



Подход **vmware**
by Broadcom

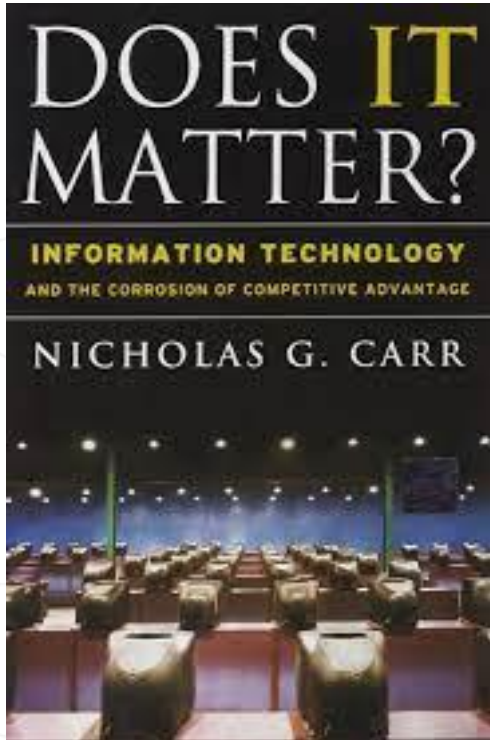
и TOP 5 важных технологий для реализации Private AI

March 2024

Alexander Starostin
VMware's Solutions Expert



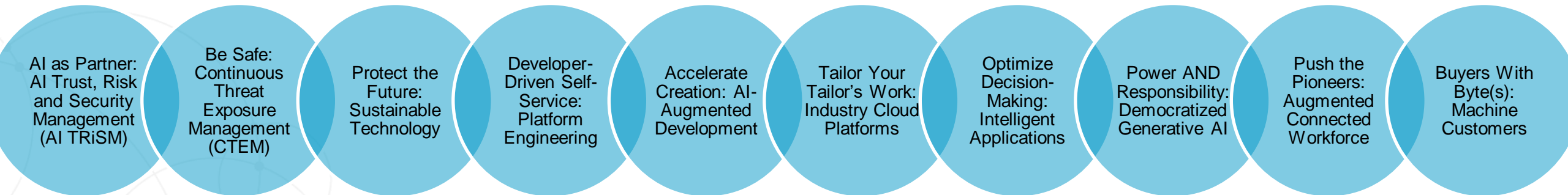
Clouds doesn't matter?



AI matter!

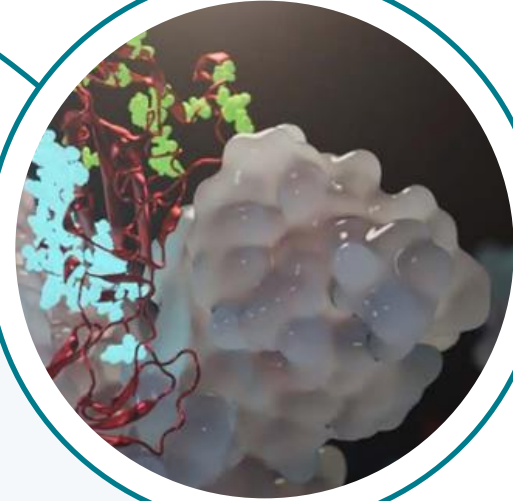


Top 10 Strategic Technology Trends 2024





AI Applications Are Transforming Every Business



4 TOP AI Cases with RAG


- **Code generation** – Improving developer productivity while preserving privacy and control of an organization's source code through solutions such as our SafeCoder offering with Hugging Face.
- **Contact centers resolution** – Improving time-to-resolution for customer support while safeguarding private internal support documentation.
- **IT operations automation** – Providing richer automation and insights into internal IT processes while safeguarding management and operational metadata.
- **Advanced information retrieval** – Streamlining document search, summation, and content creation while preserving an organization's unique intellectual property as well as its unique voice in how it communicates with customers and partners.

Retrieval Augmented Generation (RAG) - is a framework that combines elements of retrieval-based and generation-based approaches in natural

Hot News

- «
- Safeguards on general purpose artificial intelligence
 - Limits on the use of biometric identification systems by law enforcement
 - Bans on social scoring and AI used to manipulate or exploit user vulnerabilities
 - Right of consumers to launch complaints and receive meaningful explanations
- «

European Parliament in AI Act



News





European Parliament

[Homepage](#) [Press room](#) [Agenda](#) [FAQ](#) [Election Press Kit](#)


[Press room](#) / Artificial Intelligence Act: MEPs adopt landmark law

Artificial Intelligence Act: MEPs adopt landmark law

[Press Releases](#) [PLENARY SESSION](#) [IMCO](#) [LIBE](#) Today



- Safeguards on general purpose artificial intelligence
- Limits on the use of biometric identification systems by law enforcement
- Bans on social scoring and AI used to manipulate or exploit user vulnerabilities
- Right of consumers to launch complaints and receive meaningful explanations



Why Businesses Struggle with AI



Developers/ Data Scientists

- Need to experiment quickly
- Complicated to scale AI apps in production
- Ticket-based infrastructure slows development



IT Operator

- Existing infrastructure performance is insufficient for AI apps
- Shadow-IT AI silos make it challenging to manage resources
- Enterprise-class resiliency, security, and governance is difficult

About Broadcom



New story of



Broadcom Announces Successful Acquisition of VMware

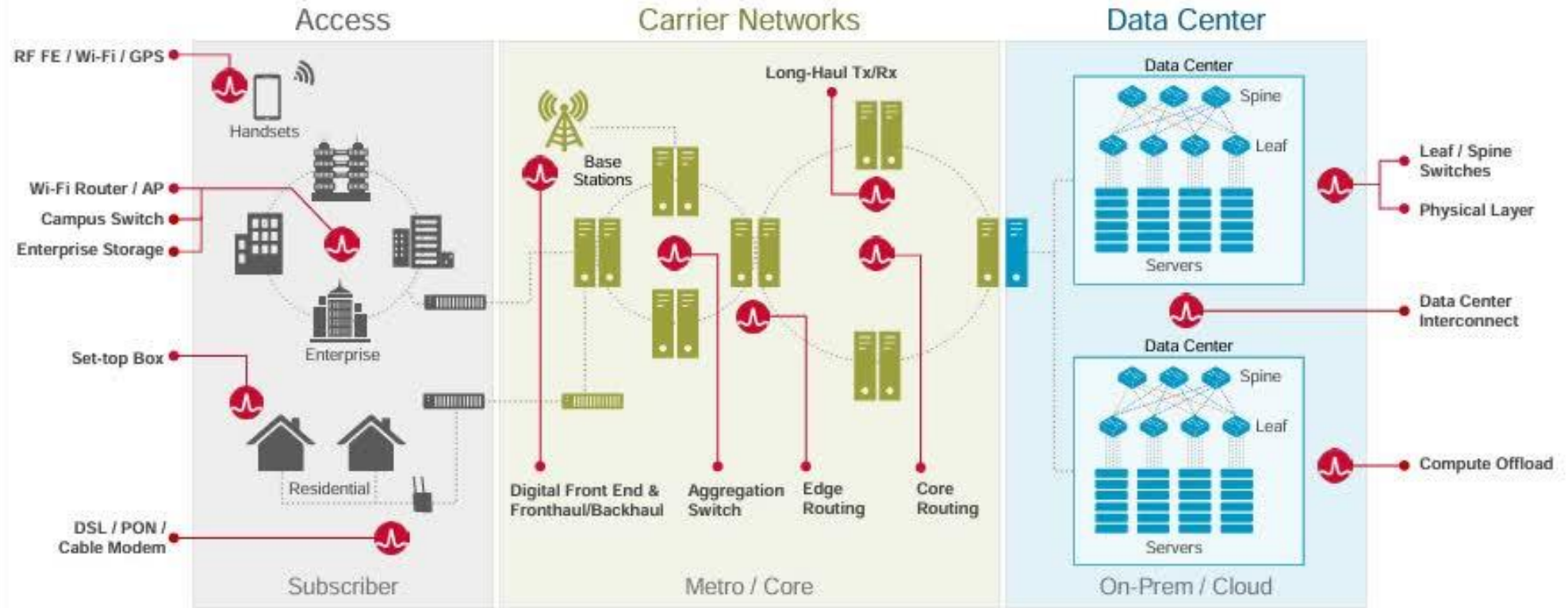
Hock Tan
President and CEO, Broadcom

[Read the Blog](#)



Who are you, Mr Broadcom?

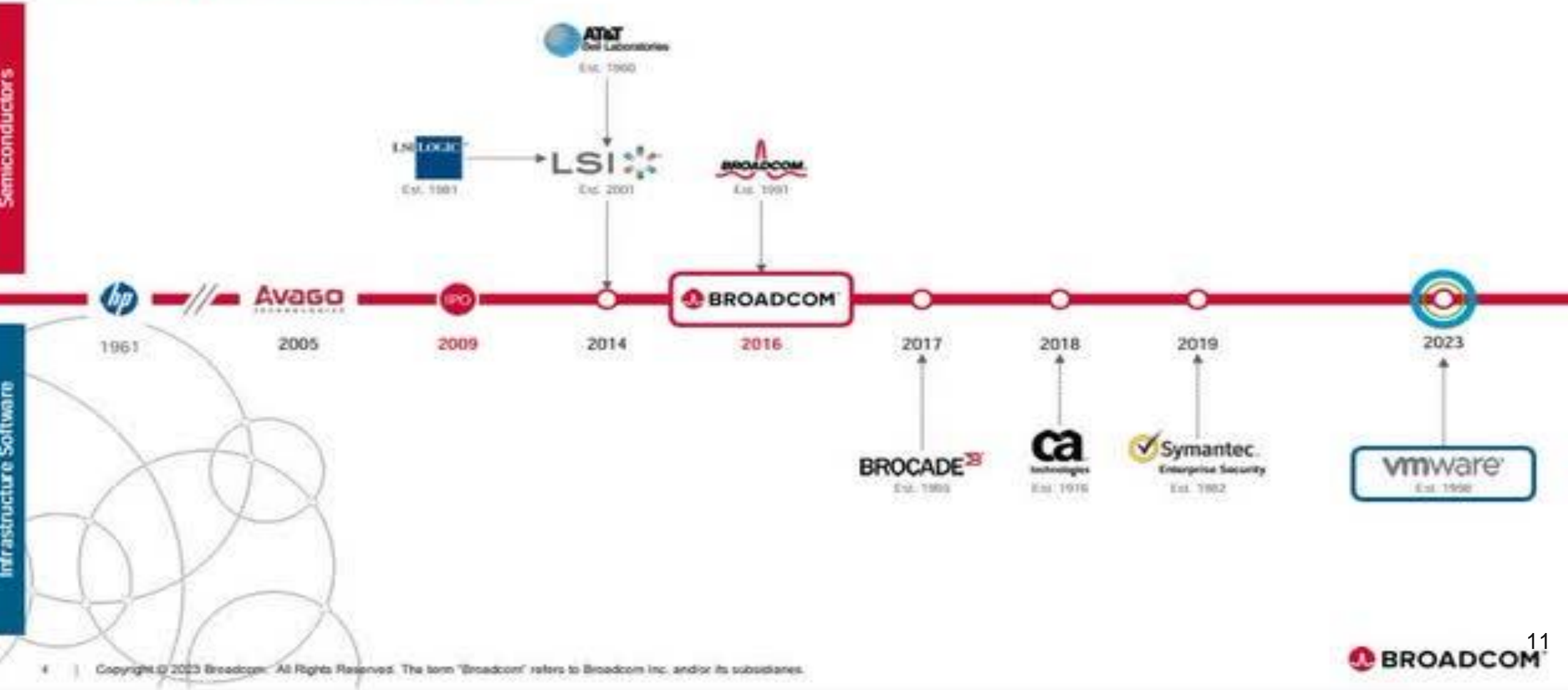
Connecting Everything® Across the Ecosystem



99.9% of All Internet Traffic Crosses at Least One Broadcom Chip

Who are you, Mr Broadcom?

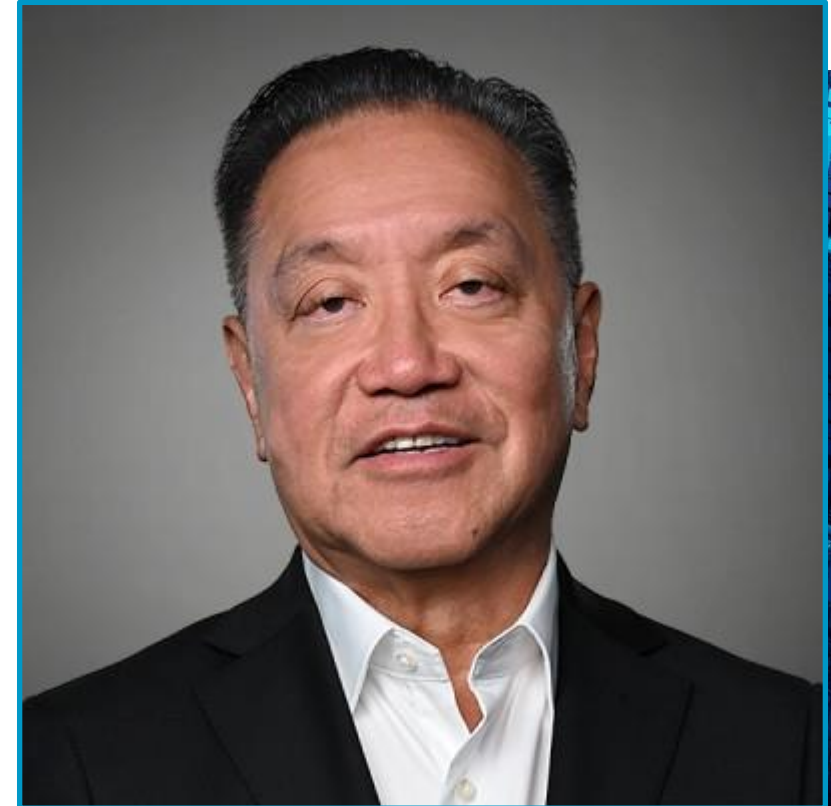
Heritage of Innovation



AI matter for Broadcom

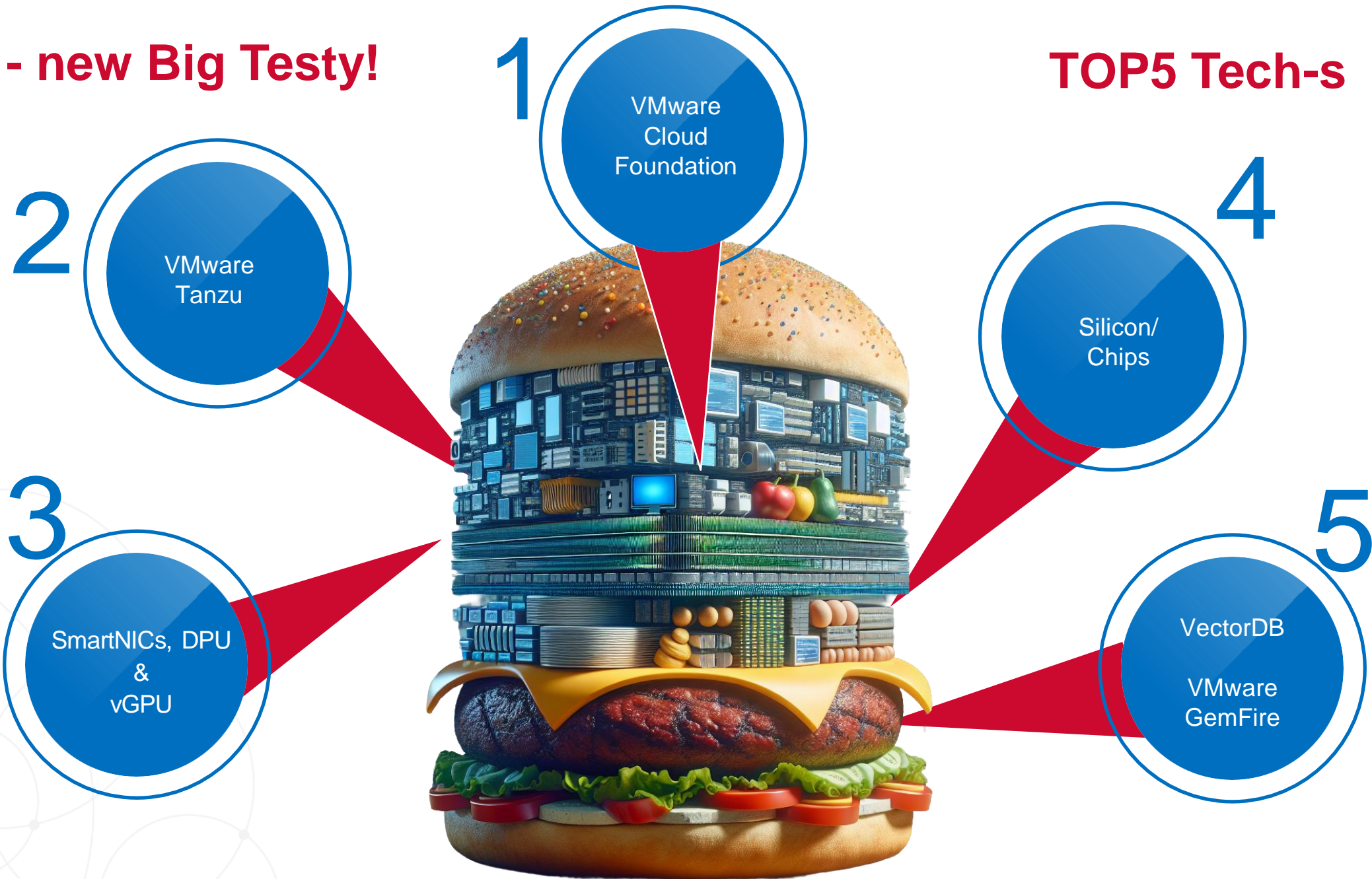
"At VMware Explore last August, VMware and NVIDIA entered into a partnership called **VMware Private AI Foundation** which enables **VCF to run GPUs**. This allows customers to deploy their AI models on-prem and wherever they do business, without having to sacrifice privacy and control of their data. **We are seeing this capability drive strong demand for VCF, from enterprises seeking to run their growing AI workloads on-prem.**"

