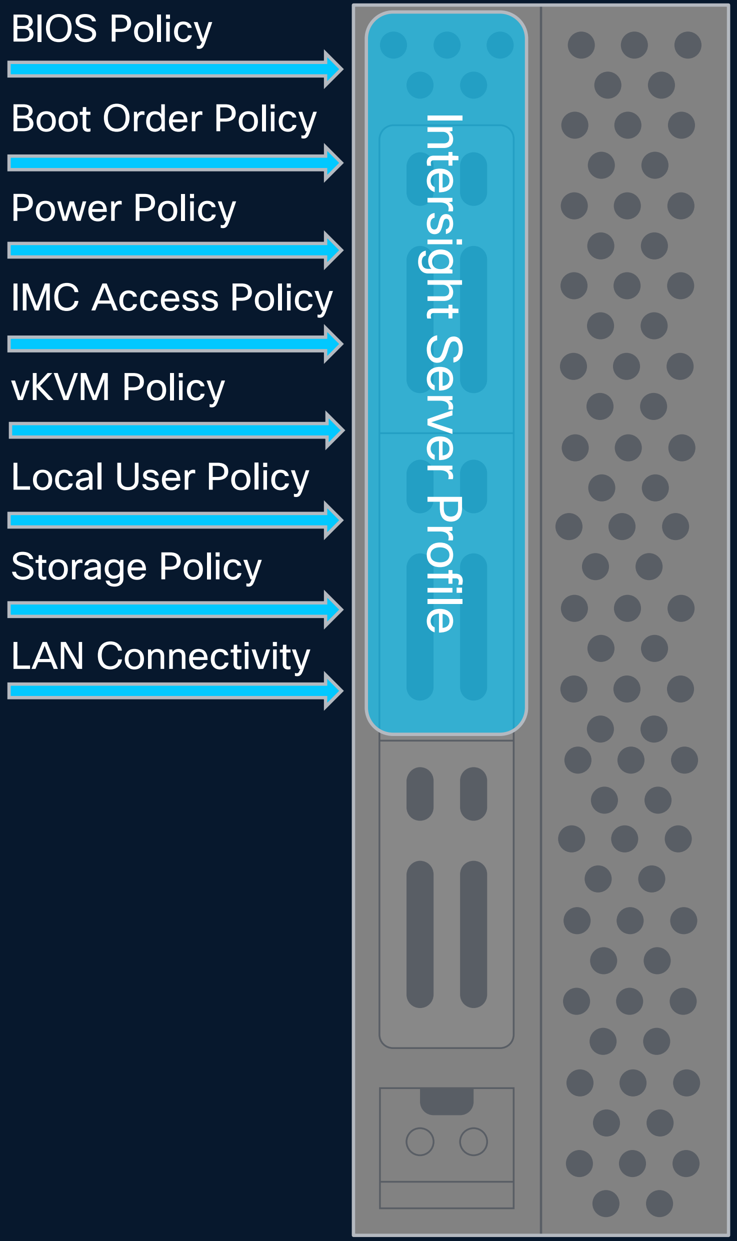
Intersight AI Bridge comes with embedded scripts to automatically deploy AI use-cases using Docker containers.

Server Profile

UCS Server configuration in Intersight is simplified with full hardware abstraction.

Each server is configured with various policies detailling the configuration (BIOS, Boot, Disk, etc. and identity (IP, MAC, etc.) inside a **Server Profile.**

The **BIOS Policy** is customized to **handle AI workloads efficiently**.



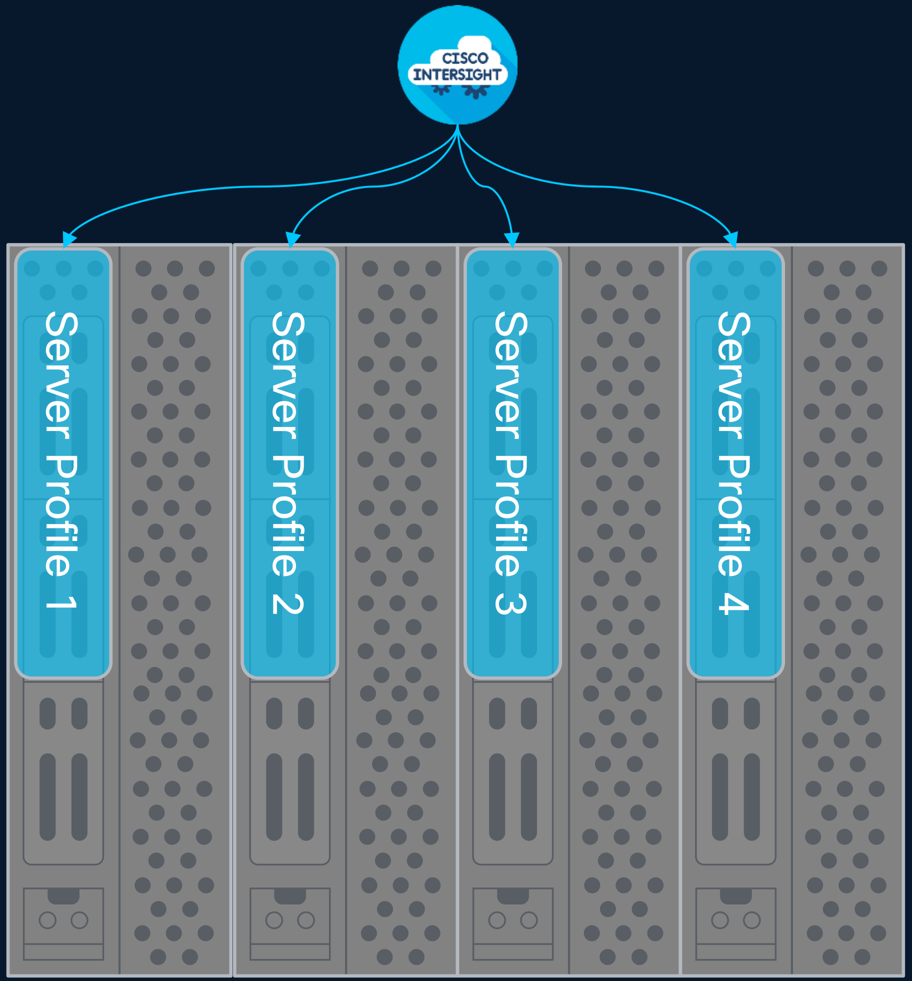
Server Profile

UCS Server configuration in Intersight is simplified with full hardware abstraction.

Each server is configured with various policies detailling the configuration (BIOS, Boot, Disk, etc. and identity (IP, MAC, etc.) inside a **Server Profile.**

The **BIOS Policy** is customized to **handle AI workloads efficiently**.

