

Adria Forum 2025

On premise AI with Cloud Foundation 9 L100

Marko Žuvanić
Territory Sales

Bruno Šunjić
System Engineer

14.10.2025.



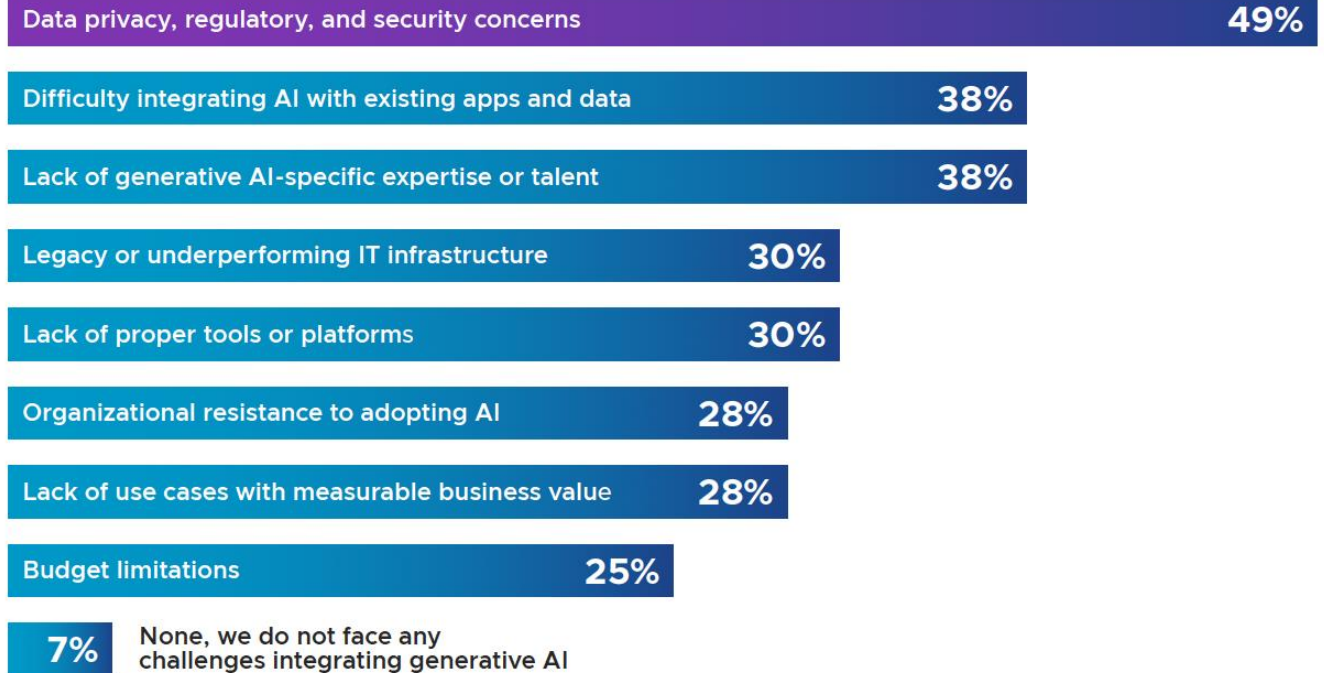
Private Cloud Outlook 2025 - The Cloud Reset

<https://www.vmware.com/docs/private-cloud-outlook-2025>

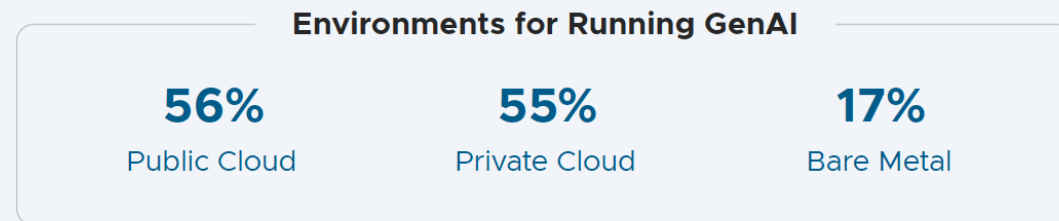
The Rise of Generative AI

Organizations are eager to harness GenAI. Only **2% report having no plans to adopt GenAI**. The remaining 98% are somewhere on the adoption curve, with **77% already running pilots or live deployments**.

Use cases being explored fall evenly across digital assistants, predictive analytics, and customer service initiatives. Still, **28% cite a shortage of use cases with clear, measurable business value as a barrier to GenAI adoption**.

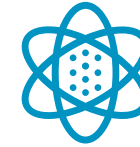


Primary Challenges for Adopting GenAI



Environments that organizations are using or planning to use for generative AI

PREDICTIVE AI



Data Scientists



Predictive Analytics



SPECIALIZED AI MODEL

Impact of Generative AI in the Enterprise

PREDICTIVE AI



Data Scientists



Predictive Analytics



SPECIALIZED AI MODEL

GENERATIVE AI



Marketing



Supply Chain



Sales



Human Resources



Customer Operations



Research and Development



Legal



Software Development



Manufacturing



Procurement



IT



Finance



NATURAL LANGUAGE INTERACTION

APIs



LARGE LANGUAGE MODELS

Impact of Generative AI in the Enterprise

GENERATIVE AI



NATURAL LANGUAGE INTERACTION

APIs

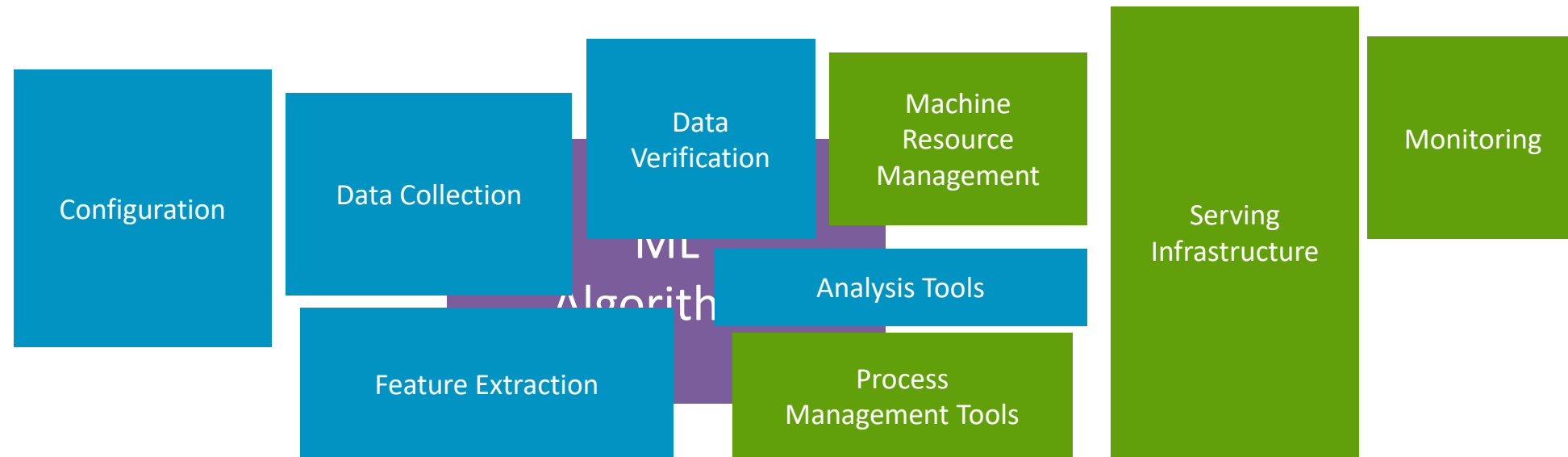
✓
LARGE LANGUAGE MODELS


\$4.4T
**annual
economic
value**

Source: McKinsey, The economic potential of generative AI:
The next productivity frontier, June 2023