Hock Tan,
President and CEO at Broadcom

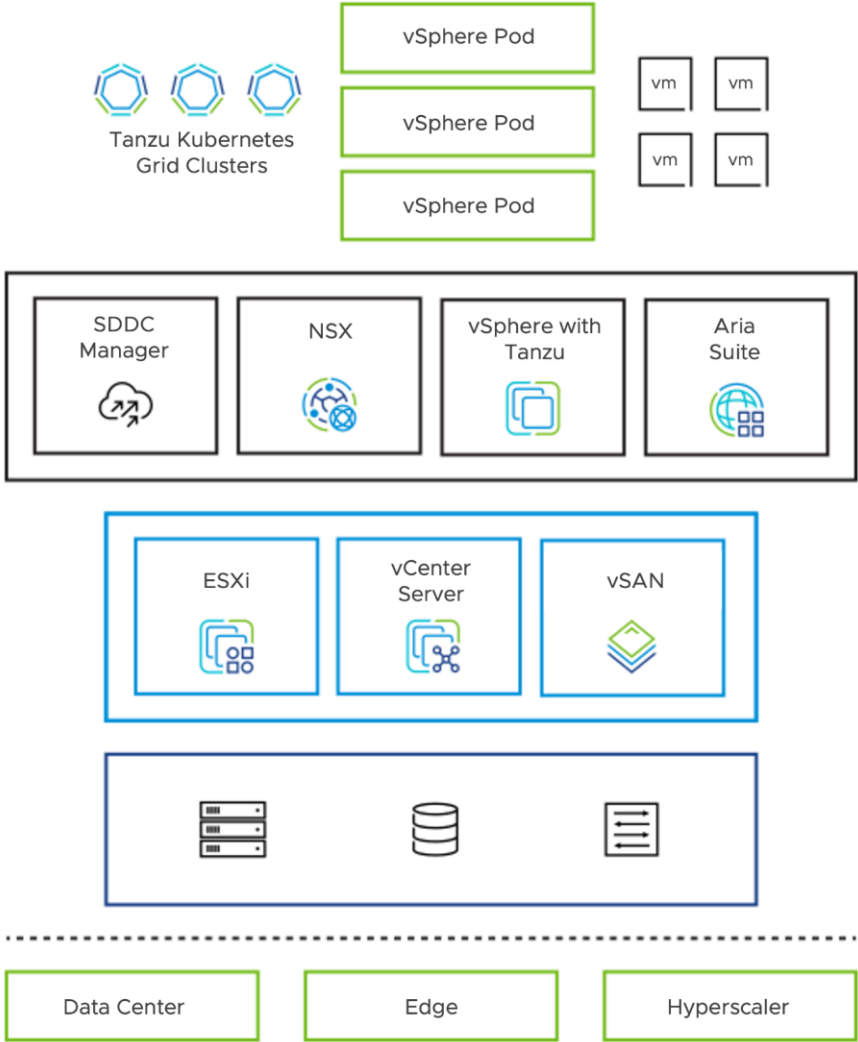
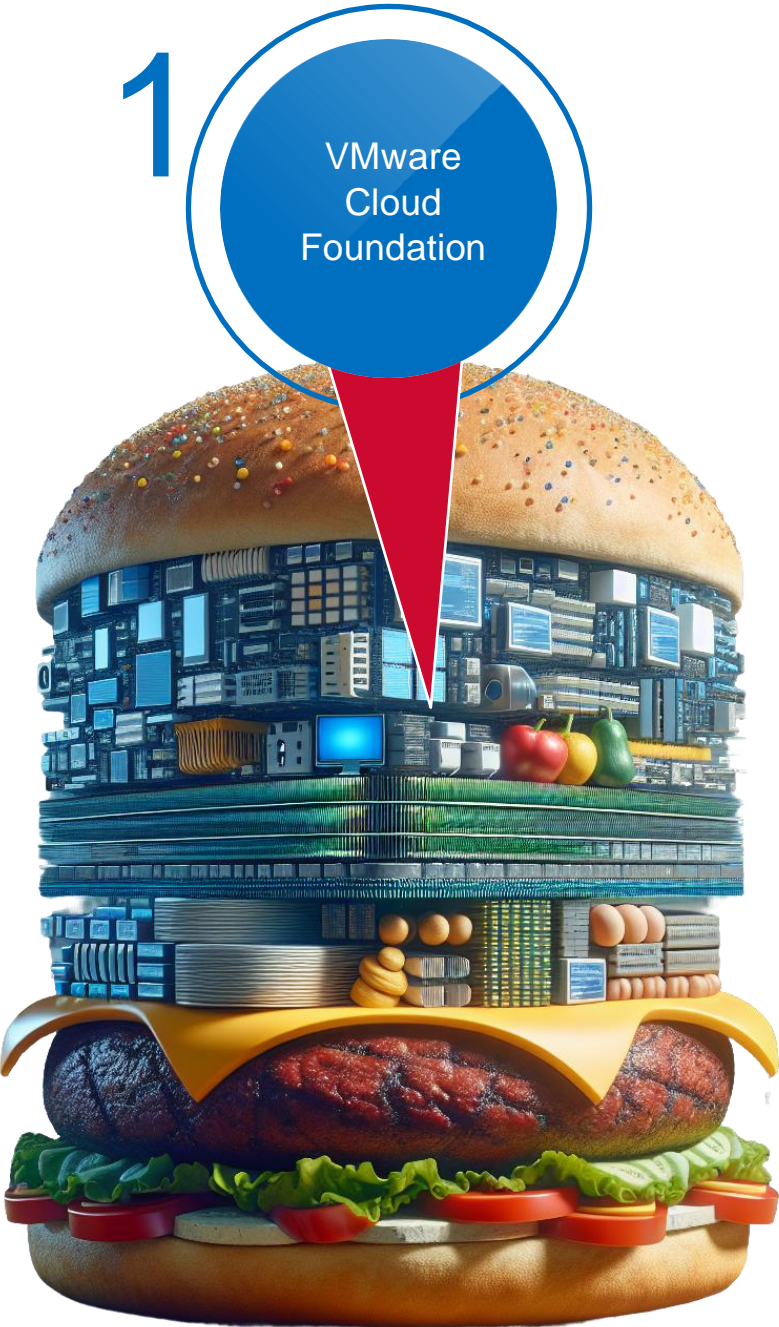


AI - new Big Testy!

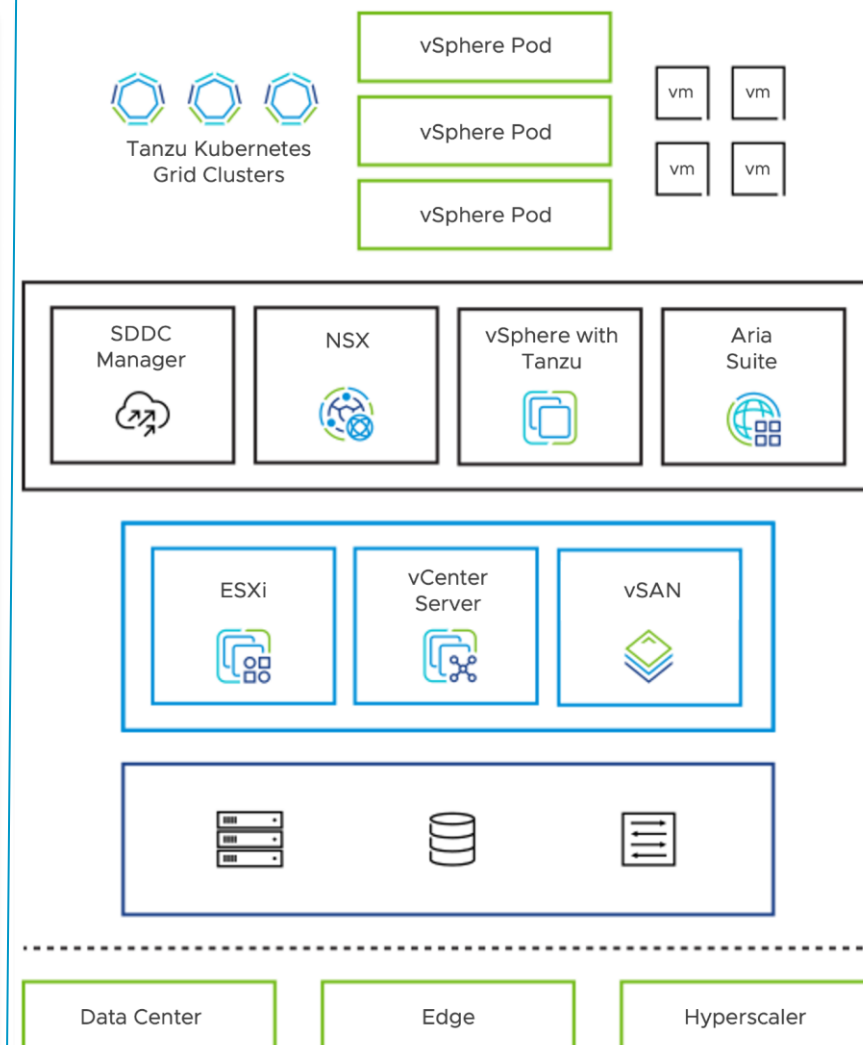
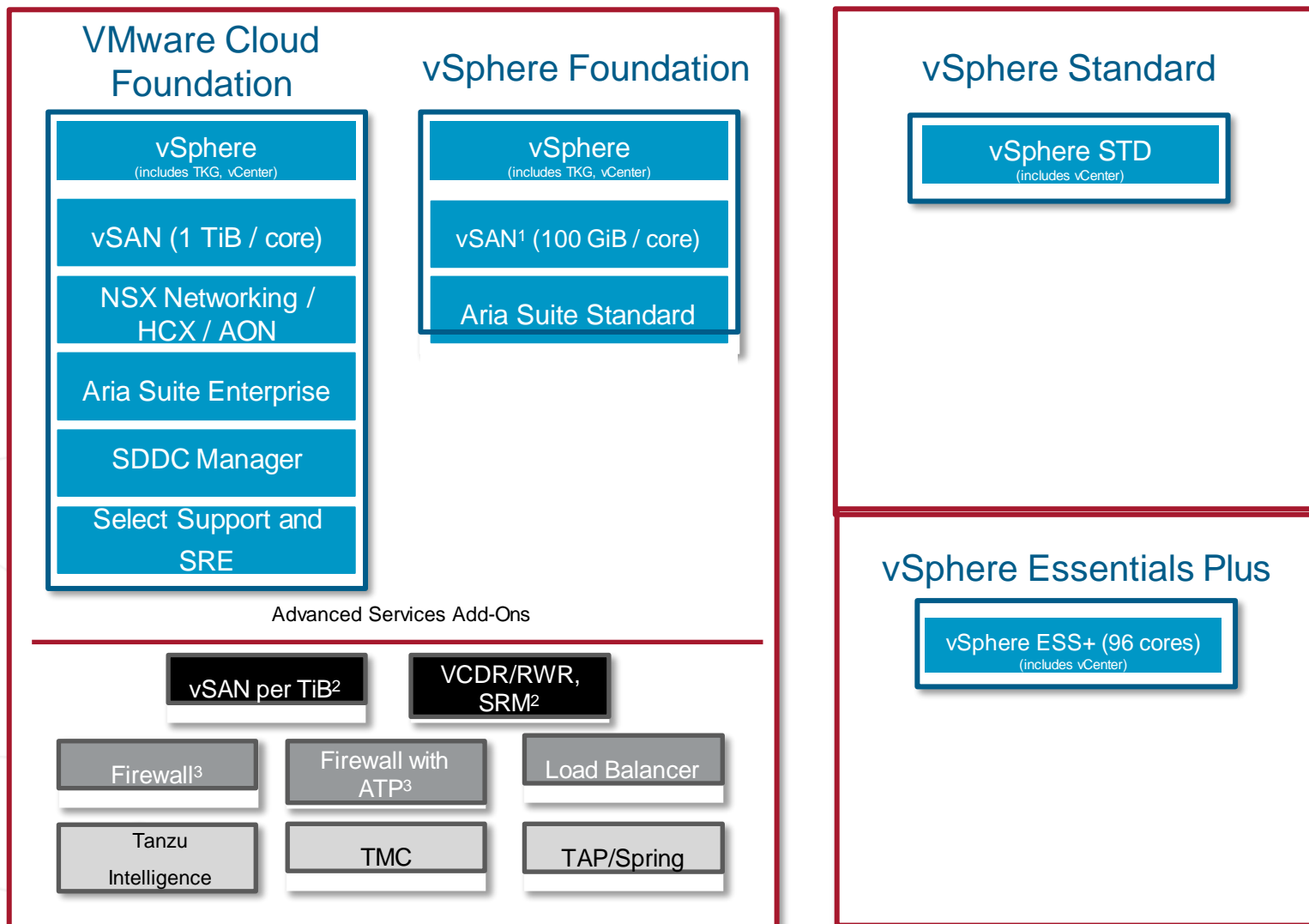
TOP5 Tech-s