Thanks to Intersight, every policy created can be **reuse for every server – less manual work – more compliance**



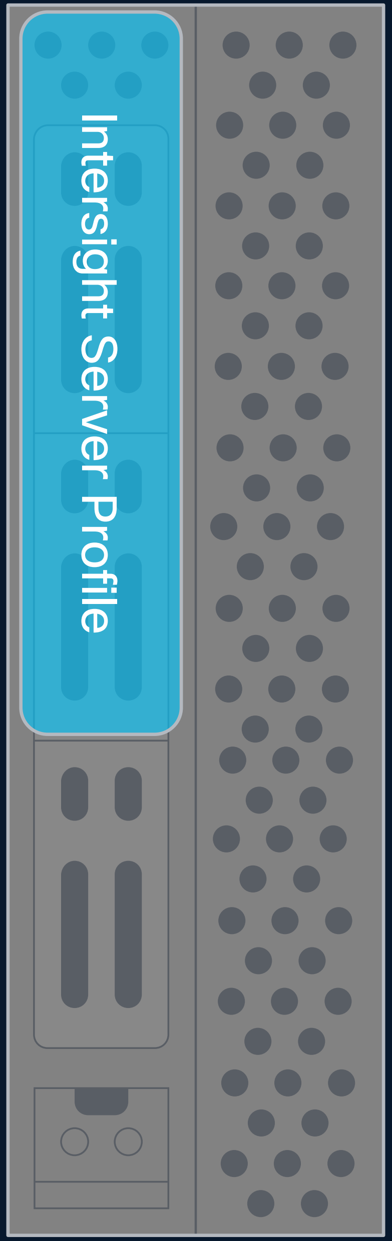
Step 1 : Deploy Server Config

The first step is using Intersight API to **create, deploy and activate** a configuration on the server.

Automated process breakdown:

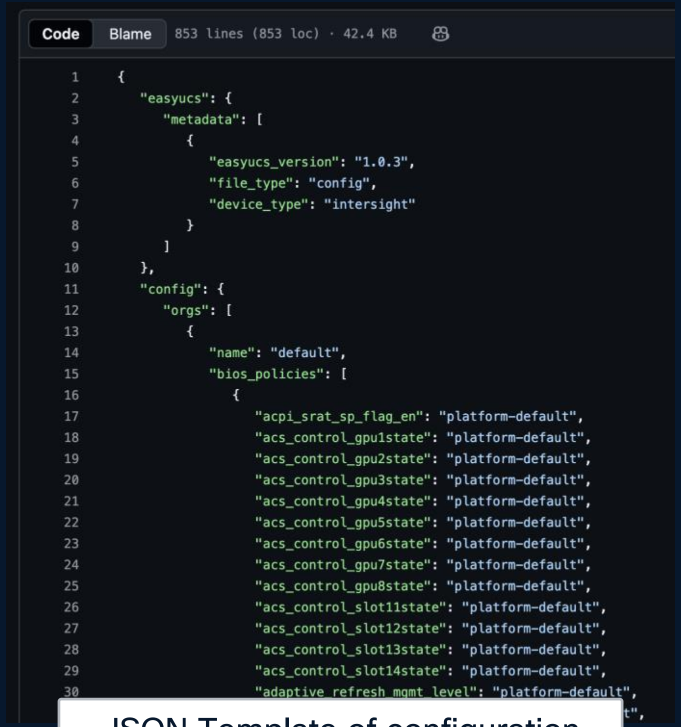
- Shape a configuration based on the Cisco Validated Designs
- Create the configuration (policies and profile) with Intersight
- Assign and Deploy the Server Profile to the server

➜ The server is now configured



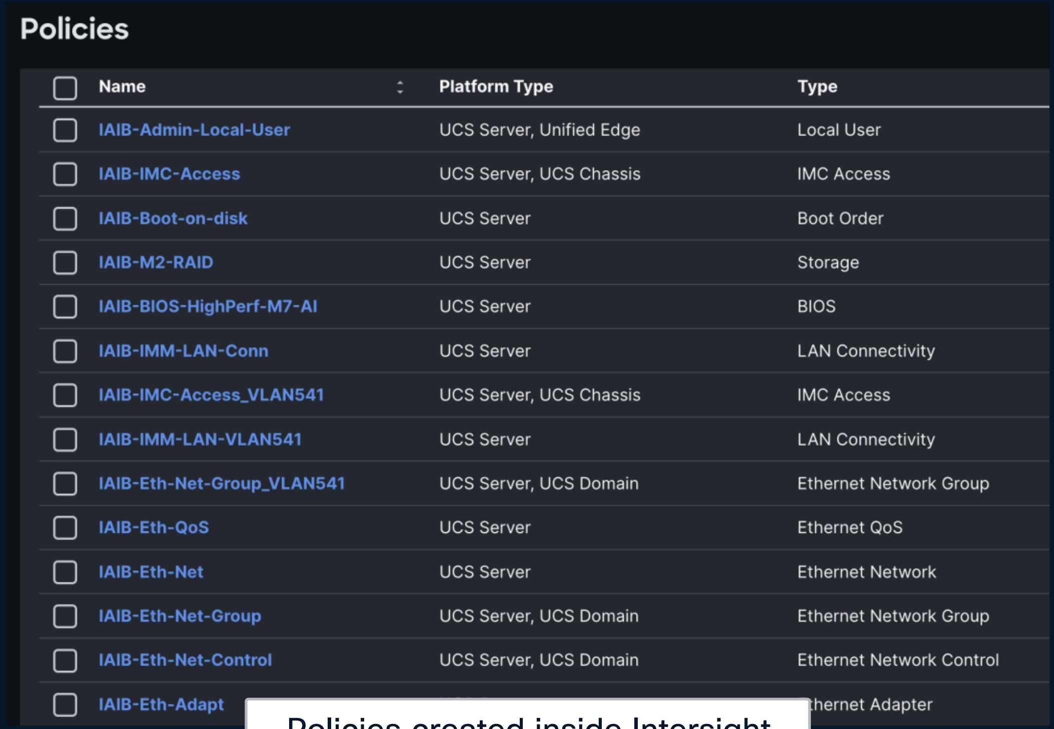
Step 1 : Deep Dive

- JSON Template is created based on the **CVDs**
- Intersight Python SDK is used to deploy JSON config through **EasyUCS** (GitHub project)
- All the **configuration Policies** and **Server Profile** are created inside Intersight and deployed on desired Server



