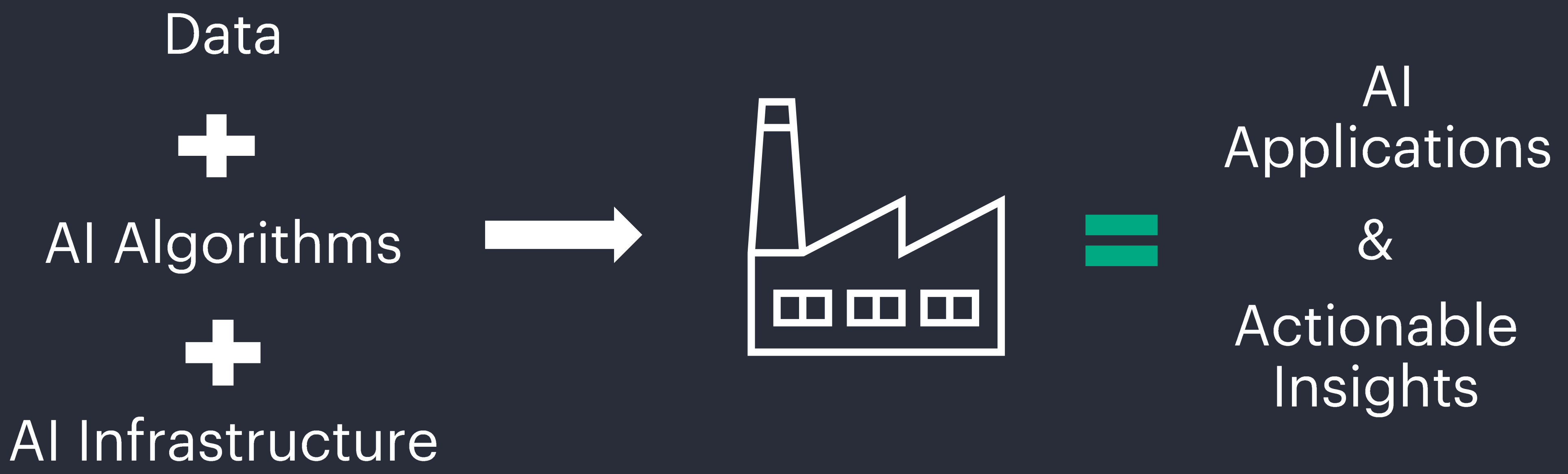
建構未來 AI 工廠：基於可擴展架構，創造實質商業成果

范欽輝 Chin-Hui Fan, HPE 技術規劃處 副總經理

October 16, 2025

AI Factory 的重要目標為具備高度可擴展性、可重複性、韌性及安全性



Data

AI Algorithms

AI Infrastructure

AI Applications

&

Actionable Insights

AI Factory 為生產一種新型商品 : AI 應用服務
支援大型組織所需的龐大規模的算力運作
提供類雲端操作體驗，並可支援各類AI 工作負載

AI Factory 建置時須考量的功能

Service Catalog

提供各式應用服務目錄
Catalogs for different services offering

Accelerated Deployment

需快速部署模型與應用，簡化繁瑣設定及維運
Ability to deploy and access resources on the go

Orchestration & Integrations

自動化協調與整合
Integrations with different cloud platforms / ISV software platforms

Multi-Tenancy & RBAC

多租戶資源與權限管理
Logical and physical isolation of resources between tenants

Monitoring & Management

系統、資源監控與管理
Real-time monitoring of GPU utilization and performance

Pay as You Go

紀錄並統計用量
Metering and Billing based on usage

任務導向的 HPE AI Factory 解決方案

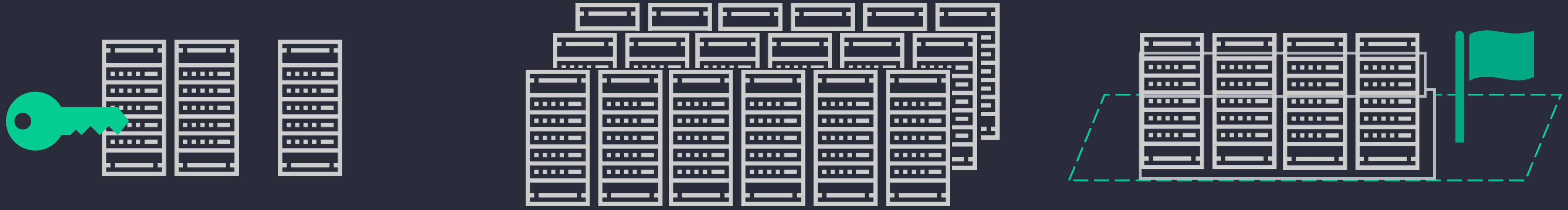
for every AI ambition, across clouds, cores and countries

Turnkey AI factory
Enterprises

AI factory at scale
Model builders & SP's

Sovereign AI factory
Governments, public sector

Common control plane: HPE Morpheus and HPE OpsRamp



Turnkey, engineered systems

Customized, validated solutions

Infrastructure | Software | Services | Ecosystem | Sustainability

任務導向的 HPE AI Factory 解決方案

for every AI ambition, across clouds, cores and countries

Turnkey AI factory
General Enterprises

- Demand rapid ROI
- NVIDIA Software-preference
- Inference & tuning
- Air-cooled

AI factory at scale
Model builders & SP's

- Tailor to your scale and scenario
- Services integrated software stack
- Model dev, training & inference
- Direct liquid & air-cooled

Sovereign AI factory
Governments, public sector

- Independence from others
- Strict data sovereignty
- Model dev, training & inference
- Direct liquid & air-cooled

Common control plane: HPE Morpheus and HPE OpsRamp

Turnkey, engineered systems

Customized, validated solutions

Infrastructure | Software | Services | Ecosystem | Sustainability

AI Factory 的核心理念

Full-Stack 全方位整合

Role Based 協同作業

Adaptable 彈性與模組化

Turnkey AI Factory

Enterprise AI infrastructure simplified

Turnkey AI factory | AI factory at scale | Sovereign AI factory

Instant AI productivity

Secure and unified data access

End-to-end AI software platform

Use cases

Development

Computer vision

Agentic AI

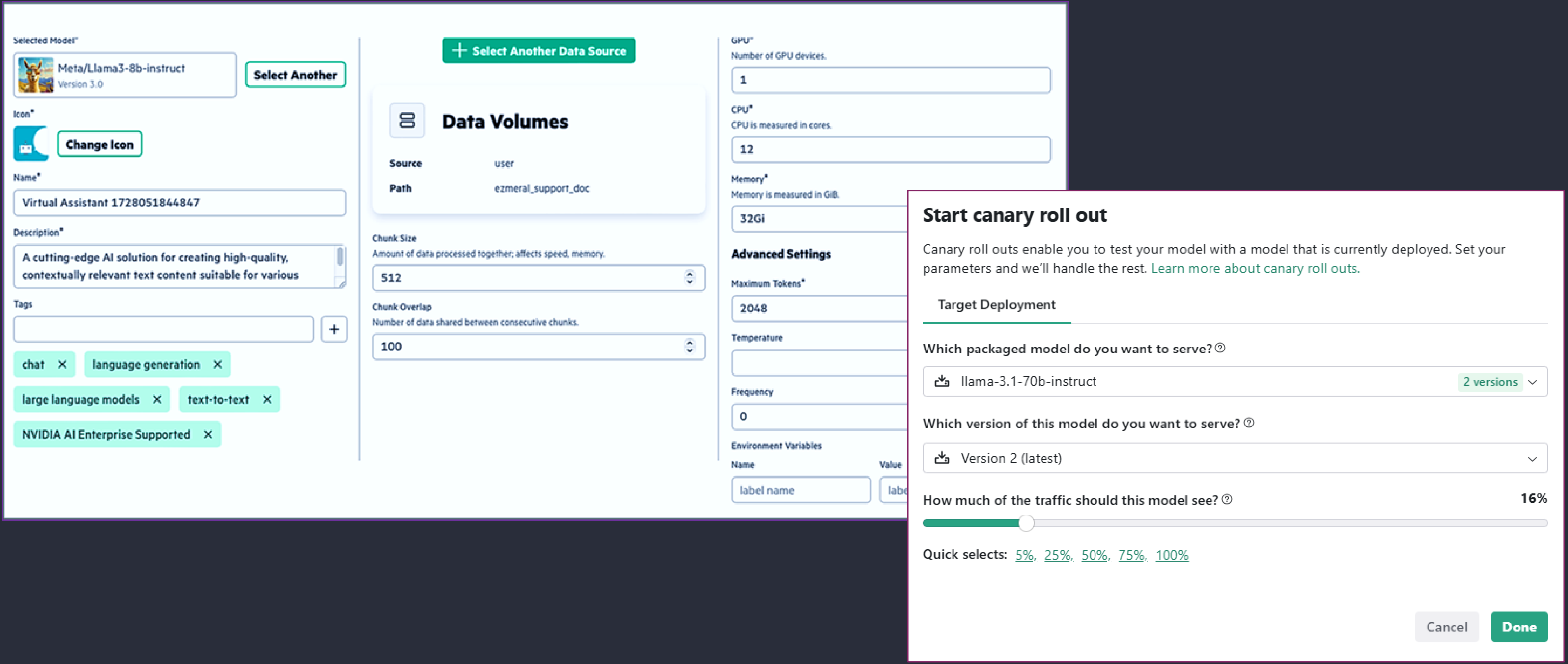
Physical AI

RAG

全方位整合可降低相容性風險、監控資源使用效率、維持一致性的維運管理，實現快速部署與應用

Low-code simplicity

建立 RAG 應用與模型上架



Selected Model*

Meta/Llama3-8b-instruct
Version 3.0

Icon*

Name*

Virtual Assistant 1728051844847

Description*

A cutting-edge AI solution for creating high-quality, contextually relevant text content suitable for various

Tags

chat ✕ | language generation ✕ | large language models ✕ | text-to-text ✕ | NVIDIA AI Enterprise Supported ✕

Data Volumes

Source	user
Path	ezmeral_support_doc

Chunk Size
Amount of data processed together; affects speed, memory.
512

Chunk Overlap
Number of data shared between consecutive chunks.
100

GPU*
Number of GPU devices.
1

CPU*
CPU is measured in cores.
12

Memory*
Memory is measured in GiB.
32Gi

Advanced Settings

Maximum Tokens*
2048

Temperature

Frequency
0

Environment Variables

Name: label name
Value: labe

Start canary roll out

Canary roll outs enable you to test your model with a model that is currently deployed. Set your parameters and we'll handle the rest. Learn more about canary roll outs.

Target Deployment

Which packaged model do you want to serve? ⓘ
llama-3.1-70b-instruct — 2 versions

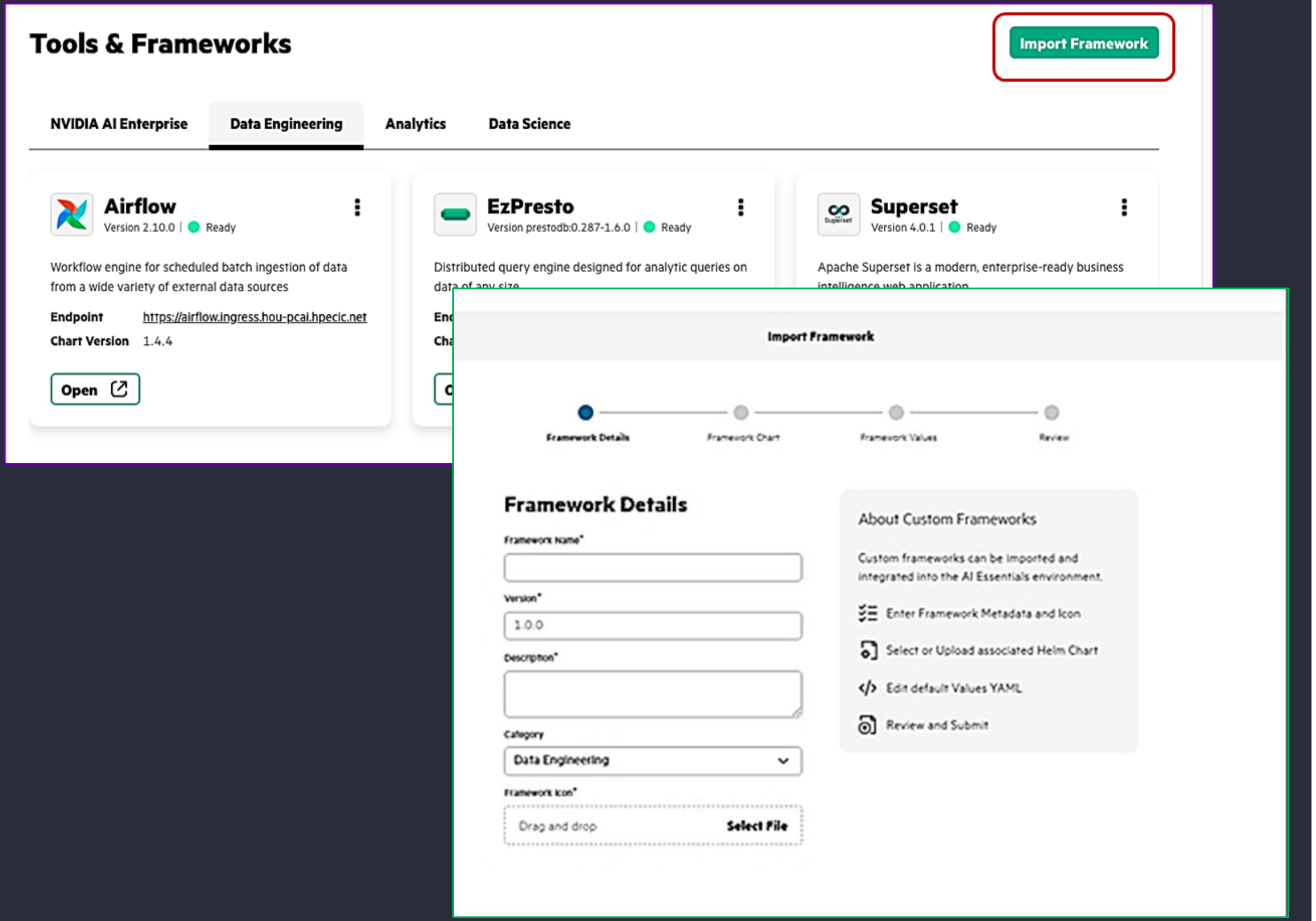
Which version of this model do you want to serve? ⓘ
Version 2 (latest)

How much of the traffic should this model see? ⓘ 16%

Quick selects: 5%, 25%, 50%, 75%, 100%

Bring Your Own Applications

NVAIE Blueprint



Tools & Frameworks

Import Framework

NVIDIA AI Enterprise | Data Engineering | Analytics | Data Science

Airflow
Version 2.10.0 | ● Ready

Workflow engine for scheduled batch ingestion of data from a wide variety of external data sources