Platfrom first



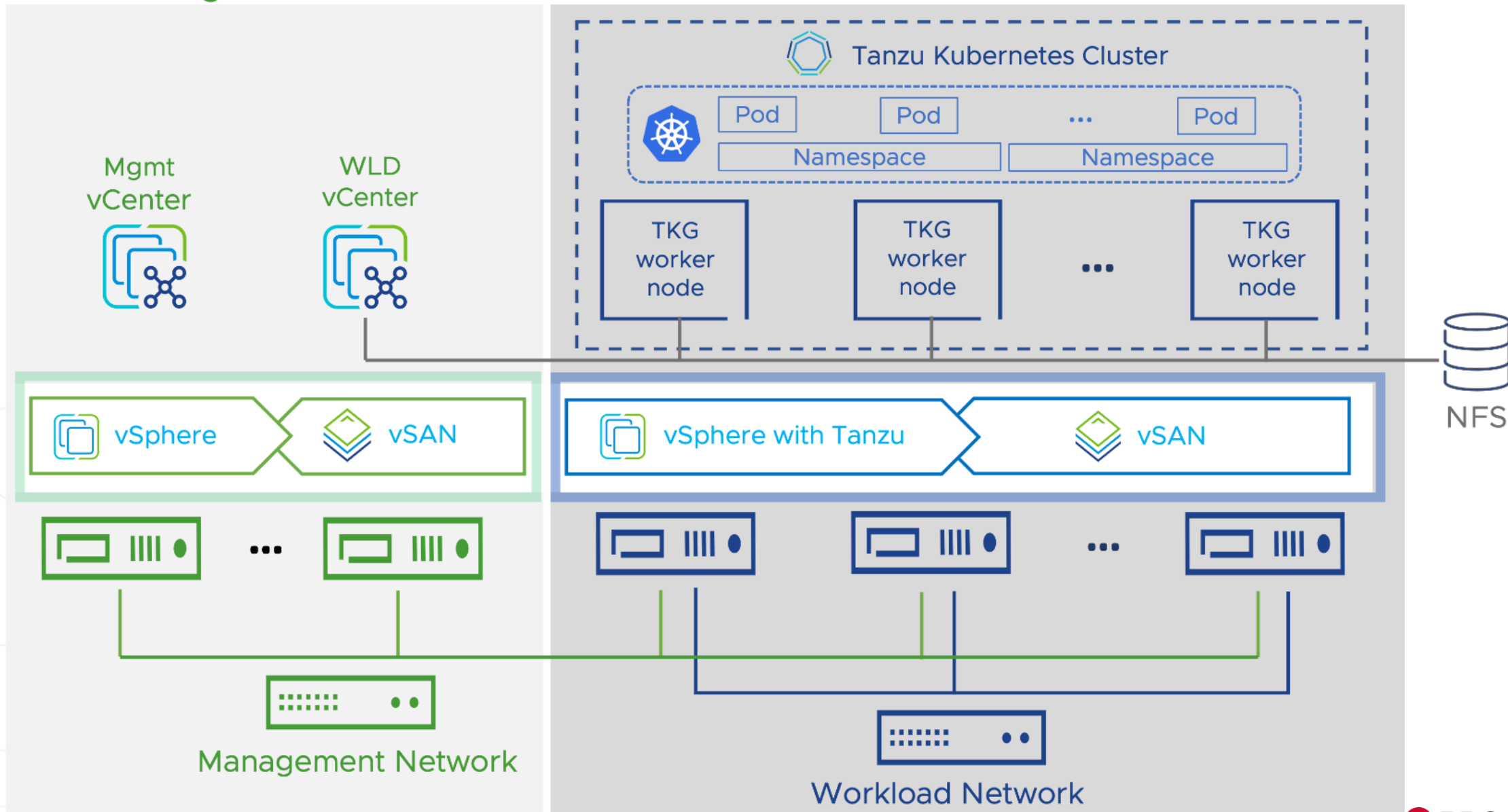
Базовые наборы бизнес-юнита VCF



¹ Available with vSphere Foundation software upgrade ² Add-on for VCF and vSphere Foundation only ³ Add-on for VCF only