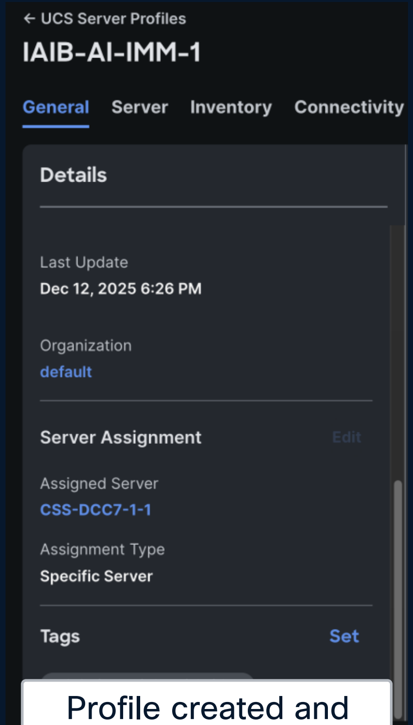
```
{
    "easyucs": {
        "metadata": [
            {
                "easyucs_version": "1.0.3",
                "file_type": "config",
                "device_type": "intersight"
            }
        ]
    },
    "config": {
        "orgs": [
            {
                "name": "default",
                "bios_policies": [
                    {
                        "acpi_srat_sp_flag_en": "platform-default",
                        "acs_control_gpu1state": "platform-default",
                        "acs_control_gpu2state": "platform-default",
                        "acs_control_gpu3state": "platform-default",
                        "acs_control_gpu4state": "platform-default",
                        "acs_control_gpu5state": "platform-default",
                        "acs_control_gpu6state": "platform-default",
                        "acs_control_gpu7state": "platform-default",
                        "acs_control_gpu8state": "platform-default",
                        "acs_control_slot11state": "platform-default",
                        "acs_control_slot12state": "platform-default",
                        "acs_control_slot13state": "platform-default",
                        "acs_control_slot14state": "platform-default",
                        "adaptive_refresh_mgmt_level": "platform-default",
                                                                        t",
```

JSON Template of configuration


Policies

□	Name	Platform Type	Type
□	IAIB-Admin-Local-User	UCS Server, Unified Edge	Local User
□	IAIB-IMC-Access	UCS Server, UCS Chassis	IMC Access
□	IAIB-Boot-on-disk	UCS Server	Boot Order
□	IAIB-M2-RAID	UCS Server	Storage
□	IAIB-BIOS-HighPerf-M7-AI	UCS Server	BIOS
□	IAIB-IMM-LAN-Conn	UCS Server	LAN Connectivity
□	IAIB-IMC-Access_VLAN541	UCS Server, UCS Chassis	IMC Access
□	IAIB-IMM-LAN-VLAN541	UCS Server	LAN Connectivity
□	IAIB-Eth-Net-Group_VLAN541	UCS Server, UCS Domain	Ethernet Network Group
□	IAIB-Eth-QoS	UCS Server	Ethernet QoS
□	IAIB-Eth-Net	UCS Server	Ethernet Network
□	IAIB-Eth-Net-Group	UCS Server, UCS Domain	Ethernet Network Group
□	IAIB-Eth-Net-Control	UCS Server, UCS Domain	Ethernet Network Control
□	IAIB-Eth-Adapt		thernet Adapter

Policies created inside Intersight



IAIB-AI-IMM-1

Details

Last Update
Dec 12, 2025 6:26 PM

Organization
default

Server Assignment

Assigned Server
CSS-DCC7-1-1

Assignment Type
Specific Server

Tags

Profile created and assigned in Intersight

Demo

Step 2 : OS Installation

The second step is also using Intersight API to install the OS on the server with the « **Install OS** » Intersight feature.

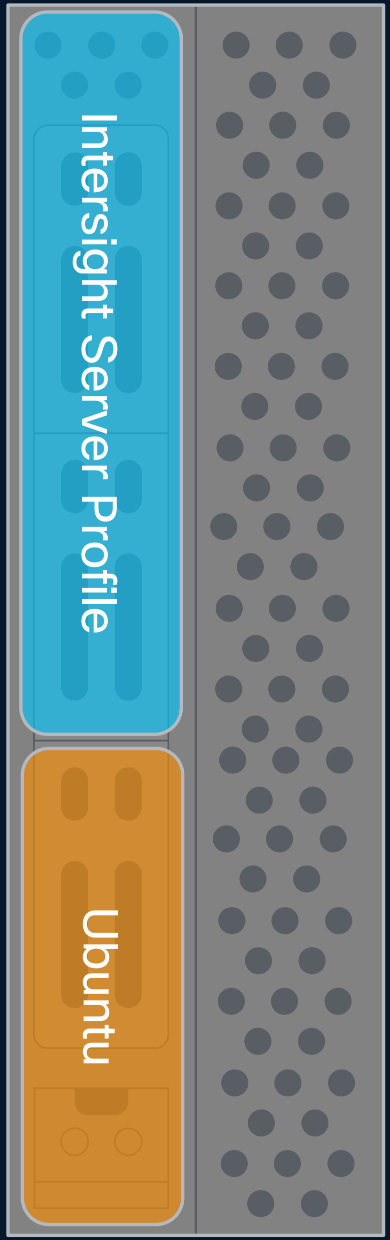
Manual process breakdown:

- Download the desired OS ISO file
- Download the recommended SCU (Server Configuration Utility) ISO file

Automated process breakdown:

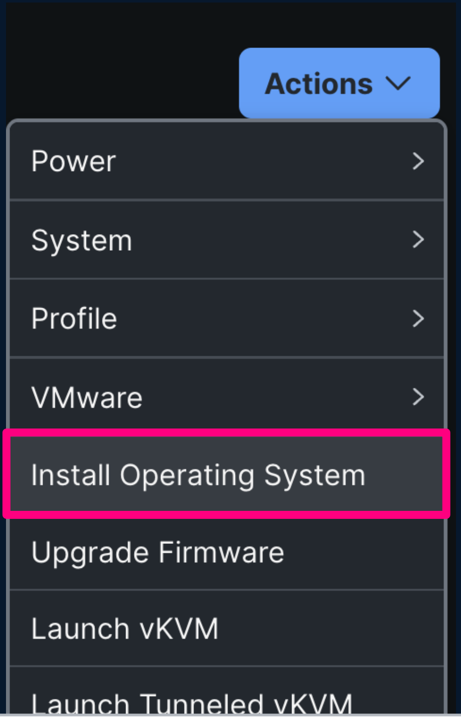
- Create an OS configuration based on recommandations
- Install the OS through Intersight « Install OS » feature