Complexity of AI Environments

Infrastructure is Half of the Challenge



-  **Corp IT / Infrastructure Domain**
-  **Data Scientists / ML Engineer Domain**

Based on: Hidden Technical Debt in Machine Learning Systems, Scully, D., et al.
https://proceedings.neurips.cc/paper_files/paper/2015/file/86df7dcfd896fcdf2674f757a2463eba-Paper.pdf

Generative AI Terminology

Key definitions

Traditional or Predictive AI

Generative AI

Large Language Models (LLMs)

Training

Fine-Tuning

Customizing

Retrieval Augmented Generation (RAG)

Inferencing



LLM - Generative AI that is focused on generating text. Examples- GPT-5 (ChatGPT), LLaMA, Claude, Mistral

Training - Educating an AI model from scratch with a lot of data. Requires high costs, massive datasets and high infrastructural resources
Example- Train a LLM from scratch on millions or trillions of data points

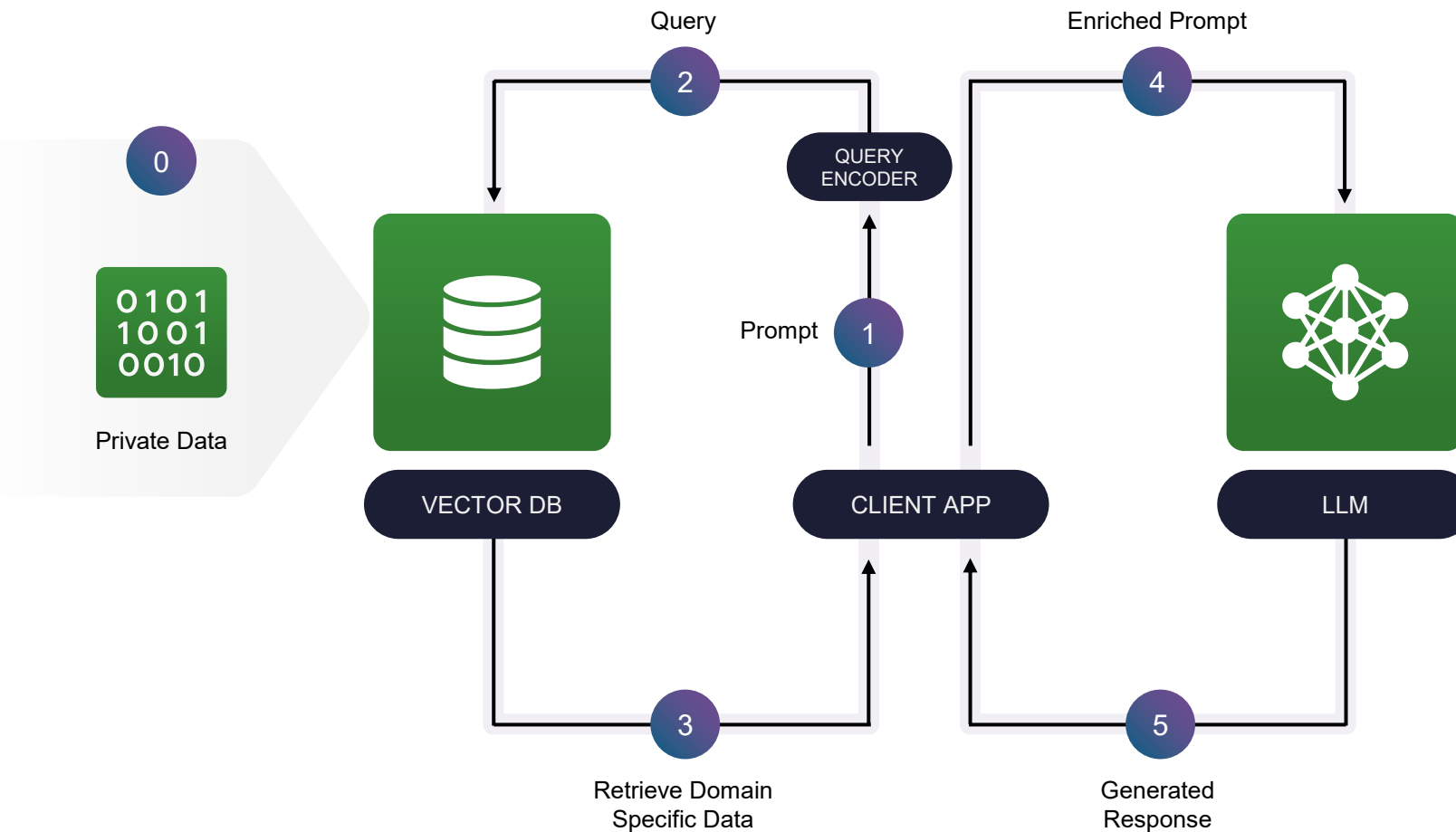
Fine-Tuning - Take a pretrained model and re-train it with your domain data
Example- Bank fine tunes an LLM with suspicious activity reports, regulatory filings and internal policy documents to better understand fraud patterns

Customizing - Adapting an AI model for your needs with better instructions, databases and more, without retraining it. e.g. RAG

Inferencing - Deploy a trained AI model to make predictions or generate outputs based on new input data

Retrieval Augmented Generation (RAG)

Intelligence + data



0 Data Loading
Feed your company's private data into the database

1 Prompt
The user generates the prompt
"How does vSAN ESA handle snapshots?"

2 Query
The prompt is forwarded to the Vector DB which looks for documents, articles, any type of information related to "How does vSAN ESA handle snapshots"

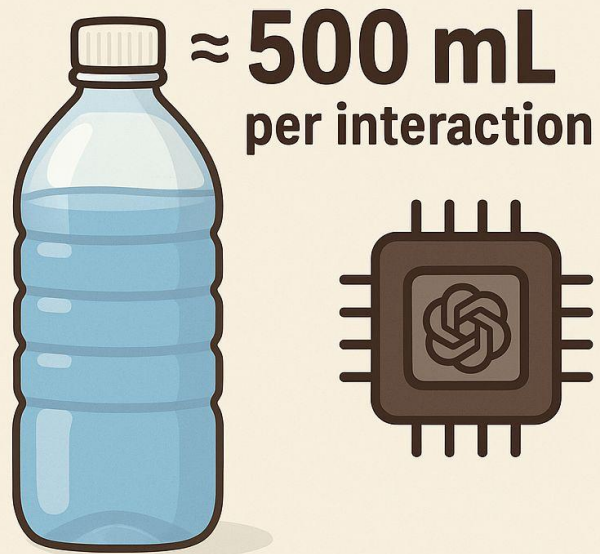
3 Retrieve Domain Specific Data
Information related to "How does vSAN ESA handle snapshots" is sent back to the client application

4 Merging Retrieval with Generation
The retrieved information is combined with the user prompt and sent as a new, enriched prompt to the Large Language Model (LLM)