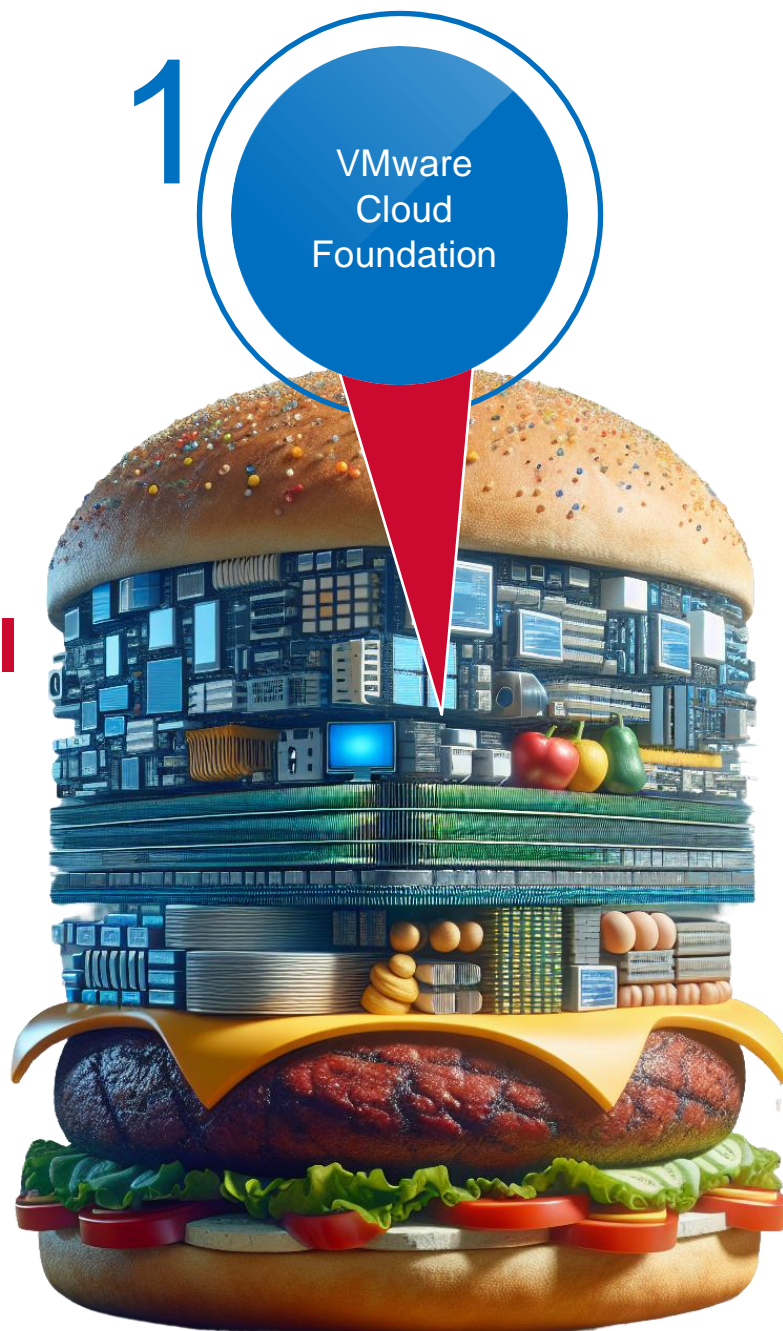
VCF Management Cluster

Workload Cluster



Platform first

+ Integration with AI community



Private AI
коллаборации

- для vSphere
Foundation и vCloud
Foundation

- несколько
вендоров/стеков
решений

Экосистема Private AI

VMware Private AI Open Ecosystem



Falcon



Llama 2



Mistral



MPT



StarCoder



WizardML



anyscale

cnvrg.io



DOMINO



DKube



FedML



Hugging Face



Kubeflow



NVIDIA
NEMO



oneAPI



PyTorch



run:ai



Weights & Biases

watsonx

HCL

IBM

kyndryl

NTT DATA



Dell Technologies



Hewlett Packard
Enterprise

Lenovo

AMD

intel



NVIDIA



VMware Cloud Foundation

AI-Ready Enterprise Platform with NVIDIA



Data Center AI Training
and Inference



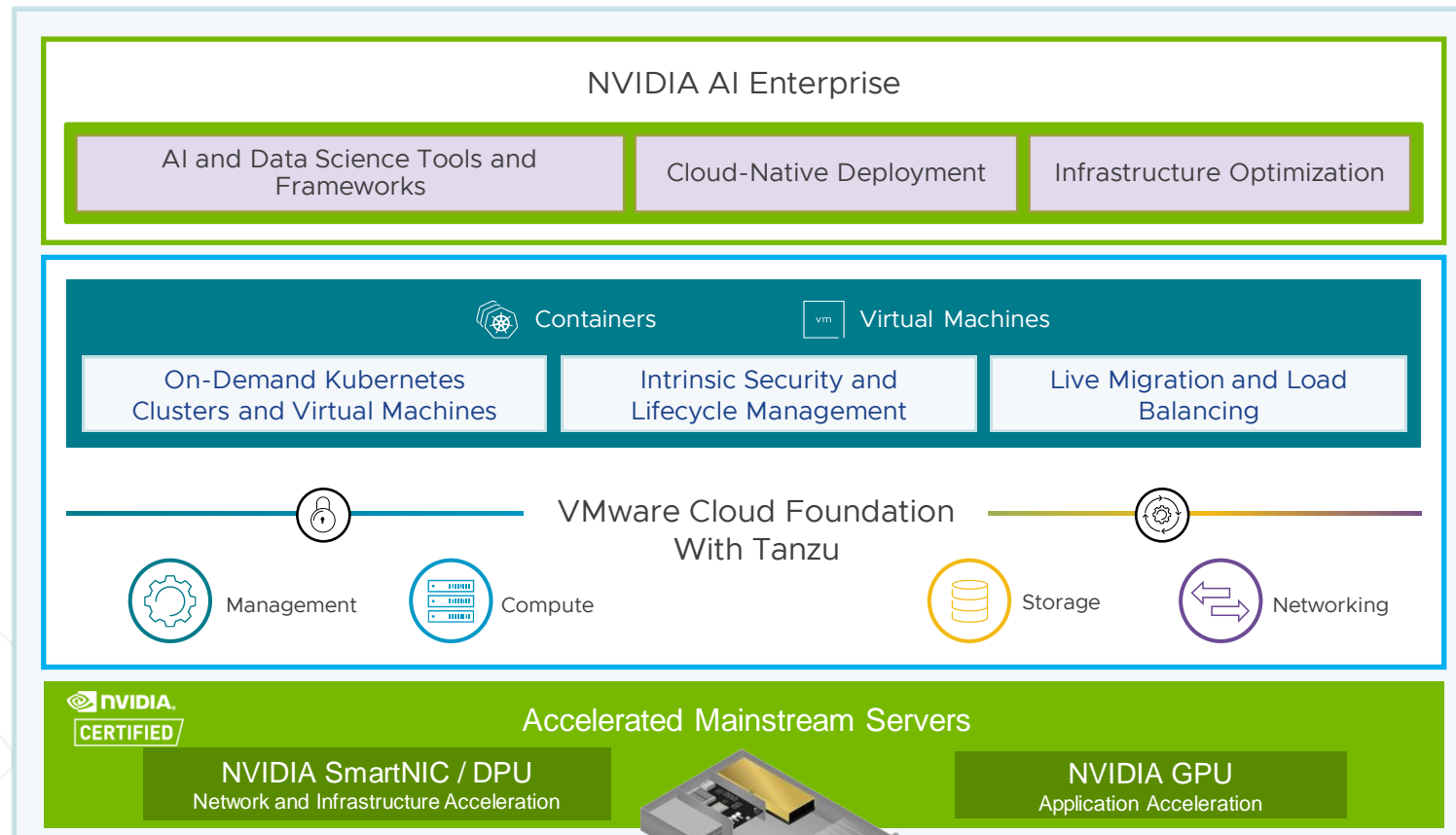
Edge AI
Inference



Data Analytics and
Machine Learning



Data Scientist
Developer
AI Researcher



Edge



Hybrid Cloud



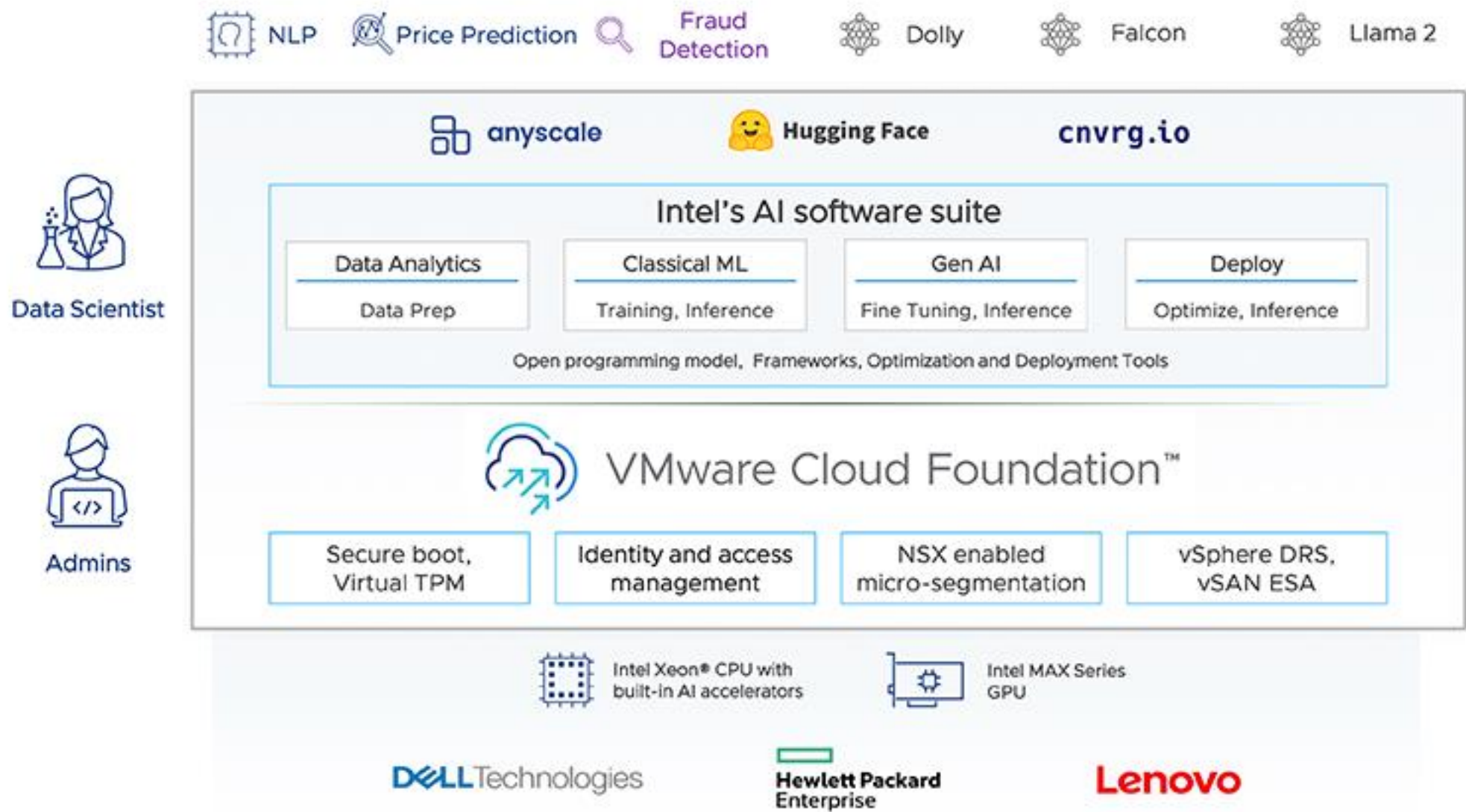
Private Cloud



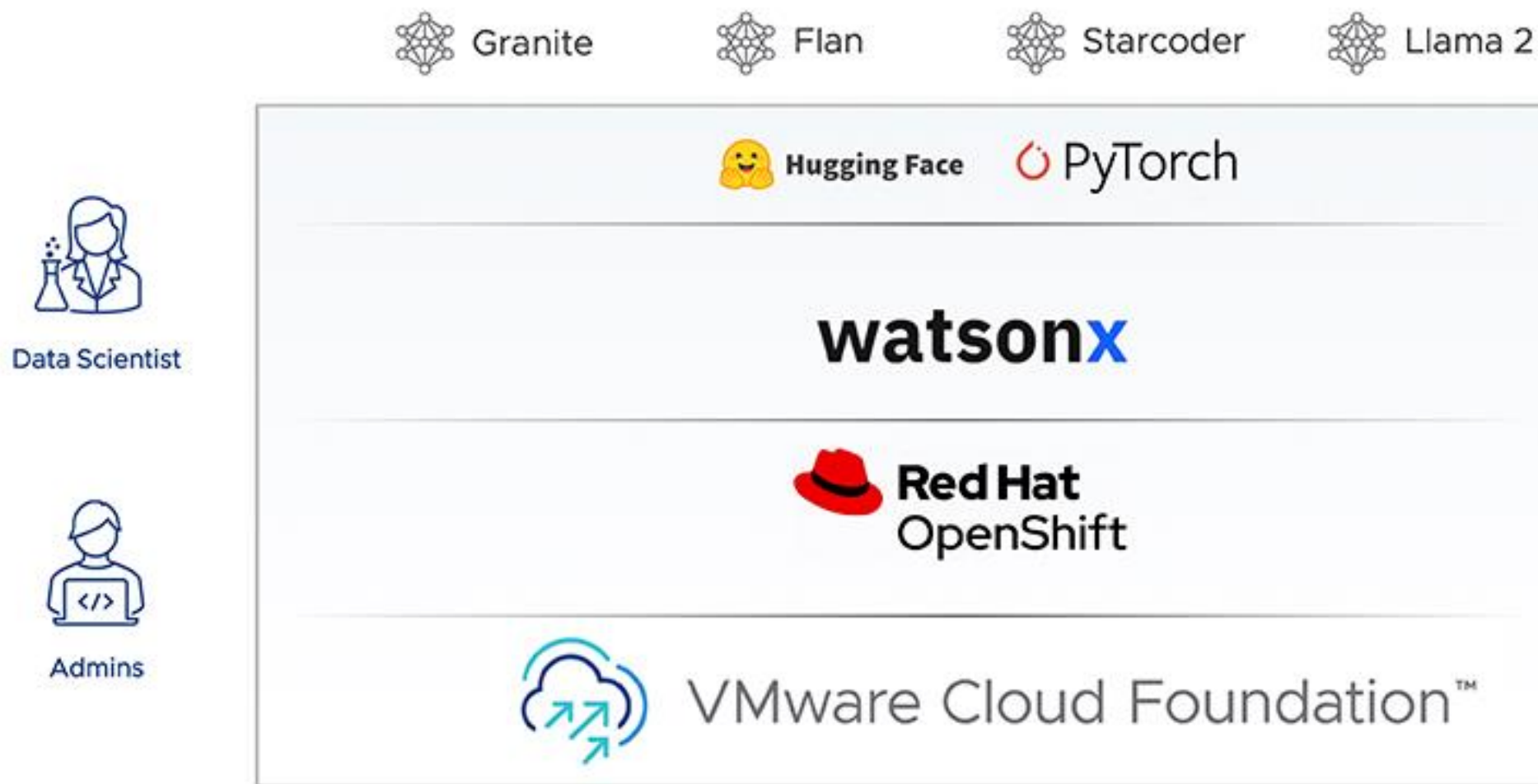
IT Admin

* Future capability not available with VCF 4.4

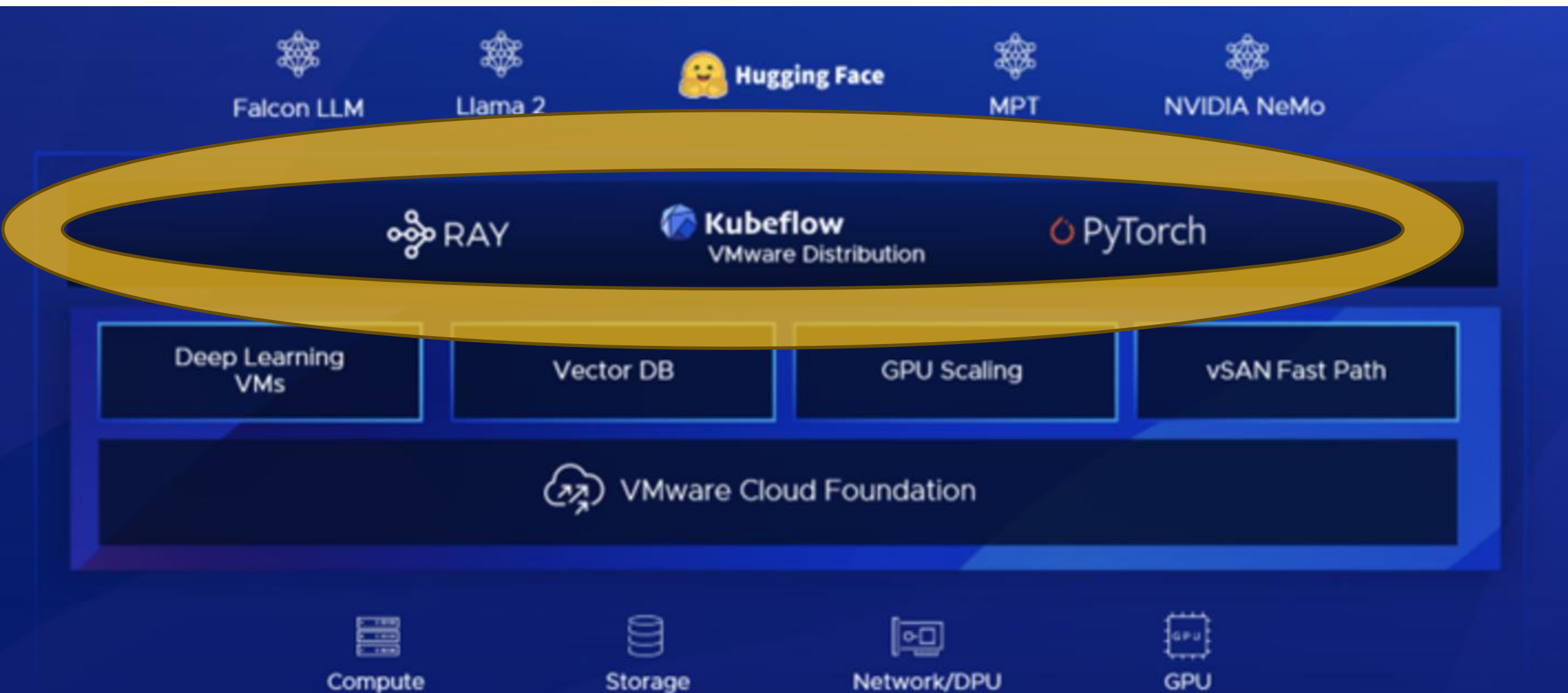
VMware Private AI with Intel



VMware Private AI with IBM



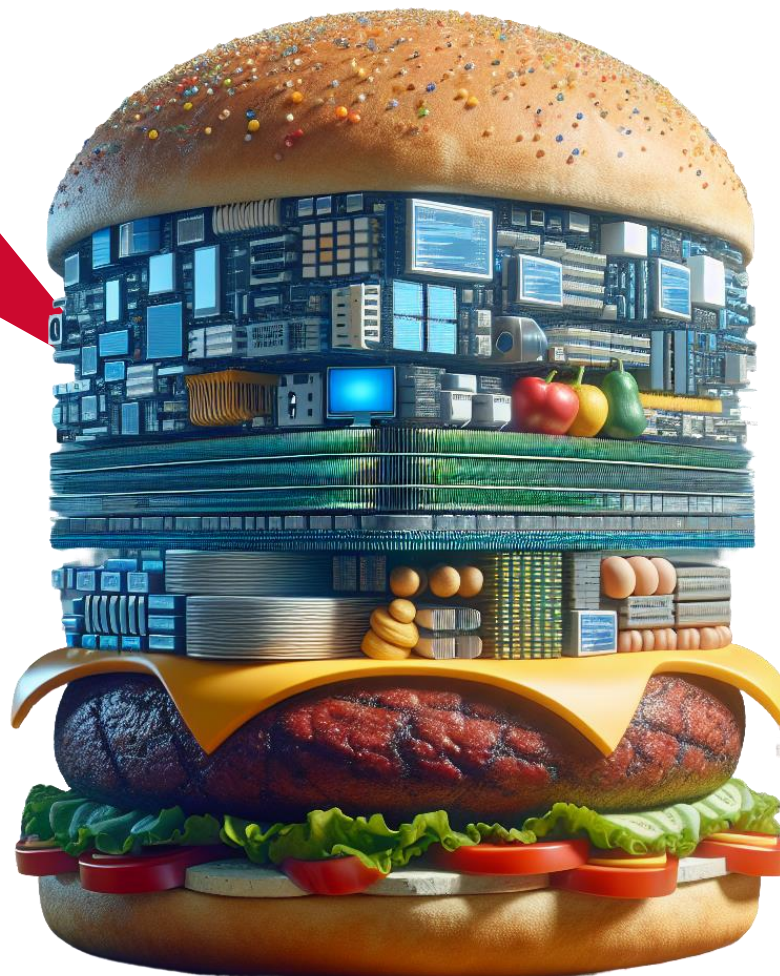
Open Solution High-Level Architecture Overview



Big Testy!

2

VMware
Tanzu

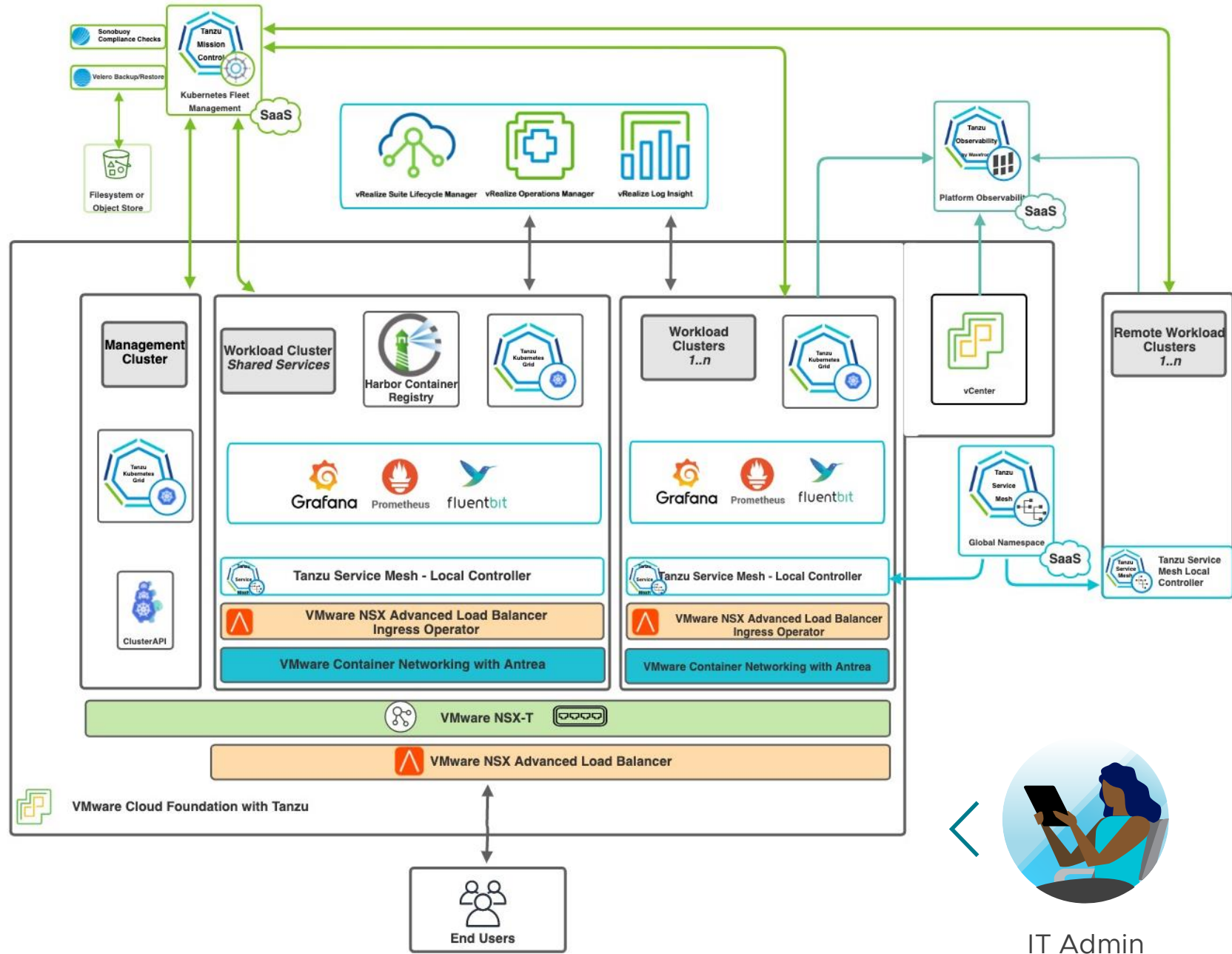


Cloud Native
Apps/микросервисная
архитектура внутри
привычной
корпоративной
инфраструктуры
виртуализации

Tanzu Application Platform



Data Scientist
Developer
AI Researcher



Tanzu with K8S for AI/ML workloads

The screenshot displays the vSphere Client interface. On the left, a navigation pane shows a hierarchy of environments: **aiml-hol**, **genai**, **tkc-2** (containing **tkc-2-9kb8r-k6d5s**, **tkc-2-gpuwork-kbp8p-557676d5d6-vc7bx**, and **tkc-2-nogpuworker-sczx7-d9db46d6-rf5qz**), **tkc1gpu** (containing **tkc1gpu-gpuworkers-lx2n6-dfcff5f4-g9b8c**, **tkc1gpu-pf2r2-64xgt**, **tkc1gpu-pf2r2-bblhb**, and **tkc1gpu-pf2r2-p2l6r**), **machine-learning**, **ml-tkg-01** (containing **ml-tkg-01-gpuworkers-zkrwr-9fd94744b-t9tkp**, **ml-tkg-01-nml5s-5zfb6**, **ml-tkg-01-nml5s-8kpzm**, **ml-tkg-01-nml5s-bqlmq**, **ml-tkg-01-non-gpuworkers-7fz4v-6865c47d94-9xnt9**, and **ml-tkg-01-non-gpuworkers-7fz4v-6865c47d94-hd74x**), **ml-mysql**, **tanzu-eval1**, and **tanzu-eval10**. The **tkc1gpu-gpuworkers-lx2n6-dfcff5f4-g9b8c** VM is selected.

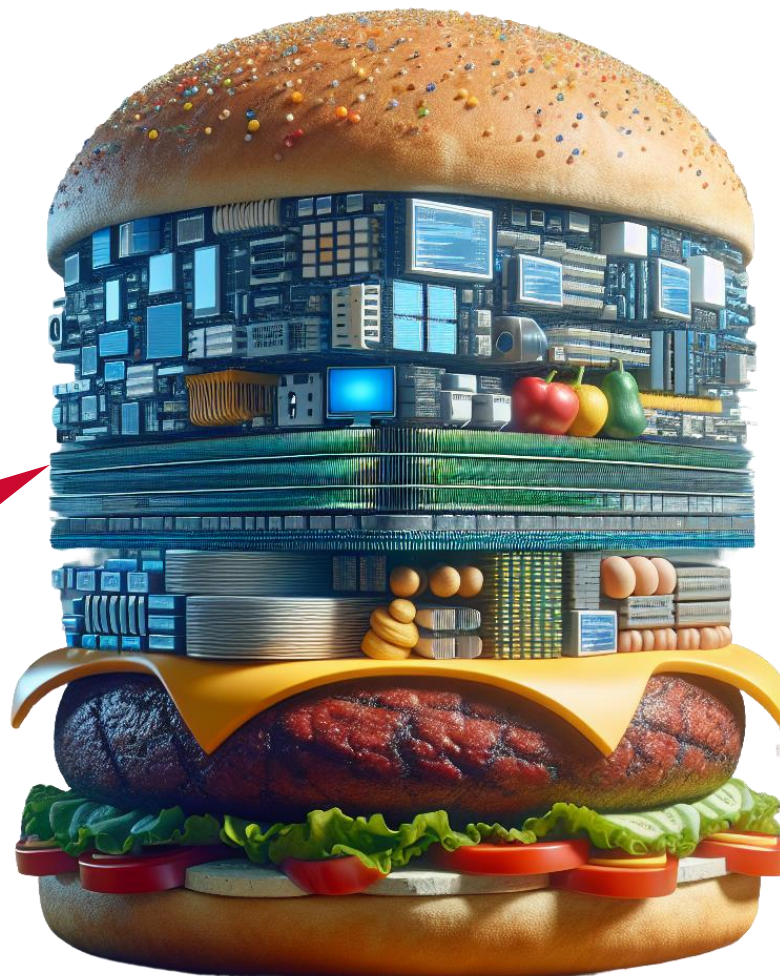
The main panel shows details for the selected VM:

- Summary** (selected):
 - Guest OS**: Includes a screenshot of the VM console and buttons for **LAUNCH REMOTE CONSOLE** and **LAUNCH WEB CONSOLE**.
 - Virtual Machine Details**:
 - Power Status**: Powered On
 - Guest OS**: Ubuntu Linux (64-bit)
 - VMware Tools**: Running, version:11360 (Guest Managed)
 - Managed By**: WCP Service
 - DNS Name (1)**: tkc1gpu-gpuworkers-lx2n6-dfcff5f4-g9b8c
 - IP Addresses (13)**: 172.16.18.110, fe80::250:56ff:feac:da35, AND 11 MORE
 - Encryption**: Not encrypted
 - PCI Devices**:
 - grid_a100-40c** (NVIDIA GRID vGPU)
 - Related Objects**:
 - Cluster**: cluster
 - Host**: w2.bs.dmz-c2701.isvlab.vm
 - Tags**