➜ **The OS is installed on the server**

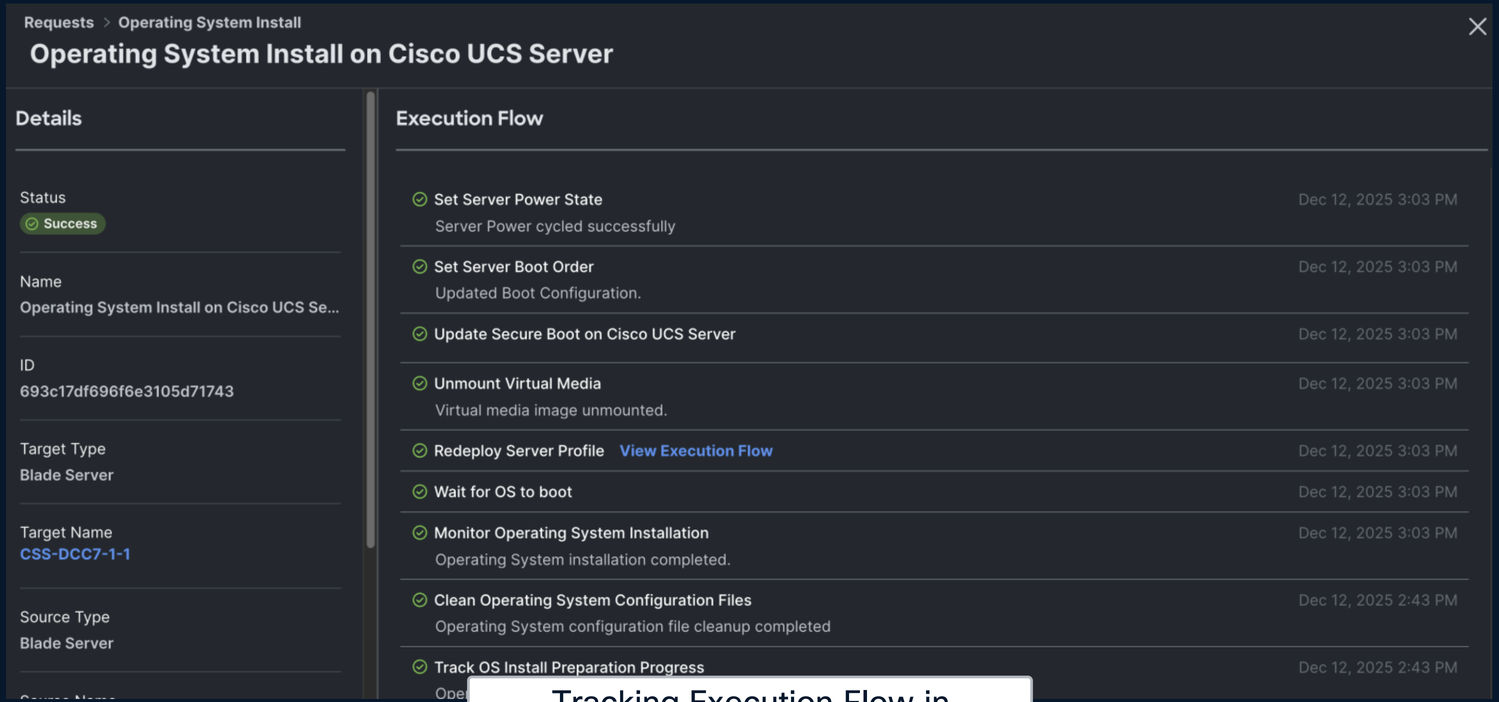


Step 2 : Deep Dive

- « **Install Operating System** » is a feature of Intersight to deploy automatically an OS through a user-friendly wizard
- Intersight **Python SDK** is used to do the process within a Python script instead of GUI
- Process can take some time and can be monitored using the UI



Install Operating System option in Intersight UI



Requests > Operating System Install

Operating System Install on Cisco UCS Server

Details

Status
✓ Success

Name
Operating System Install on Cisco UCS Se...

ID
693c17df696f6e3105d71743

Target Type
Blade Server

Target Name
CSS-DCC7-1-1

Source Type
Blade Server

Execution Flow

✓ Set Server Power State — Dec 12, 2025 3:03 PM
Server Power cycled successfully

✓ Set Server Boot Order — Dec 12, 2025 3:03 PM
Updated Boot Configuration.

✓ Update Secure Boot on Cisco UCS Server — Dec 12, 2025 3:03 PM

✓ Unmount Virtual Media — Dec 12, 2025 3:03 PM
Virtual media image unmounted.

✓ Redeploy Server Profile View Execution Flow — Dec 12, 2025 3:03 PM

✓ Wait for OS to boot — Dec 12, 2025 3:03 PM

✓ Monitor Operating System Installation — Dec 12, 2025 3:03 PM
Operating System installation completed.

✓ Clean Operating System Configuration Files — Dec 12, 2025 2:43 PM
Operating System configuration file cleanup completed

✓ Track OS Install Preparation Progress — Dec 12, 2025 2:43 PM
Ope

Tracking Execution Flow in Intersight

Demo

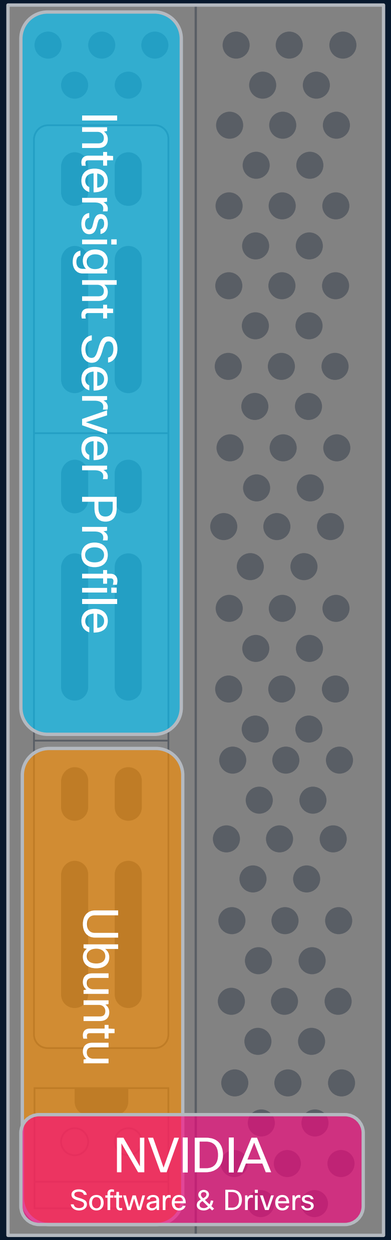
Step 3 : Setup system & NVIDIA software

The third step is about **deploying the packages, software and drivers** to use the GPUs.

Automated process breakdown:

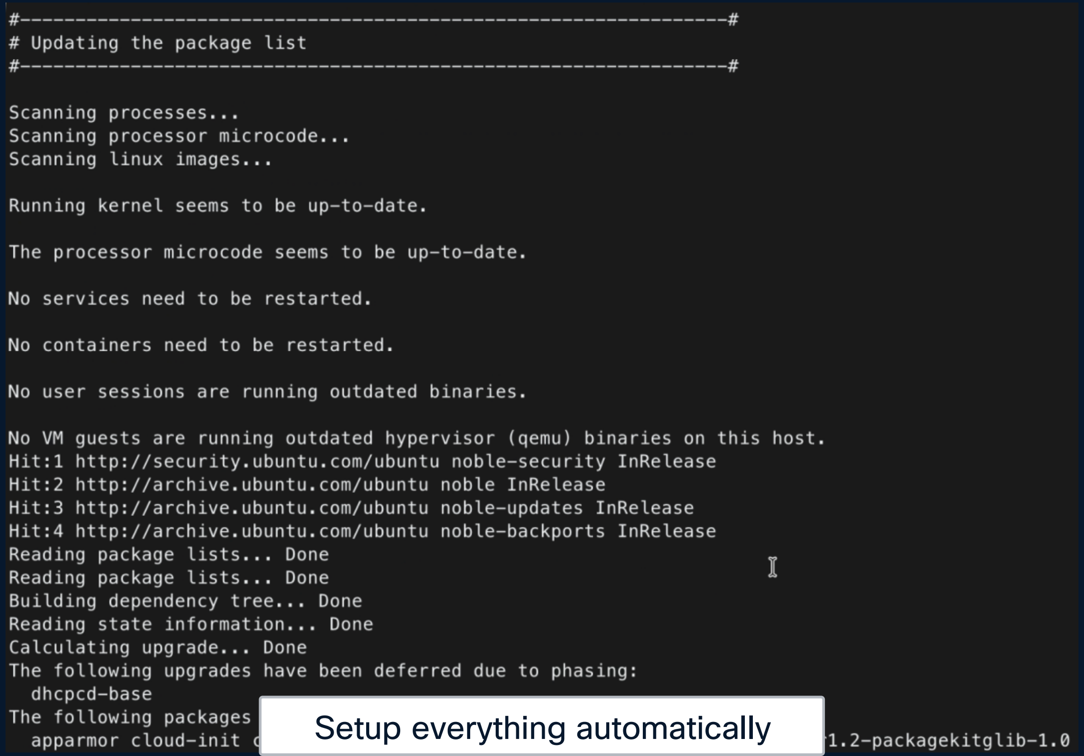
- Install up to date packages
- Install NVIDIA drivers
- Install Docker and NVIDIA container toolkit
- Install NVIDIA CUDA toolkit
- Install Python & Hugging Face Hub
- Reboot the server to apply all the changes

➜ The software stack is ready on the OS



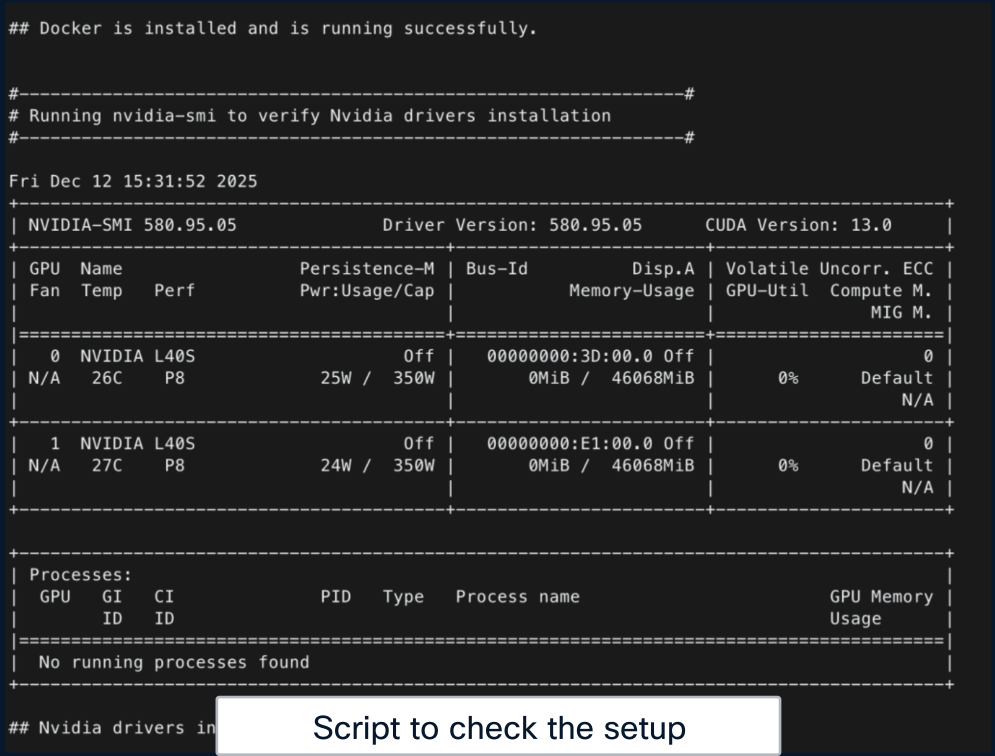
Step 3 : Deep Dive

- Setup shell script will download and setup everything needed to use the GPUs, it includes:
  - **CUDA** (Compute Unified Device Architecture) is a toolkit from NVIDIA to allow developers to use GPUs for processing in addition to graphics
  - **Hugging Face Hub** is a Python package that provides an interface to Hugging Face and access thousands of pre-trained models



```
#-----------------------------------------------------------------#
# Updating the package list
#-----------------------------------------------------------------#

Scanning processes...
Scanning processor microcode...
Scanning linux images...

Running kernel seems to be up-to-date.

The processor microcode seems to be up-to-date.

No services need to be restarted.

No containers need to be restarted.

No user sessions are running outdated binaries.

No VM guests are running outdated hypervisor (qemu) binaries on this host.
Hit:1 http://security.ubuntu.com/ubuntu noble-security InRelease
Hit:2 http://archive.ubuntu.com/ubuntu noble InRelease
Hit:3 http://archive.ubuntu.com/ubuntu noble-updates InRelease
Hit:4 http://archive.ubuntu.com/ubuntu noble-backports InRelease
Reading package lists... Done
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
Calculating upgrade... Done
The following upgrades have been deferred due to phasing:
  dhcpcd-base
The following packages
  apparmor cloud-init c                                    1.2-packagekitglib-1.0
```

Setup everything automatically



```
## Docker is installed and is running successfully.


#-----------------------------------------------------------------#
# Running nvidia-smi to verify Nvidia drivers installation
#-----------------------------------------------------------------#

Fri Dec 12 15:31:52 2025
+-----------------------------------------------------------------------------------------+
| NVIDIA-SMI 580.95.05              Driver Version: 580.95.05      CUDA Version: 13.0     |
+-----------------------------------------+------------------------+----------------------+
| GPU  Name                 Persistence-M | Bus-Id          Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util  Compute M. |
|                                         |                        |               MIG M. |
|=========================================+========================+======================|
|   0  NVIDIA L40S                    Off |   00000000:3D:00.0 Off |                    0 |
| N/A   26C    P8             25W /  350W |       0MiB /  46068MiB |      0%      Default |
|                                         |                        |                  N/A |
+-----------------------------------------+------------------------+----------------------+
|   1  NVIDIA L40S                    Off |   00000000:E1:00.0 Off |                    0 |
| N/A   27C    P8             24W /  350W |       0MiB /  46068MiB |      0%      Default |
|                                         |                        |                  N/A |
+-----------------------------------------+------------------------+----------------------+

+-----------------------------------------------------------------------------------------+
| Processes:                                                                              |
|  GPU   GI   CI              PID   Type   Process name                        GPU Memory |
|        ID   ID                                                               Usage      |
|=========================================================================================|
|  No running processes found                                                             |
+-----------------------------------------------------------------------------------------+

## Nvidia drivers in
```