Endpoint https://airflow.ingress.hou-pcai.hpecic.net
Chart Version 1.4.4

Open

EzPresto
Version prestodb:0.287-1.6.0 | ● Ready

Distributed query engine designed for analytic queries on data of any size

Superset
Version 4.0.1 | ● Ready

Apache Superset is a modern, enterprise-ready business intelligence web application

Import Framework

Framework Details — Framework Chart — Framework Values — Review

Framework Details

Framework Name*

Version*
1.0.0

Description*

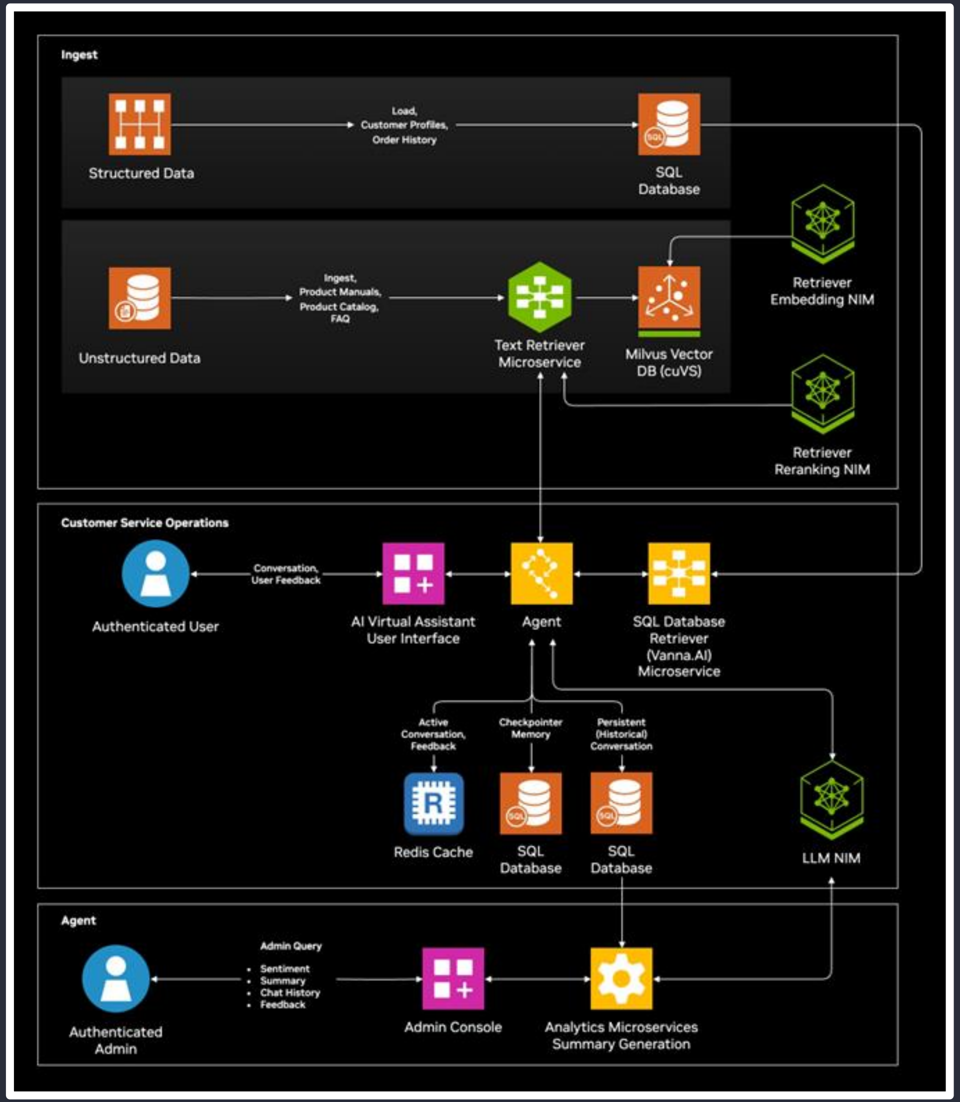
Category
Data Engineering

Framework Icon*
Drag and drop | Select File

About Custom Frameworks

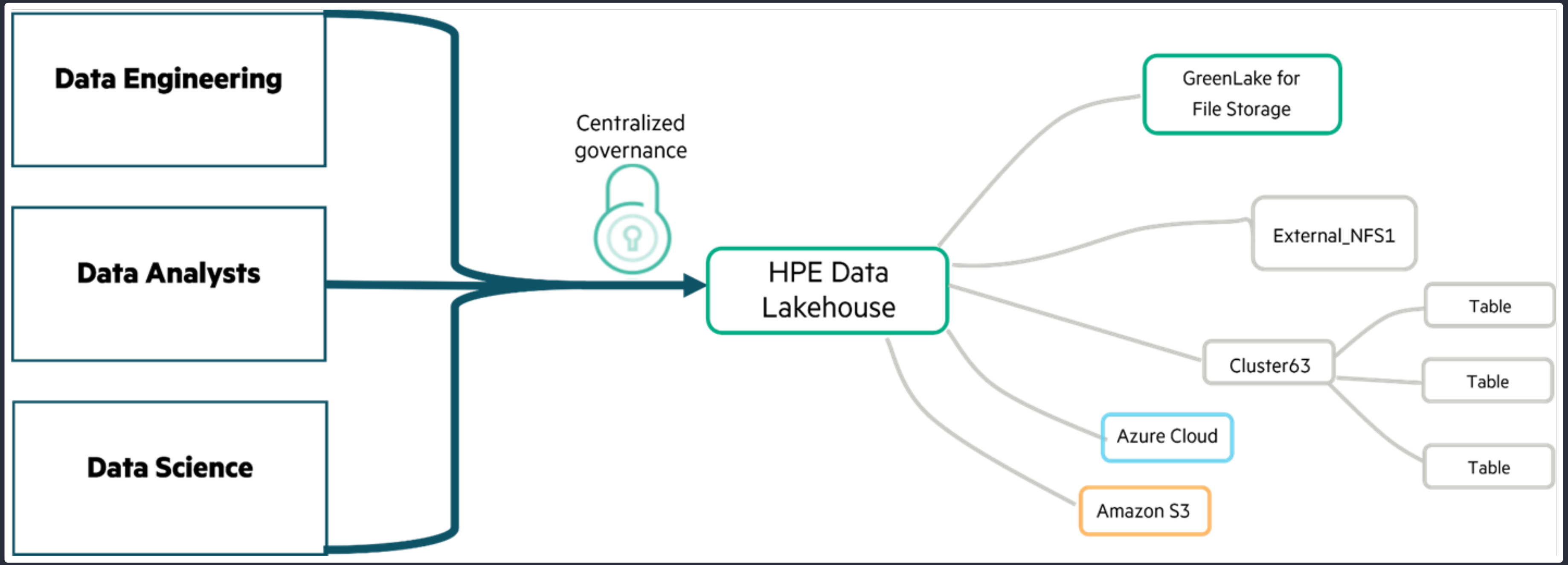
Custom frameworks can be imported and integrated into the AI Essentials environment.

- Enter Framework Metadata and Icon
- Select or Upload associated Helm Chart
- Edit default Values YAML
- Review and Submit


Ingest
Structured Data
Load, Customer Profiles, Order History
SQL Database
Retriever Embedding NIM
Unstructured Data
Ingest, Product Manuals, Product Catalog, FAQ
Text Retriever Microservice
Milvus Vector DB (cuVS)
Retriever Reranking NIM
Customer Service Operations
Authenticated User
Conversation, User Feedback
AI Virtual Assistant User Interface
Agent
SQL Database Retriever (Vanna.AI) Microservice
Active Conversation, Feedback
Checkpointer Memory
Persistent (Historical) Conversation
Redis Cache
SQL Database
SQL Database
LLM NIM
Agent
Authenticated Admin
Admin Query
Sentiment
Summary
Chat History
Feedback
Admin Console
Analytics Microservices Summary Generation

簡化資料存取
Unified data Lakehouse

Turnkey AI factory | AI factory at scale | Sovereign AI factory



Global Namespace

Present external iceberg format data sources (S3, NFS) as federated resource to HPE AI Essentials

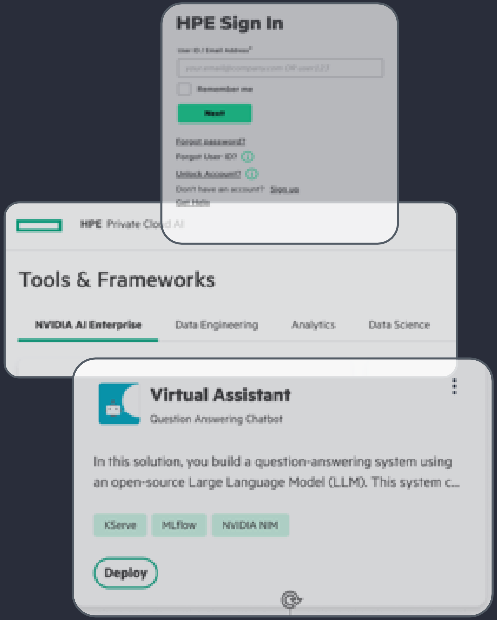
Use your preferred analytics engines

BUILT ON APACHE ICEBERG OPEN API

Turnkey AI factory | AI factory at scale | Sovereign AI factory

Turnkey AI Factory

HPE Private Cloud AI 基礎架構-四種配置規格



In this solution, you build a question-answering system using an open-source Large Language Model (LLM). This system c...

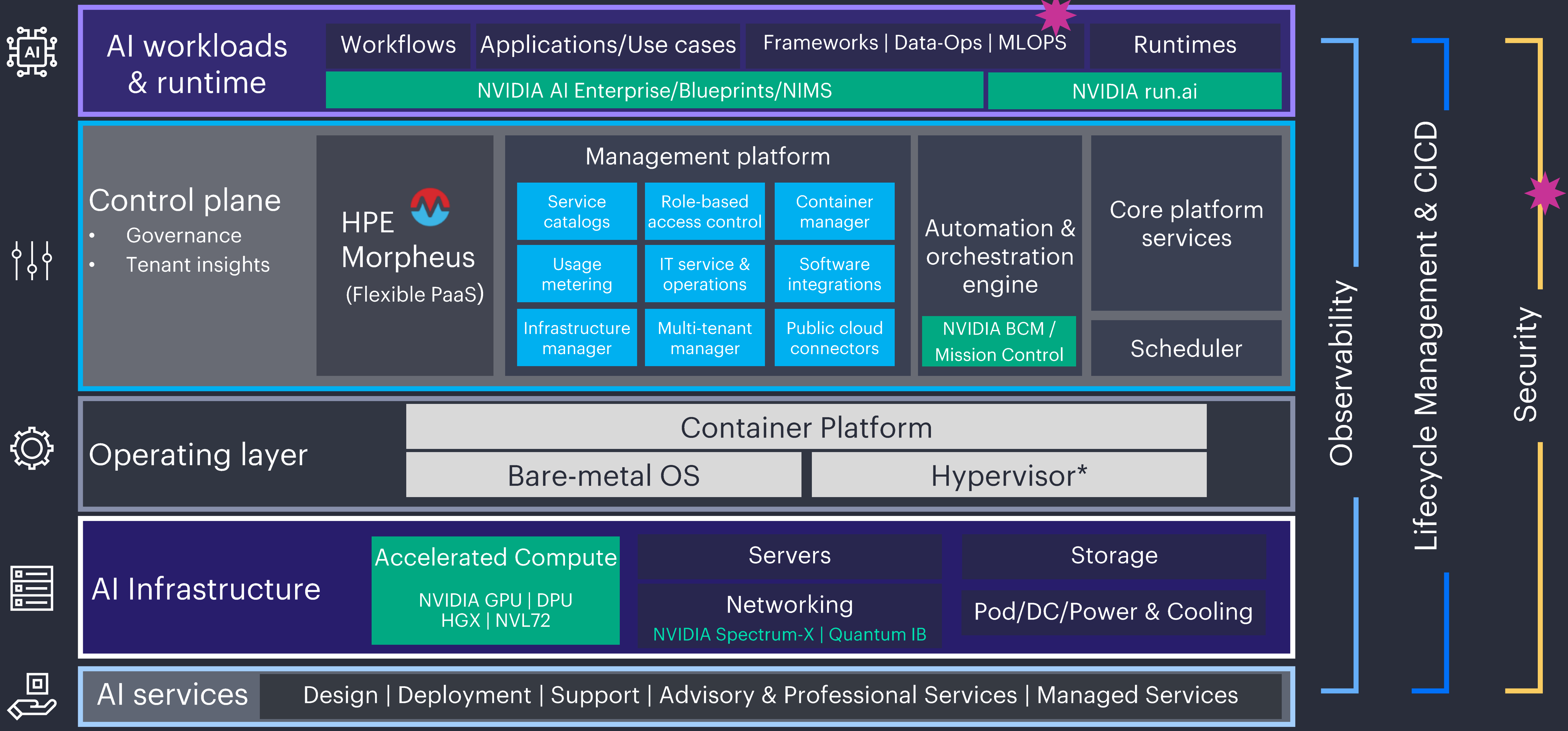
	All-in-one Node AI Sandbox	All-in-one Rack	All-in-one Rack	Scalable by Rack
	Developer System **(1 x DL380a Gen11)**	**Small** **(1 or 2x DL380a Gen12)**	**Medium** **(2 x DL380a Gen12)**	**Large** **(2x DL380a Gen12)**
Compute	2 NVIDIA H100 NVL GPU's	4 /8 NVIDIA RTX 6000 GPUs	8 NVIDIA H200NVL GPUs	16 NVIDIA H200NVL GPUs
Storage	32 TB Integrated	109 TB file storage in rack	109 TB file storage in rack	217TB file storage in rack
Networking	Customer Network	400GbE NVIDIA Networking	400GbE NVIDIA Networking	400GbE NVIDIA Networking
Power	Up to 2.2 kW	10 kW per rack	13 KW per rack	17 KW per rack

Optional 8~16 GPU expansion racks for Small, Med & Large

Unified experience through HPE GreenLake Cloud

At-scale AI Factory Solution

Turnkey AI factory | AI factory at scale | Sovereign AI factory



AI workloads & runtime

Workflows | Applications/Use cases | Frameworks | Data-Ops | MLOPS ✱ | Runtimes

NVIDIA AI Enterprise/Blueprints/NIMS | NVIDIA run.ai

Control plane
- Governance
- Tenant insights

HPE Morpheus (Flexible PaaS)

Management platform

Service catalogs | Role-based access control | Container manager

Usage metering | IT service & operations | Software integrations

Infrastructure manager | Multi-tenant manager | Public cloud connectors

Automation & orchestration engine

NVIDIA BCM / Mission Control

Core platform services

Scheduler

Operating layer

Container Platform

Bare-metal OS | Hypervisor*

AI Infrastructure

Accelerated Compute

NVIDIA GPU | DPU
HGX | NVL72

Servers

Networking

NVIDIA Spectrum-X | Quantum IB

Storage

Pod/DC/Power & Cooling

AI services

Design | Deployment | Support | Advisory & Professional Services | Managed Services