Smart IO for AI

3

SmartNICs, DPU
&
vGPU

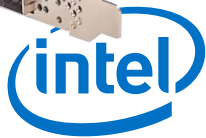
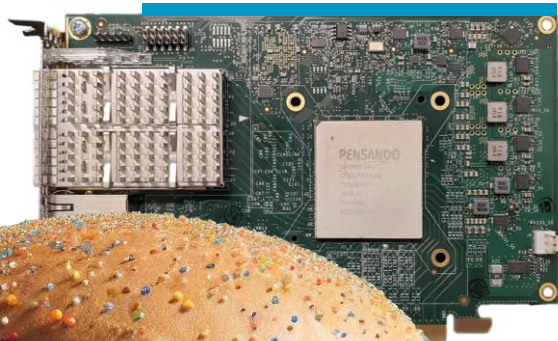
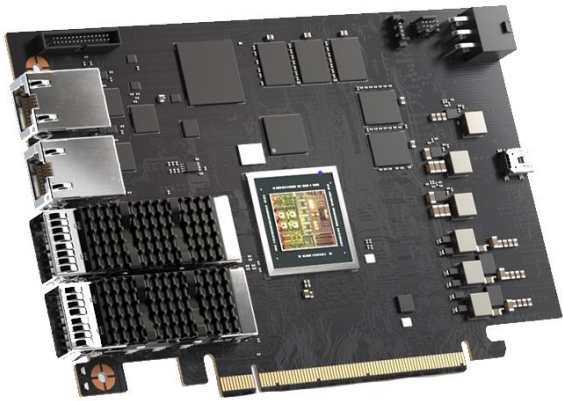


Акселерация AI/ML на
уровне «железа»

- Digital Processing Unit (DPU)
- Graphics Processing Unit (GPU)

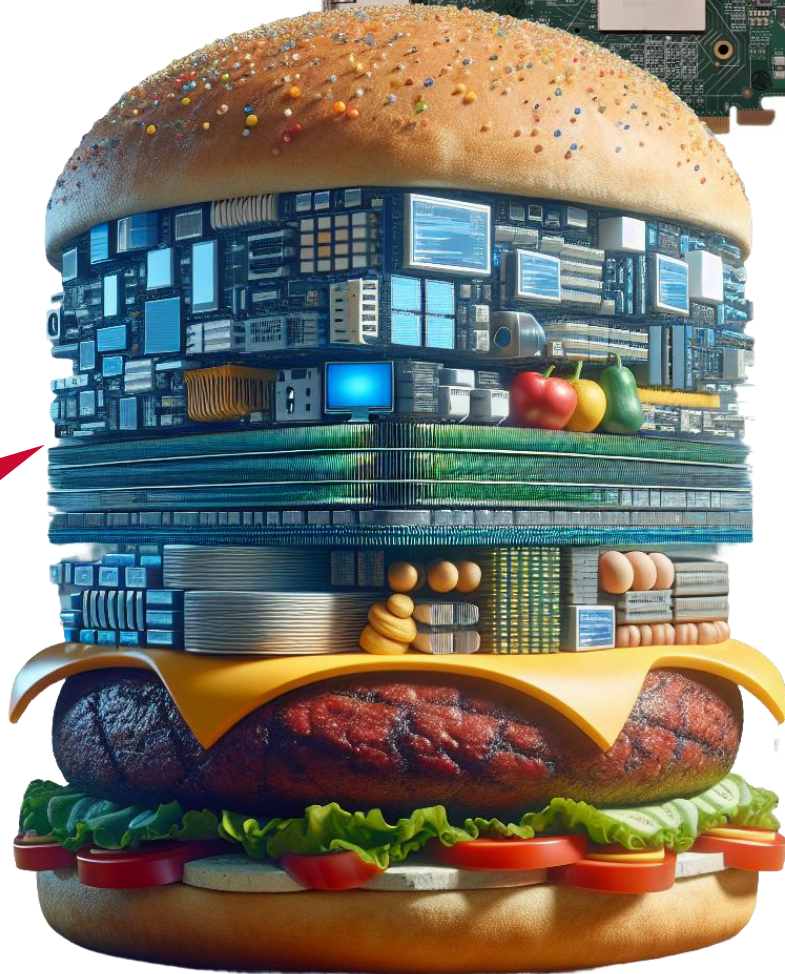
Smart NICs & DPU

PENSANDO



3

SmartNICs, DPU
&
vGPU



Работа VMware с
несколькими
поставщиками
SmartNIC/DPU

Underhood of SmartNIC

VMware ESXio 8.0.1 (VMKernel Release Build 21495797)

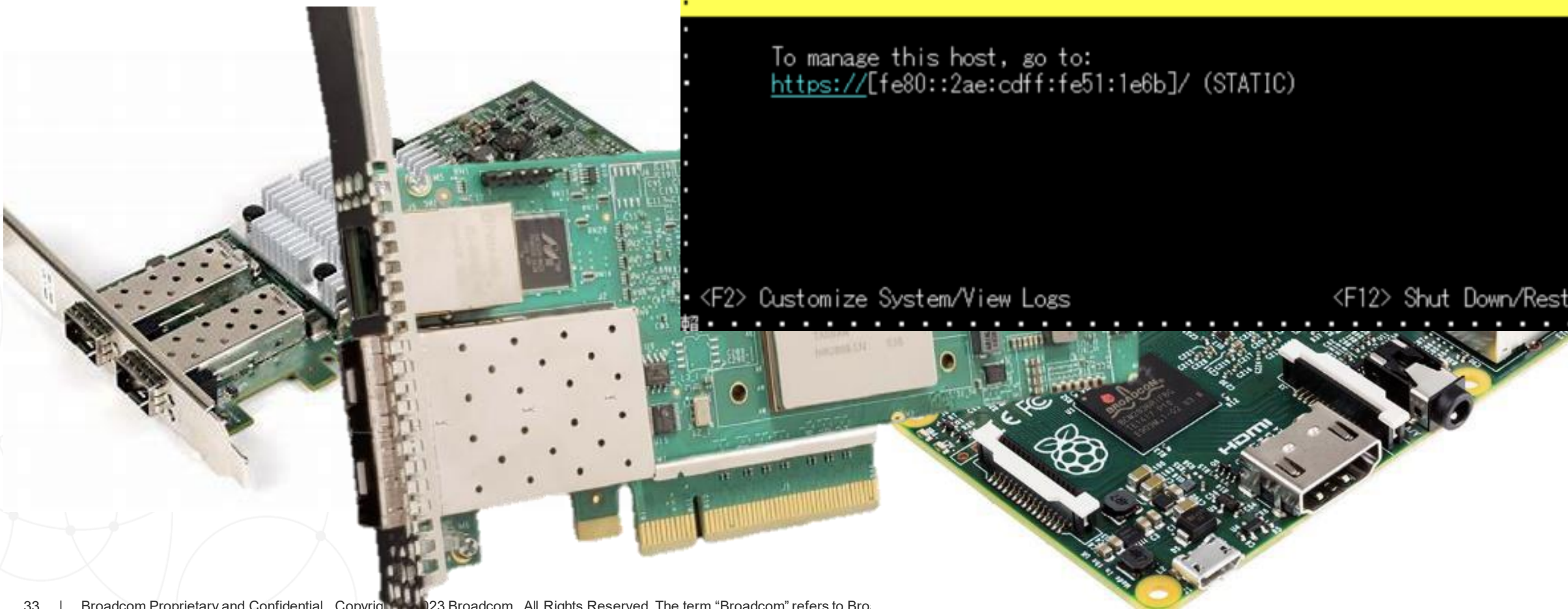
Pensando PCA DSP DSC25v2 10/25G 2p SFP28 32G SPL

ARM Limited Cortex-A72 r1p0
24 GiB Memory

To manage this host, go to:
[https://\[fe80::2ae:cdff:fe51:1e6b\]/](https://[fe80::2ae:cdff:fe51:1e6b]/) (STATIC)

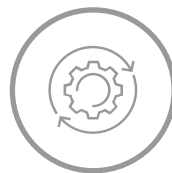
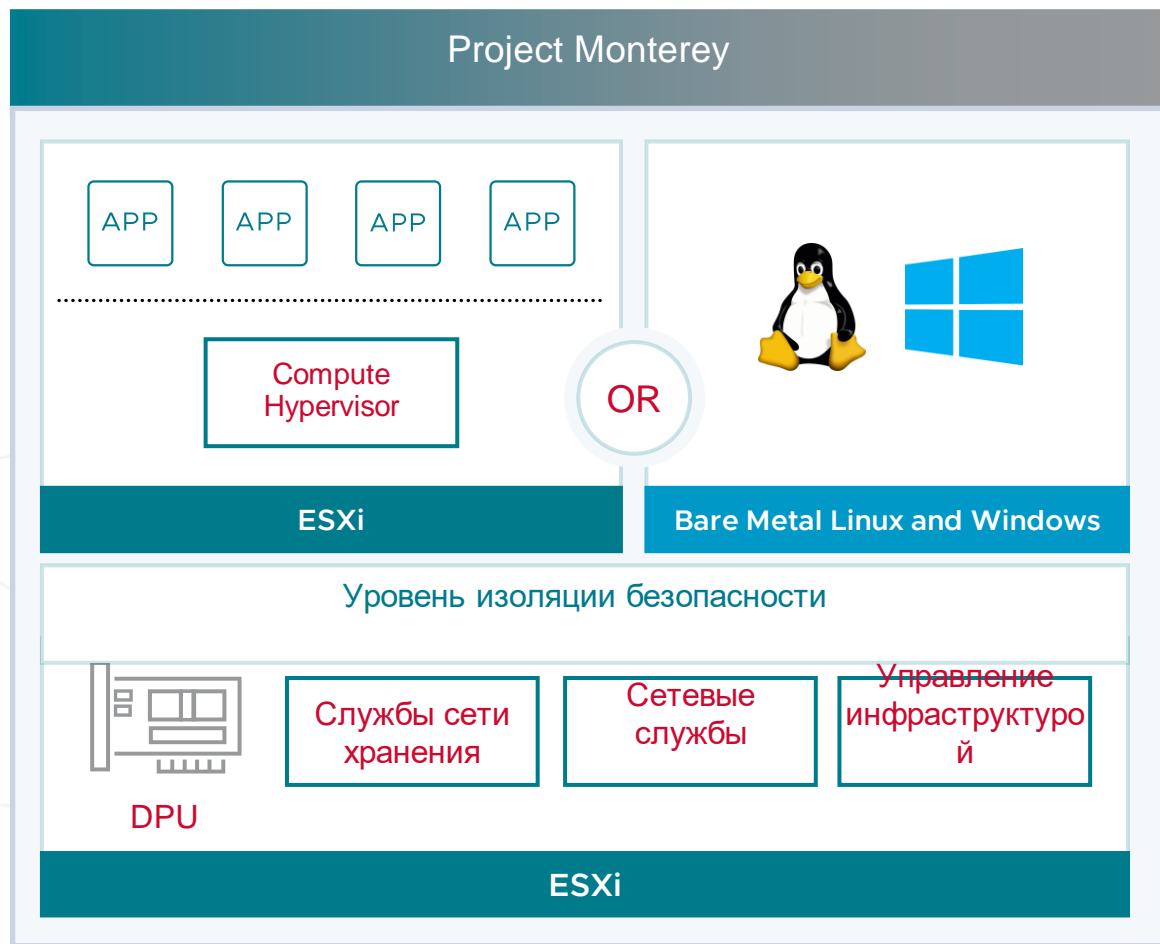
<F2> Customize System/View Logs

<F12> Shut Down/Restart



Vmware's Project Monterey

Расширение инфраструктуры виртуализации до распределенной структуры управления



Единая, безопасная модель управления между рабочими нагрузками



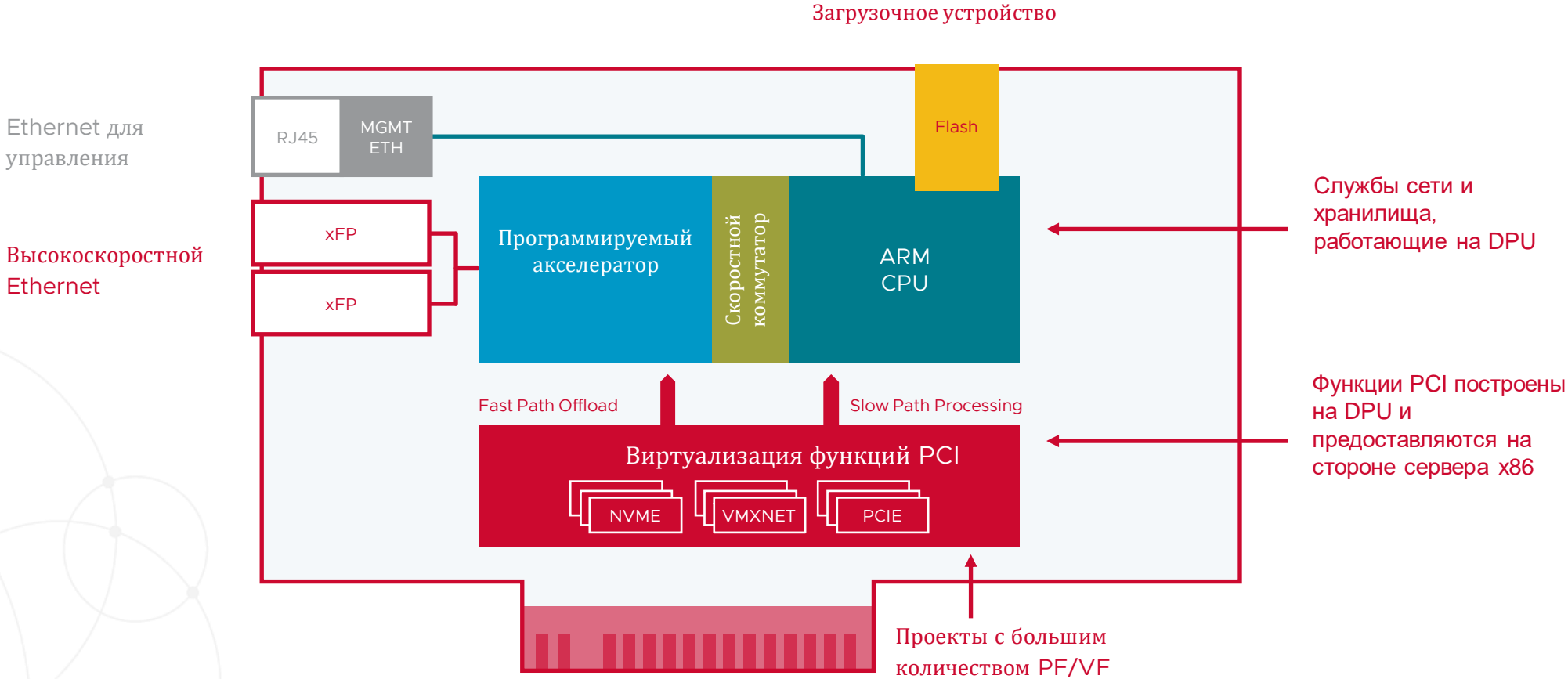
Отдельные домены рабочей нагрузки и инфраструктуры



Разгрузка обработки инфраструктурных служб (сеть+хранение) в DPU

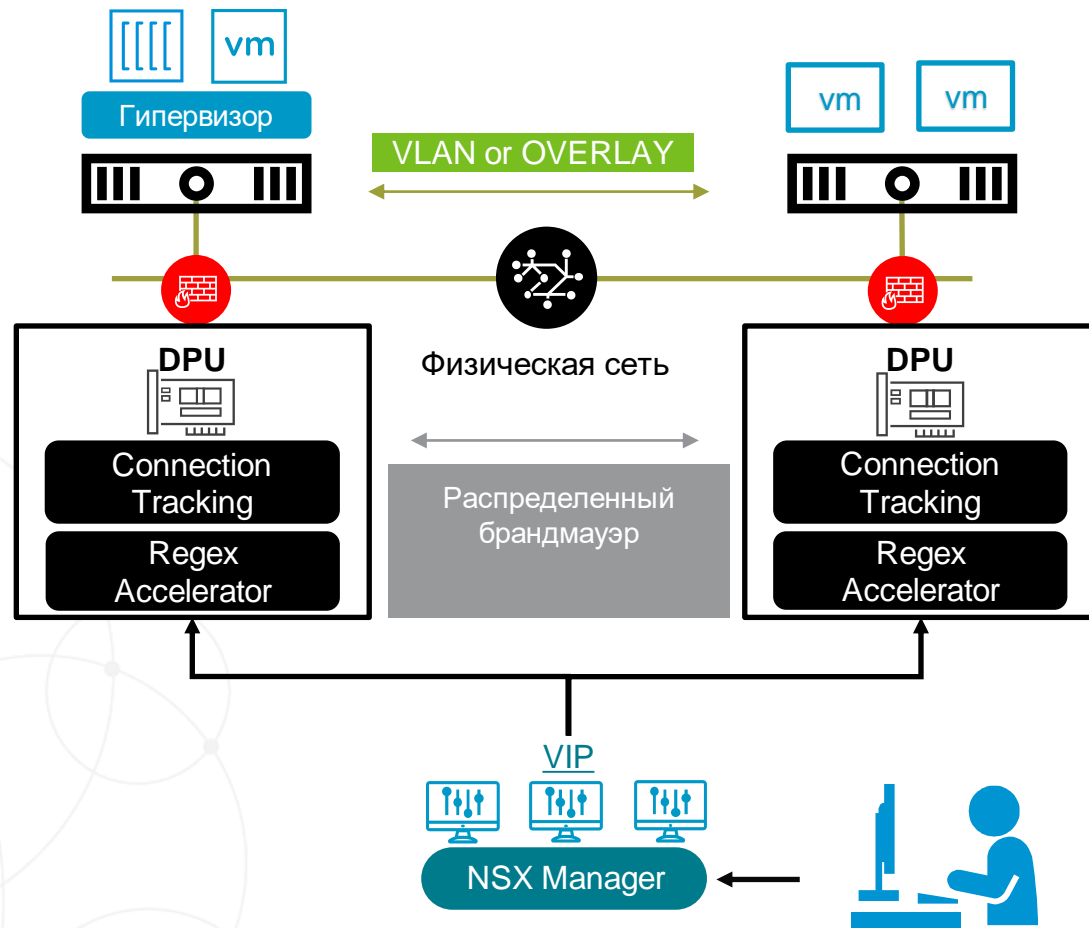
Структура SmartNIC/DPU

Общая архитектура DPU



DPU

& VMware NSX (SDN)