Script to check the setup

Demo

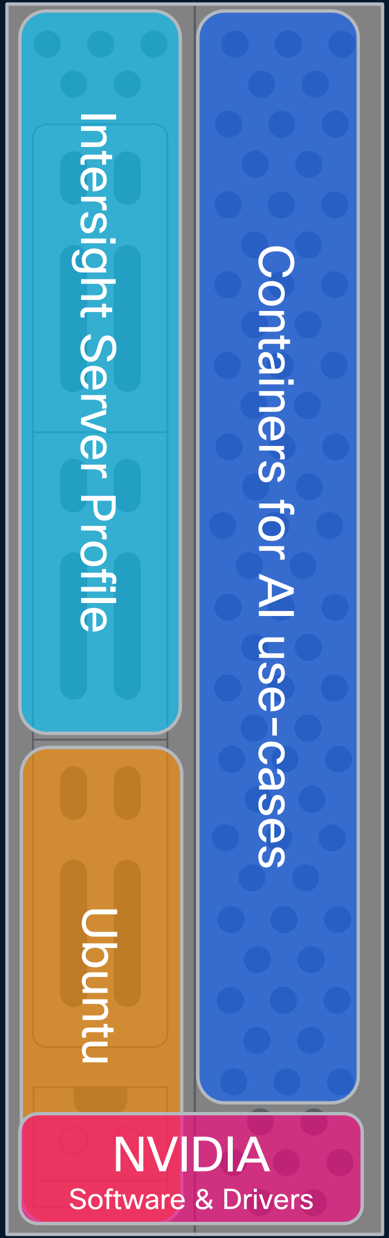
Step 4 : Validate with AI use-cases

The last step is about **choosing from a list of scenarios** the AI use-cases that best fit the usage needed.

Automated process breakdown:

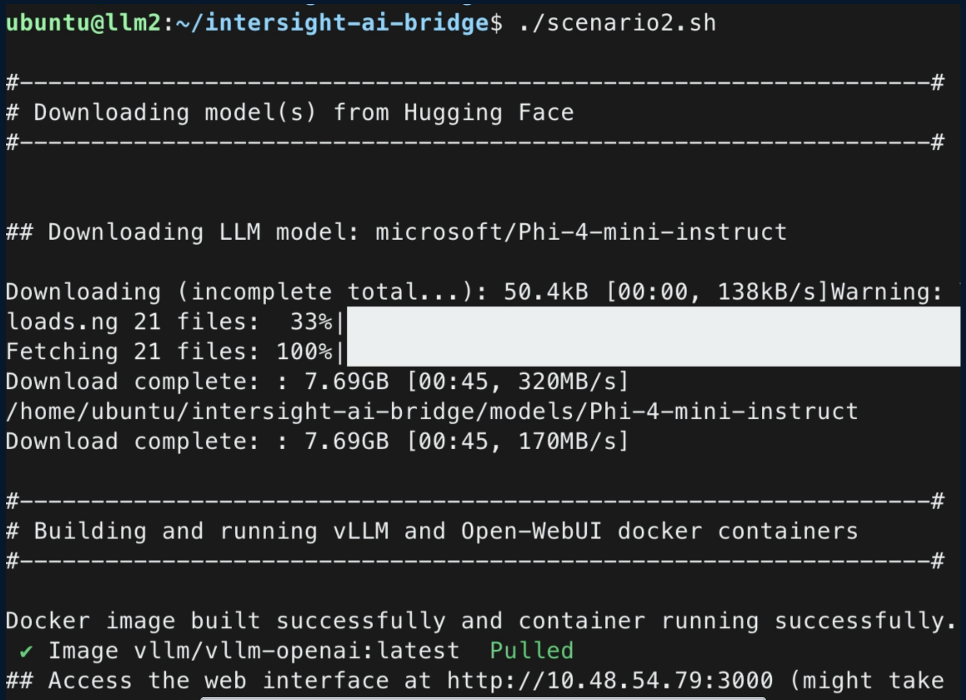
- Download the recommended LLM (Large Language Model) on Hugging Face
- Use Docker compose on the pre-established scenarios
- Deploy container

➜ The GPUs are now being used for AI use-cases



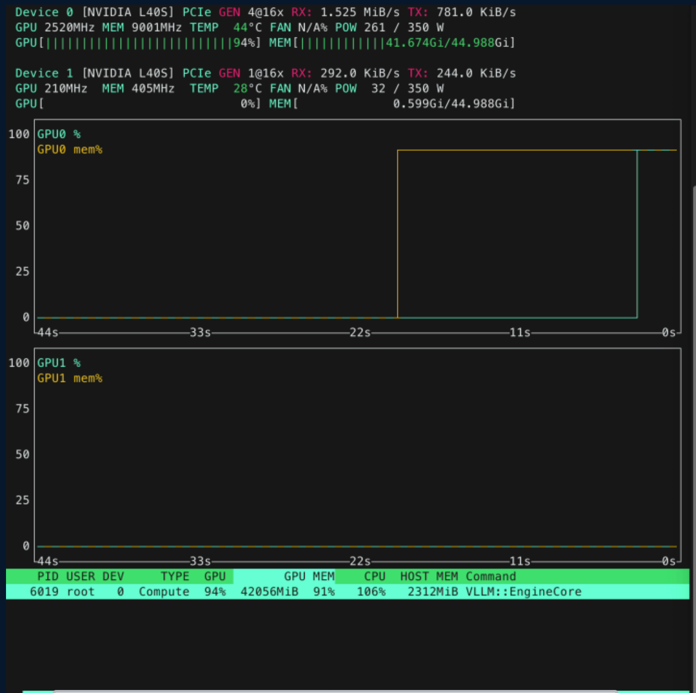
Step 4 : Deep Dive

- **Pre-assembled scenarios using GPUs** for AI use-cases are available with shell scripting
  - Use **HuggingFace Hub** to download the LLM and Docker to **host and deploy** the workload
  - Deploy **vLLM**: a fast and easy-to-use library for LLM inference and serving with API capabilities



