





Future Driven Infrastructure

ขับเคลื่อนธุรกิจด้วยโซลูชันคลาวด์บน Data Center แห่งใหม่



Teerawich Charoenyardprasert
Senior Cloud Solution Architect
teerawich@readyidc.com





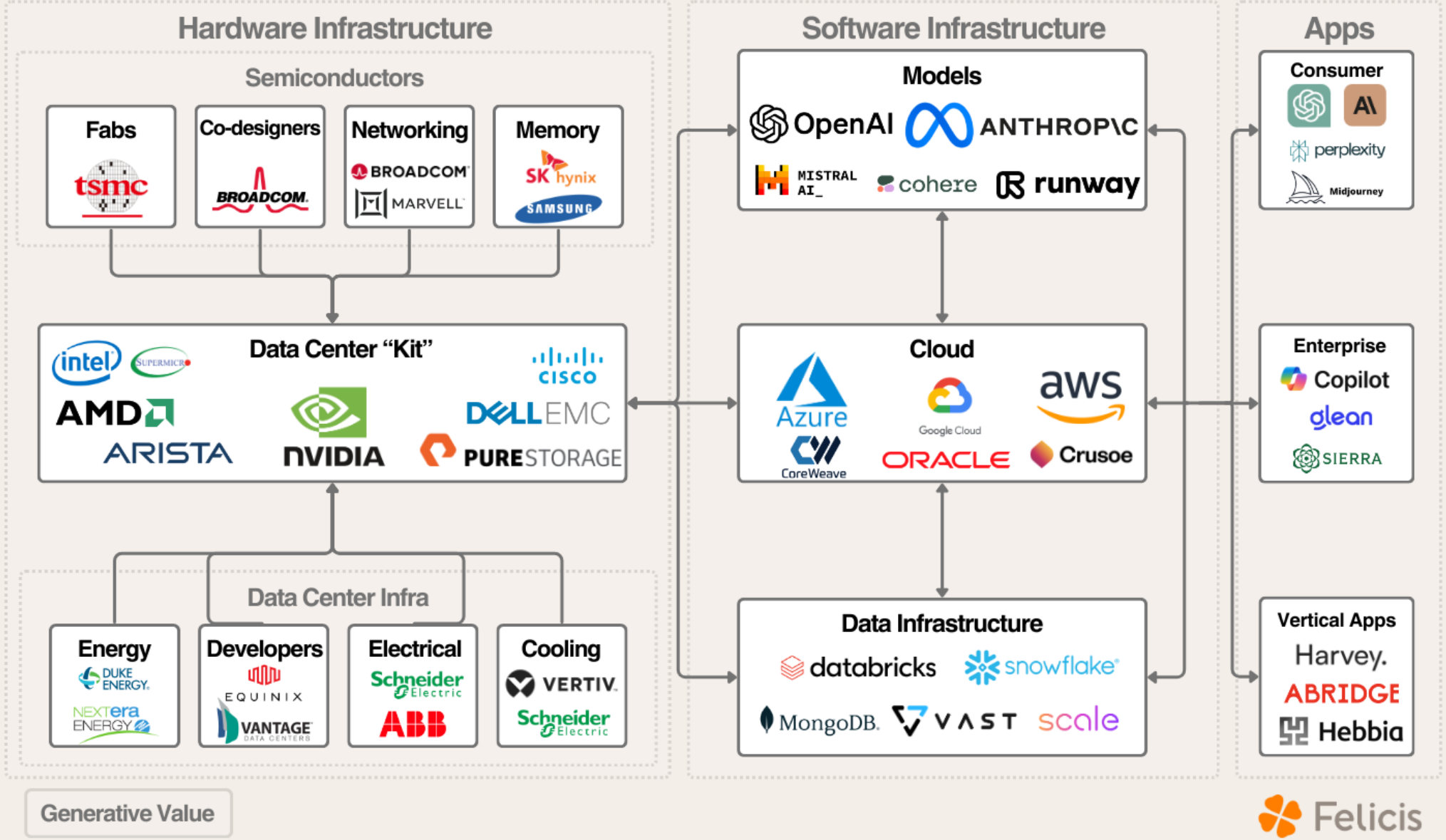
EDGE DATA CENTER
ReadyIDC Co.,Ltd.
Internet Service Provider (ISP)
Location : ศรีสμαν, นนทบุรี
North Bangkok