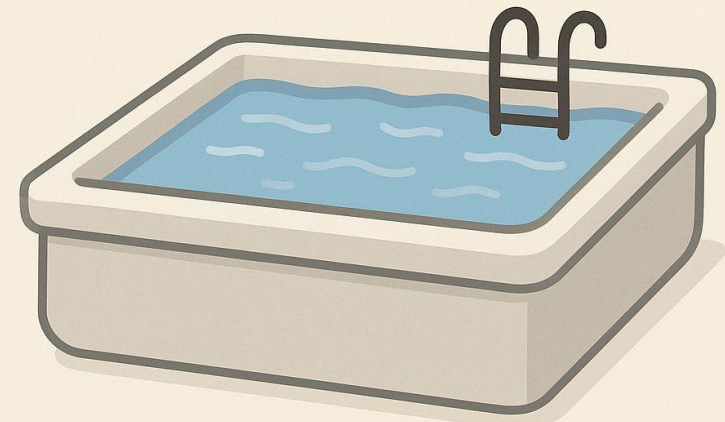
5 Generated Response
The LLM uses the enriched prompt (original prompt + retrieved information) and its creative writing skills to create a unique response

We often think of AI as “in the cloud” – but it’s deeply physical.
It’s powered by energy, cooled by water, and grounded in real-world infrastructure

WATER CONSUMPTION FOR AN AI PROMPT



WATER CONSUMPTION FROM GPT USAGE



1,073,5 liters

Professor Shaolei Ren at the University of California, Riverside

https://www.researchgate.net/publication/370202417_The_Environmental_Impact_of_AI_A_Case_Study_of_Water_Consumption_by_Chat_GPT



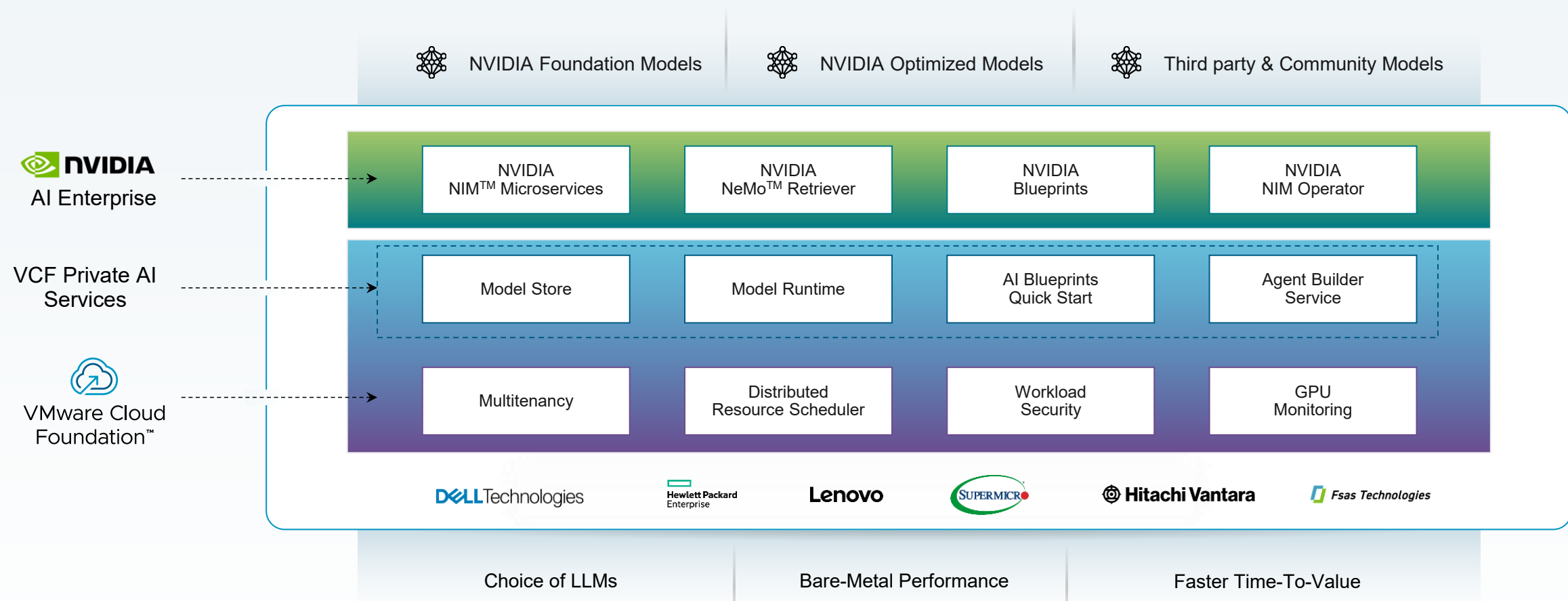
VMware Private AI

An architectural approach that balances the business gains from AI with the privacy and compliance needs of the organization.

VCF is now an AI-native platform

Private AI Services are now a part of VCF Subscription

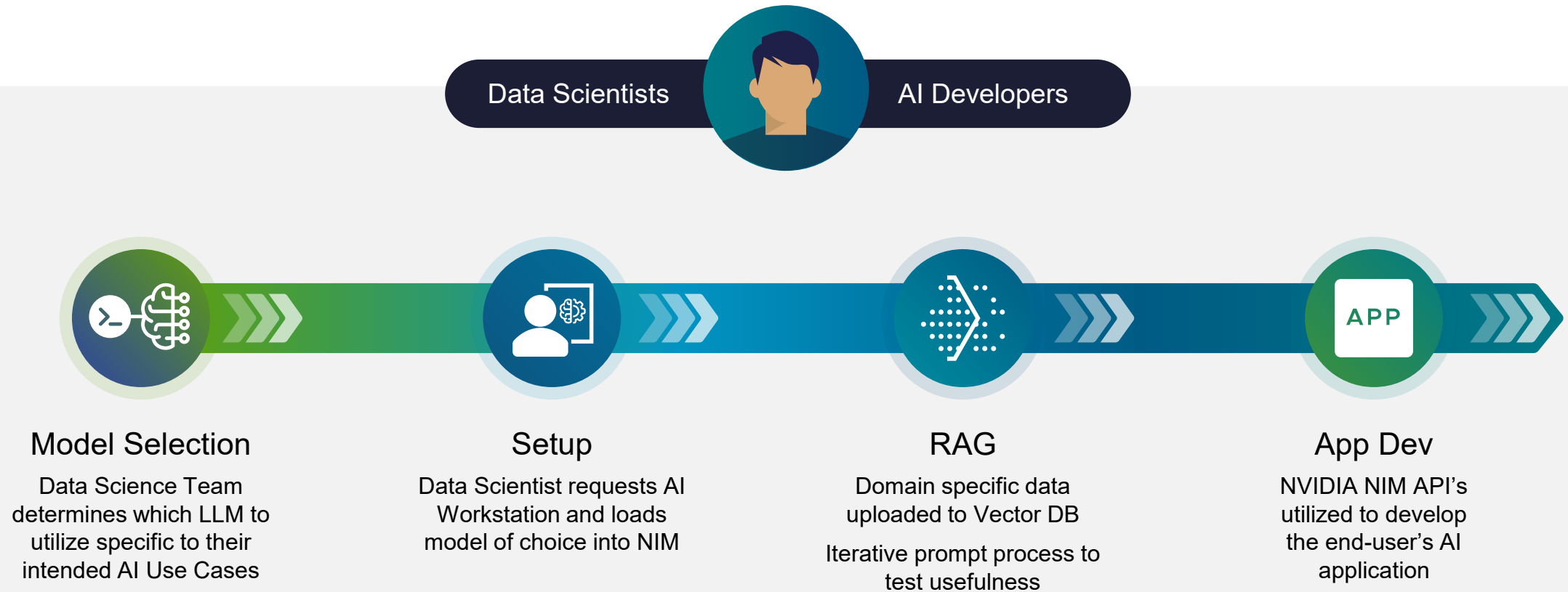
VMware Private AI Foundation with NVIDIA



Why VCF for Private AI?