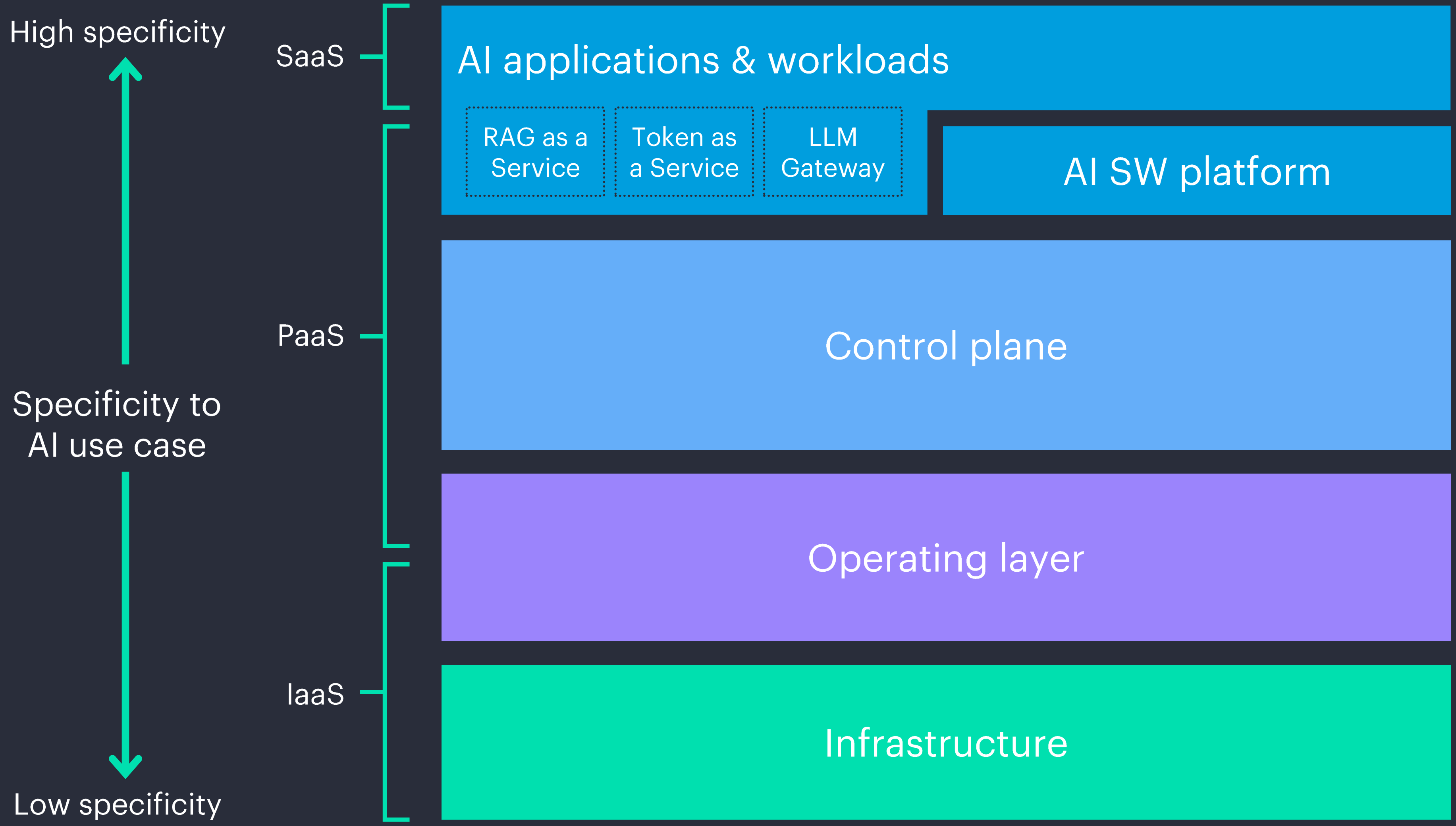
Observability

Lifecycle Management & CICD

Security ✱

✱ Optional Modules

AI Factory at scale – 邏輯化服務階層



High specificity

Specificity to AI use case

Low specificity

SaaS

PaaS

IaaS

AI applications & workloads

RAG as a Service

Token as a Service

LLM Gateway

AI SW platform

Control plane

Operating layer

Infrastructure

Tier 4 AI SaaS
- AI SW platform
- LLM as a Service
- Token as a Service
- Custom AI apps

Tier 3 AI PaaS
- Multi-tenant manager
- Self-service catalog
- E2E automation

Tier 2 IaaS
- GPU as a Service
- Managed Kubernetes / VMs

Tier 1 AI Foundation
- Bare-metal GPU cluster
- Validated performance and reliability

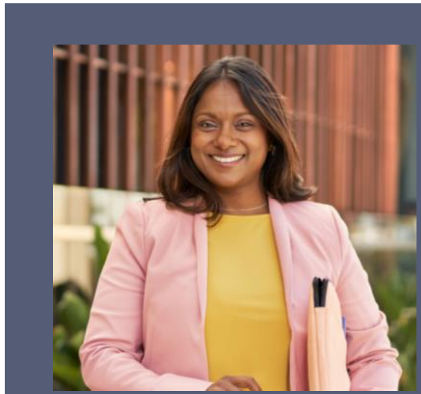
垂直整合，加速協同作業

Infrastructure administrator

- Responsible for allocating infrastructure
- Manage available resources
- Create a tenant/workgroup
- Provision IP address spaces for each tenant

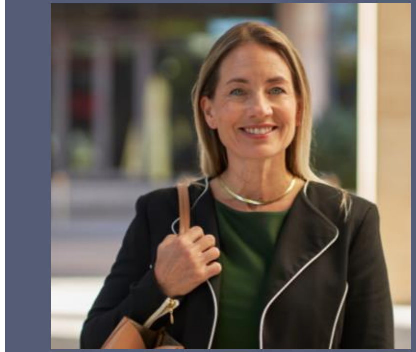
MLOps engineer

- Responsible for curating AI tools
- Creates projects and assigns users to projects
- Sets resources quotas on a project
- Reports on project utilization



Tenant administrator

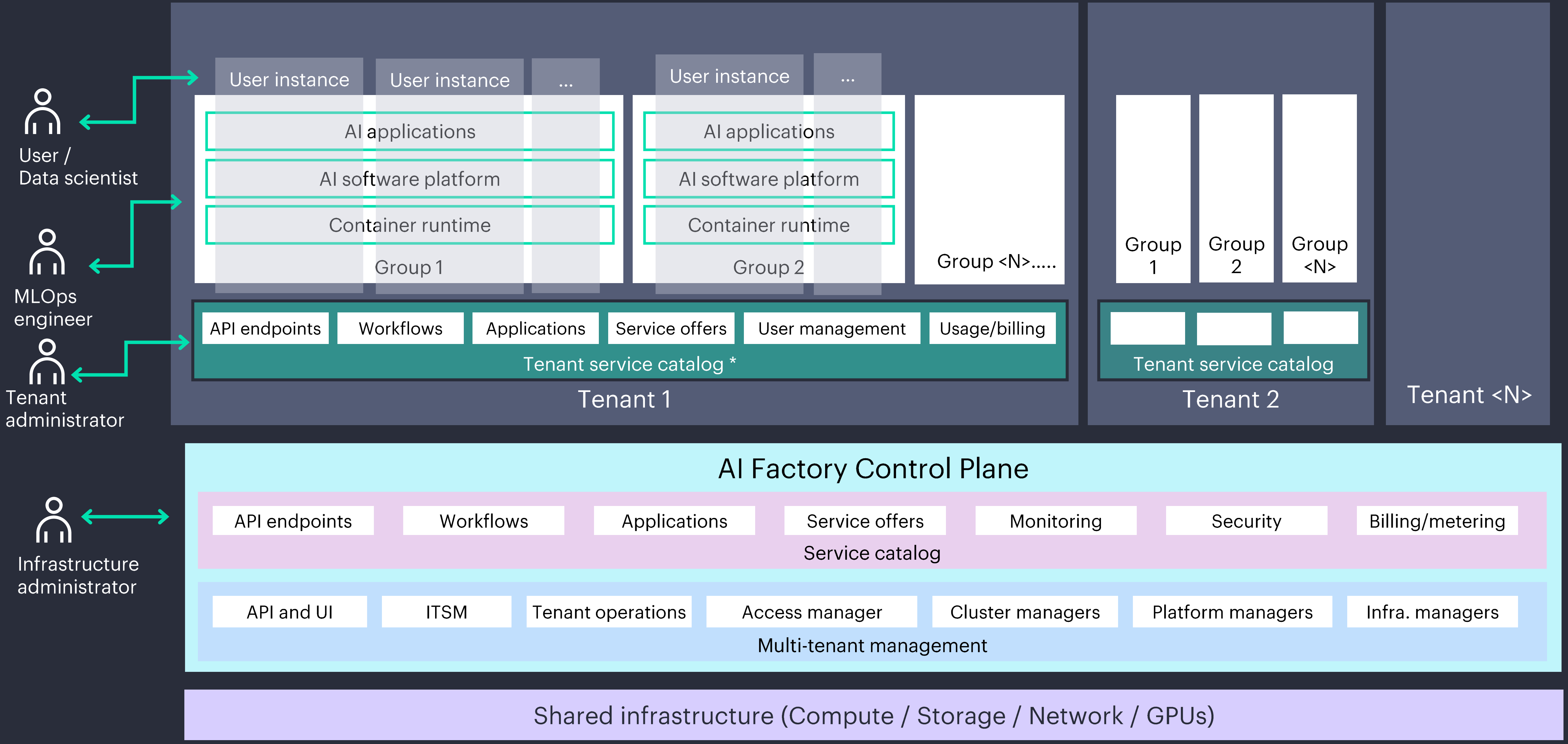
- Responsible for managing tenant assigned resources
- Creates one or more platforms using allocated resources
- Creates a workspace and assigns users to these workspaces



Data scientist

- Uses the tools and apps provided by the platform
- Executes an experiment within the platform
- Deploys AI applications and workloads

垂直整合，加速協同作業



User instance

User instance

...

User instance

...

AI applications

AI software platform

Container runtime

Group 1

AI applications

AI software platform

Container runtime

Group 2

Group <N>.....

Group 1

Group 2

Group <N>

User / Data scientist

MLOps engineer

Tenant administrator

API endpoints | Workflows | Applications | Service offers | User management | Usage/billing

Tenant service catalog *

Tenant 1

Tenant service catalog

Tenant 2

Tenant <N>

Infrastructure administrator

AI Factory Control Plane

API endpoints | Workflows | Applications | Service offers | Monitoring | Security | Billing/metering

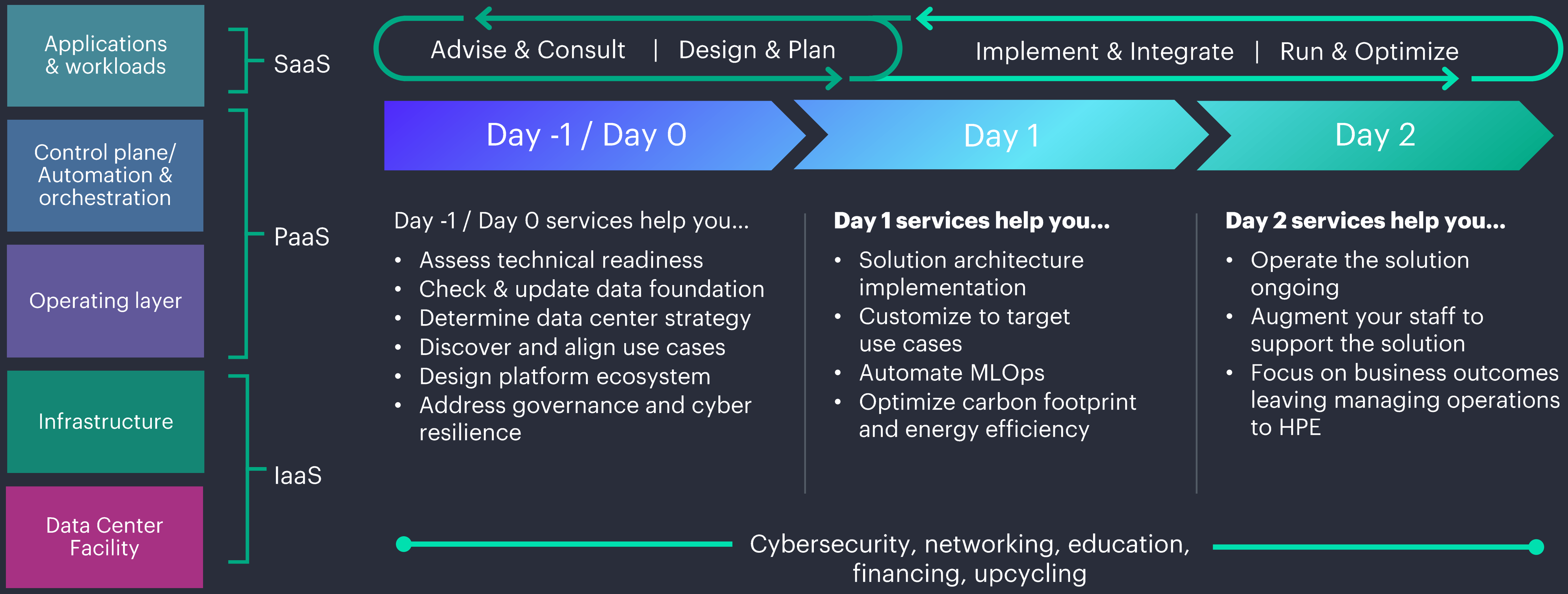
Service catalog

API and UI | ITSM | Tenant operations | Access manager | Cluster managers | Platform managers | Infra. managers

Multi-tenant management

Shared infrastructure (Compute / Storage / Network / GPUs)

共同建立維運能量



Applications & workloads

Control plane/ Automation & orchestration

Operating layer

Infrastructure

Data Center Facility

SaaS

PaaS

IaaS

Advise & Consult | Design & Plan

Implement & Integrate | Run & Optimize

Day -1 / Day 0

Day 1

Day 2

Day -1 / Day 0 services help you…

- Assess technical readiness
- Check & update data foundation
- Determine data center strategy
- Discover and align use cases
- Design platform ecosystem
- Address governance and cyber resilience

Day 1 services help you…

- Solution architecture implementation
- Customize to target use cases
- Automate MLOps
- Optimize carbon footprint and energy efficiency

Day 2 services help you…

- Operate the solution ongoing
- Augment your staff to support the solution
- Focus on business outcomes leaving managing operations to HPE