Tier III
ISO 27001:2022, 9001
CSA STAR



READYIDC

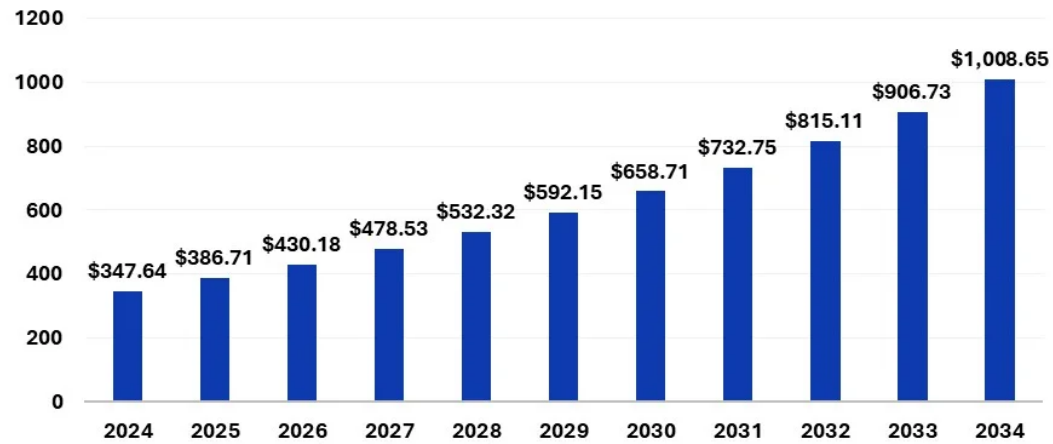
The AI Value Chain



Data Center & AI Market Size

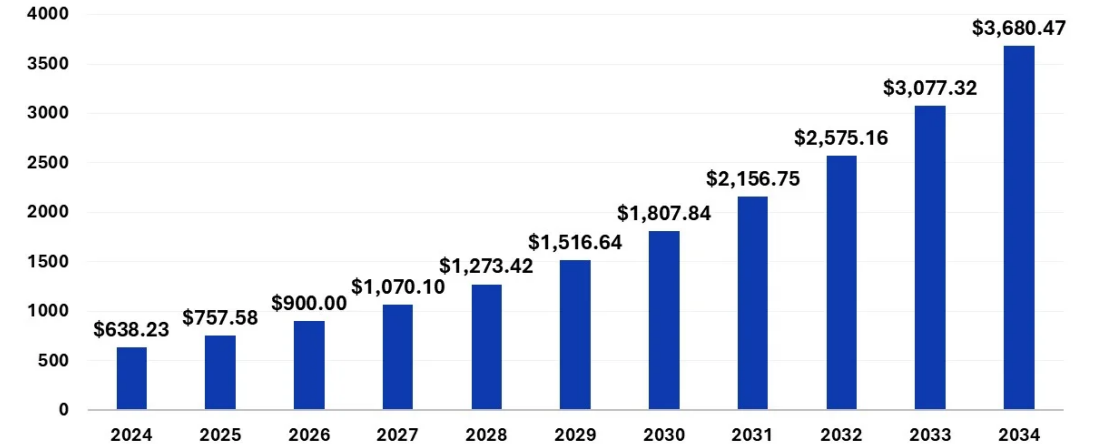


Data Center Market Size 2024 to 2034 (USD Billion)

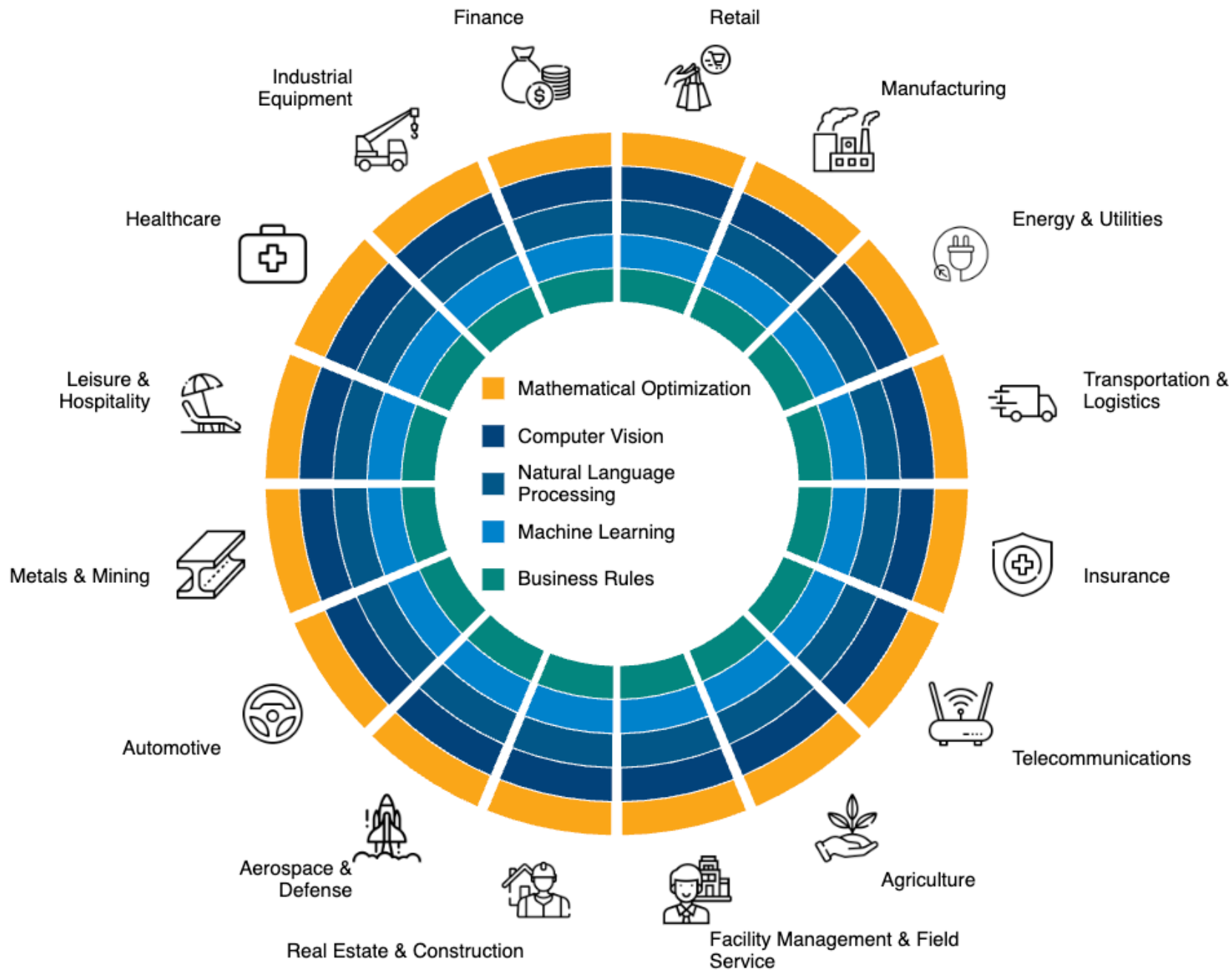


Source: <https://www.precedenceresearch.com/data-center-market>

Artificial Intelligence (AI) Market Size 2024 to 2034 (USD Billion)