VLAN, Оверлей, режим EVPN

Распределенная маршрутизация
L3

Распределенный брандмауэр L4

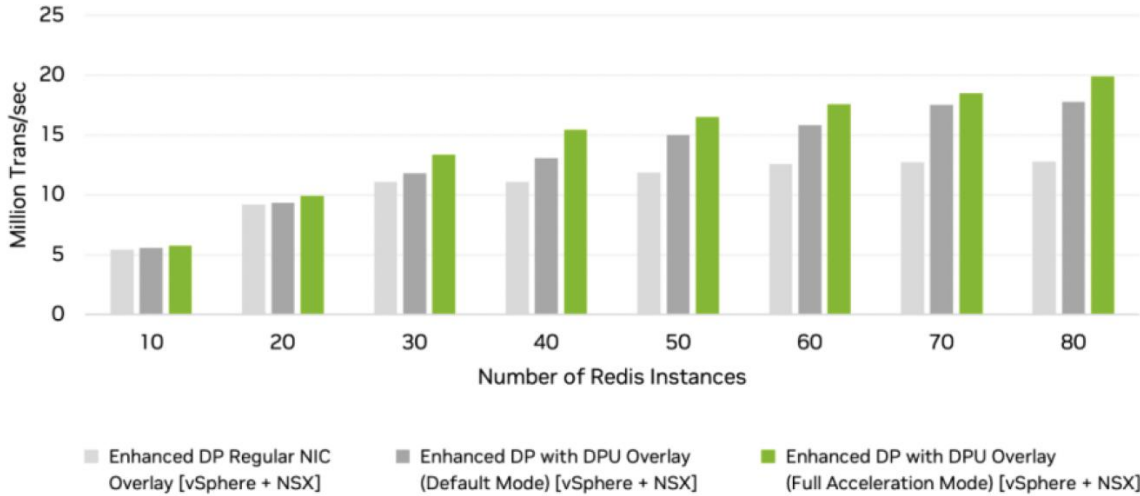
Функции безопасности L7

Распределенная подсистема
балансировки нагрузки L4

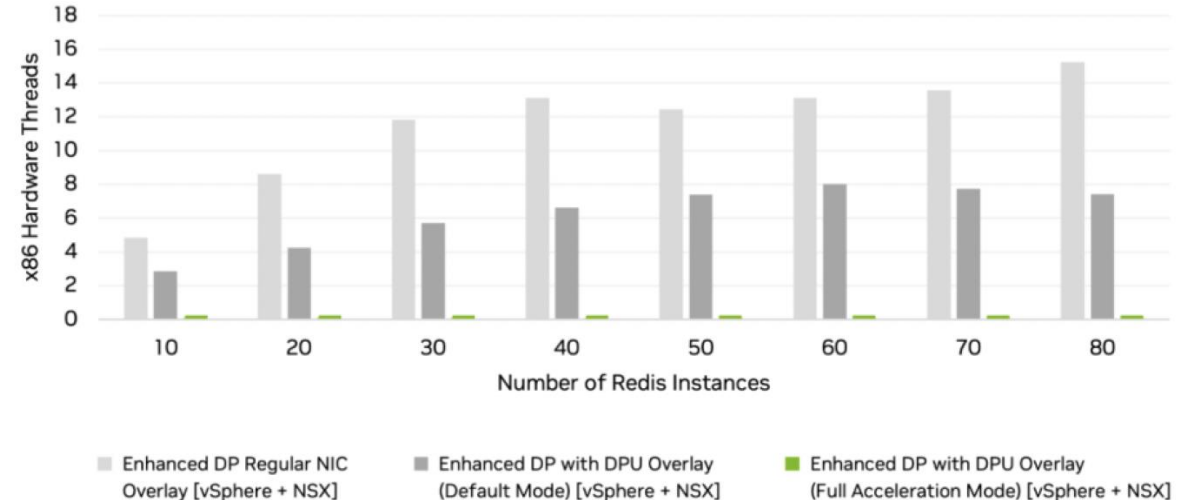
IPFIX, Mirroring, Traceflow и т. Д.

NVIDIA社からも検証結果でています！

Millions of Redis Transactions/sec



x86 CPU Cores Used for Network Traffic Processing



<https://resources.nvidia.com/en-us-accelerated-networking-resource-library/nvidia-vmware-redis>

<https://blogs.vmware.com/performance/2023/04/https://resources.nvidia.com/en-us-accelerated-networking-resource-library/nvidia-vmware-redis-s-of-using-nvidia-dpu-with-vsphere-8.html>

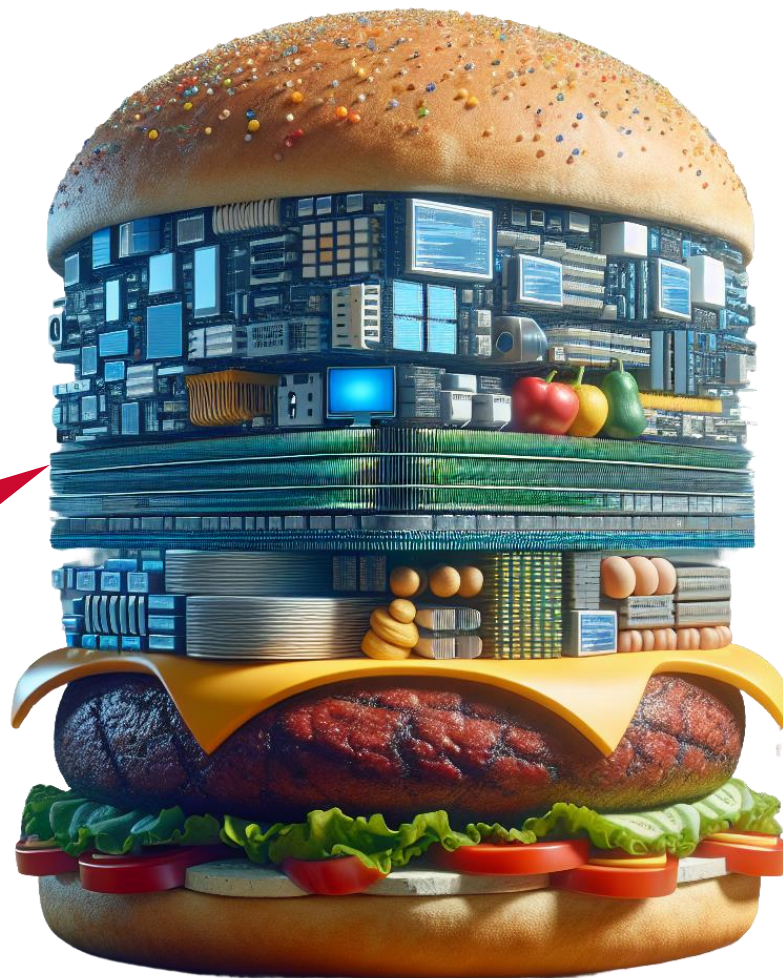
One of the most useful metrics used to understand the potential CPU savings from a DPU for a target use case is to measure the number of l-cores consumed for network processing on ESXi when using a regular NIC. This can be obtained by summing up the used metric reported for all the EnsNetWorld processes on ESXi using the following command:

```
net-stats -A -t WwQqihVcE -i 10 | grep "EnsNetWorld"
```


vGPU

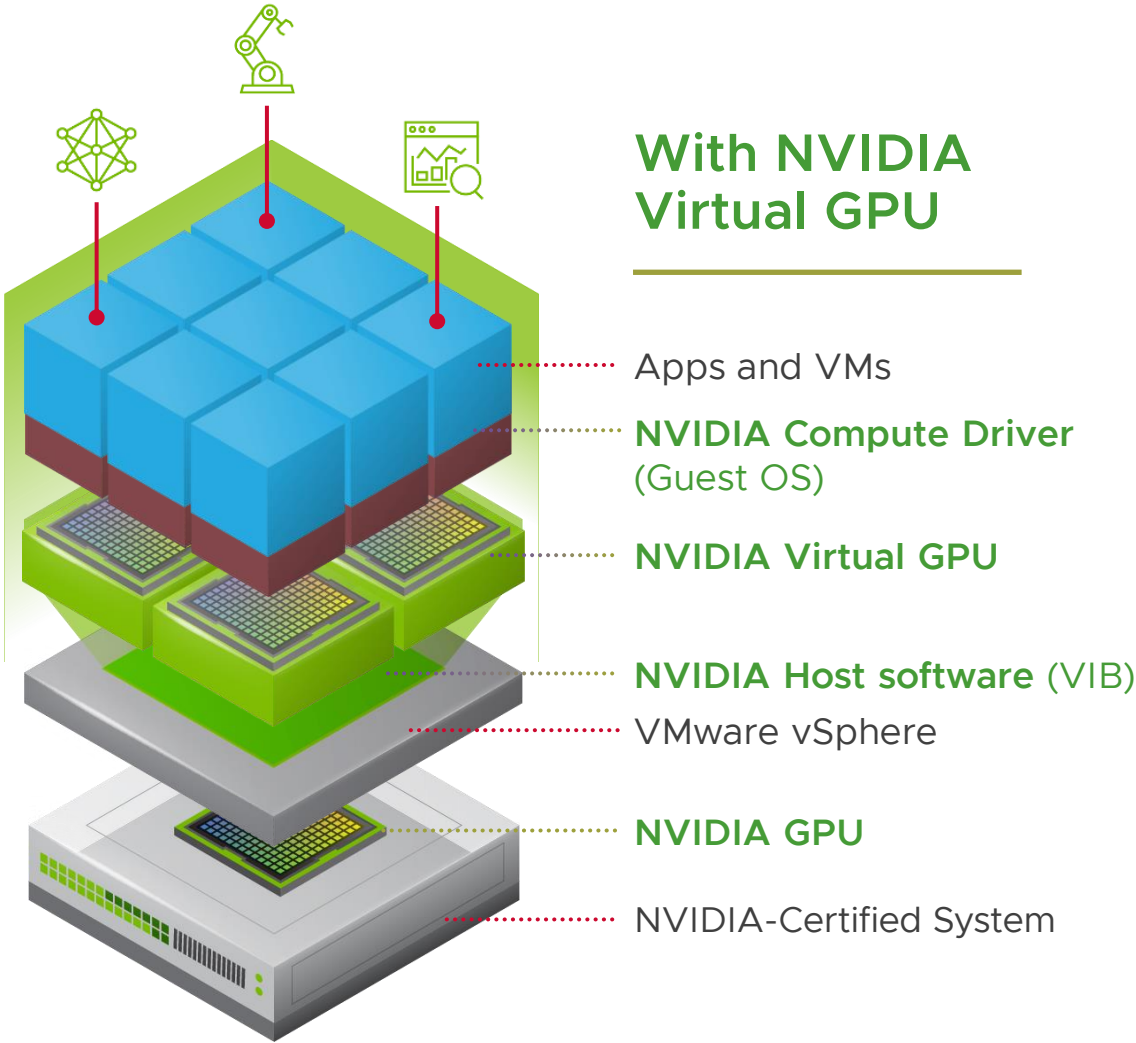
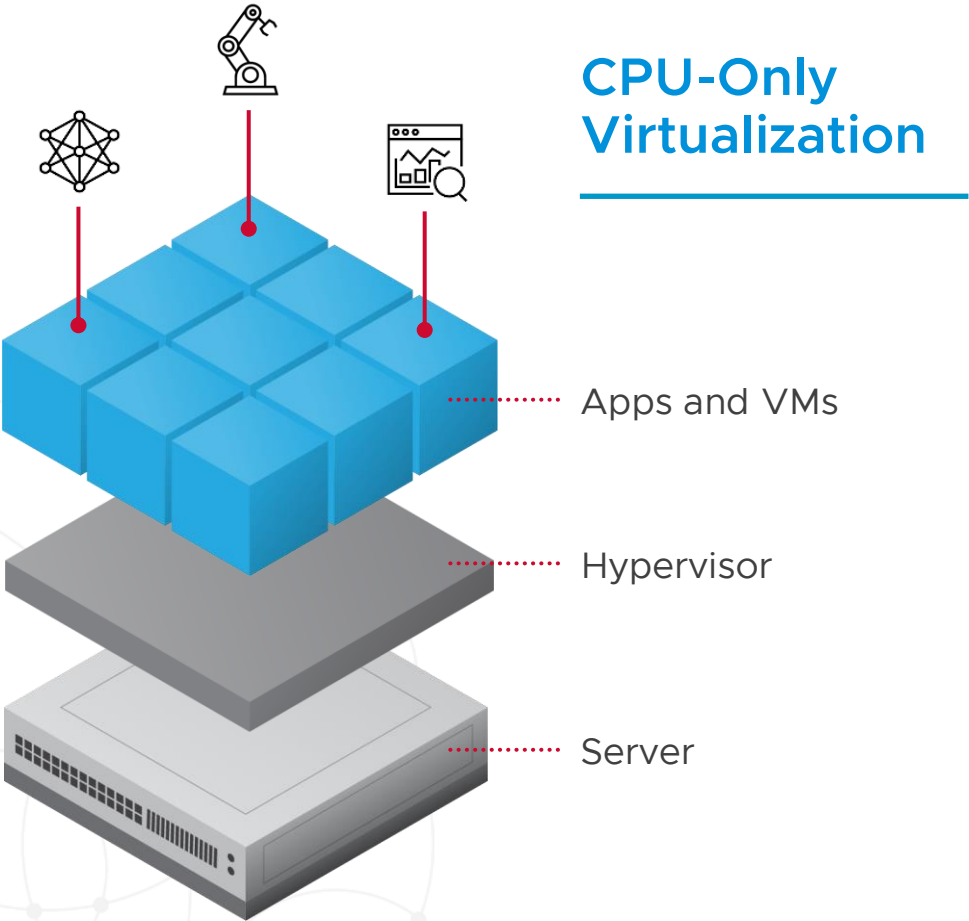
3