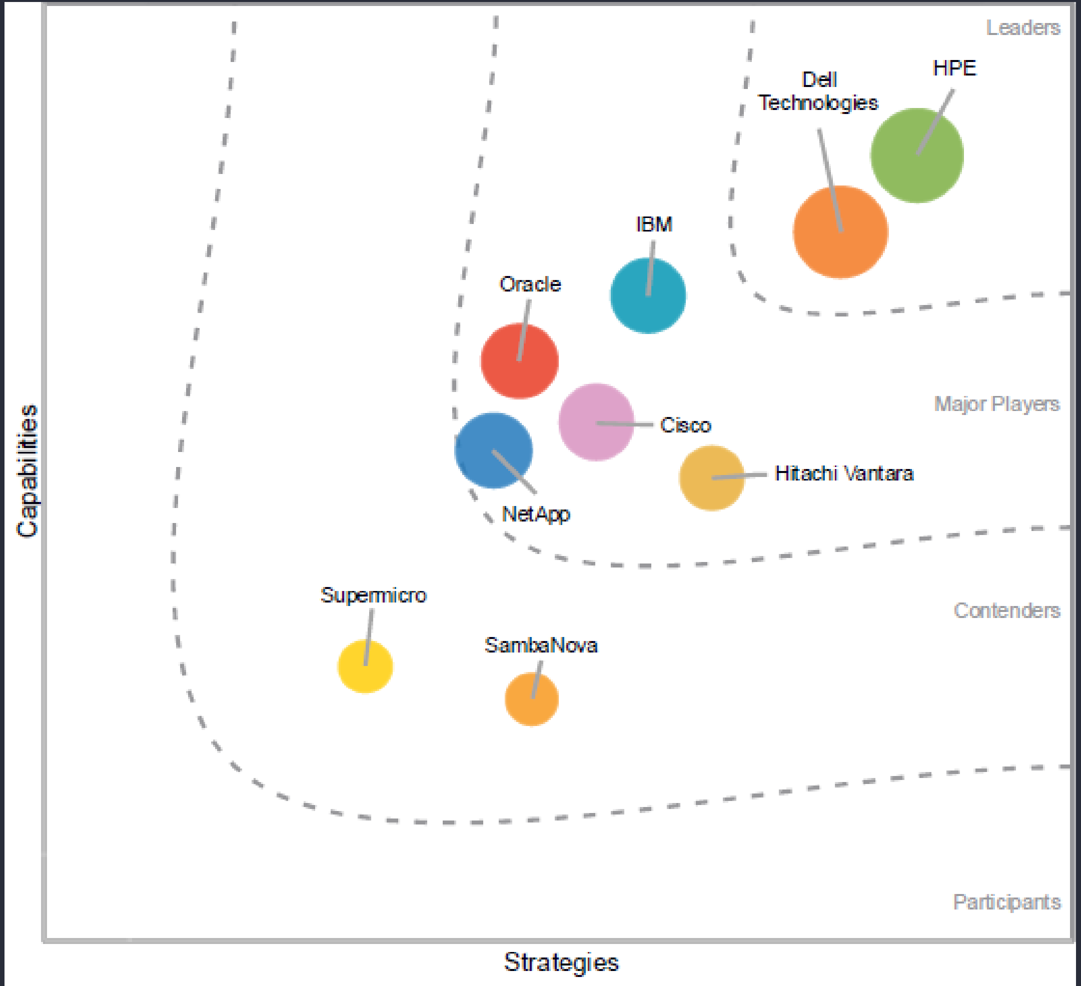
Cybersecurity, networking, education, financing, upcycling

IDC Marketscape: Worldwide Private AI Infrastructure Systems 2025 Vendor Assessment (August 2025)

HPE is the clear leader.

"As AI innovation accelerates, private AI infrastructure systems are emerging as important options for customers that want faster deployment of complete, optimized, fit-for-purpose stacks in dedicated on premises or collocated facilities."

-Mary Johnston Turner
IDC Research VP, Digital Infrastructure Strategies
WW Infrastructure Research



Source: IDC MarketScape: Worldwide Private AI Inrastrcture Systems 2025 Vendor Assessment

HPE DISCOVER 2025

HPE Cooling Technologies

The Reality of Direct Liquid Cooling: One Size Does Not Fit All

D. I. Tsai, Solution Architect, HPC & AI, APAC

Oct 16, 2025

HPE has delivered the three world’s fastest, verified supercomputers

#1
EL CAPITAN
Lawrence Livermore National Laboratory
U.S. DEPARTMENT OF ENERGY
NNSA
Hewlett Packard Enterprise
AMD
Red Hat
EL CAPITAN
ADVANCED SIMULATION & COMPUTING

ranked
SUPERCOMPUTER
in the world.
at 1.742 exaflops.

TOP 500 The List.

#2
CRAY
OAK RIDGE National Laboratory
U.S. DEPARTMENT OF ENERGY
Hewlett Packard Enterprise
AMD
FRONTIER

ranked
SUPERCOMPUTER
in the world.
at 1.353 exaflops.

TOP 500 The List.

#3
Argonne NATIONAL LABORATORY
U.S. DEPARTMENT OF ENERGY
intel
Hewlett Packard Enterprise
Aurora

ranked
SUPERCOMPUTER
in the world.
at 1.012 exaflops.

TOP 500 The List.

Sources: Nov 2024 Top500

Enabling Large-Scaling AI Workloads Around the Globe

Los Alamos
NATIONAL LABORATORY
EST.1943

10 EFLOPS

single-precision AI Performance
with NVIDIA GH200 superchips

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

20 EFLOPS

single-precision AI Performance
with NVIDIA GH200 superchips

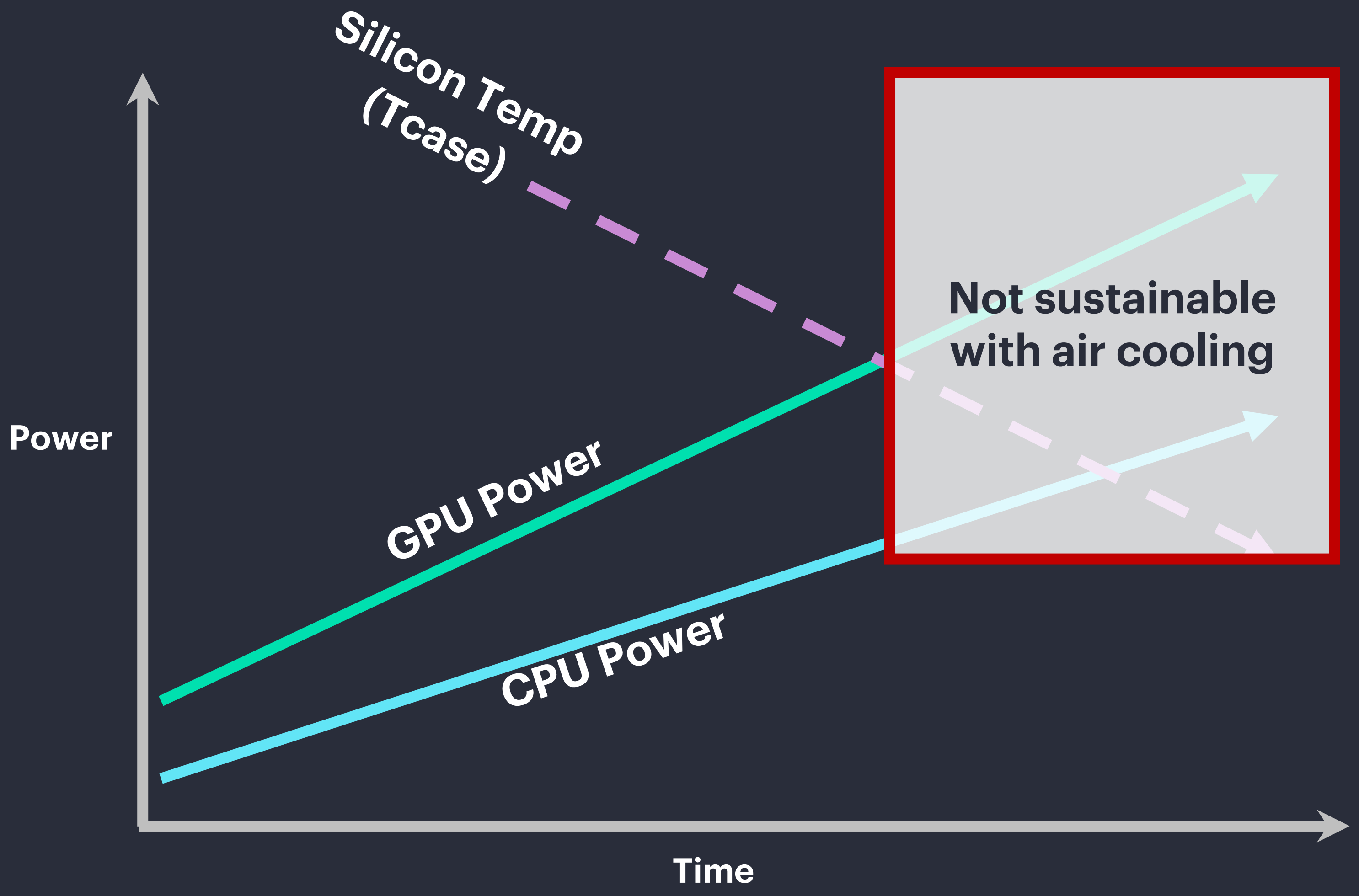
BriCS
Bristol Centre for Supercomputing

21 EFLOPS

single-precision AI Performance
with NVIDIA GH200 superchips

Cooling matters more than ever

The cooling dilemma



Not sustainable with air cooling

Why liquid cooling

Performance

Reliable top-bin CPU/GPU operation

Sustained turbo modes

Density

More servers per rack

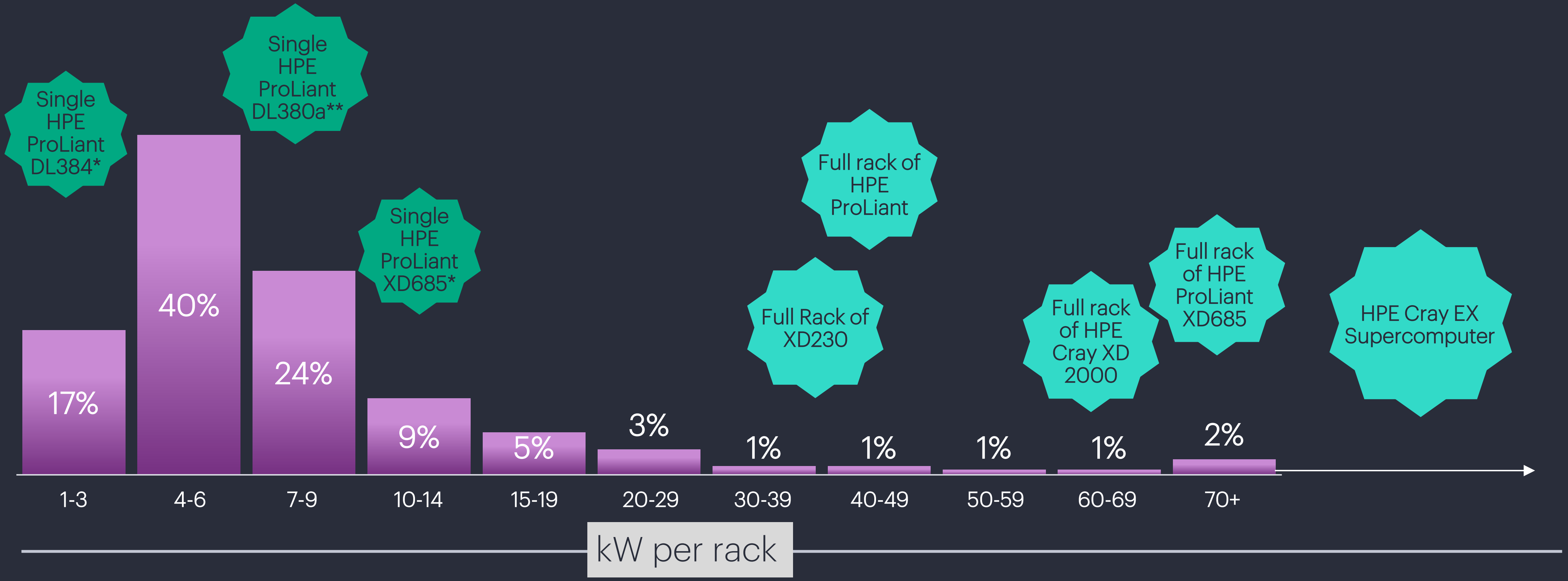
Fewer racks required

Efficiency

More effective heat capture

Lower cooling power required

Power trends

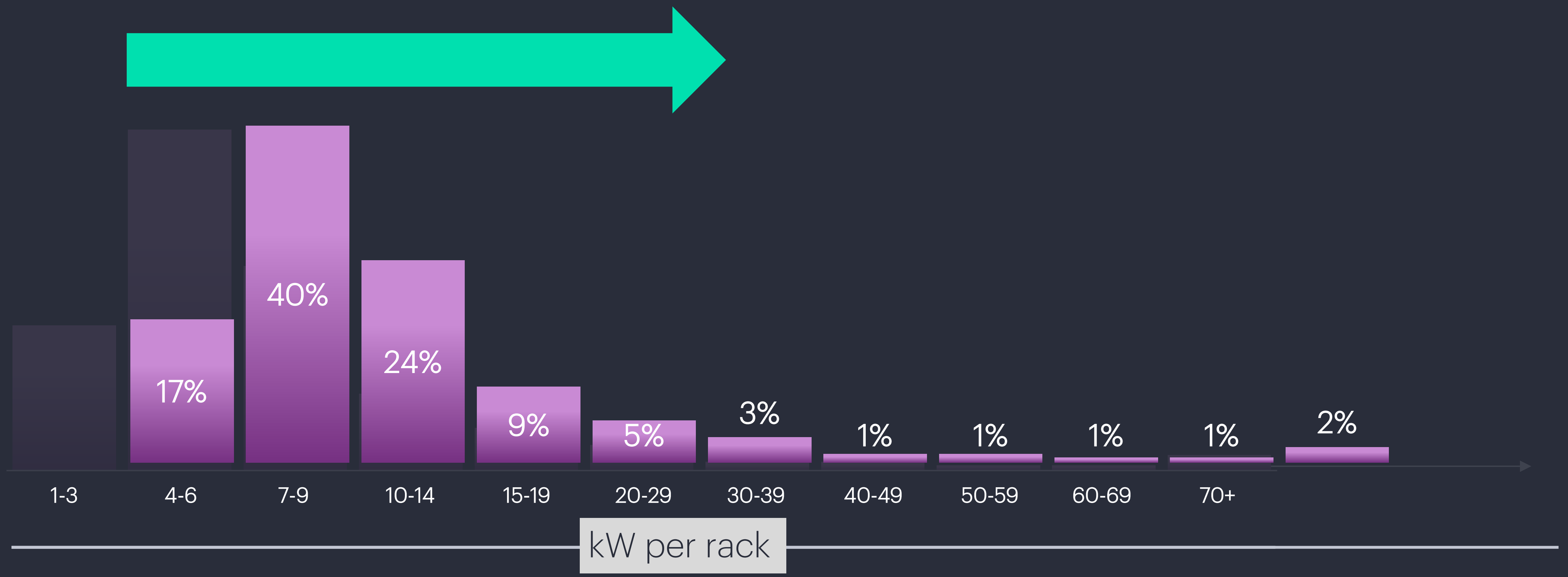


Data developed from Uptime Institute Survey of IT and Data Center Managers- 2022 and 2024