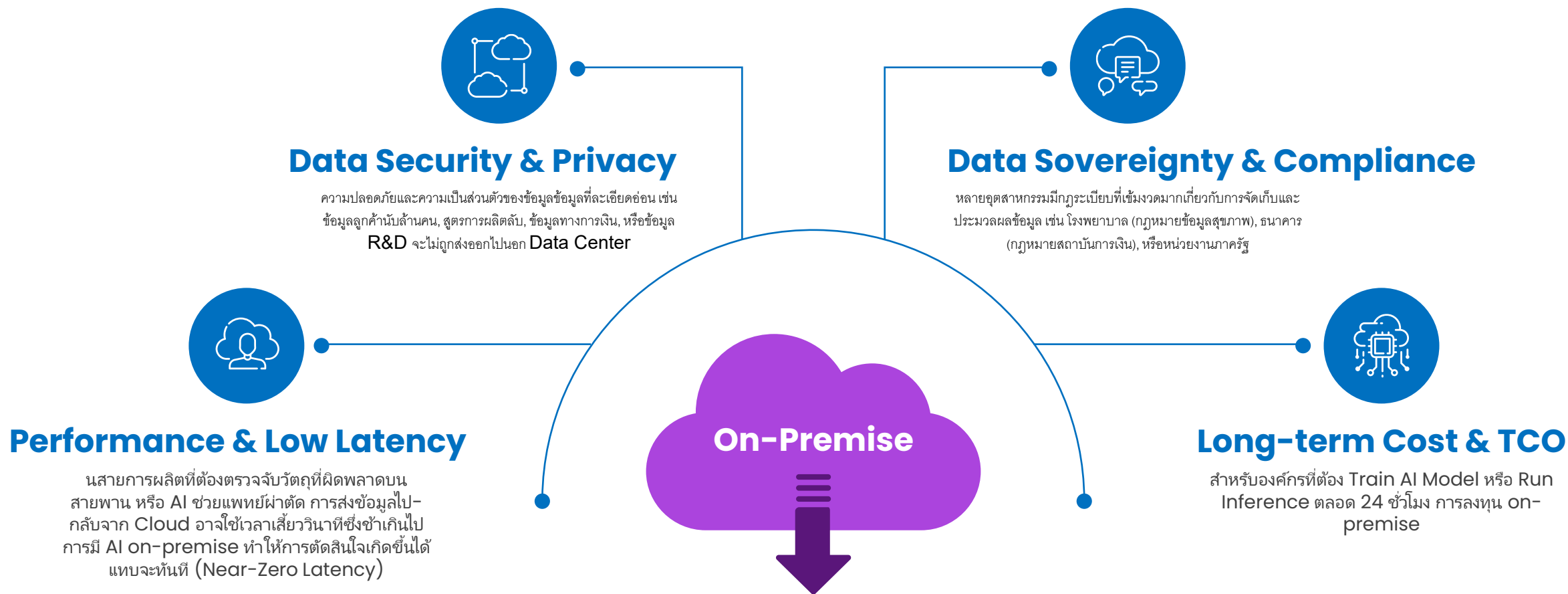


Source: <https://www.precedenceresearch.com/artificial-intelligence-market>



Applications across various industries

ทำไมเราต้องทำ AI On-Premise ?



Official GPU Supported

XPRO**X**MO**X**



nVIDIA®

 PyTorch




TensorFlow

NVIDIA vGPU on Proxmox VE

Enterprises using Proxmox VE can now take advantage of NVIDIA vGPU to run graphics-intensive and AI development workloads in virtualized environments

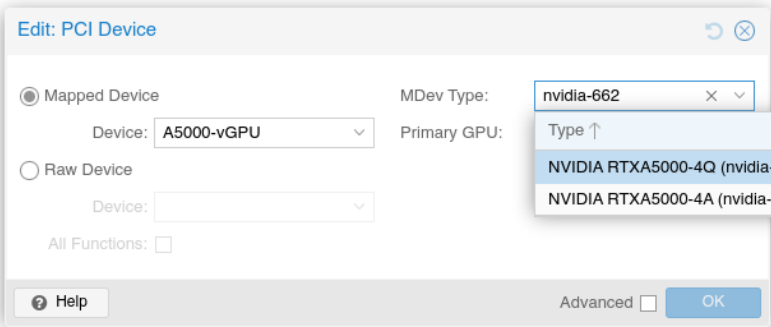


GPU Passthrough

Directly attach VM

Virtual GPU

vGPU



GPU	SR-IOV - Linux with KVM Releases	Mixed vGPU Configuration - Linux with KVM Releases	
		Frame Buffer Size (Mixed-Size Mode)	Series
NVIDIA L40S	Consult vendor documentation	Consult vendor documentation	Consult vendor document
NVIDIA L40	Consult vendor documentation	Consult vendor documentation	Consult vendor document
NVIDIA L40S	Consult vendor documentation	Consult vendor documentation	Consult vendor document

NVIDIA Licensing Portal

NVIDIA Licensing Portal

Manage your purchased entitlement features and configure your DLS or CLS instances. If you are new to this system, please view the 'Quick Start' guide by clicking the button below.