A Day in the Life of a Generative AI App Developer



Teams of Data Scientists

Working in parallel on Models



Complexity of AI Environments

Infrastructure is half of the challenge



Corp IT | Infrastructure Domain



Process Automation/Management Tools

Monitoring



Serving Infrastructure



Machine Resource Management





Data Scientists | ML Engineer Domain

ML Models



Feature Extraction

Data Verification



Data Collection



Analysis Tools

Configuration



Deployment Options



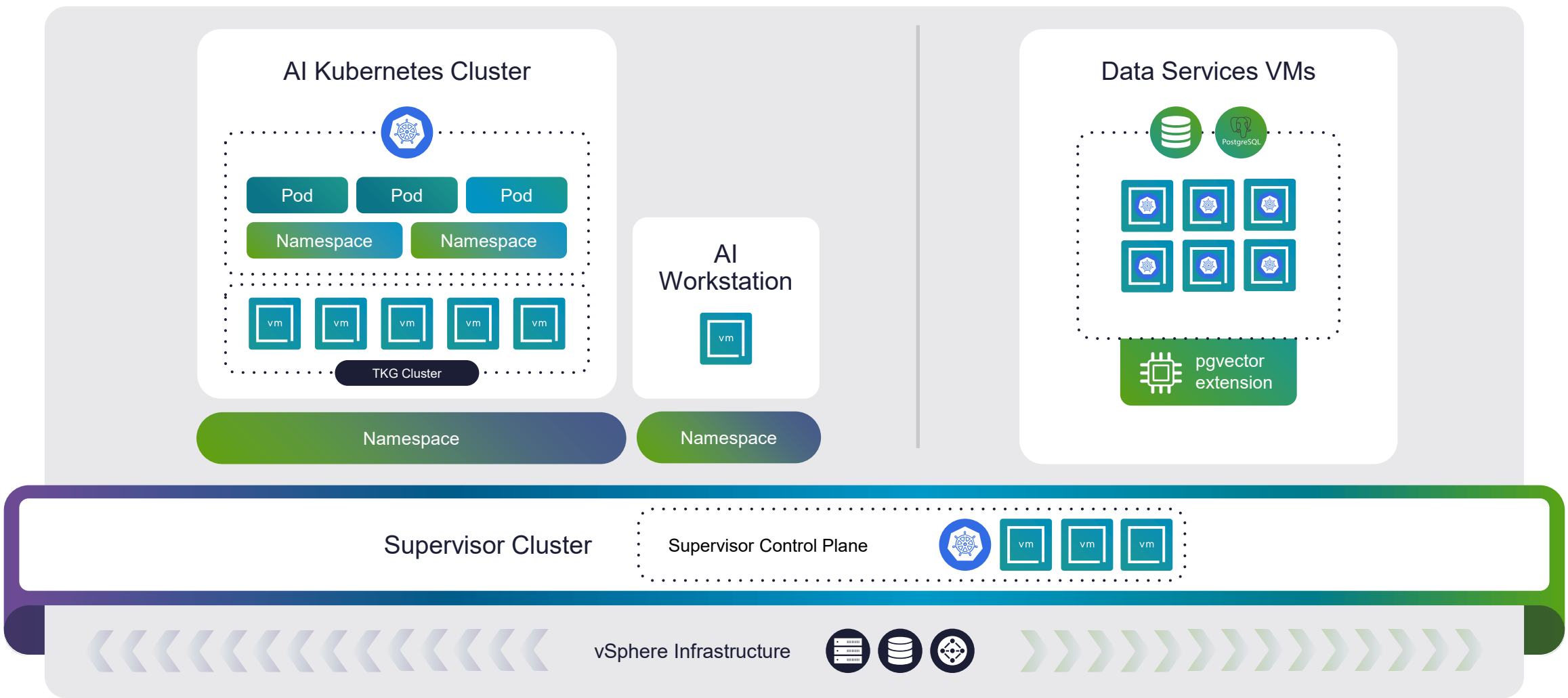
VMware Cloud Foundation

VMware Private AI Foundation with NVIDIA - Add-on VMware components