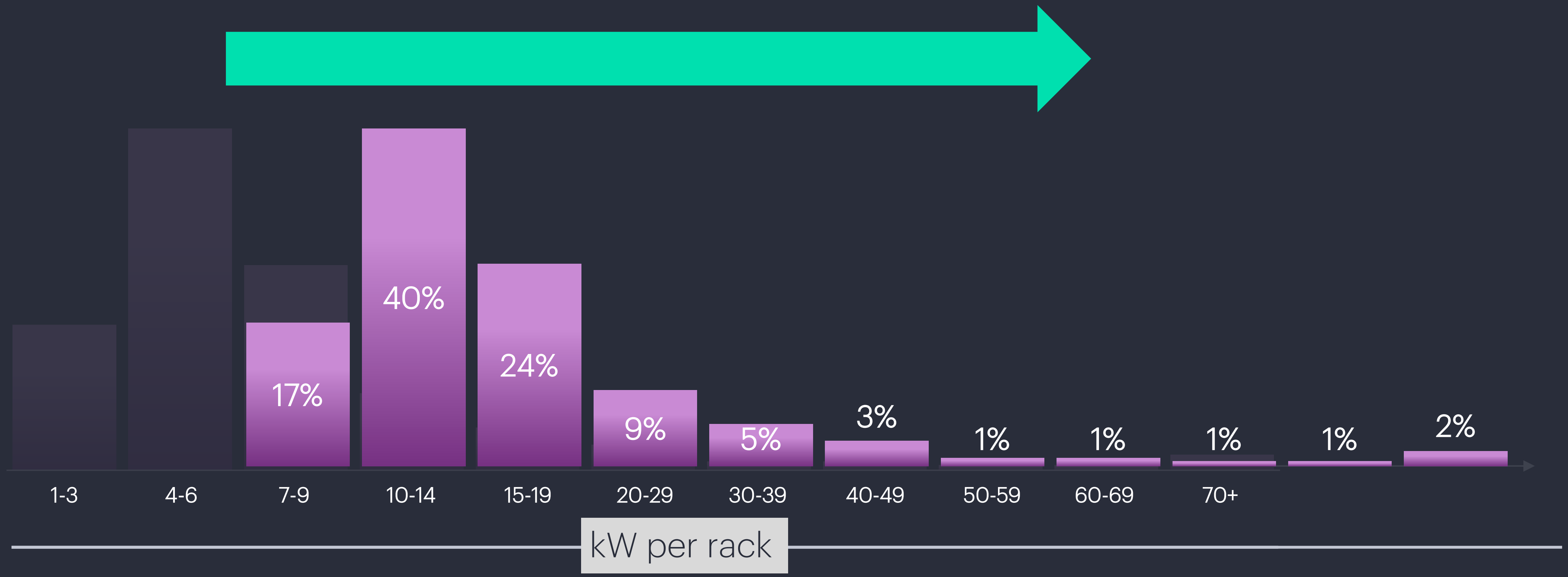
**Approximate Values Assuming 80% Max Load and Full Fans*

*** Depending on configuration.*

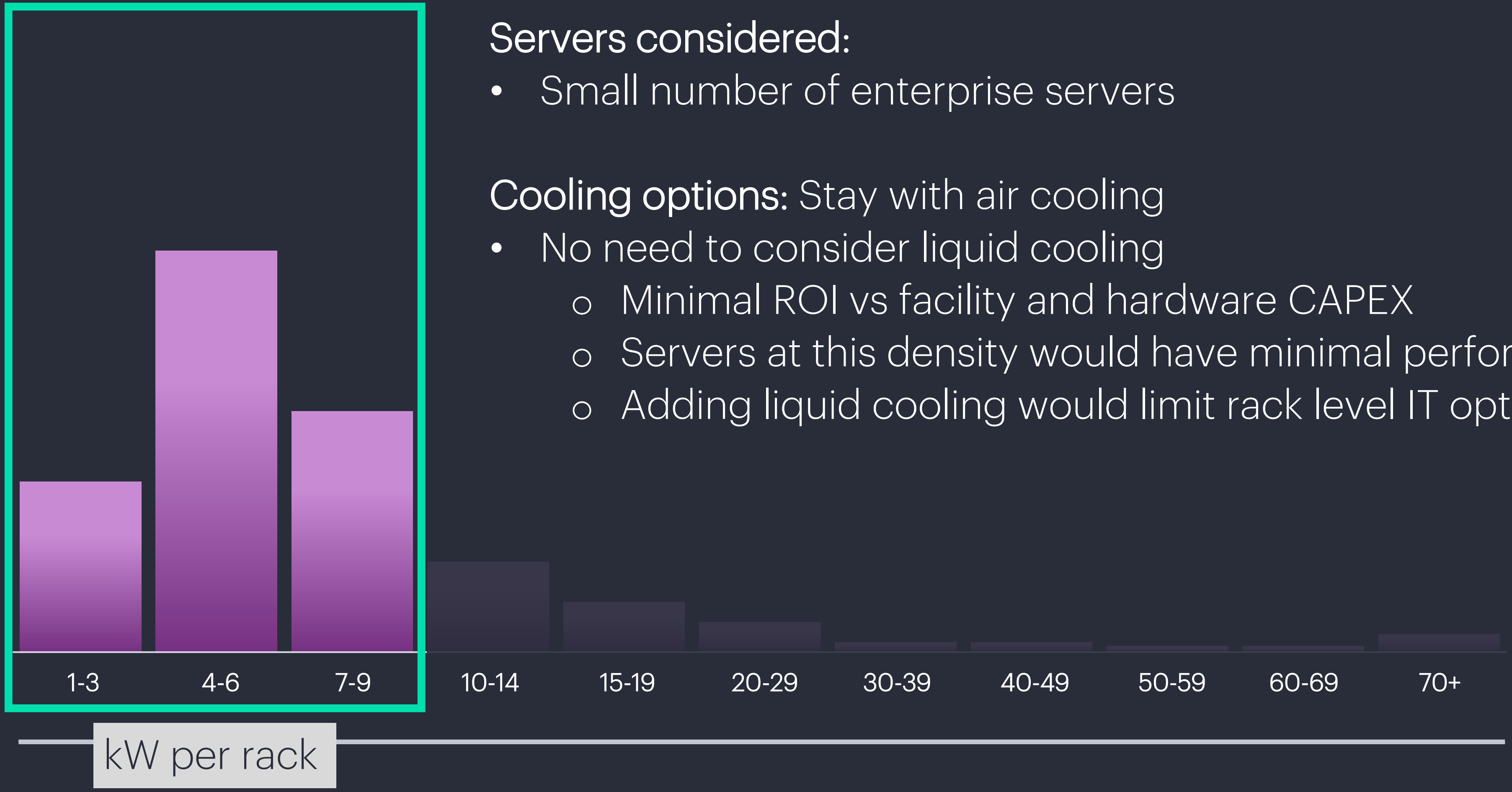
Power trends…next 1-3 years



Power trends…next 3-5 years



Cooling considerations for 1 to 9kW Racks



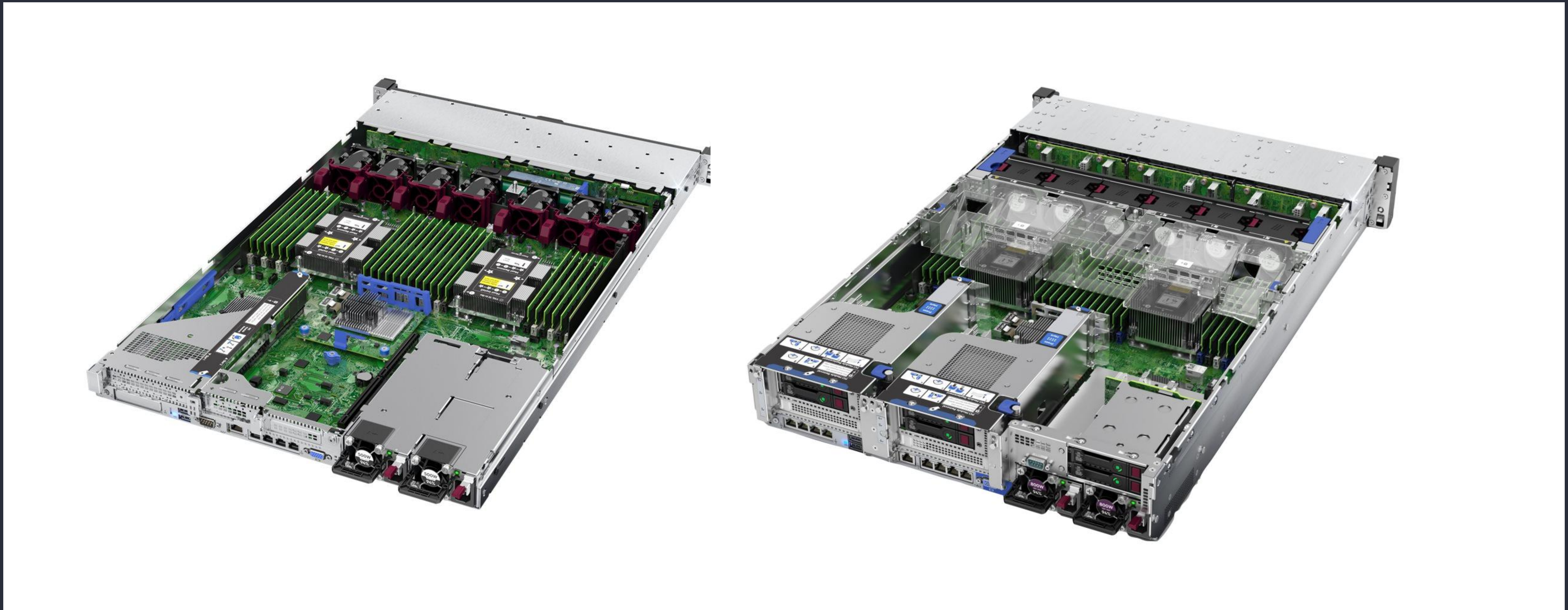
Servers considered:

- Small number of enterprise servers

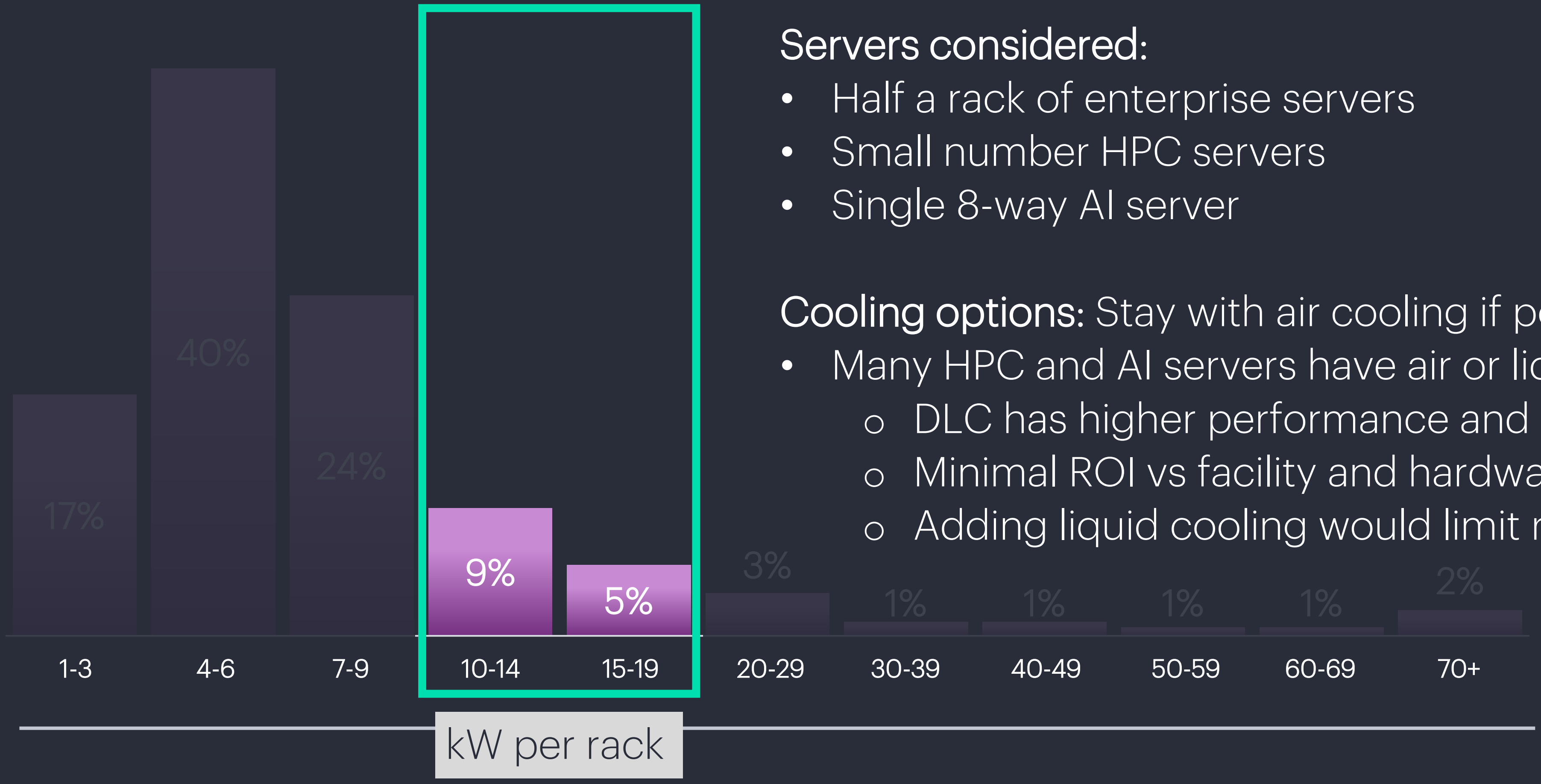
Cooling options: Stay with air cooling

- No need to consider liquid cooling
  - Minimal ROI vs facility and hardware CAPEX
  - Servers at this density would have minimal performance benefits
  - Adding liquid cooling would limit rack level IT options