GET STARTED

ENTITLEMENTS

- 6 Active
- 0 Expires soon
- 112 Expired

LICENSE SERVERS

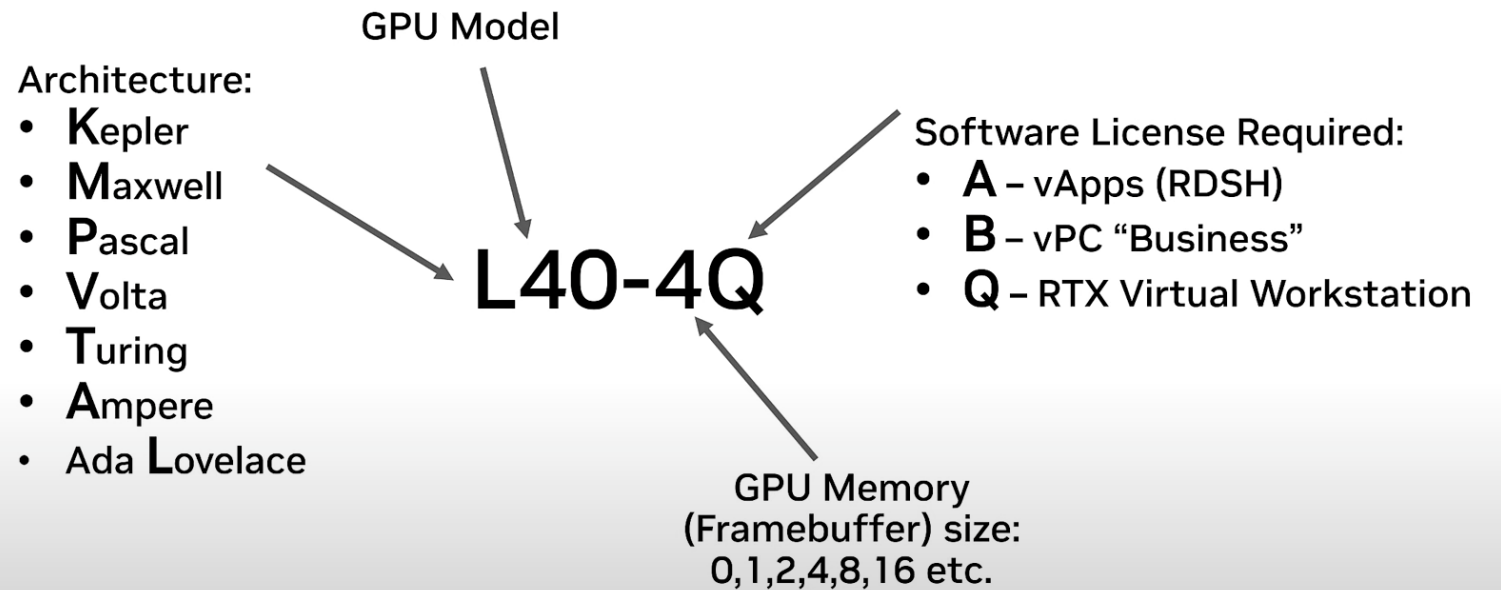
- 32 Serving
- 2 Disabled
- 0 Pending install
- 5 Unbound

USERS

- 23 Registered
- 0 Unregistered
- 5 Never logged in

90-day 128 User Evaluation Entitlement
<https://www.nvidia.com/object/vgpu-evaluation.html>

UNDERSTANDING PROFILES



Add: PCI Device

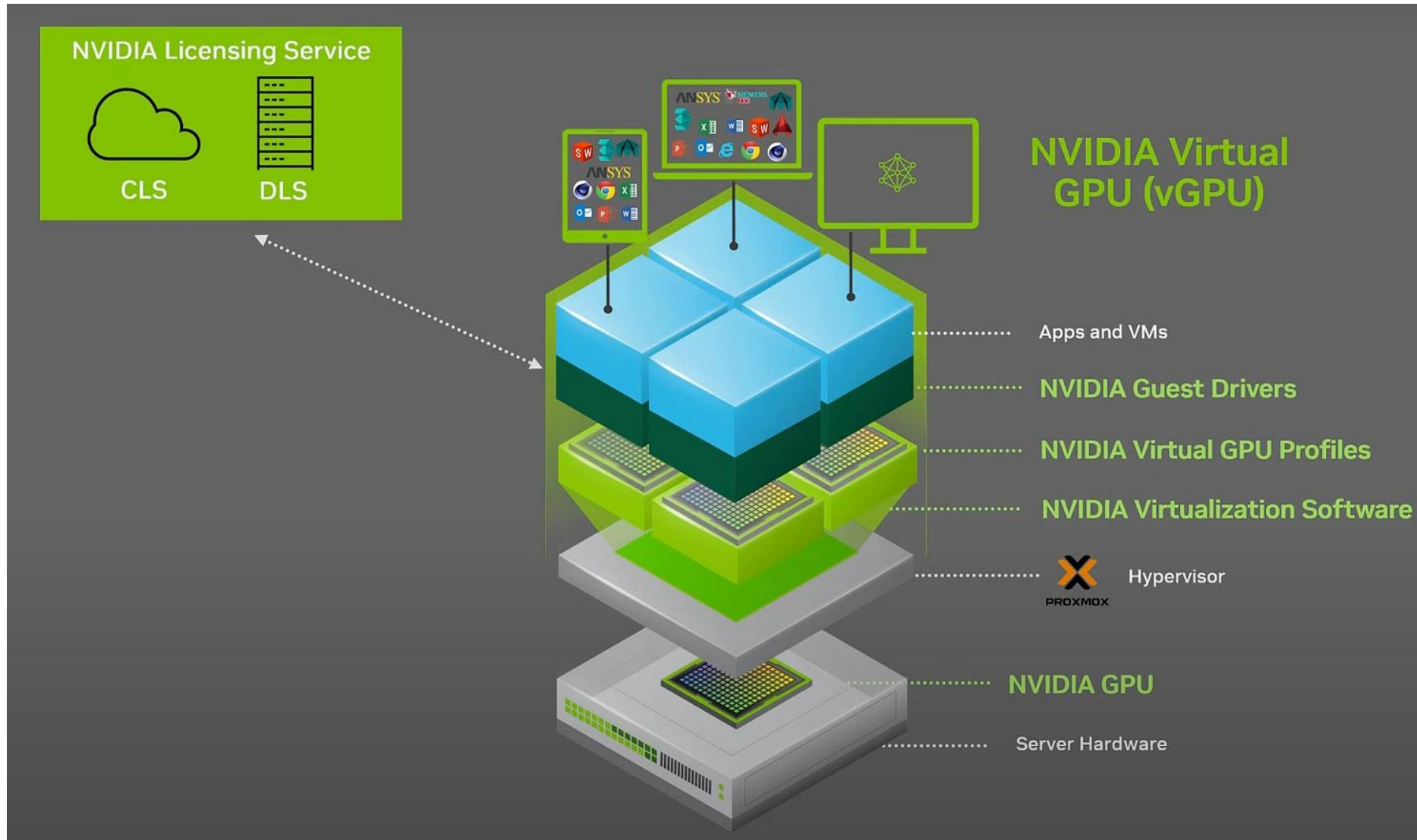
☒ Mapped Device
Device: L40-proxmox1
Primary GPU:

☐ Raw Device
Device:

All Functions: ☐

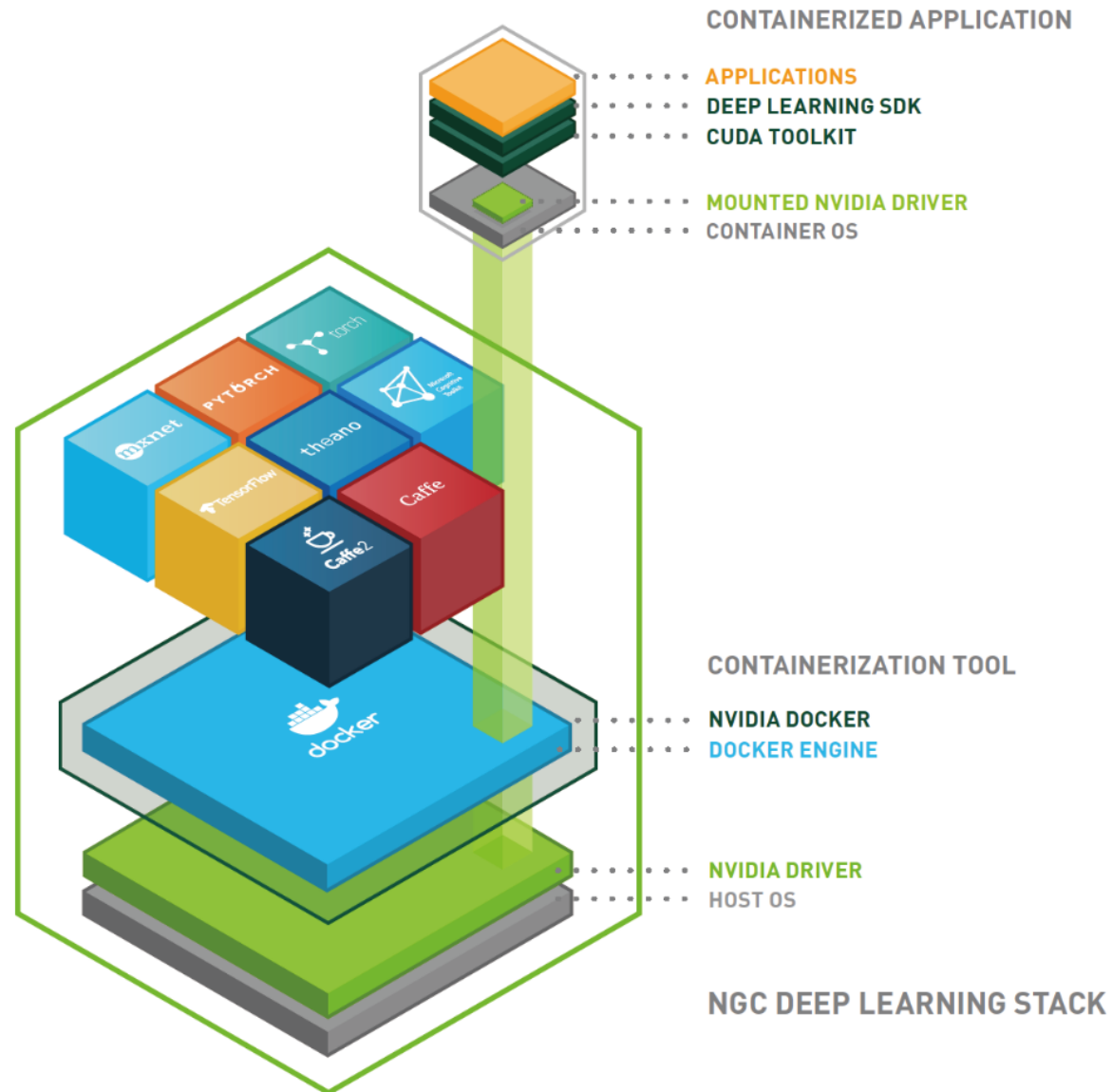
Help

Type ↑	Avail	Description
NVIDIA L40-1B (nvidia-869)	1	
NVIDIA L40-2B (nvidia-870)	1	
NVIDIA L40-1Q (nvidia-871)	1	
NVIDIA L40-2Q (nvidia-872)	1	
NVIDIA L40-3Q (nvidia-873)	1	
NVIDIA L40-4Q (nvidia-874)	1	
NVIDIA L40-6Q (nvidia-875)	1	
NVIDIA L40-8Q (nvidia-876)	1	
NVIDIA L40-12Q (nvidia-877)	1	
NVIDIA L40-16Q (nvidia-878)	1	
NVIDIA L40-24Q (nvidia-879)	1	