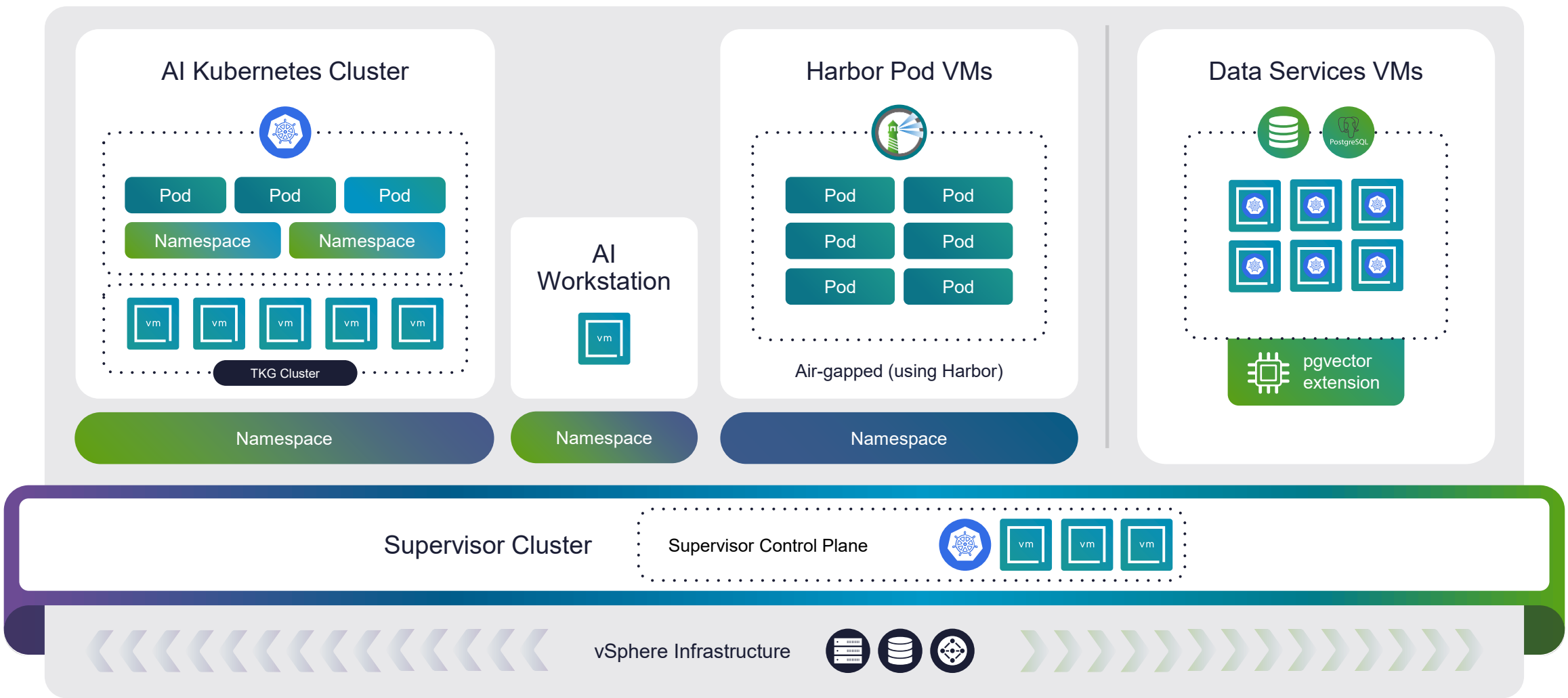


1 IaaS Control Plane formerly known as vSphere with Tanzu

VMware Private AI Foundation with NVIDIA - Deployments

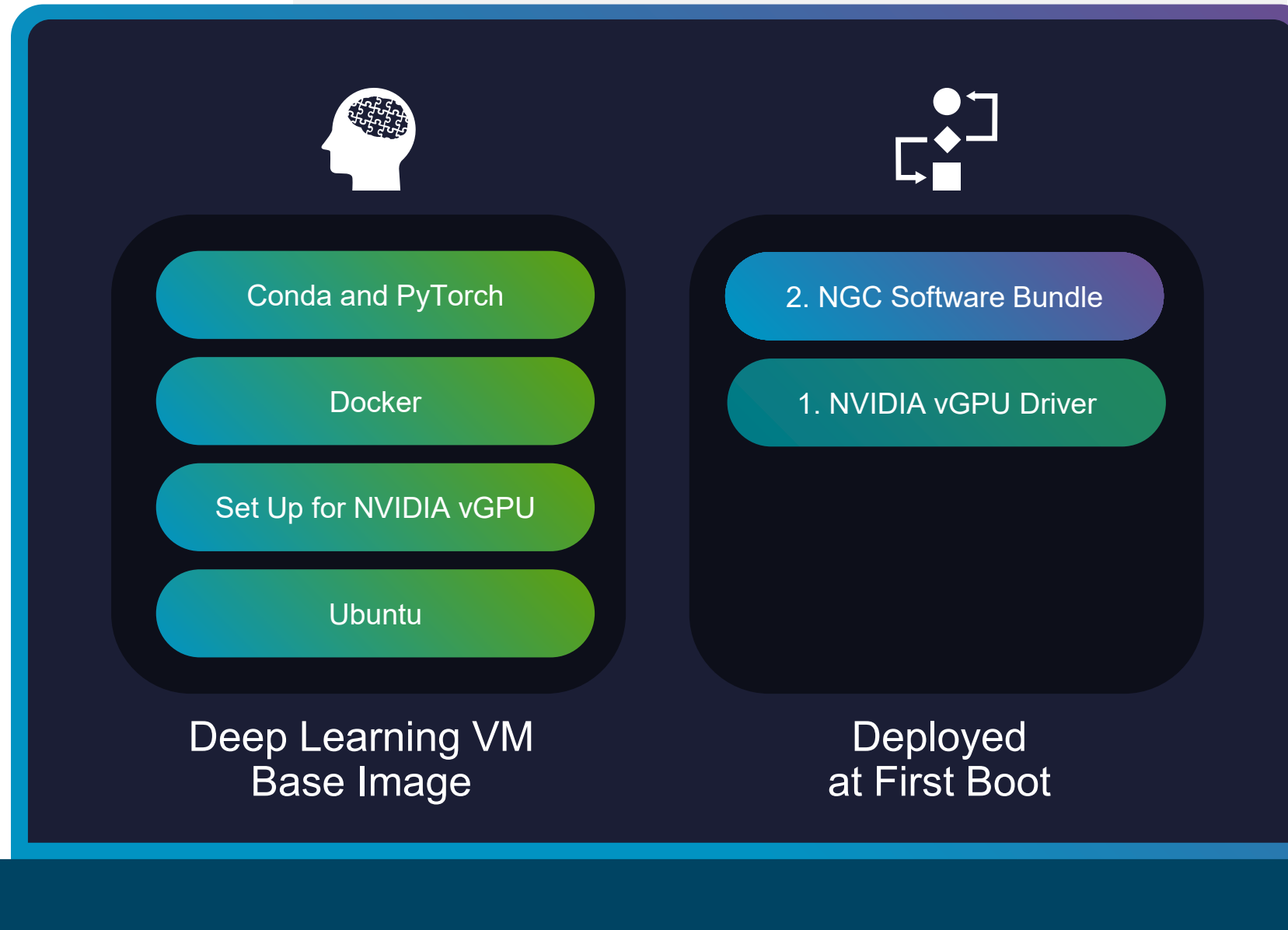


VMware Private AI Foundation with NVIDIA



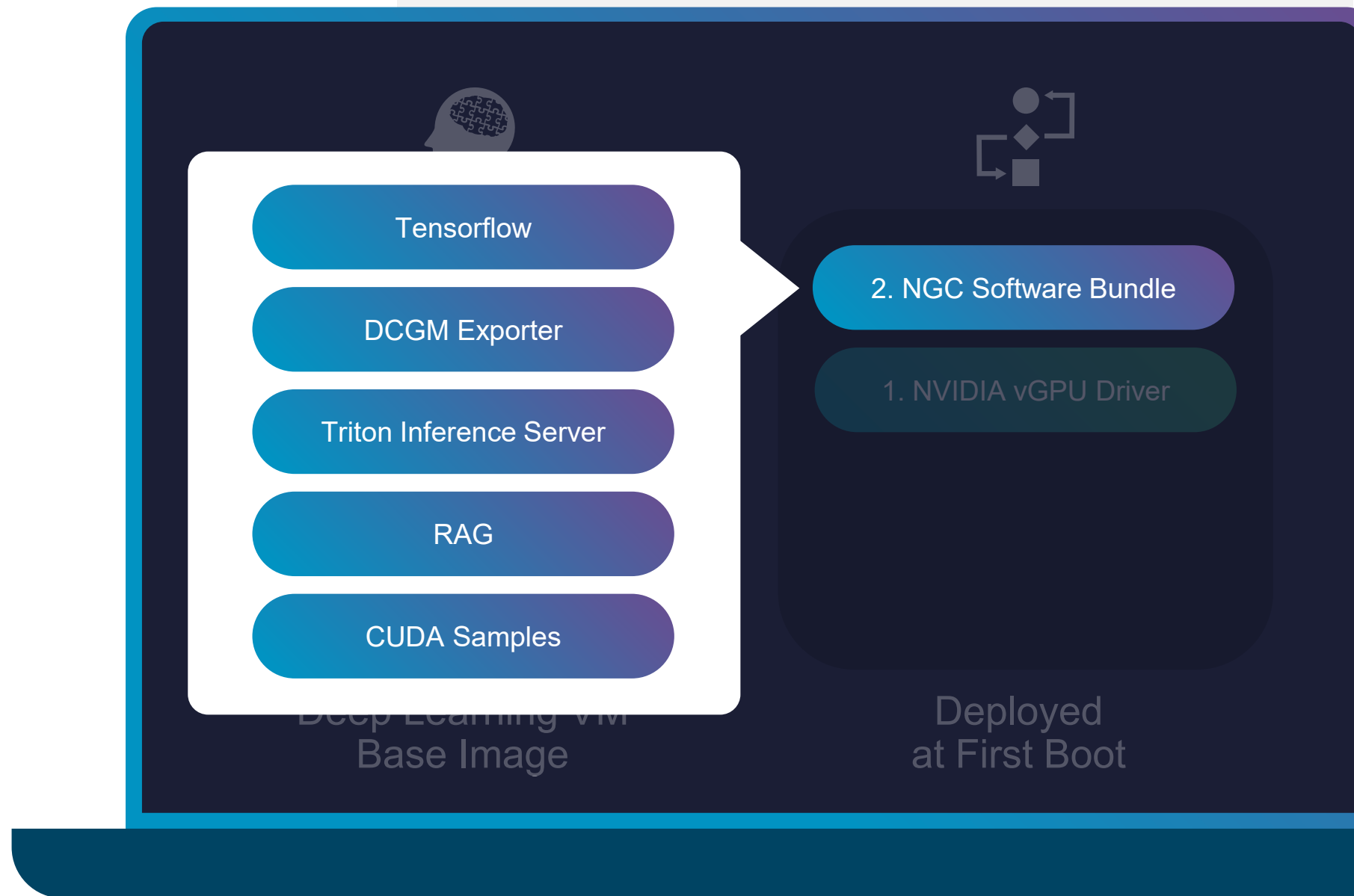
AI Workstation

Deep learning
virtual machine