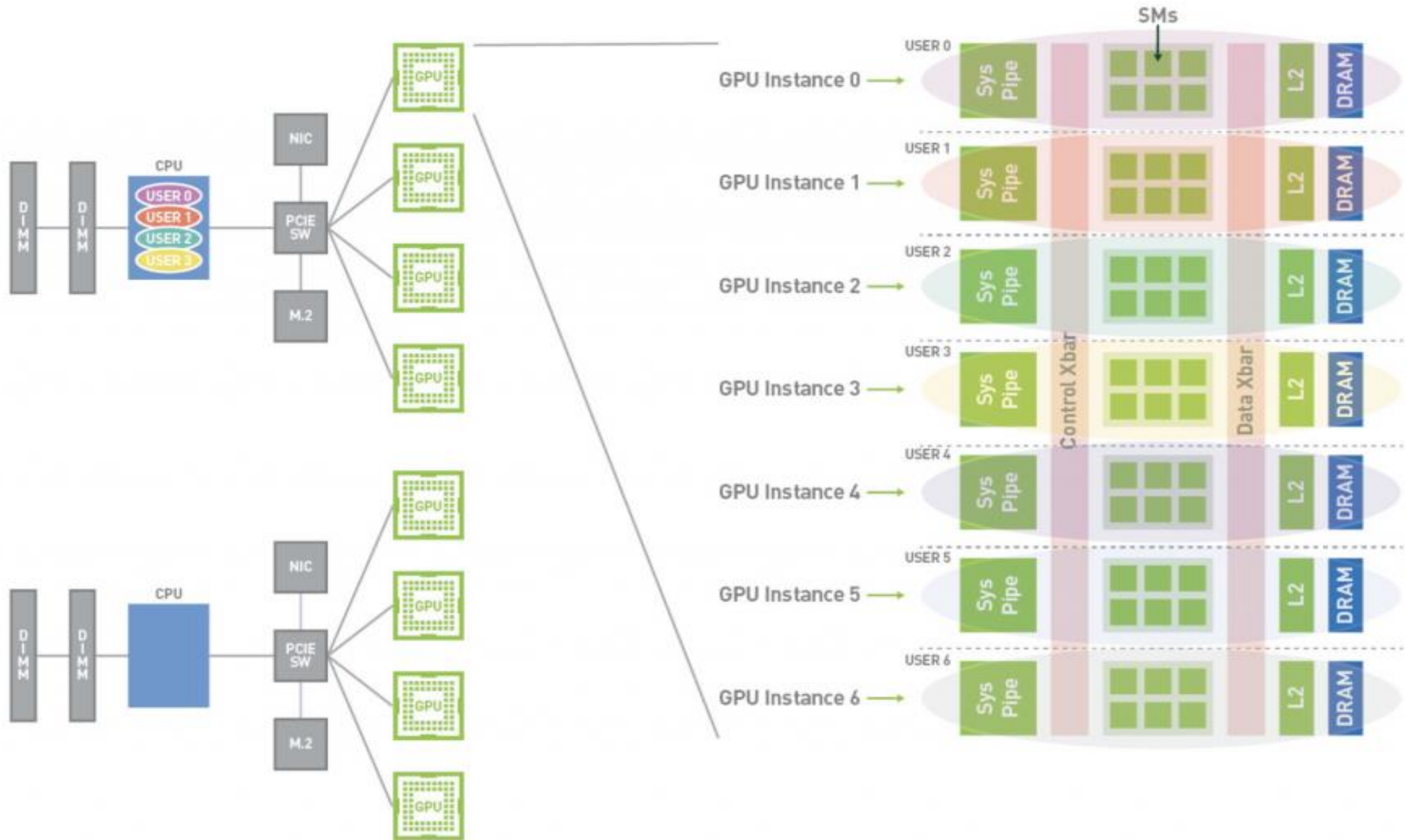
SmartNICs, DPU
&
vGPU



GPU Virtualization Accelerates Compute Workloads

VMware Cloud Foundation + NVIDIA AI-Ready Enterprise Suite





MIG

Multi-Instance GPU

Supported GPUs

MIG is supported on GPUs starting with the NVIDIA Ampere generation (i.e. GPUs with compute capability ≥ 8.0). The following table provides a list of supported GPUs:

Table 1. Supported GPU Products					
Product	Architecture	Microarchitecture	Compute Capability	Memory Size	Max Number of Instances
H100-SXM5	Hopper	GH100	9.0	80GB	7
H100-PCIE	Hopper	GH100	9.0	80GB	7
H100-SXM5	Hopper	GH100	9.0	94GB	7
H100-PCIE	Hopper	GH100	9.0	94GB	7
H100 on GH200	Hopper	GH100	9.0	96GB	7
A100-SXM4	NVIDIA Ampere	GA100	8.0	40GB	7
A100-SXM4	NVIDIA Ampere	GA100	8.0	80GB	7
A100-PCIE	NVIDIA Ampere	GA100	8.0	40GB	7
A100-PCIE	NVIDIA Ampere	GA100	8.0	80GB	7
A30	NVIDIA Ampere	GA100	8.0	24GB	4

Supported Configurations

Supported deployment configurations with MIG include

- Bare-metal, including [containers](#) and [Kubernetes](#)
- GPU pass-through virtualization to Linux guests on top of supported hypervisors
- vGPU on top of supported hypervisors

Edit Settings | thunder9

Virtual Hardware | VM Options

ADD NEW DEVICE

> CPU 8

> Memory 128 GB

> Hard disk 1 200 GB

> SCSI controller 0 LSI Logic Parallel

> Network adapter 1 APPS-1607 ☒ Connect...

> New PCI device NVIDIA GRID vGPU grid_a100-8c

☐ DirectPath IO ☐ Dynamic DirectPath IO ☒ NVIDIA GRID vGPU

NVIDIA GRID vGPU Profile grid_a100-8c

grid_a100-8c

grid_a100-7-40c

grid_a100-5c

grid_a100-4c

grid_a100-40c

grid_a100-4-20c

grid_a100-3-20c

grid_a100-20c

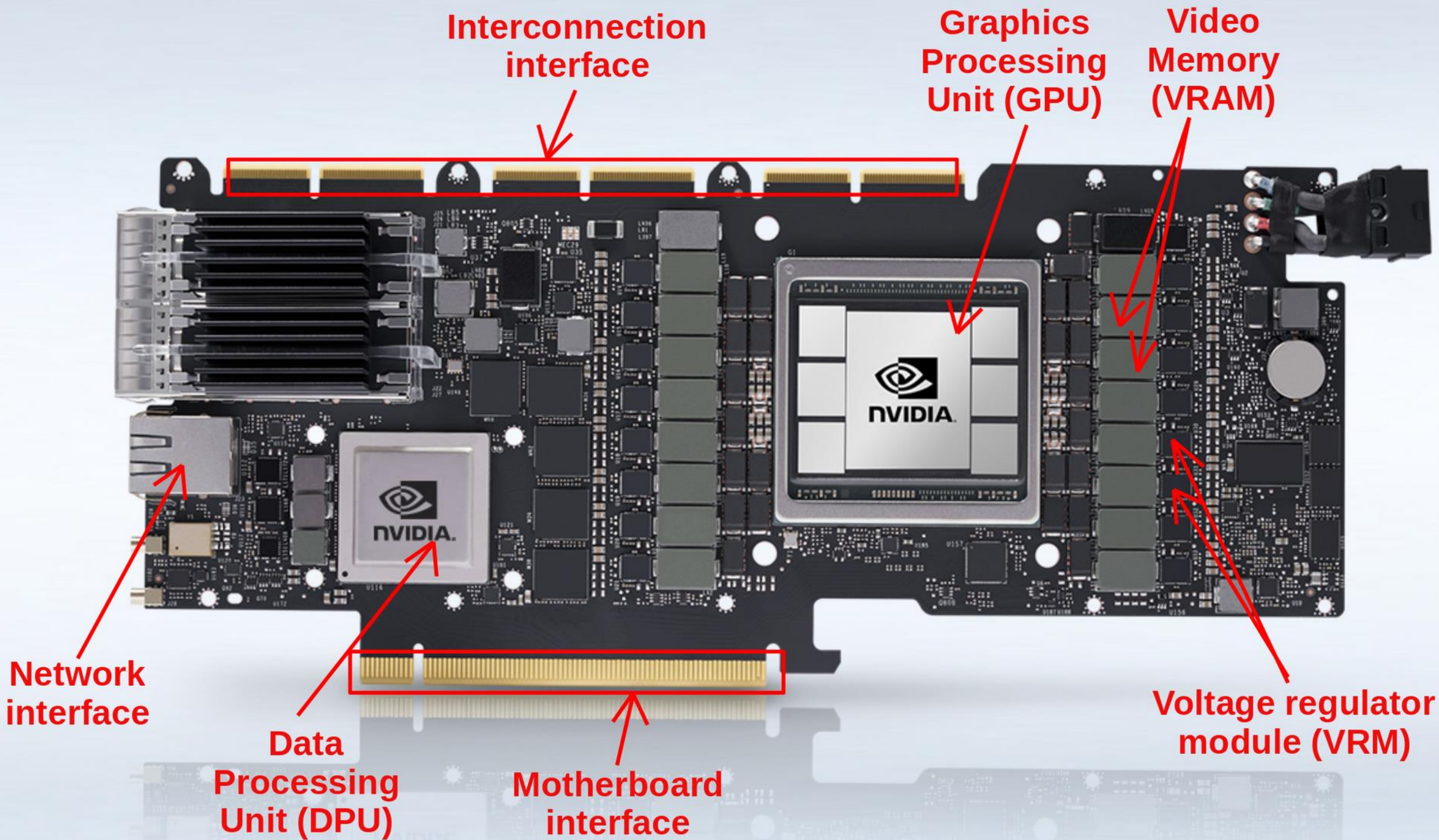
grid_a100-2-10c

grid_a100-10c

grid_a100-1-5c

undefined

CANCEL



Interconnection
interface

Graphics
Processing
Unit (GPU)

Video
Memory
(VRAM)

Network
interface

Data
Processing
Unit (DPU)

Motherboard
interface

Voltage regulator
module (VRM)

Nvidia A100 40GB

5GB	5GB	5GB	5GB	5GB	5GB	5GB	5GB
	1 compute	1 compute	1 compute	1 compute	1 compute	1 compute	1 compute

5GB	5GB	5GB	5GB	5GB	5GB	5GB	5GB
	1g.5gb	GPU Instance <ul style="list-style-type: none">• Fixed partition of memory and compute• Fixed amount of “other” GPU Engines (depending on size)					
	1 compute	1 compute	1 compute	1 compute	1 compute	1 compute	1 compute

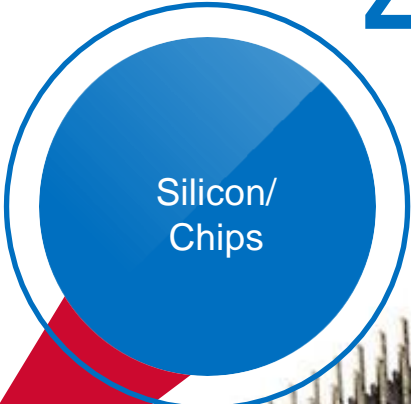
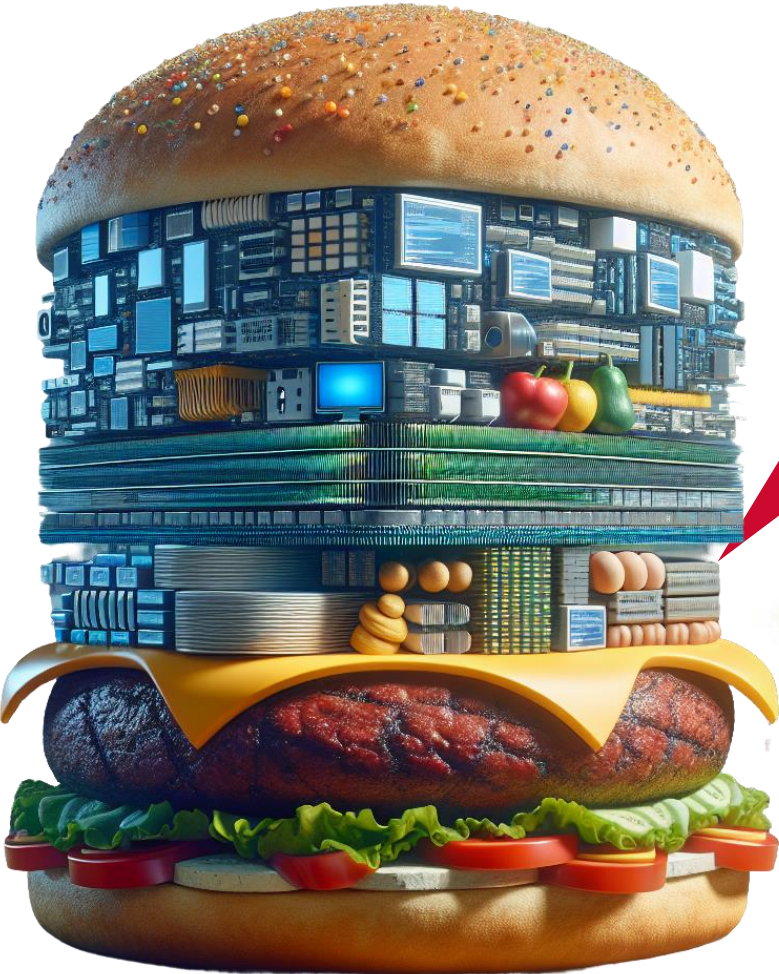
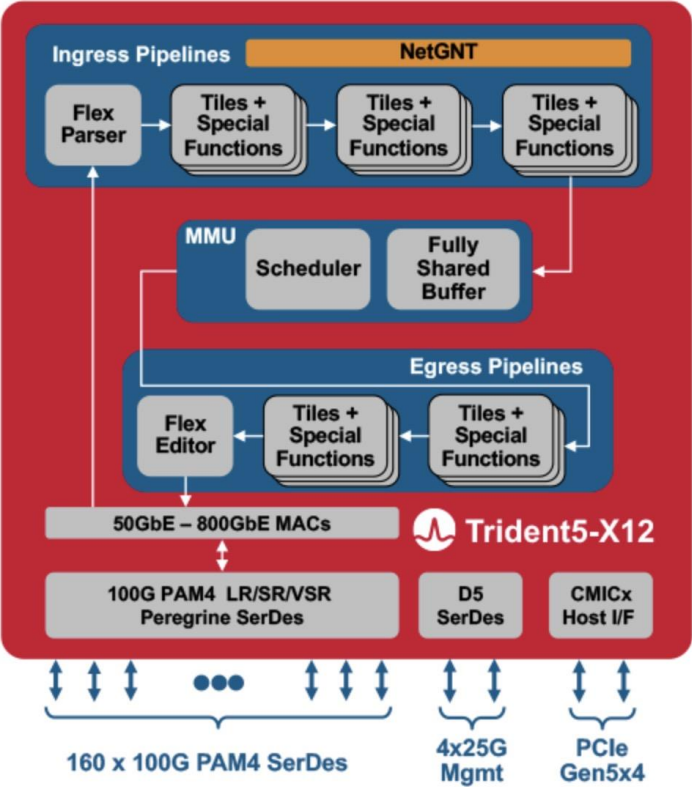
NVIDIA A100 (40GB)

- 8 x 5GB Memory Slices
- 7 Compute Slices

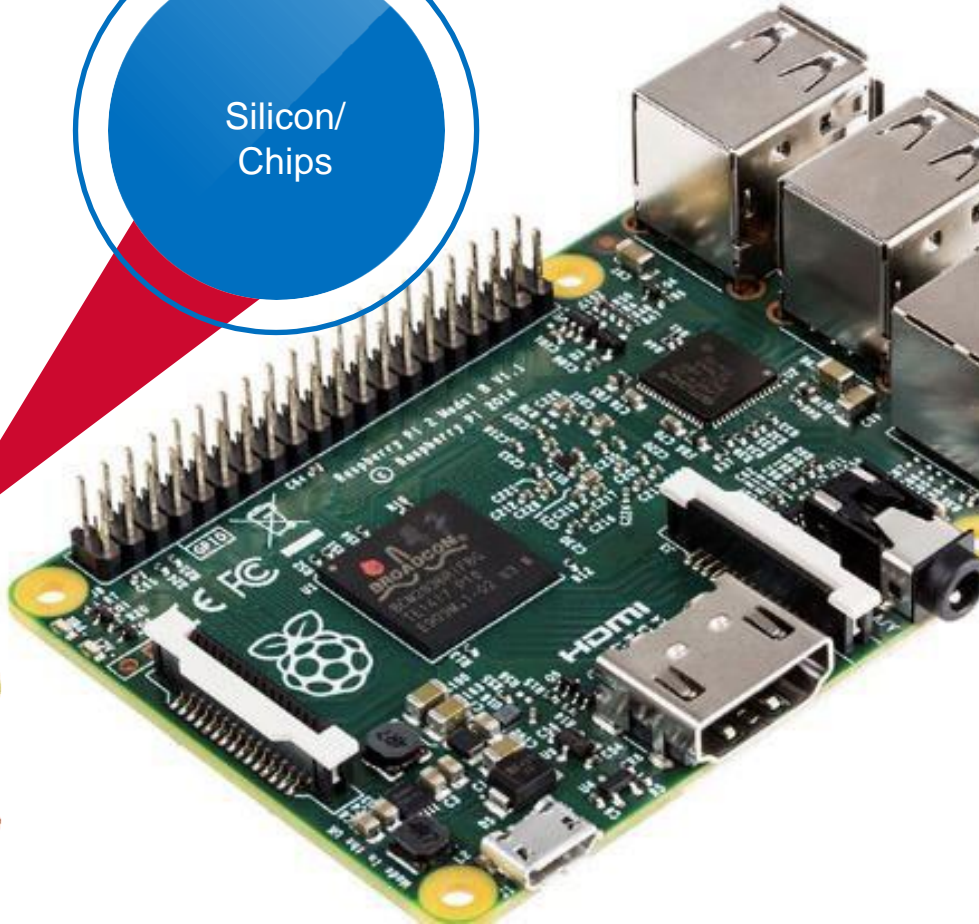
5GB	5GB	5GB	5GB	5GB	5GB	5GB	5GB
	4g.20gb						
	1 compute	1 compute	1 compute	1 compute	1 compute	1 compute	1 compute
5GB	5GB	5GB	5GB	5GB	5GB	5GB	5GB
	1c.4g.20gb						
	1 compute	1 compute	1 compute	1 compute	1 compute	1 compute	1 compute