Cooling considerations for 1 to 9kW Racks



Cooling considerations for 10 to 19kW Racks



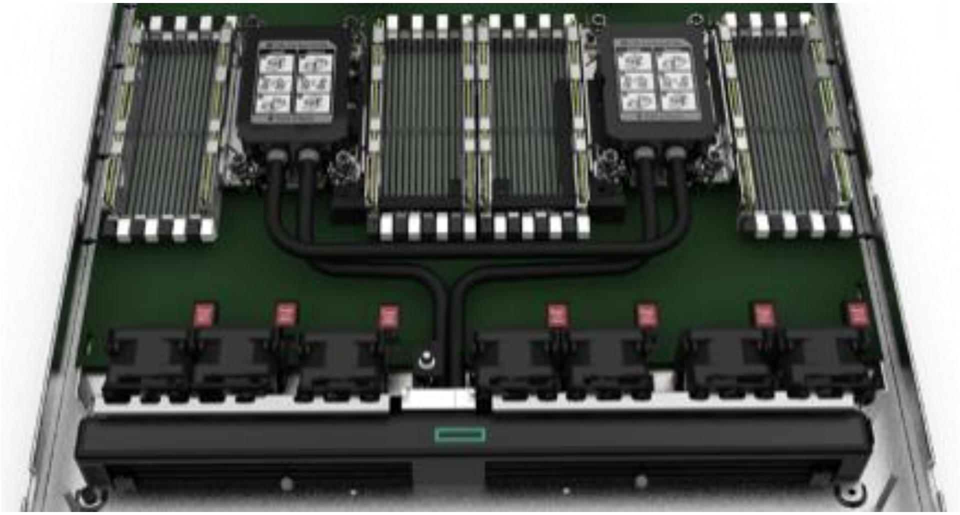
Servers considered:

- Half a rack of enterprise servers
- Small number HPC servers
- Single 8-way AI server

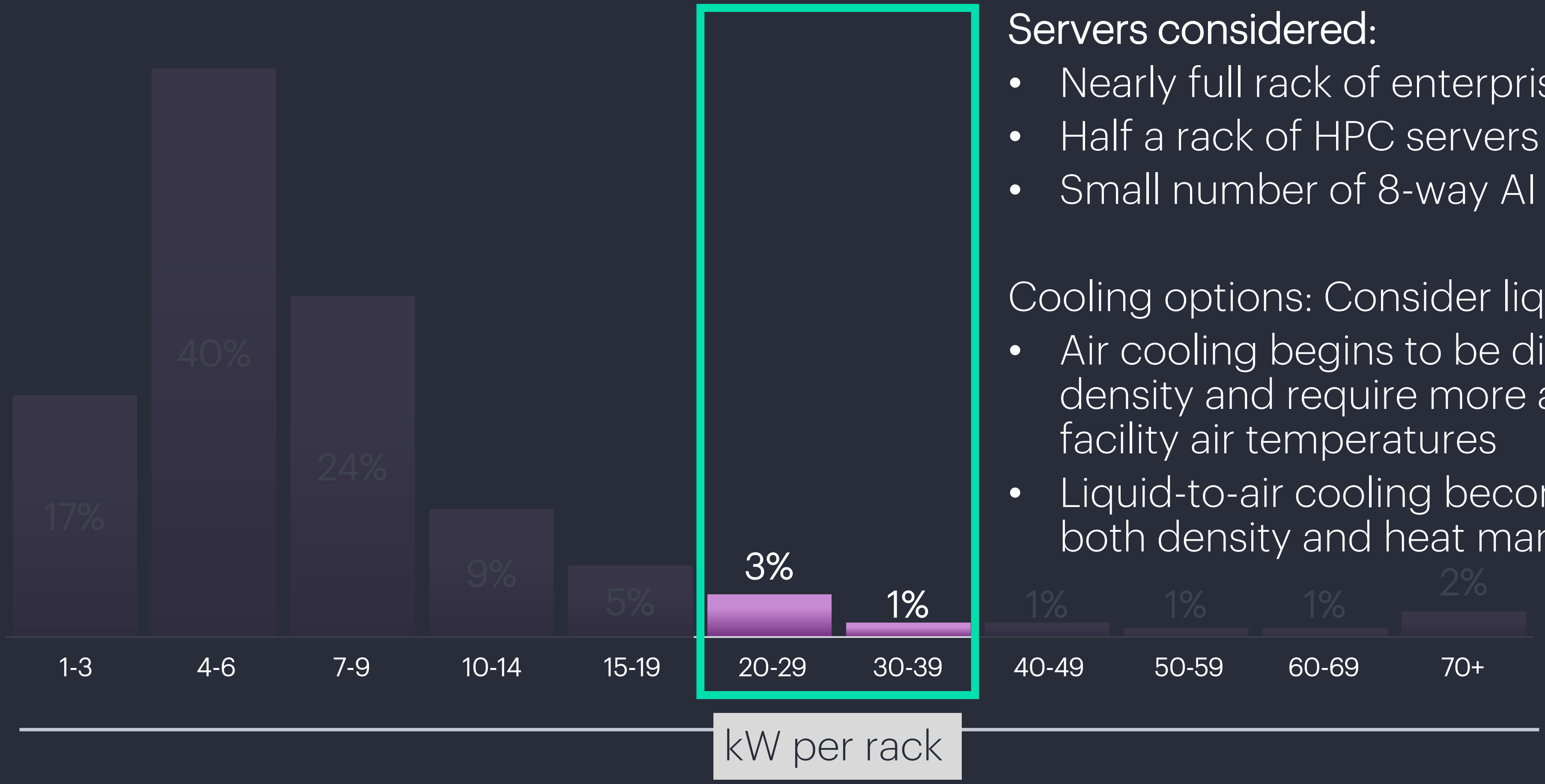
Cooling options: Stay with air cooling if possible

- Many HPC and AI servers have air or liquid cooling options.
  - DLC has higher performance and density
  - Minimal ROI vs facility and hardware CAPEX
  - Adding liquid cooling would limit rack level IT options

Cooling considerations for 10 to 19kW Racks



Cooling considerations for 20 to 39kW Racks



Servers considered:

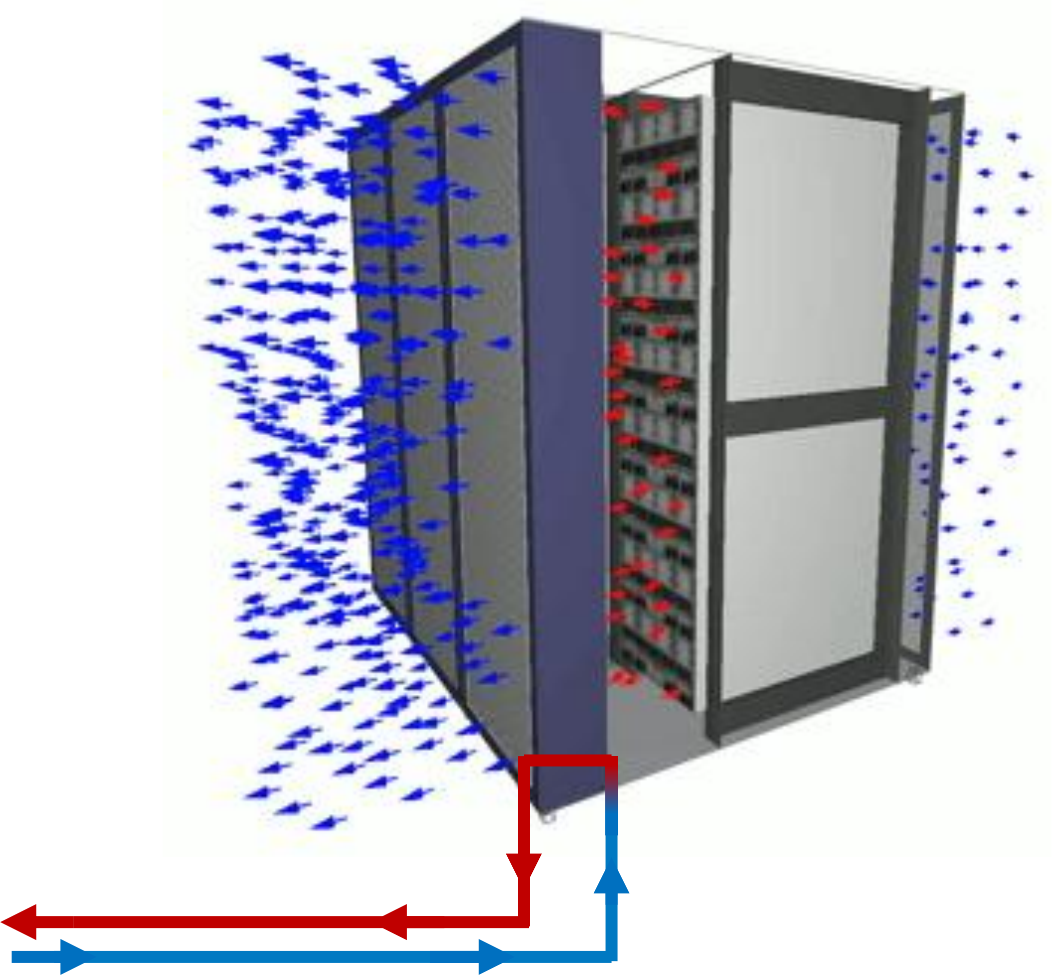
- Nearly full rack of enterprise servers
- Half a rack of HPC servers
- Small number of 8-way AI servers

Cooling options: Consider liquid to air cooling

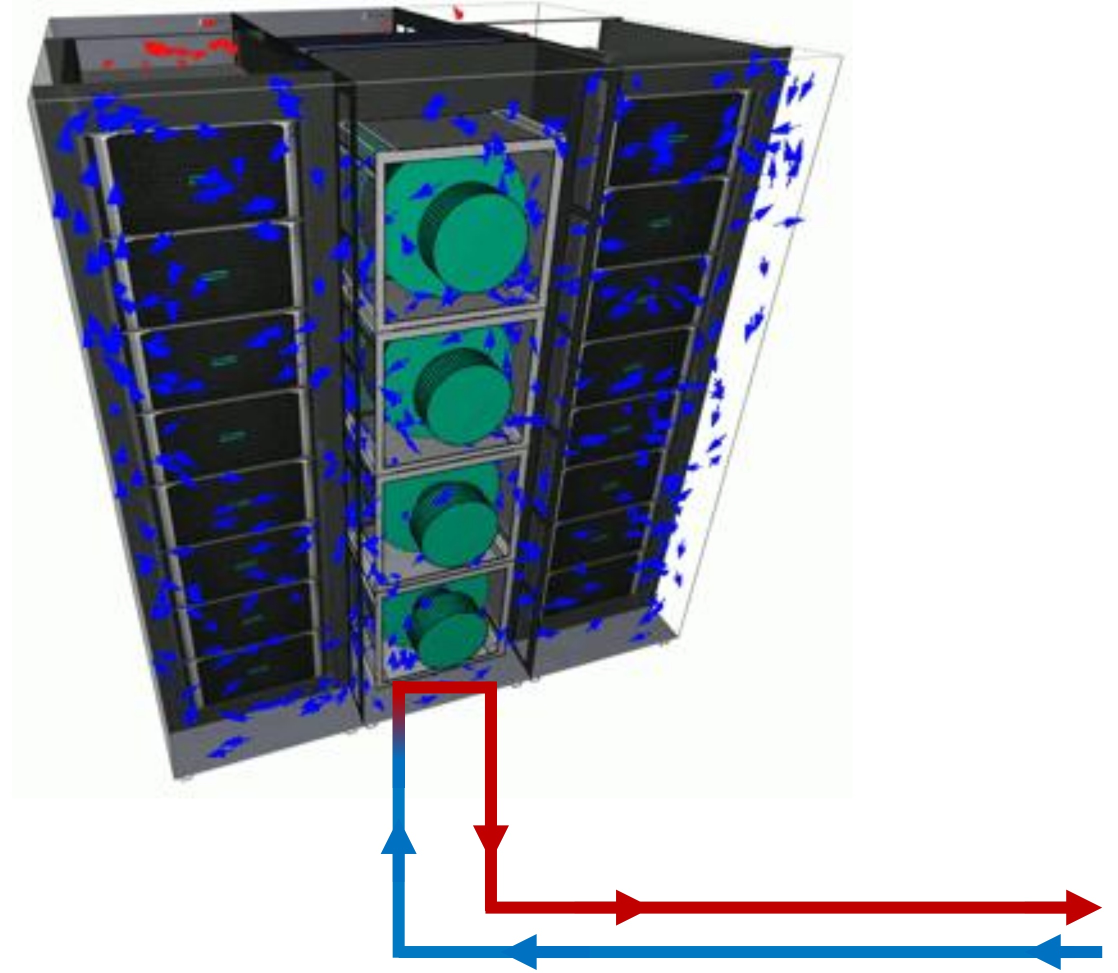
- Air cooling begins to be difficult to cool at this density and require more air flow and colder facility air temperatures
- Liquid-to-air cooling becomes more ideal for both density and heat management

Cooling considerations for 20 to 39kW Racks

Rear Door Heat Exchanger (RDHX)



Adaptive Rack Cooling Systems (ARCS)



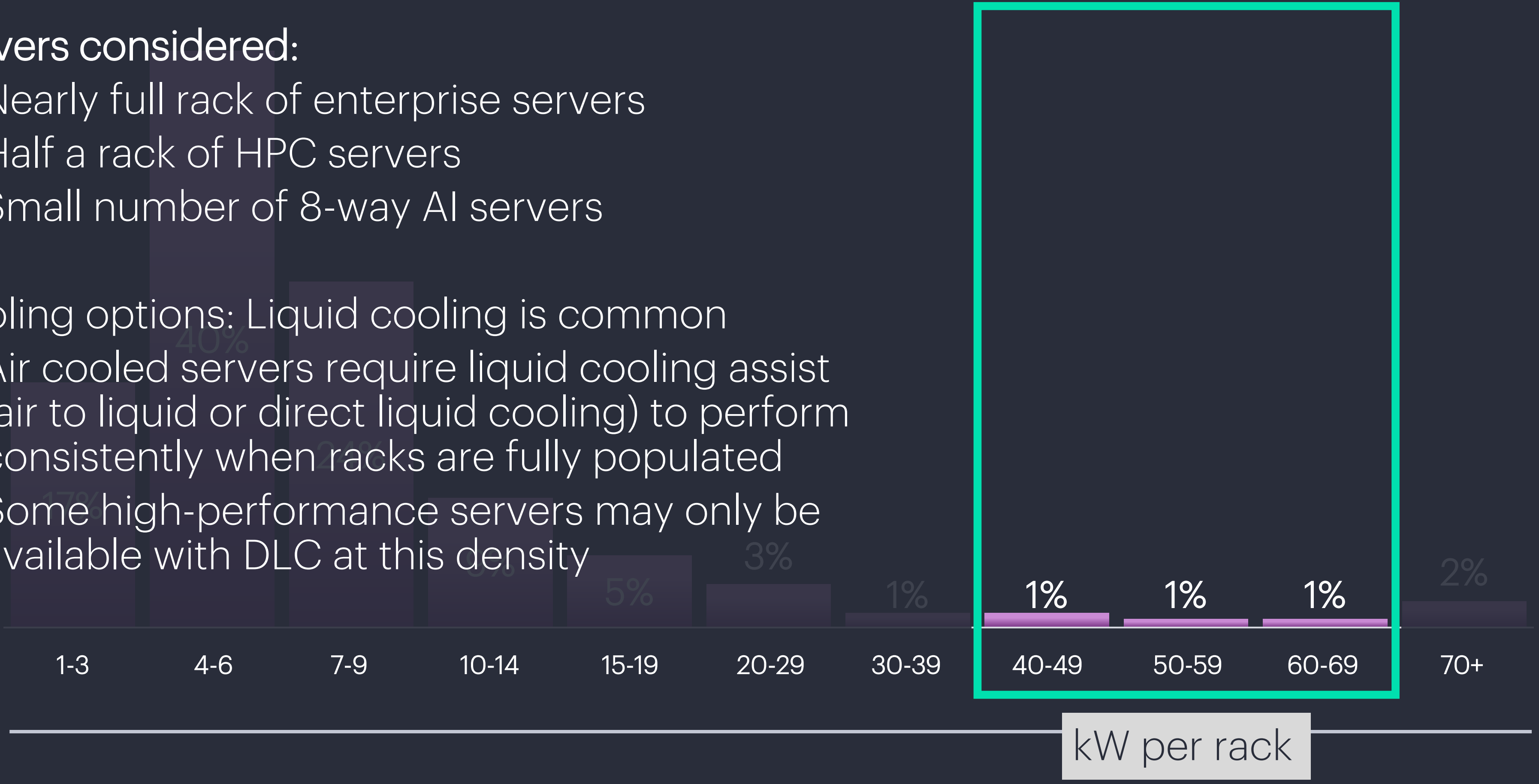
Cooling considerations for 40 to 69kW Racks