AI Workstation

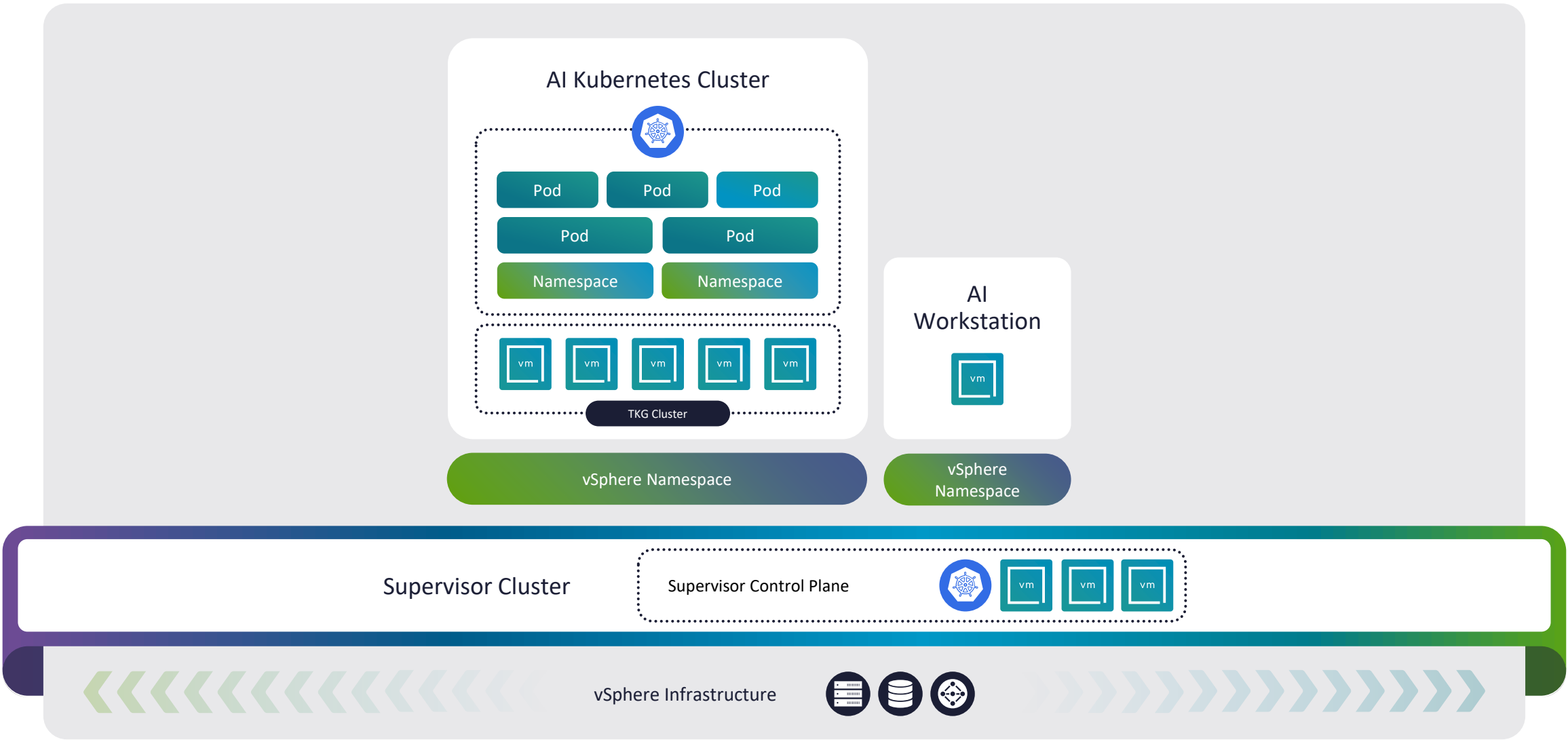
Deep learning
virtual machine



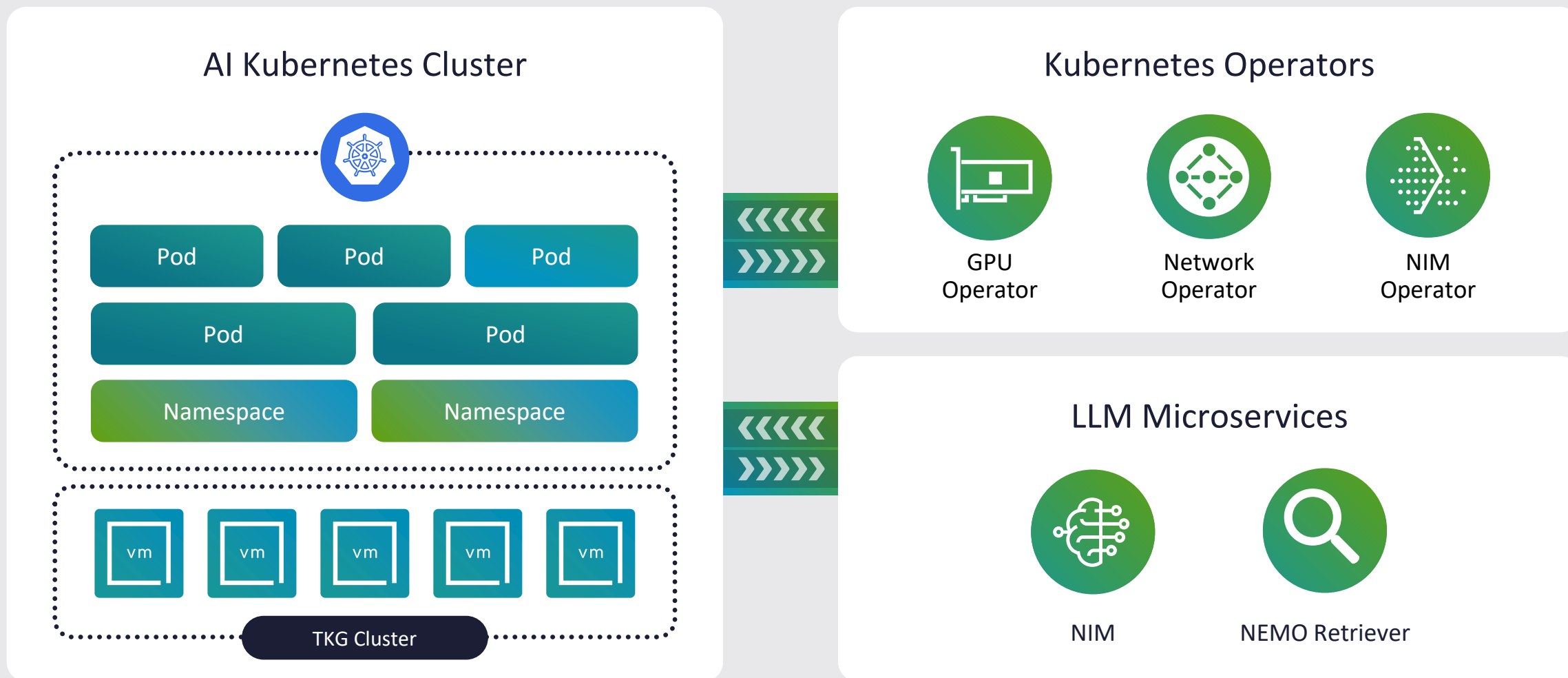
Deploy an AI Kubernetes Cluster



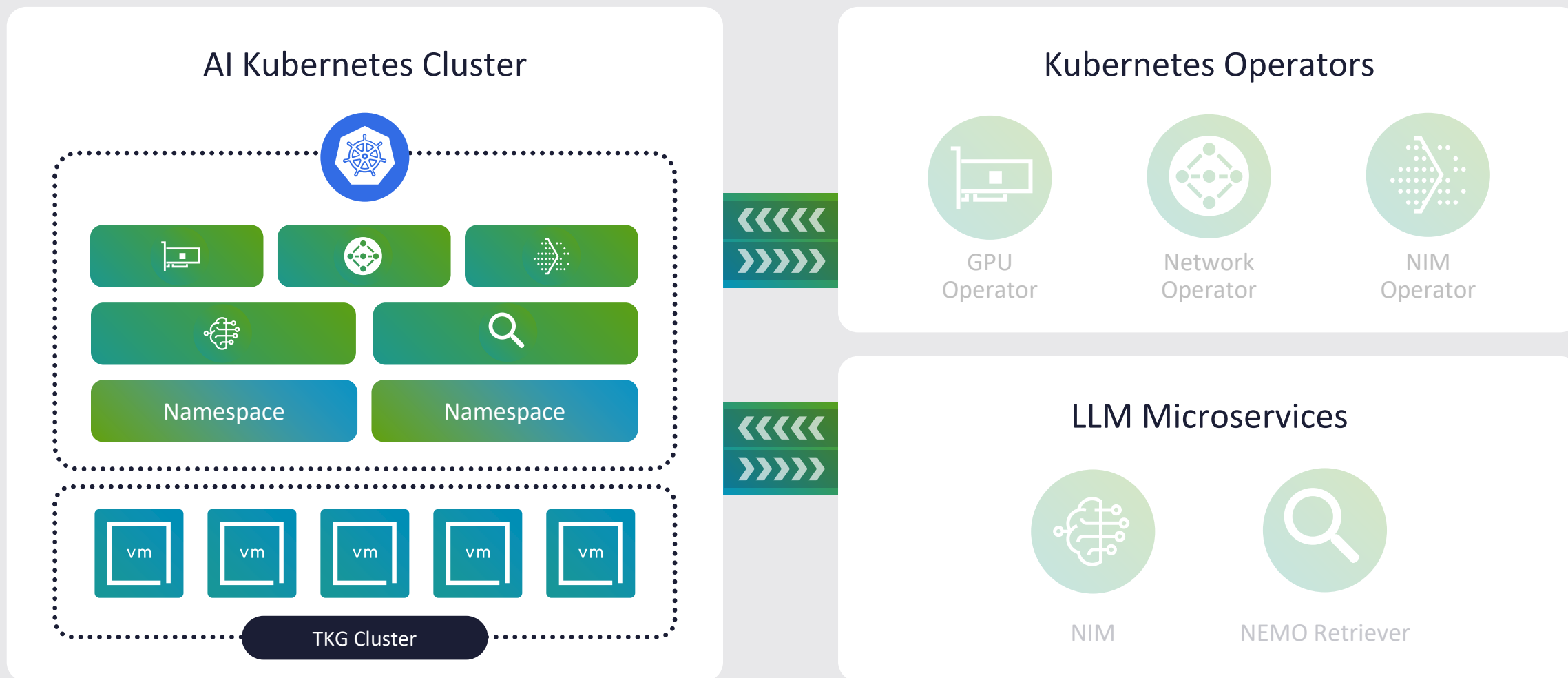
Deploying Microservices



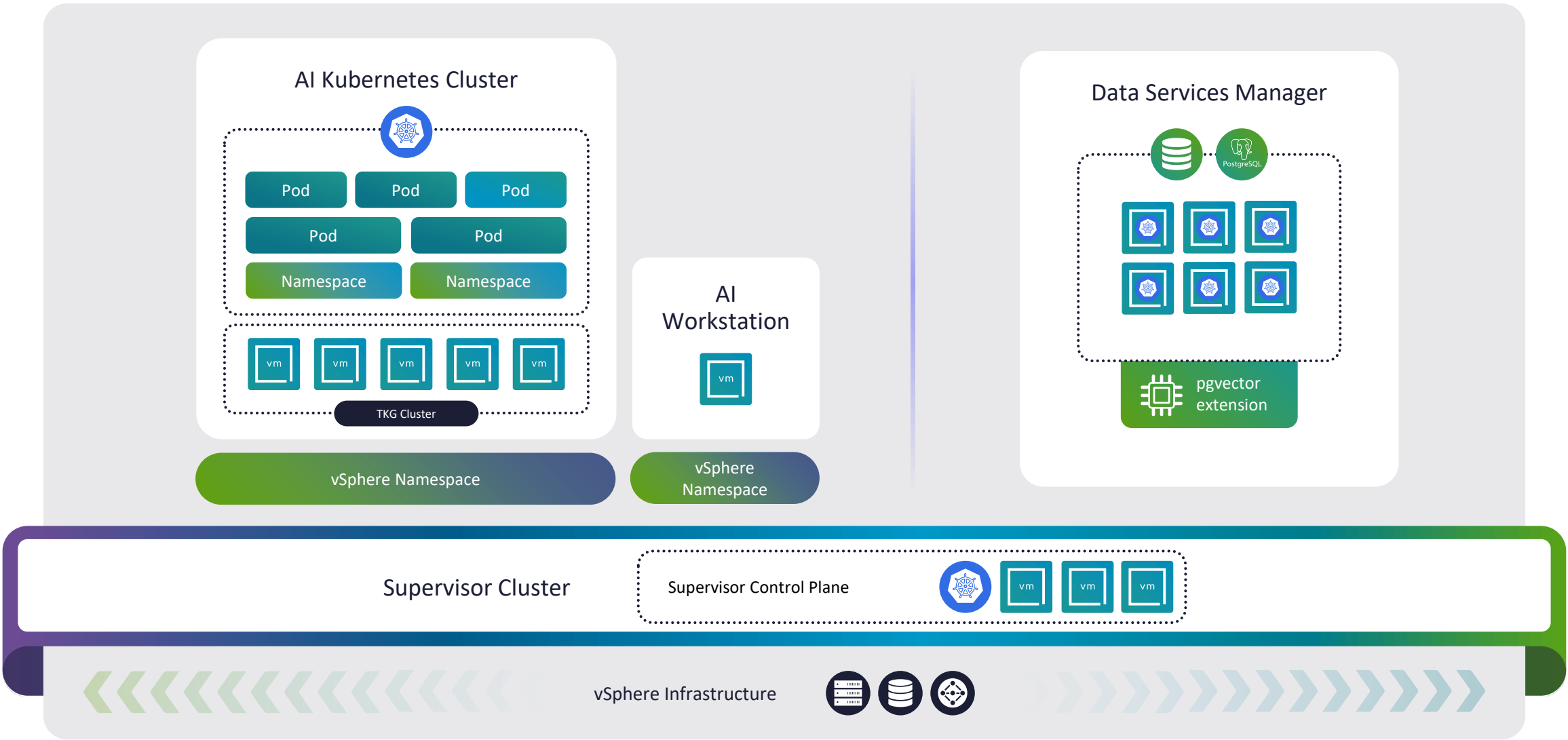
Operators and LLM Microservices (Containers)