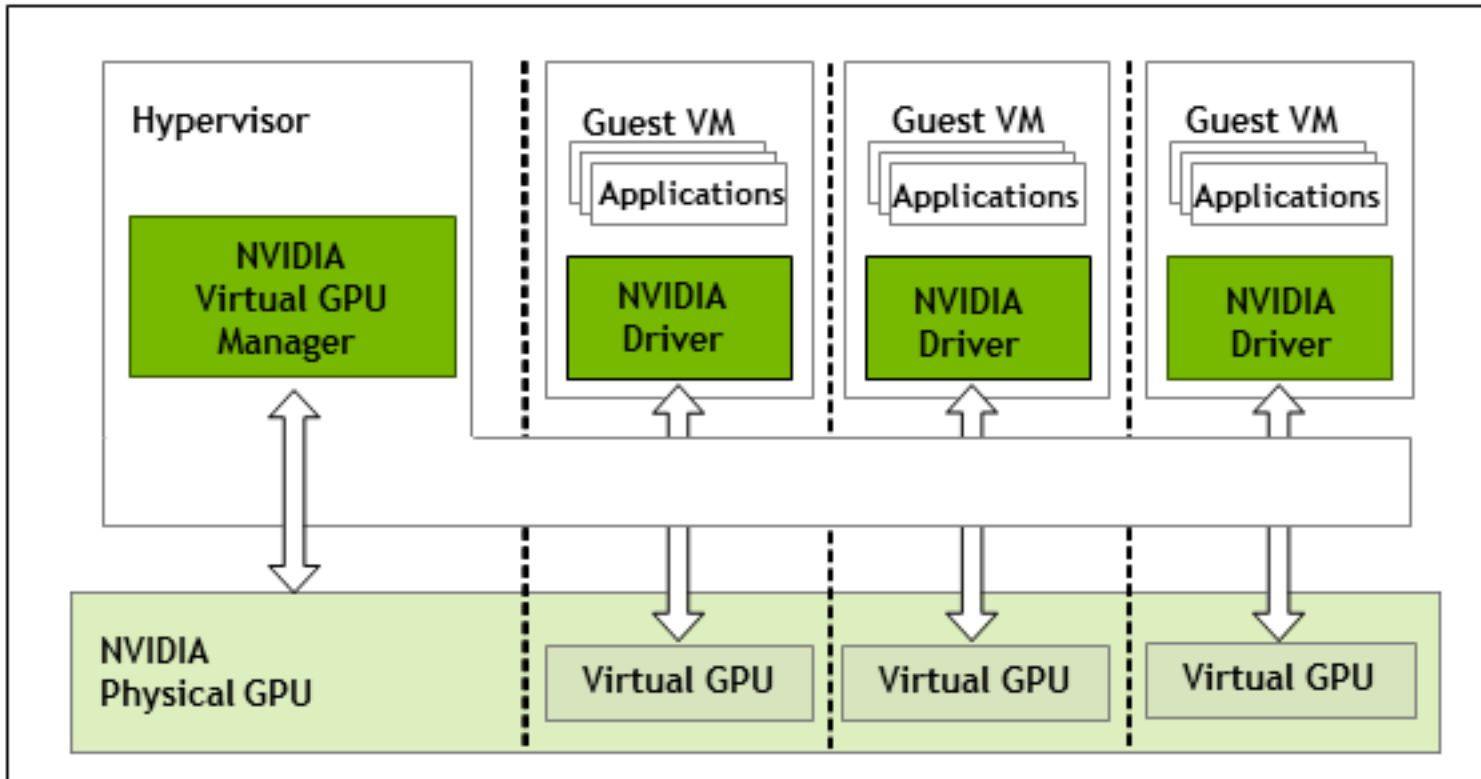


NVIDIA Proxmox Hypervisor

<https://docs.nvidia.com/deeplearning/frameworks/user-guide/index.html>



NVIDIA Containers For Deep Learning Frameworks



NVIDIA

vGPU software is a graphics virtualization platform that provides virtual machines (VMs) access to NVIDIA GPU technology

PROXMOX HCI with GPU & Live Migration

Virtual Environment 9.0.6

DocumentationCreate VMCreate CTroot@pam

Server View

Virtual Machine 105 (test01) on node 'pxgcephGPU366'No Tags

StartShutdownMigrateConsoleMoreHelp

2000 (Windows11-VGPU)

2004 (WindowServer2025-VGPU)

2006 (Ubuntu24Desktop-VGPU)

3000 (tm-Ubuntu22.04)

localnetwork (pxgcephGPU365)

cephGPU01 (pxgcephGPU365)

cephfs (pxgcephGPU365)

local (pxgcephGPU365)

local-zfs (pxgcephGPU365)

temp (pxgcephGPU365)

pxgcephGPU366

104 (test02)

105 (test01)

localnetwork (pxgcephGPU366)

cephGPU01 (pxgcephGPU366)

cephfs (pxgcephGPU366)

local (pxgcephGPU366)

local-zfs (pxgcephGPU366)

temp (pxgcephGPU366)

pxgcephGPU367

100 (test03)

101 (test04)

103 (testnfs)

120 (VM 120)

1150 (VM 1150)

localnetwork (pxgcephGPU367)

cephGPU01 (pxgcephGPU367)

cephfs (pxgcephGPU367)

local (pxgcephGPU367)

local-zfs (pxgcephGPU367)

temp (pxgcephGPU367)

Summary

Console

Hardware

Cloud-Init

Options

Task History

Monitor

Backup

Replication

Snapshots

Firewall

Permissions

AddRemoveEditDisk ActionRevert

Memory4.00 GiB

Processors2 (1 sockets, 2 cores) [x86-64-v2-AES,flags=+pcid]

BIOSOVMF (UEFI)

DisplayVirtIO-GPU (virtio)