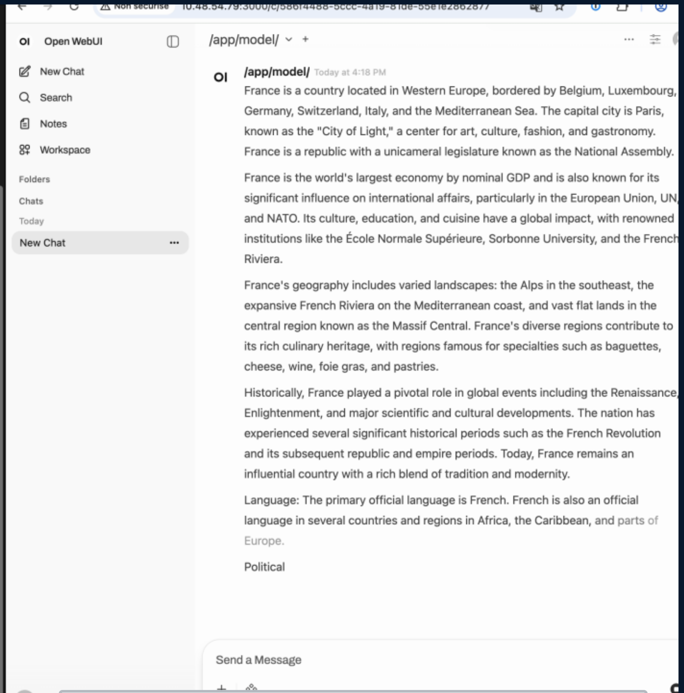
```
ubuntu@llm2:~/intersight-ai-bridge$ ./scenario2.sh

#------------------------------------------------------------------#
# Downloading model(s) from Hugging Face
#------------------------------------------------------------------#


## Downloading LLM model: microsoft/Phi-4-mini-instruct

Downloading (incomplete total...): 50.4kB [00:00, 138kB/s]Warning:
loads.ng 21 files:  33%|
Fetching 21 files: 100%|
Download complete: : 7.69GB [00:45, 320MB/s]
/home/ubuntu/intersight-ai-bridge/models/Phi-4-mini-instruct
Download complete: : 7.69GB [00:45, 170MB/s]

#------------------------------------------------------------------#
# Building and running vLLM and Open-WebUI docker containers
#------------------------------------------------------------------#

Docker image built successfully and container running successfully.
 ✔ Image vllm/vllm-openai:latest  Pulled
## Access the web interface at http://10.48.54.79:3000 (might take
 ✔ Network vll
```

Launching scenario 2



```
Device 0 [NVIDIA L40S] PCIe GEN 4@16x RX: 1.525 MiB/s TX: 781.0 KiB/s
GPU 2520MHz MEM 9001MHz TEMP  44°C FAN N/A% POW 261 / 350 W
GPU[||||||||||||||||||||||||||94%] MEM[||||||||||||41.674Gi/44.988Gi]

Device 1 [NVIDIA L40S] PCIe GEN 1@16x RX: 292.0 KiB/s TX: 244.0 KiB/s
GPU 210MHz  MEM 405MHz  TEMP  28°C FAN N/A% POW  32 / 350 W
GPU[                           0%] MEM[             0.599Gi/44.988Gi]

  PID USER DEV     TYPE  GPU        GPU MEM    CPU  HOST MEM Command
 6019 root   0  Compute  94%  42056MiB  91%   106%   2312MiB VLLM::EngineCore

F2Setu
```

Monitoring GPU Utilization



/app/model/ Today at 4:18 PM

France is a country located in Western Europe, bordered by Belgium, Luxembourg, Germany, Switzerland, Italy, and the Mediterranean Sea. The capital city is Paris, known as the "City of Light," a center for art, culture, fashion, and gastronomy. France is a republic with a unicameral legislature known as the National Assembly.

France is the world's largest economy by nominal GDP and is also known for its significant influence on international affairs, particularly in the European Union, UN, and NATO. Its culture, education, and cuisine have a global impact, with renowned institutions like the École Normale Supérieure, Sorbonne University, and the French Riviera.

France's geography includes varied landscapes: the Alps in the southeast, the expansive French Riviera on the Mediterranean coast, and vast flat lands in the central region known as the Massif Central. France's diverse regions contribute to its rich culinary heritage, with regions famous for specialties such as baguettes, cheese, wine, foie gras, and pastries.

Historically, France played a pivotal role in global events including the Renaissance, Enlightenment, and major scientific and cultural developments. The nation has experienced several significant historical periods such as the French Revolution and its subsequent republic and empire periods. Today, France remains an influential country with a rich blend of tradition and modernity.

Language: The primary official language is French. French is also an official language in several countries and regions in Africa, the Caribbean, and parts of Europe.

Political

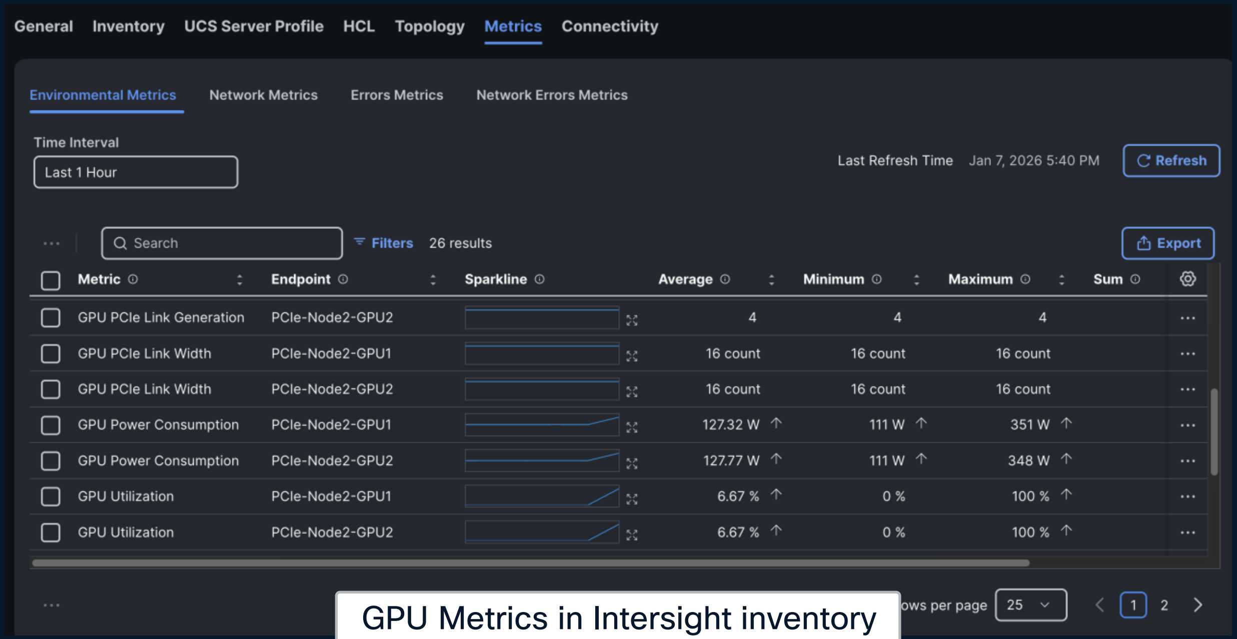
Chatbot using Web UI and LLM’s API

Demo

Intersight observability on GPU

Intersight can **display and explore the metrics** related to the **GPUs**:

- GPU Clockspeed
- GPU Memory Clockspeed
- GPU Power Consumption
- GPU Utilization
- And more!