Silicon/Chips for Compute&Network

Broadcom Trident 5-X12 with 800GbE and AI features an integrated neural network engine called NetGNT (Networking General-purpose Neural-network Traffic-analyzer)

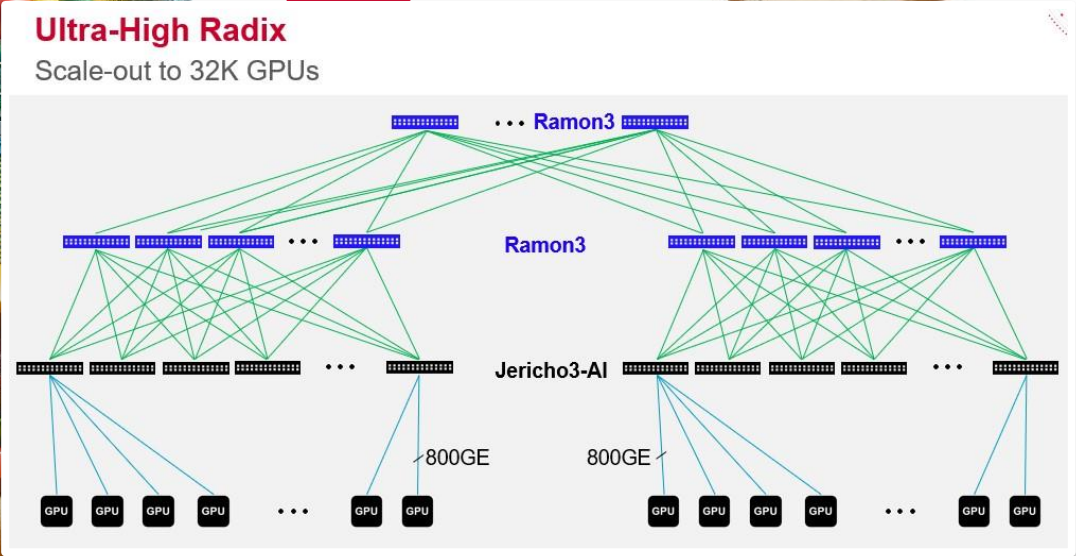
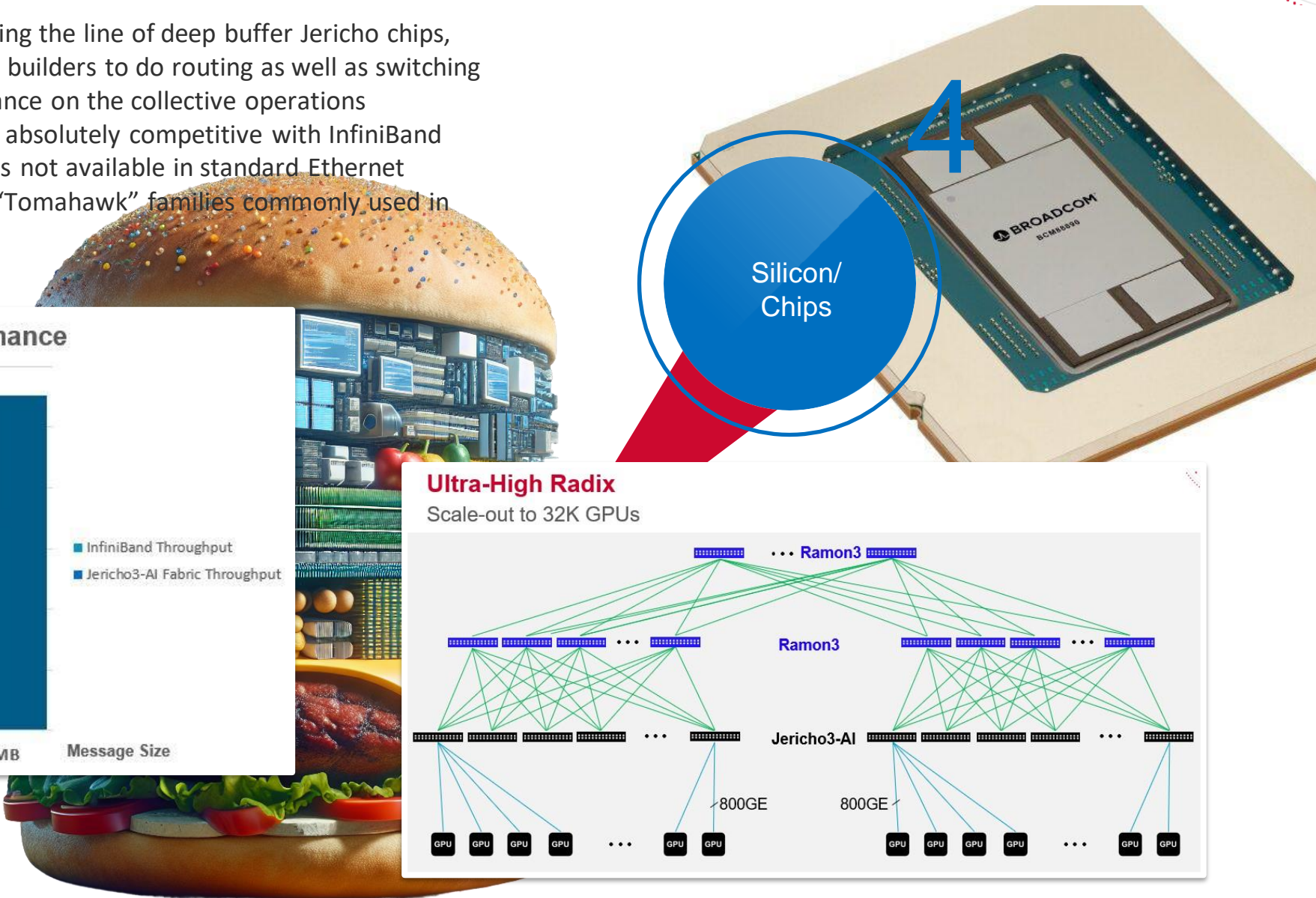
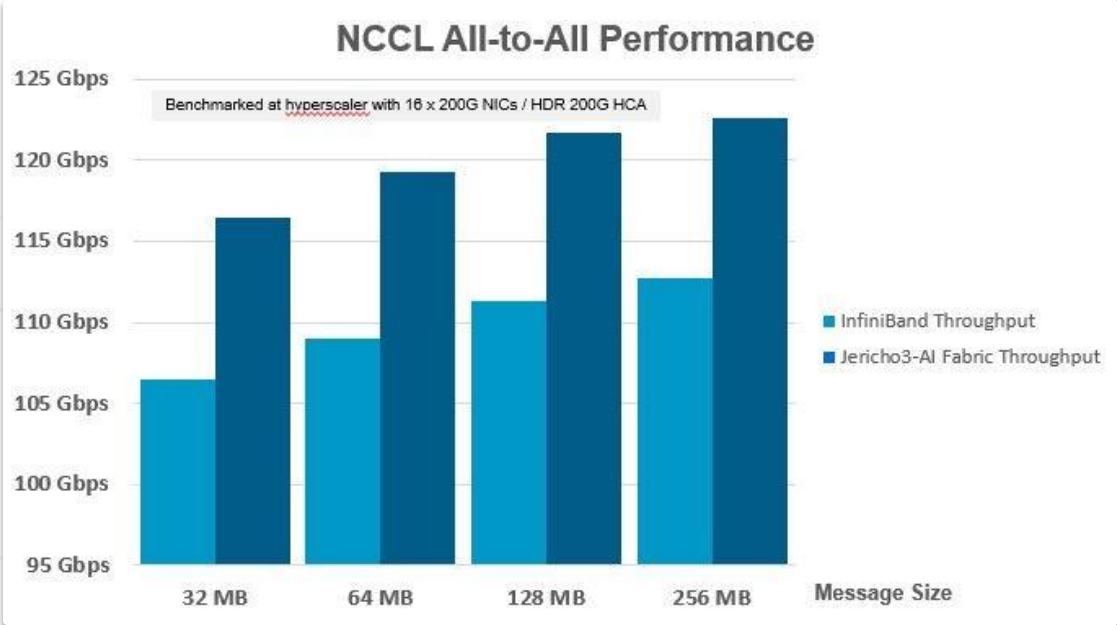


4

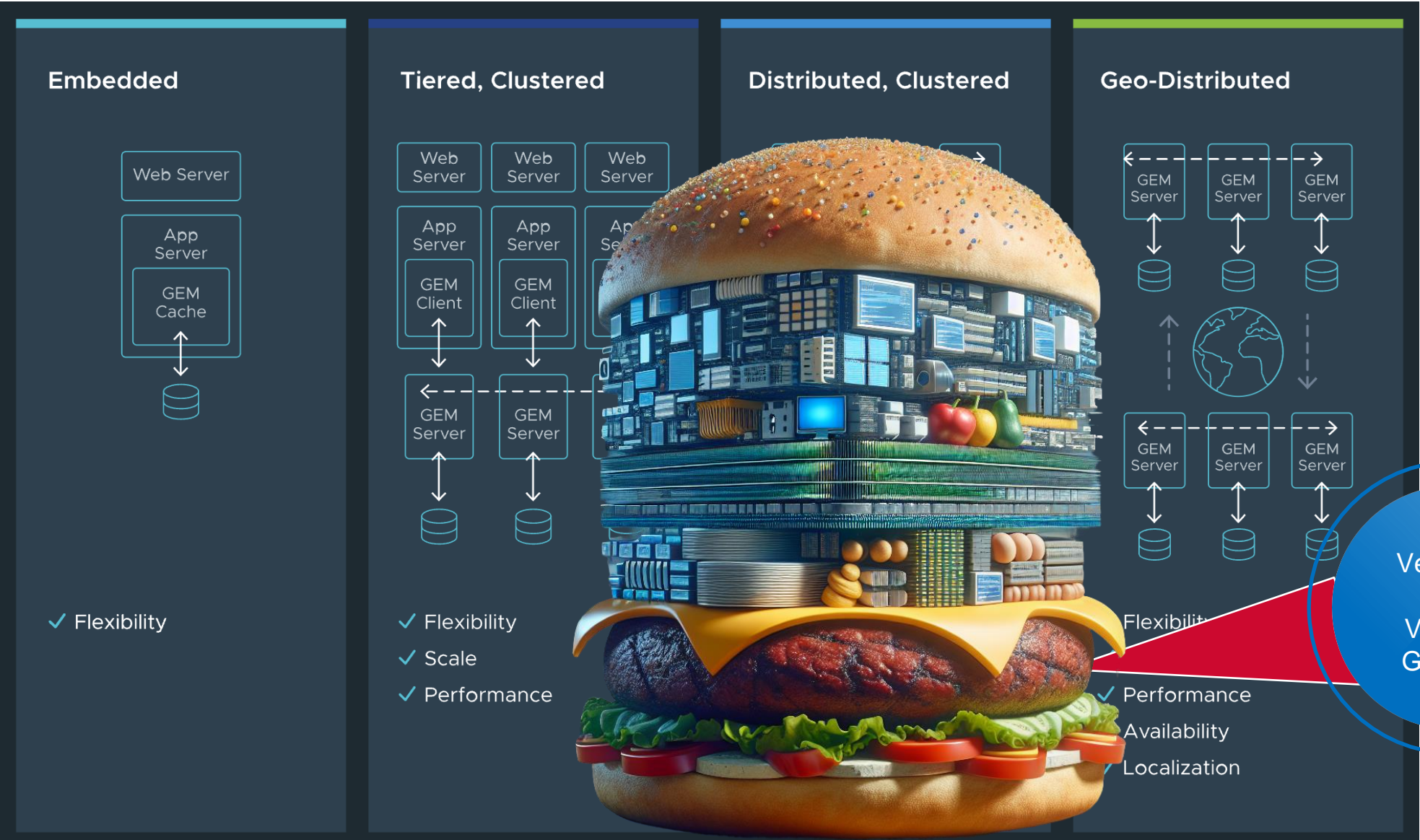


Silicon/Chips for AI

With the **Jericho3-AI chips**, Broadcom is reworking the line of deep buffer Jericho chips, which often are used by hyperscalers and cloud builders to do routing as well as switching functions, and giving them the kind of performance on the collective operations commonly used in AI and HPC that makes them absolutely competitive with InfiniBand for AI workloads and that gives them capabilities not available in standard Ethernet ASICs, including those in its own “Trident” and “Tomahawk” families commonly used in the datacenters of all scales.



Vector DB best for ML



5

VectorDB
VMware
GemFire

Tanzu GemFire

Tanzu[®]
by Broadcom

Векторная DB имеет преимущества в области генеративного искусственного интеллекта и ML:

1. Высокая производительность и масштабируемость
2. Распределенная архитектура
3. Низкая задержка
4. Управление данными в реальном времени
5. Согласованность данных
6. Поддержка транзакций
7. Гибкость и интеграция



Effortless elastic scaling



Real-time event processing



Multisite replication



High availability and business continuity



Predictable low latency



Hidden gem discovery



Polyglot language support



Security that's built in, not bolted on



Cloud ready

VMware Data Solutions

Infrastructure for running modern apps and backing services with consistent, conformant Kubernetes everywhere.



GemFire

Fast In-Memory data store for Caching (Redis compatible), Transactional and NoSQL

I need a fast data store



SQL

Relational MySQL or Postgres database for Transactional or Analytic data processing

I need to replatform a relational database



Greenplum

Massively Parallel Processing (MPP) Postgres for Big Data store for analytics, Machine Learning and Artificial Intelligence

drive analytic value of out tons of existing data



Rabbit MQ

High throughput broker for reliable messaging delivery

I need reliable messaging delivery



I need flexible and manageable data integrations

Spring + Steeltoe + Data Data services, connectors and integration orchestration for data pipelines (ex: ETL, streaming, etc.)



Data Management
Management for VMware Data Solutions instances

Features

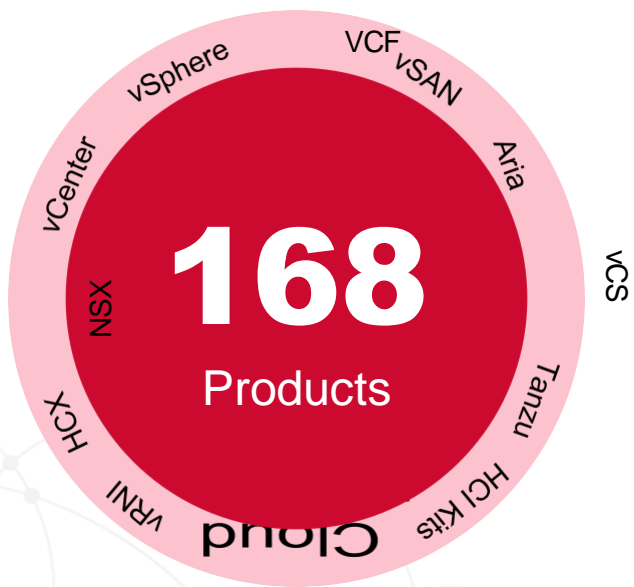
- ✓ **Cloud deployed backing-services**
- ✓ **Self Provisioning**
- ✓ **On-Premise and Multi-Cloud**
- ✓ **Scaling**
- ✓ **HA - Fault Tolerant**
- ✓ **Secured access**
- ✓ **Based on open source**
- ✓ **24/7 Support**

That's all?

CALL TO ACTION

More технологий для Private AI

От VMware by Broadcom ☺



NSX

Network and security virtualization platform



VMware Tanzu™
GemFire®



vSAN

Flash-optimized, vSphere-native storage




VMware vSphere+

Enterprise Workload Platform



VMware Cloud
Foundation

Multi-cloud Platform

 **BROADCOM®**
AI Ready. A u?

Thank You