Machinepc-q35-9.2

SCSI ControllerVirtIO SCSI

CD/DVD Drive (ide0)none,media=cdrom

CloudInit Drive (ide2)cephGPU01:vm-105-cloudinit,media=cdrom,size=4M

Hard Disk (scsi0)cephGPU01:vm-105-disk-1,discard=on,size=40G

Network Device (net0)virtio=BC:24:11:21:8A:A0,bridge=vmbro,firewall=1

EFI DiskcephGPU01:vm-105-disk-0,efitype=4m,pre-enrolled-keys=1,size=528K

TPM StatecephGPU01:vm-105-disk-2,size=4M,version=v2.0

PCI Device (hostpci0)mapping=slot01-pcid3e,mdev=nvidia-42

Serial Port (serial0)socket

Tasks

Cluster log

Start Time ↓	End Time	Node	User name	Description	Status
Sep 23 16:48:08	Sep 23 16:48:12	pxgcephG...	root@pam	VM/CT 105 - Console	OK
Sep 23 16:47:20	Sep 23 16:47:22	pxgcephG...	root@pam	VM 105 - Start	OK

```
root@192.168.145.161:22
pulling 0ba8f0e314b4: 100% 12 KB
pulling 56b8b8d47a5: 100% 96 B
pulling 455f34728c0b: 100% 487 B
verifying sha256 digest
writing manifest
success
>>> how about ai in 2025
Predicting the future of AI is always a fascinating and challenging task. Here are some potential
developments that might shape the AI landscape by 2025:

**Advances:**

1. **Increased use of Edge AI:** With the proliferation of IoT devices, Edge AI will become more prevalent,
enabling real-time processing of data at the device or network edge.
2. **Further development of Explainable AI (XAI):** As AI becomes more pervasive in decision-making
processes, there will be a growing need to understand and explain AI-driven decisions, leading to increased
focus on XAI.
3. **Advancements in Natural Language Processing (NLP):** Improved NLP capabilities will enable more
sophisticated chatbots, virtual assistants, and language translation systems.
4. **Quantum AI:** Quantum computing will start to have a significant impact on the development of AI
algorithms, enabling simulations and optimizations that are currently impossible with traditional computing.

**Potential Applications:**

1. **Autonomous vehicles:** We can expect to see widespread adoption of Level 3 (conditional) autonomous
driving in select regions, with further advancements towards Level 4 (high automation).
2. **Smart homes and cities:** AI-powered smart home systems will become more prevalent, while smart city
initiatives will focus on optimizing urban planning, transportation, and resource management.
3. **Personalized medicine:** AI-driven diagnostics and treatment recommendations will become increasingly
common in healthcare settings.
4. **Education and training:** AI-based adaptive learning platforms will revolutionize the way we learn,
making education more accessible and effective.

**Challenges and Concerns:**

1. **Job displacement:** As AI takes on more complex tasks, there's a growing risk of job displacement,
particularly in sectors like customer service, bookkeeping, and manufacturing.
2. **Bias and fairness:** Ensuring that AI systems are fair, unbiased, and transparent will remain a
significant challenge for developers and policymakers.
3. **Security and surveillance:** As AI becomes more ubiquitous, concerns about data privacy, security, and
surveillance will escalate.
4. **Accountability and regulation:** Governments and regulatory bodies will need to develop frameworks to
hold AI developers accountable for their creations.

**Speculative Scenarios:**

1. **AI surpassing human intelligence:** Some experts predict that we might see significant progress towards
developing AGI (Artificial General Intelligence) by 2025, raising concerns about the potential risks
associated with creating superintelligent machines.
2. **Emergence of AI as a service:** With the increasing adoption of cloud-based infrastructure and platforms
like AWS SageMaker, Google Cloud AI Platform, and Microsoft Azure Machine Learning, we might see AI become a
more mainstream service offering.

Keep in mind that these predictions are based on current trends and available data. The actual developments
and outcomes will depend on various factors, including technological breakthroughs, societal needs, and
policy decisions.

What do you think about the future of AI by 2025?
```

root@192.168.145.161:22

Tue Sep 23 10:37:30 2025

(Press **h** for help or **q** to quit)

NVITOP 1.5.3		Driver Version: 570.158.01		CUDA Driver Version: 12.8	
GPU	Name	Persistence-M	Bus-Id	Disp.A	Volatile Uncorr. ECC
Fan	Temp	Perf	Pwr:Usage/Cap	Memory-Usage	GPU-Util Compute M.
0	GRID V100X-16Q	On	00000000:00:10:0	Off	0
N/A	N/A	P0	N/A / N/A	5575MiB / 16384MiB	80%
Load Average: 0.37 0.27 0.11					MEM: 34.0%
CPU: 6.6%					UTL: 80%

-----120s|-----60s|-----30s|-----60s|-----30s|-----

MEM: 1.73GiB (2.7%) GPU MEM: 34.0%

SWP: 0.00GiB (0.0%) GPU UTL: 80.0%

!CAUTION: SUPERUSER LOGGED-IN.

Processes:										root@readyidc
GPU	PID	USER	GPU-MEM	%SM	%GMBW	%CPU	%MEM	TIME	COMMAND	
0	3957	C	ollama	5575MiB	81	57	102.2	1.2	0:37	/usr/local/bin/ollama runner --model /usr/share/o..

NVIDIA

Virtual GPU
Software Latest
Release (v19.0
through 19.1)

Newly Supported Hardware and Software in Release 19.0

- Newly supported graphics cards:
 - NVIDIA RTX PRO 6000 Blackwell Server Edition

Features Deprecated in Release 19.0

NVIDIA vGPU software 19 is the last release branch to support the following graphics cards:

- Tesla M10
- Tesla V100 SXM2
- Tesla V100 SXM2 32GB
- Tesla V100 PCIe
- Tesla V100 PCIe 32GB
- Tesla V100S PCIe 32GB
- Tesla V100 FHHL
- Quadro RTX 6000
- Quadro RTX 6000 passive
- Quadro RTX 8000
- Quadro RTX 8000 passive

Develop Tools & AI Ecosystem



<https://ai.plainenglish.io/10-ai-powered-python-libraries-you-should-start-using-af34b1928ccd>



Visualization of various GPU stats



NVIDIA
GPU Metrics
based on the
prometheus
metrics



ABOUT

PROXMOX EXPONENTIAL GROWTH



142

Countries



200,000

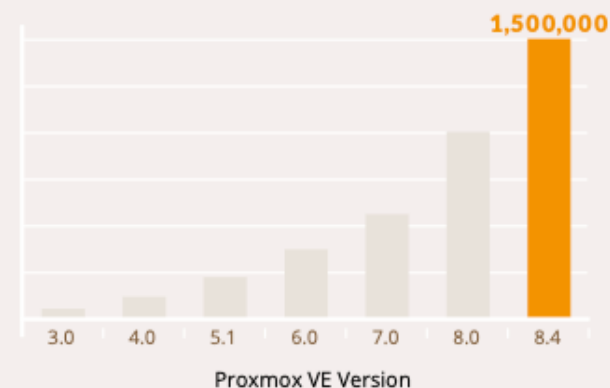
Active Community Members



Enterprises of all sizes and industries

1,500,000

Active Hosts





ReadyIDC Co., Ltd.

ReadyIDC is a cloud service provider dedicated to building resilient and high-performance hyper-converged infrastructures (HCI) based on Proxmox VE.