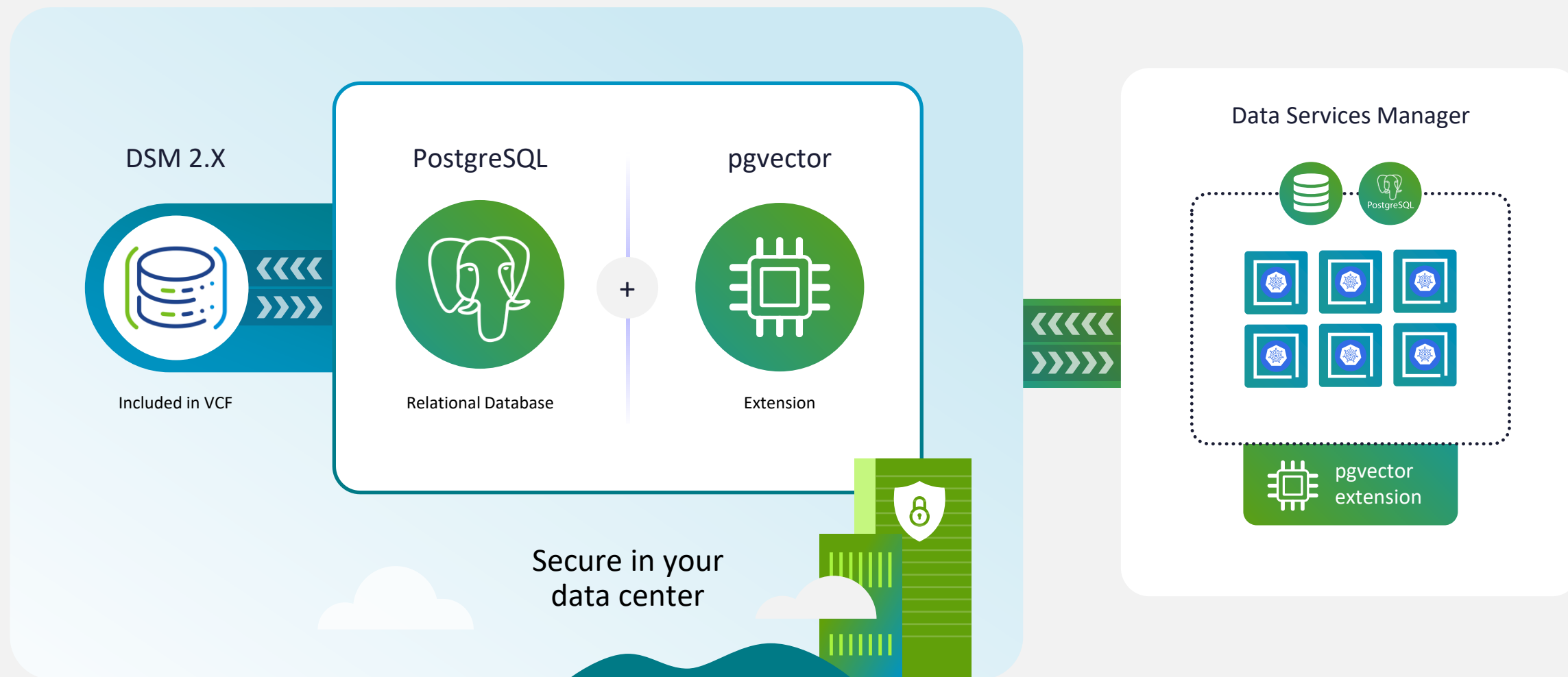
Operator Services



Operator Services



Data Services Manager Deploys PostgreSQL and pgvector

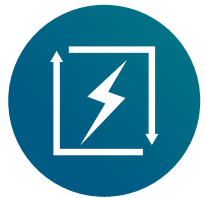


Backup and Availability

Features for vector database backup and replication



Replication scheme
with availability



Ensure it's all
backed up



Access control with
RBAC (role based
access control)



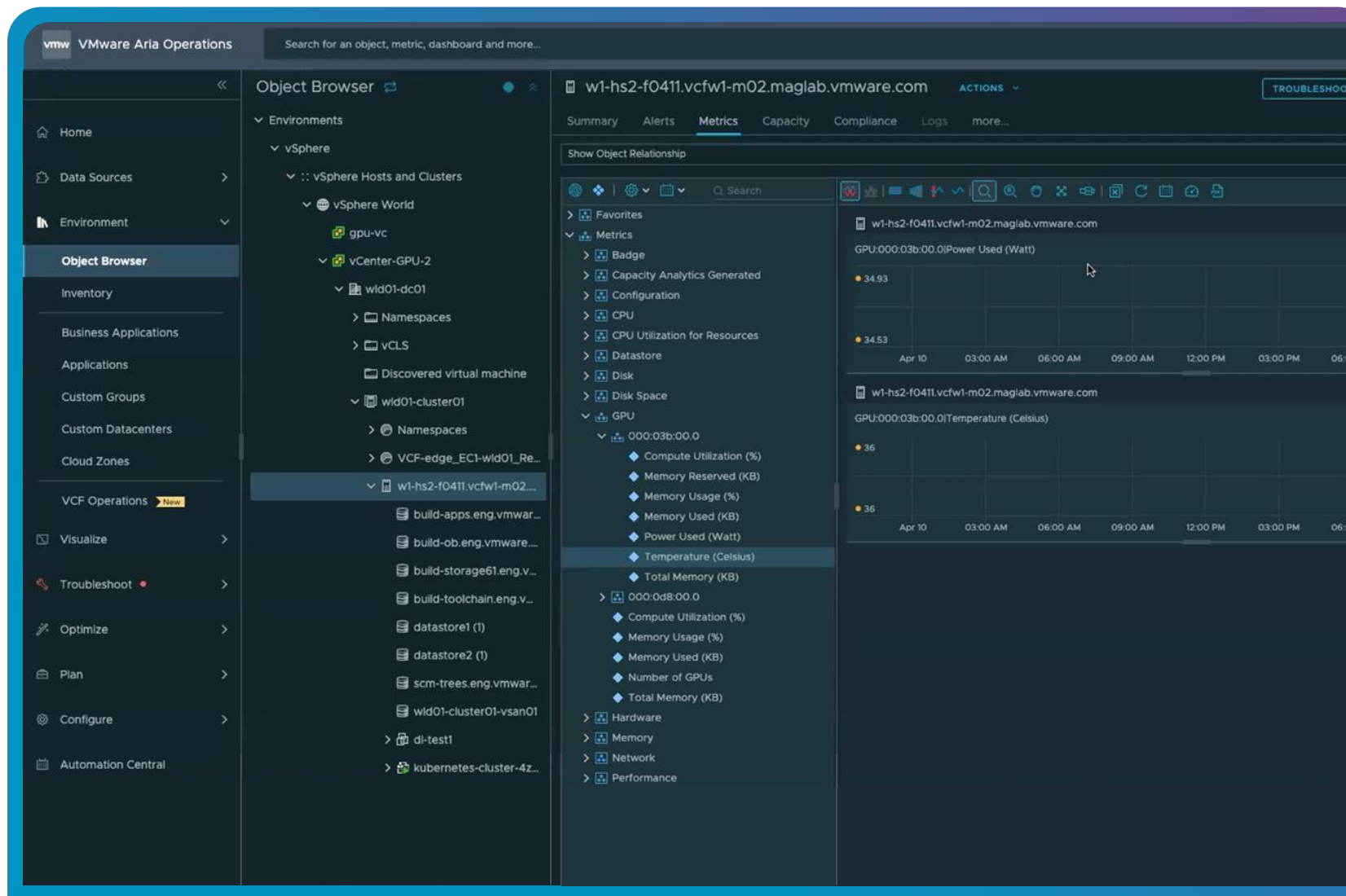
GPU Monitoring





GPU Health Monitoring

Enhanced GPU visibility with insight into compute and memory usage across hosts and clusters



Thank you for participating!

Not finished yet 😊

November & December

VMware Tech Talks winter edition

Series of afterwork gatherings with hot&cold drinks and hot topics