Servers considered:

- Nearly full rack of enterprise servers
- Half a rack of HPC servers
- Small number of 8-way AI servers

Cooling options: Liquid cooling is common

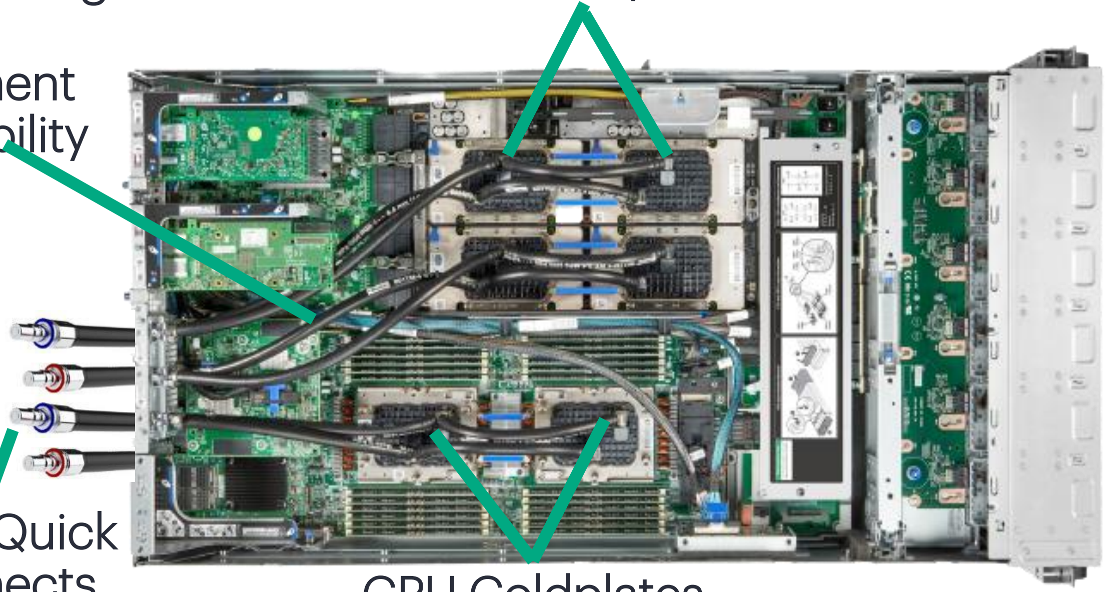
- Air cooled servers require liquid cooling assist (air to liquid or direct liquid cooling) to perform consistently when racks are fully populated
- Some high-performance servers may only be available with DLC at this density



Cooling considerations for 40 to 69kW Racks

70% server heat goes to water

30% server heat goes to air

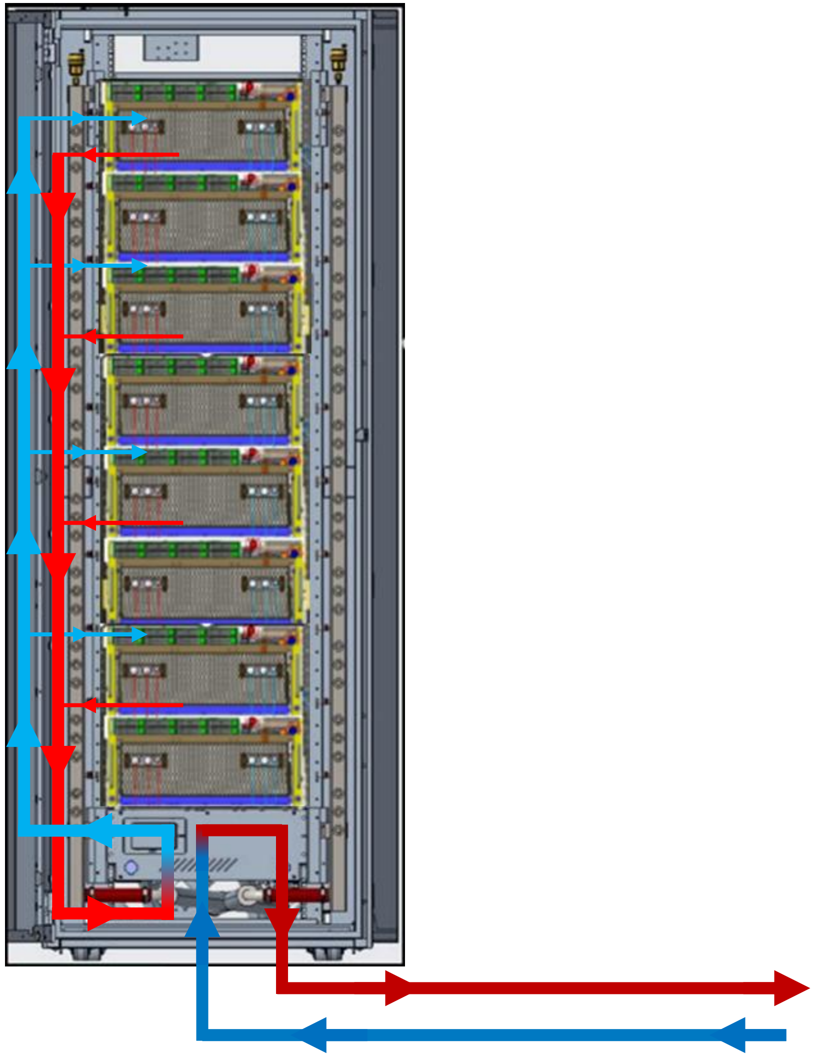


Flexible Tubing for Component Serviceability

GPU Coldplates

Dripless Quick Disconnects

CPU Coldplates



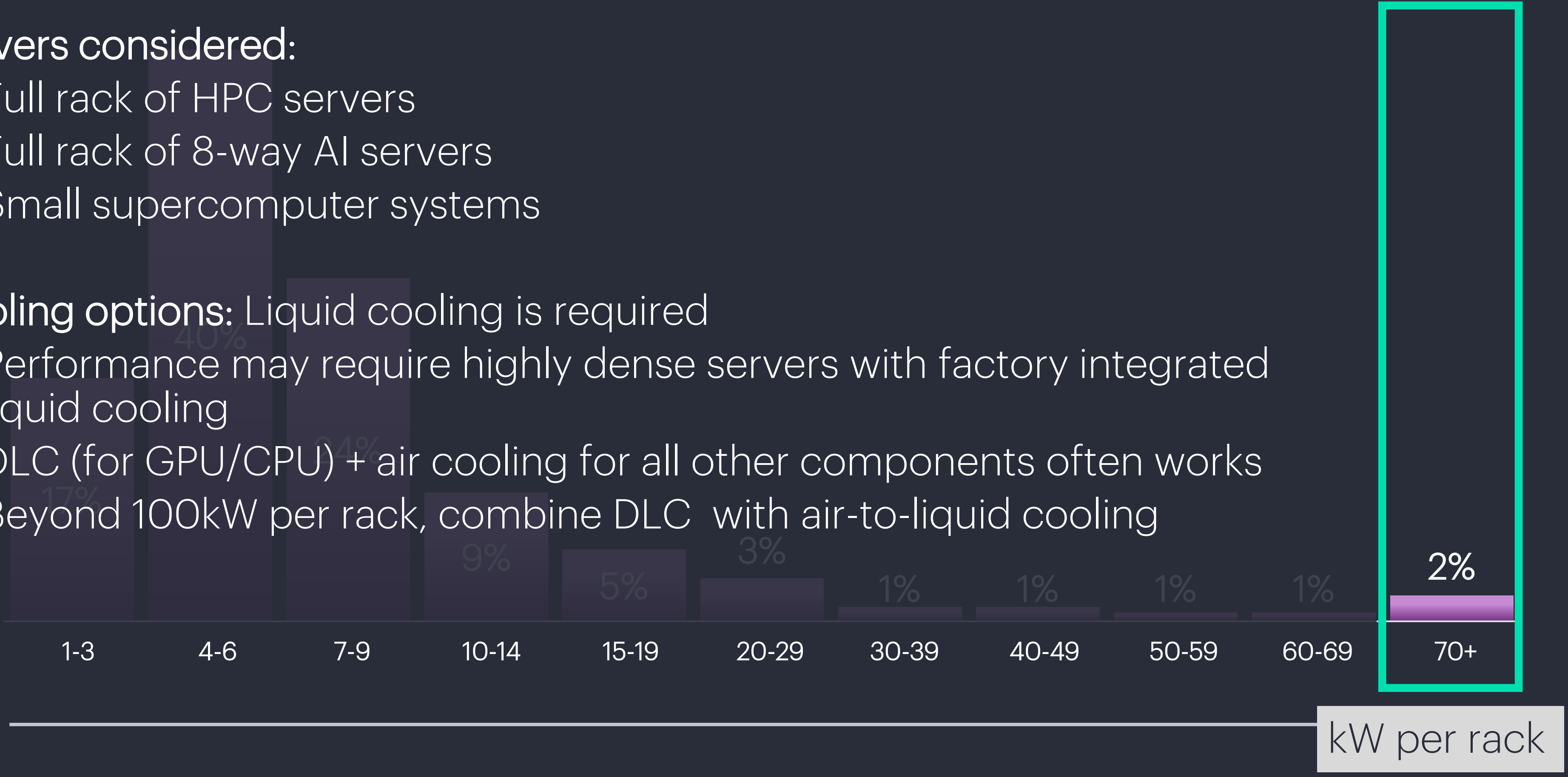
Cooling considerations for +70kW Racks

Servers considered:

- Full rack of HPC servers
- Full rack of 8-way AI servers
- Small supercomputer systems

Cooling options: Liquid cooling is required

- Performance may require highly dense servers with factory integrated liquid cooling
- DLC (for GPU/CPU) + air cooling for all other components often works
- Beyond 100kW per rack, combine DLC with air-to-liquid cooling



Cooling considerations for +70kW Racks

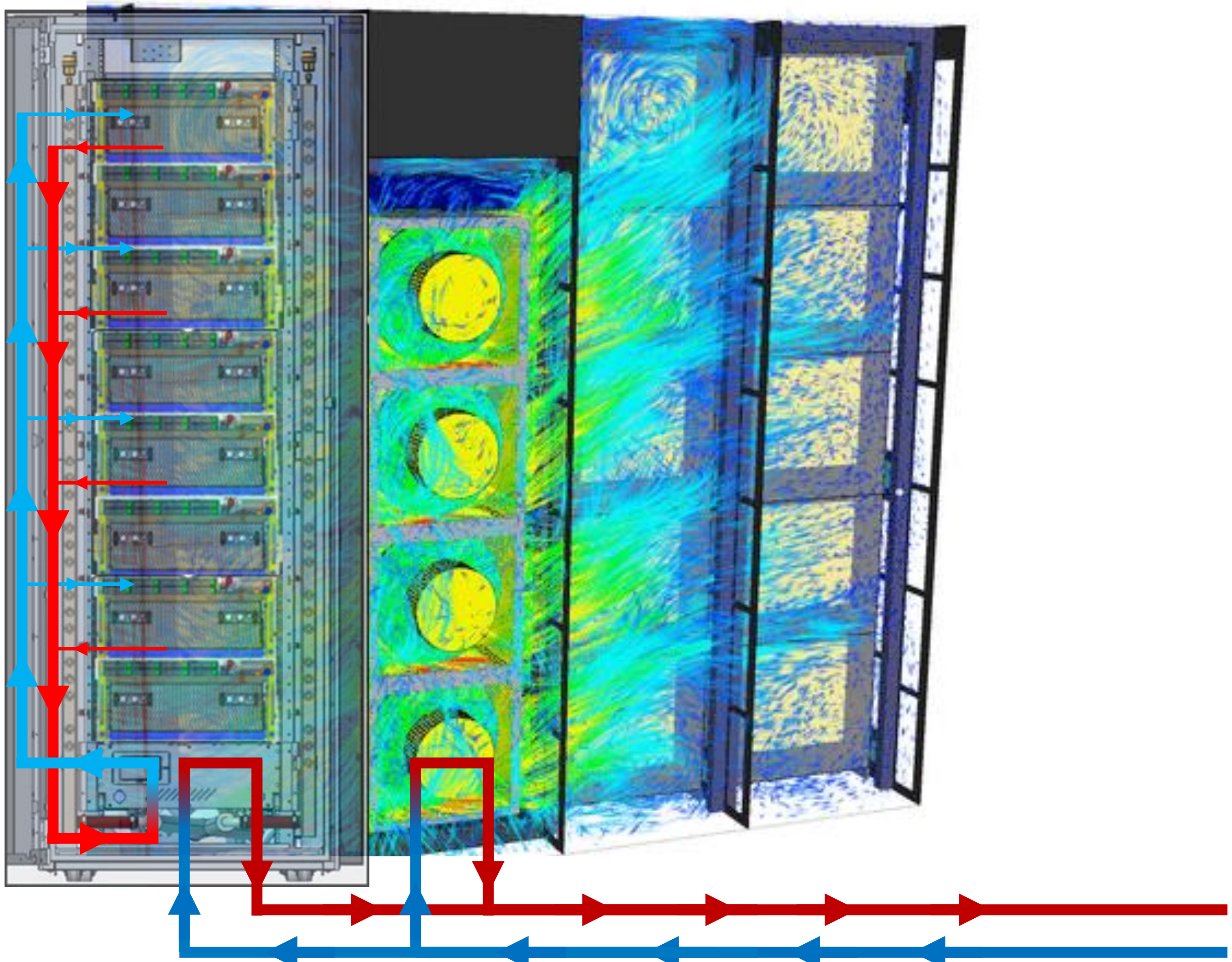
~70%

server heat goes to water (DLC)