Time Interval: Last 1 Hour

Last Refresh Time Jan 7, 2026 5:40 PM

26 results

Metric	Endpoint	Sparkline	Average	Minimum	Maximum	Sum
GPU PCIe Link Generation	PCIe-Node2-GPU2		4	4	4	
GPU PCIe Link Width	PCIe-Node2-GPU1		16 count	16 count	16 count	
GPU PCIe Link Width	PCIe-Node2-GPU2		16 count	16 count	16 count	
GPU Power Consumption	PCIe-Node2-GPU1		127.32 W ↑	111 W ↑	351 W ↑	
GPU Power Consumption	PCIe-Node2-GPU2		127.77 W ↑	111 W ↑	348 W ↑	
GPU Utilization	PCIe-Node2-GPU1		6.67 % ↑	0 %	100 % ↑	
GPU Utilization	PCIe-Node2-GPU2		6.67 % ↑	0 %	100 % ↑	

ows per page 25 ‹ 1 2 ›

GPU Metrics in Intersight inventory per server

Intersight observability on GPU

Intersight can **display and explore the metrics** related to the **GPUs**:

- GPU Clockspeed
- GPU Memory Clockspeed
- GPU Power Consumption
- GPU Utilization
- And more!

without Operating System agent

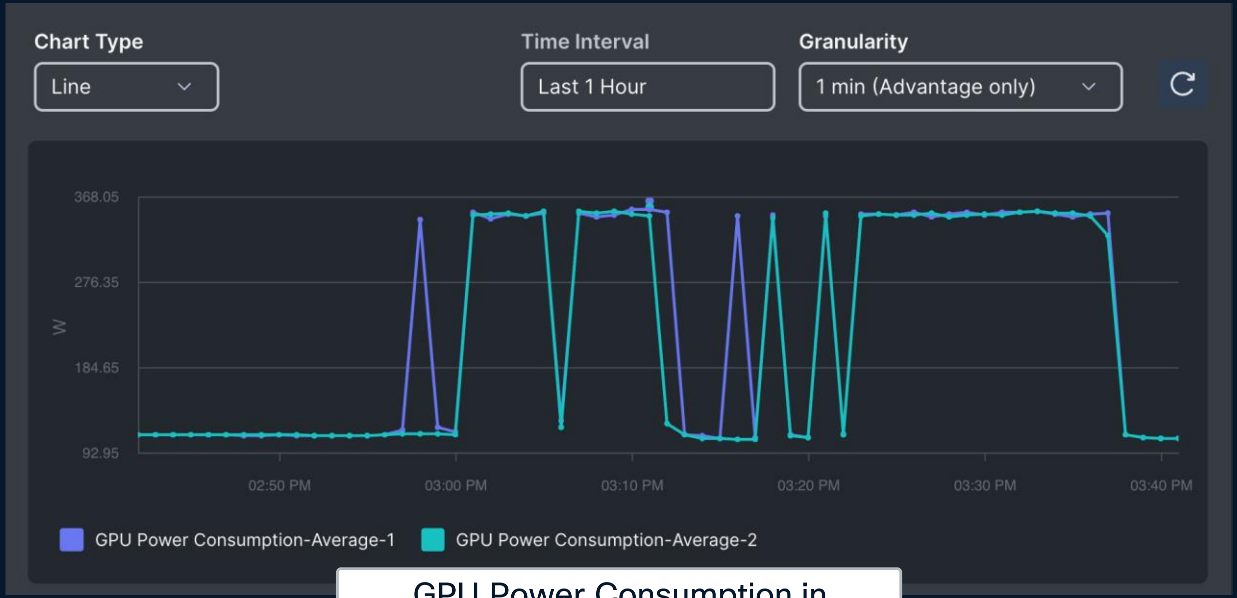


Chart Type: Line

Time Interval: Last 1 Hour

Granularity: 1 min (Advantage only)

GPU Power Consumption in Intersight Metrics Explorer

Key Takeaways - Call to Action

Every step of an AI deployment can be **automated** and **accelerated**:

- Intersight configurations and OS deployments done with **API**
- All the OS requirements achieved **automatically**
- Get GPU observability **easily with Intersight**

Intersight AI Bridge is a **DevNet** community project:

- **Free**, available on GitHub and DevNet website
- Each step can work **independently**
- Very **modular**: Easily customisable
- **Create and share** new AI scenario with the community

➜ From weeks of learning, work and experimentation to 1.5 hours of automated deployment

Webex App

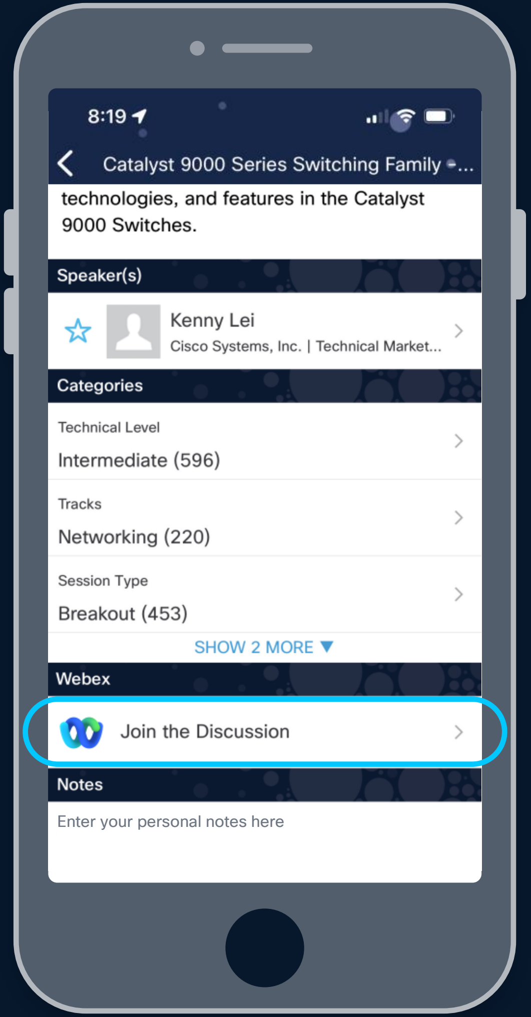
Questions?

Use Webex app to chat with the speaker after the session

How

1. Find this session in the Cisco Events app
2. Click “Join the Discussion”
3. Install the Webex app or go directly to the Webex space
4. Enter messages/questions in the Webex space

Webex spaces will be moderated by the speaker until February 27, 2026.



8:19

Catalyst 9000 Series Switching Family -...

technologies, and features in the Catalyst 9000 Switches.

Speaker(s)

Kenny Lei
Cisco Systems, Inc. | Technical Market...

Categories

Technical Level
Intermediate (596)

Tracks
Networking (220)

Session Type
Breakout (453)

SHOW 2 MORE ▼

Webex

Join the Discussion

Notes

Enter your personal notes here

https://ciscolive.ciscoevents.com/ciscolivebot/#DEVNET-2488

Complete your session surveys

Complete your surveys in the Cisco Events app.

Complete a minimum of 4 session surveys and the overall event survey to receive a unique Cisco Live t-shirt.

(from 11:30 on Thursday, while supplies last)

Continue your education

Visit the Cisco Showcase for related demos

Book your one-on-one Meet the Engineer meeting

Visit the Technical Solutions Clinics to discuss your technical questions

Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs

Visit the On-Demand Library for more sessions at CiscoLive.com/On-Demand

Contact me at: mabuelgh@cisco.com

Thank you