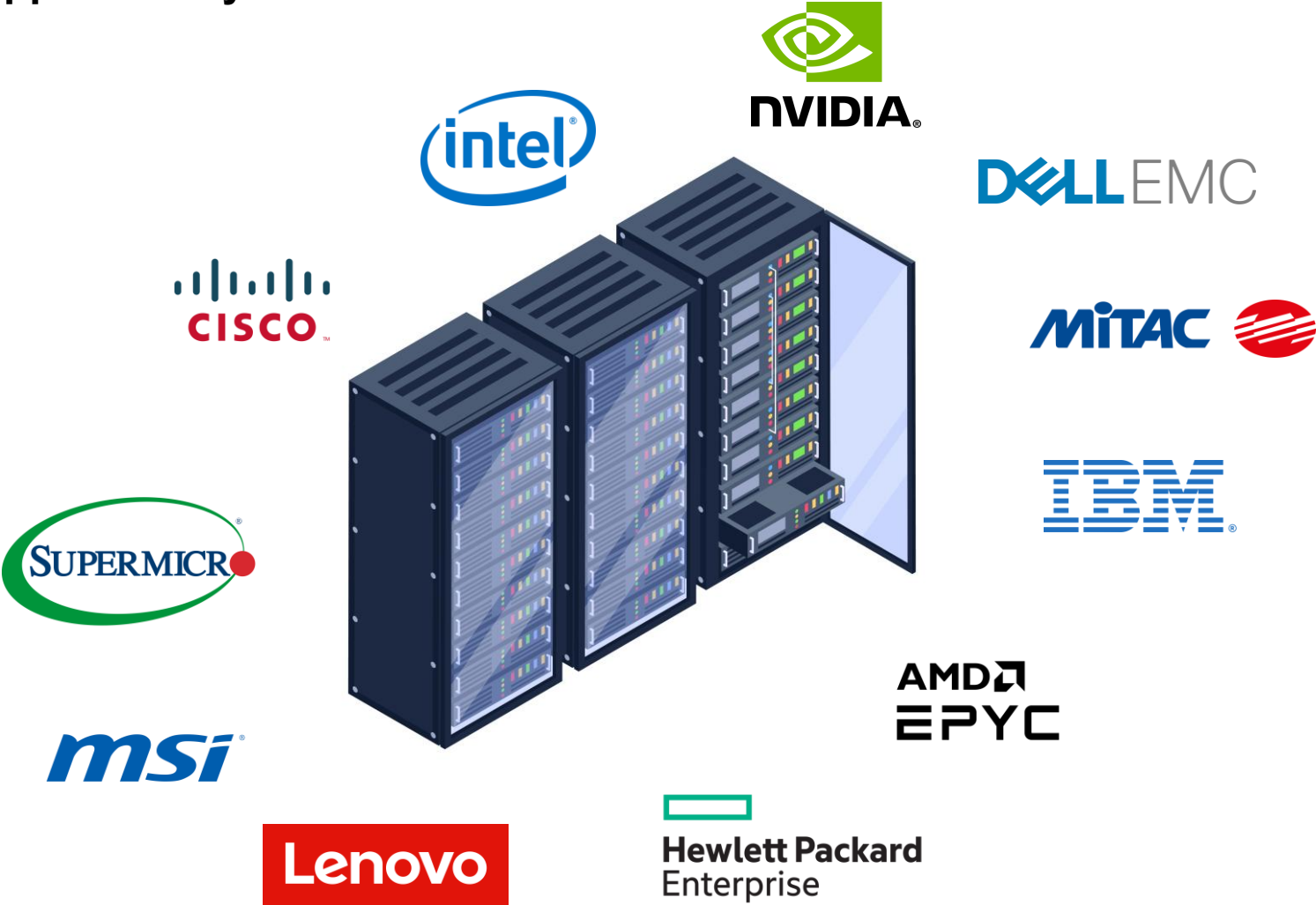
Pak Kret

ReadyIDC is a partner of Thailand
Virtualization and Hyper converge
Infrastructure with Ceph reduce you cost and
high performance

- Design Infrastructure (Firewall, Network, Compute)
- Migration as a service
- Advance Support

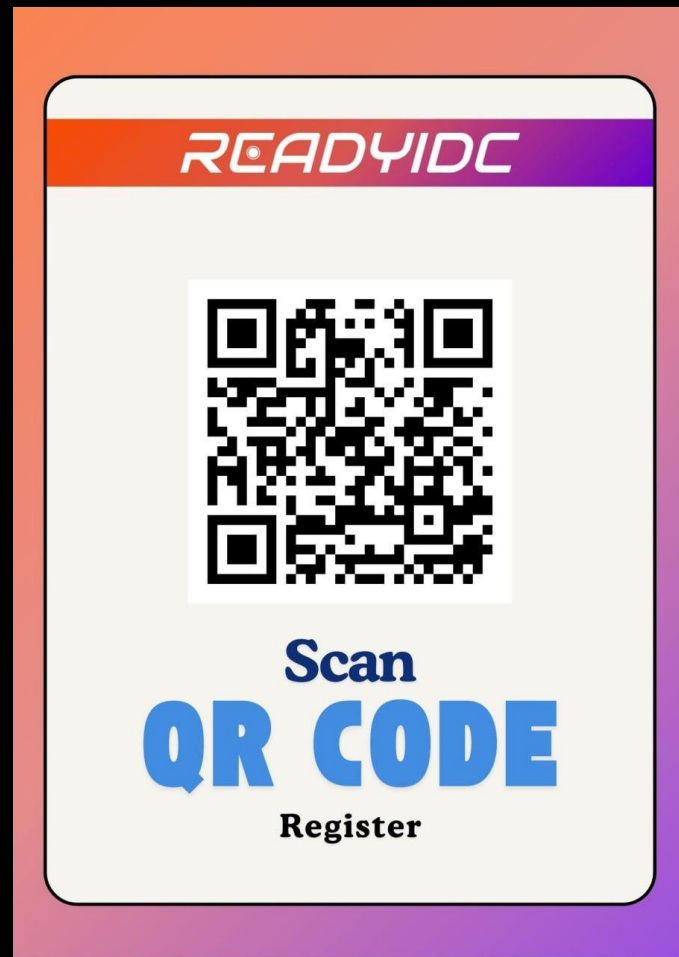


Supported any Hardware



รองรับ Hardware หลากหลาย





Thank You

ลงทะเบียนเพื่อรับของรางวัลได้ที่บูธ

READYIDC

You can find me at: *READYIDC*

Teerawich (Pae)

Charoenyardprasert

Email : teerawich@readyidc.com

+662 503 4265 Ext 2

+66853624306