+

~30%

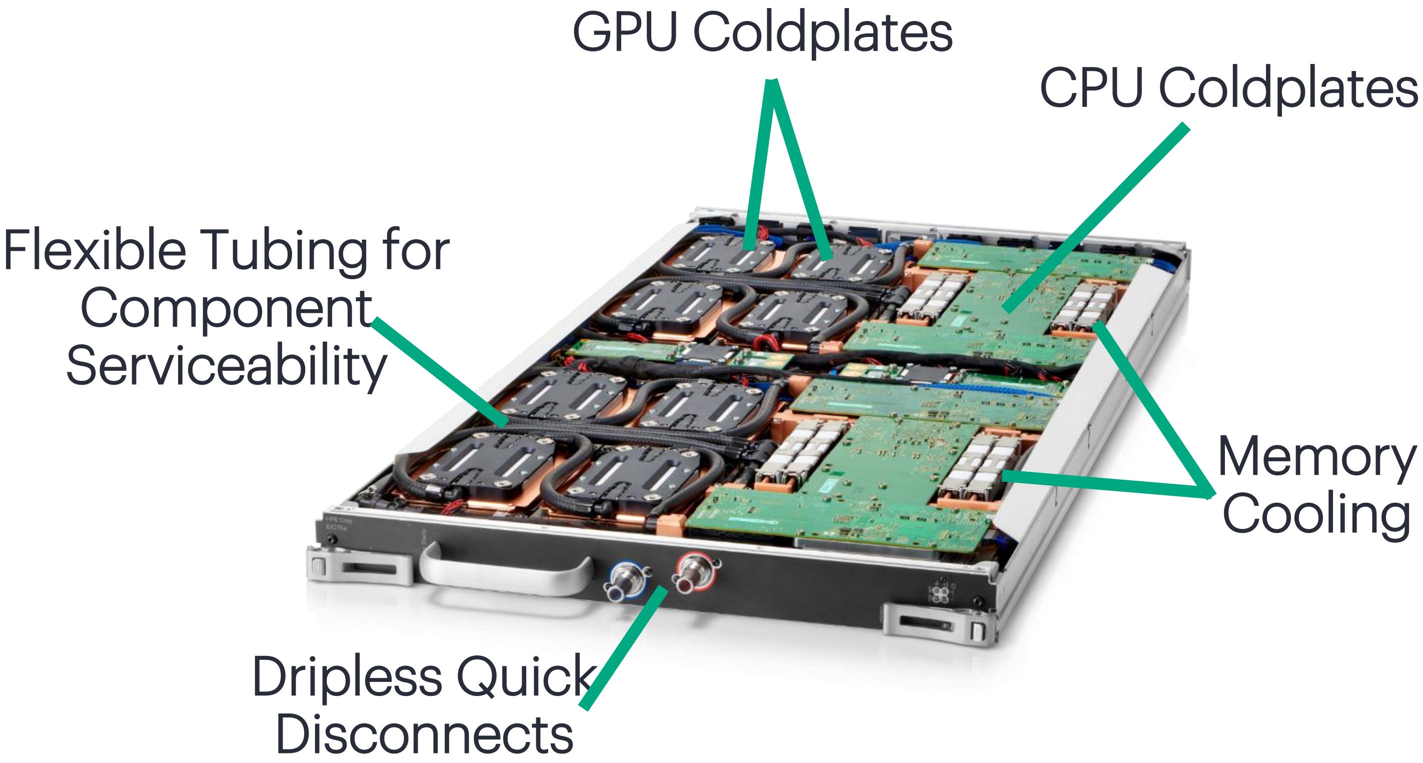
server heat goes to air ARCS or RDHX



Cooling considerations for +70kW Racks

~99% server heat goes to water

~1% server heat goes to air



GPU Coldplates

CPU Coldplates

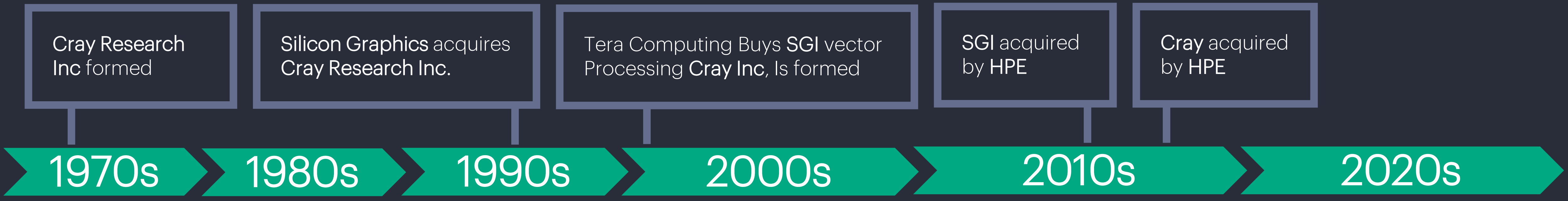
Flexible Tubing for Component Serviceability

Memory Cooling

Dripless Quick Disconnects

HPE Liquid Cooling Expertise

Legacy of liquid cooling innovation



Cray Research Inc formed

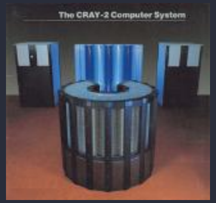
Silicon Graphics acquires Cray Research Inc.

Tera Computing Buys SGI vector Processing Cray Inc, Is formed

SGI acquired by HPE

Cray acquired by HPE

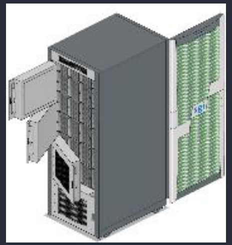
Cray 1
Refrigerant Cooled



Cray 2
Pumped Single Phase Immersion Cooled (Fluorinert)

Cray YMP
Cold Plate Pumped Fluorinert Cooled

Cray C-90
Cold Plate (Fluorinert)

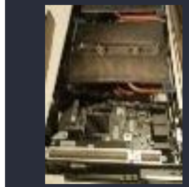


SGI ICE
Room Neutral Cooling

Cray XT
Vertical Refrigerant Cooling

HPE SGI 8600
ICE Hybrid Cell Cooling

Cray XC
Horizontal Chilled Water Cooling



HPE Apollo 8000
Liquid Cooling

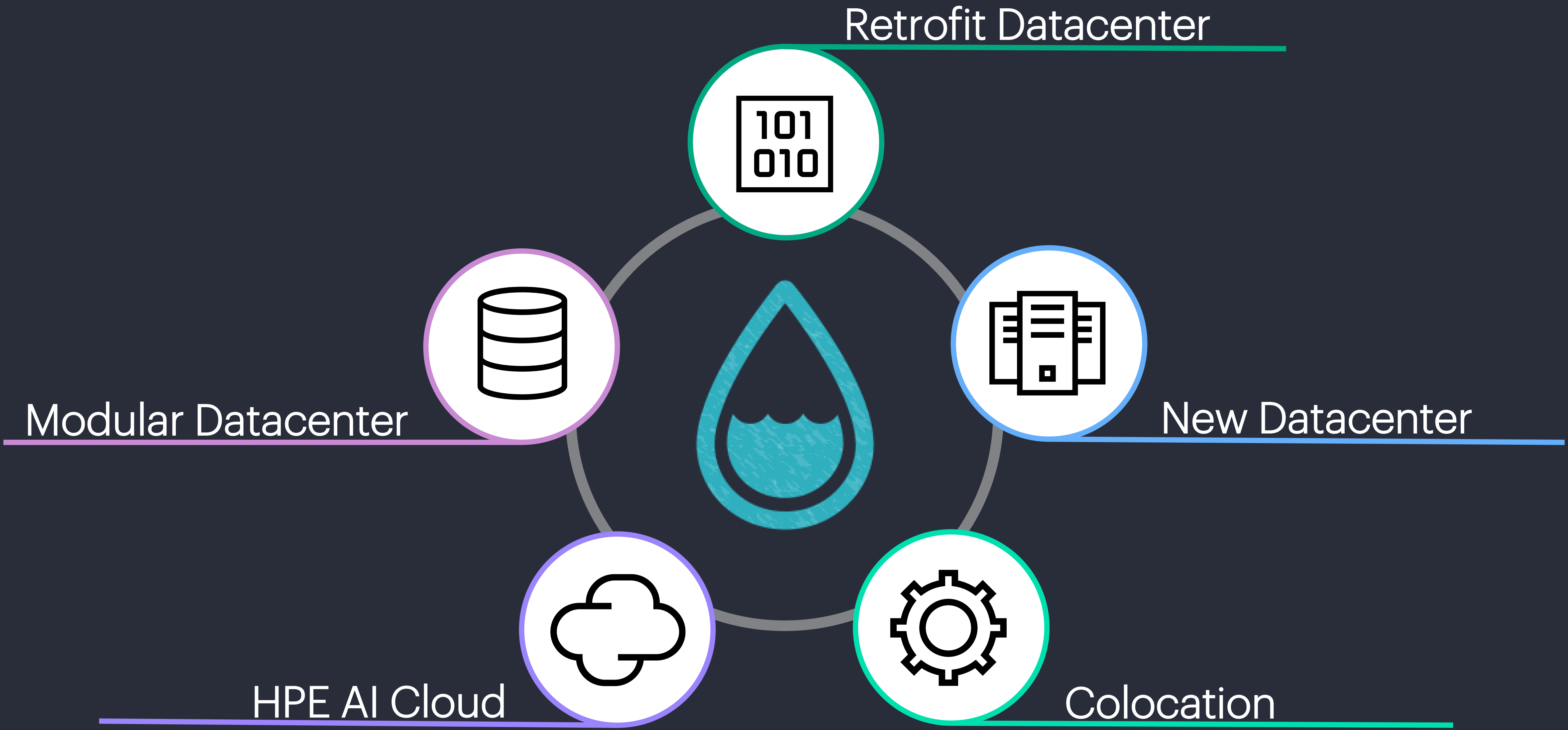
Cray EX
Fanless DLC

HPE Apollo Gen10+
DLC

HPE Cray XD and ProLiant
DLC

HPE ARCS
Room neutral cooling

Paths to enable liquid cooling



Retrofit Datacenter

Modular Datacenter

New Datacenter

HPE AI Cloud

Colocation

Case study: University artificial intelligence researchers

Goals:

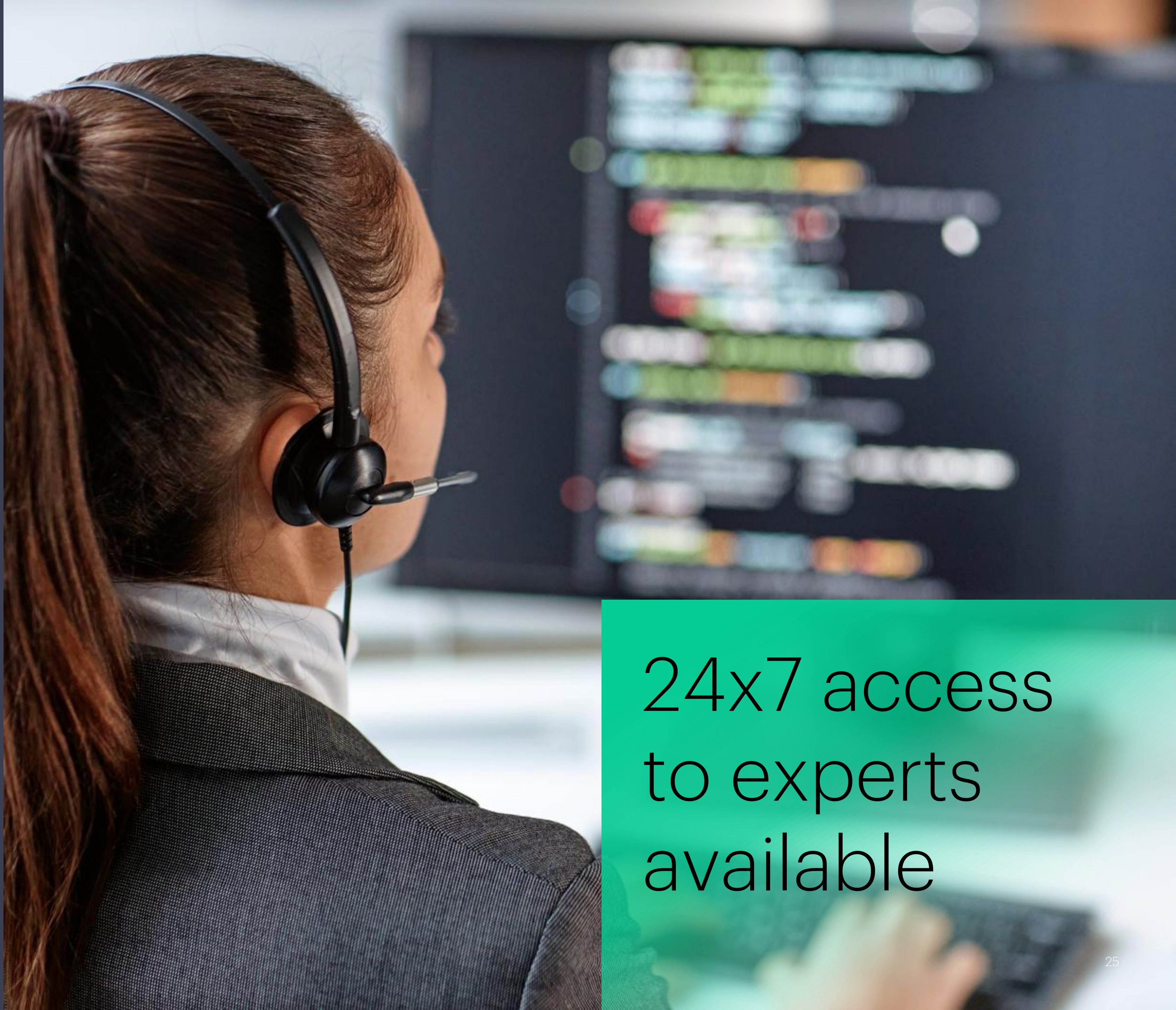
Deploy cutting edge AI technology in a short time frame

"It was 48 hours from an empty concrete pad to a data centre. And then two weeks later, the supercomputer was up and running and was reproducing all of the factory acceptance tests... a working data centre with a system that's in position 128 of the top 500 and achieving number two on the Green500."

- July 16, 2024
Professor Simon McIntosh-Smith
Director of the Bristol Centre for
Supercomputing (BriCS) at the University of Bristol

Tailored HPE Support Services

- Dedicated account manager for your entire IT environment
- Site planning and cooling performance optimization
- Yearly coolant system health checks and maintenance
- End-of-life system decommissioning



24x7 access to experts available

Thank you