

# Welcome to an Omdia Analyst Summit



# Omdia Analyst Summit: Where is AI really headed?

Time	Session	Speakers
11:00 – 11:45	Where is AI really headed? AI的发展方向在哪里?	Vlad Galabov
11:45 – 12:30	Data center rack architecture and power density out to 2030 2030年前的数据中心机柜架构与功率密度	Manoj Sukumaran
12:30 – 14:00	Break 茶歇	
14:00 – 15:00	Powering the AI data center now and in 2030 为当下和2030年的数据中心供电	Shen Wang
15:00 – 15:30	Break 茶歇	
15:30 – 16:30	Cooling the AI data center now and in 2030 为当下和2030年的数据中心散热	Jessica Nian
16:30 – 17:00	Wrap up and audience Q&A 总结与问答环节	Vlad Galabov

# Where is AI really headed?

**Vlad Galabov**

Senior Director,  
Enterprise Infrastructure Research

Copyright © 2025 TechTarget, Inc. or its subsidiaries. All rights reserved.

# Fundamentals

# More compute is fundamental for the formation of the AI economy

↑ Usefulness ↑



**AI usefulness and performance will only improve at the expense of power**

↑ Demand ↑



**Improving productivity gains will accelerate end user adoption**

↑ Supply ↑



**Vendor companies will/should not risk getting frozen out of the AI economy**

↑ Cash access ↑



**Data center investment will not be constrained by cash**

↑ Power access ↑



**Data center investment will not be constrained by power**

# More compute is fundamental for the formation of the AI economy

↑ Usefulness ↑



**AI usefulness and performance will only improve at the expense of power**

- **AI training compute requirement is doubling roughly every five months**
- **More and more AI models coming up**
- **Inference is becoming the dominant AI workload**
- **Reasoning capability is making AI inference more compute intensive**
- **Agentic AI relies on a network of multiple AI models working on a singular prompt**

Improving productivity gains will accelerate end user adoption

Vendor companies will/should not risk getting frozen out of the AI economy

Data center investment will not be constrained by cash

Data center investment will not be constrained by power

# More compute is fundamental for the formation of the AI economy

↑ Usefulness ↑



AI usefulness and performance will only improve at the expense of power

↑ Demand ↑



Improving productivity gains will accelerate end user adoption

- **AI monetization has begun**
  - **OpenAI revenue surpassed a billion dollars**
  - **Accelerated adoption is reflected in accelerated investment**
    - A fast-growing and diverse AI Neocloud market has formed
    - T1 cloud (excl. top 4 us hyperscale SPs) ramped investment
    - Government funding/direct investment is increasing
    - Enterprises prioritising AI investment amongst other priorities will/should not risk getting frozen out of the AI economy
- Data center investment will not be constrained by cash
- Data center investment will not be constrained by power

# More compute is fundamental for the formation of the AI economy



# More compute is fundamental for the formation of the AI economy

- **Data center industry is a safe place for investor cash during a volatile stock market and increased economic uncertainty**
- **Institutional investors (e.g. Blackrock) dramatically ramped their investment in the data center industry**
- **Public investment in AI computing is matching that of hyperscale cloud SPs**
- **A multitude of new investment partnerships are forming**

AI usefulness and performance will only improve at the expense of power

Improving productivity gains will accelerate end user adoption

Vendor companies will/should not risk getting frozen out of the AI economy

↑ Cash access ↑



**Data center investment will not be constrained by cash**

↑ Power access ↑



Data center investment will not be constrained by power

# More compute is fundamental for the formation of the AI economy

- **Data center operators have begun investing in self-generation capabilities**
  - Gas genset, turbine and fuel cells
- **Campus selection strategies are helping operators cope with power constraints**
  - Campus in East London vs. West London
  - Nordic countries in Europe, France, Canada, and China, amongst others, produce more power than they consume

AI usefulness and performance will only improve at the expense of power

Improving productivity gains will accelerate end user adoption

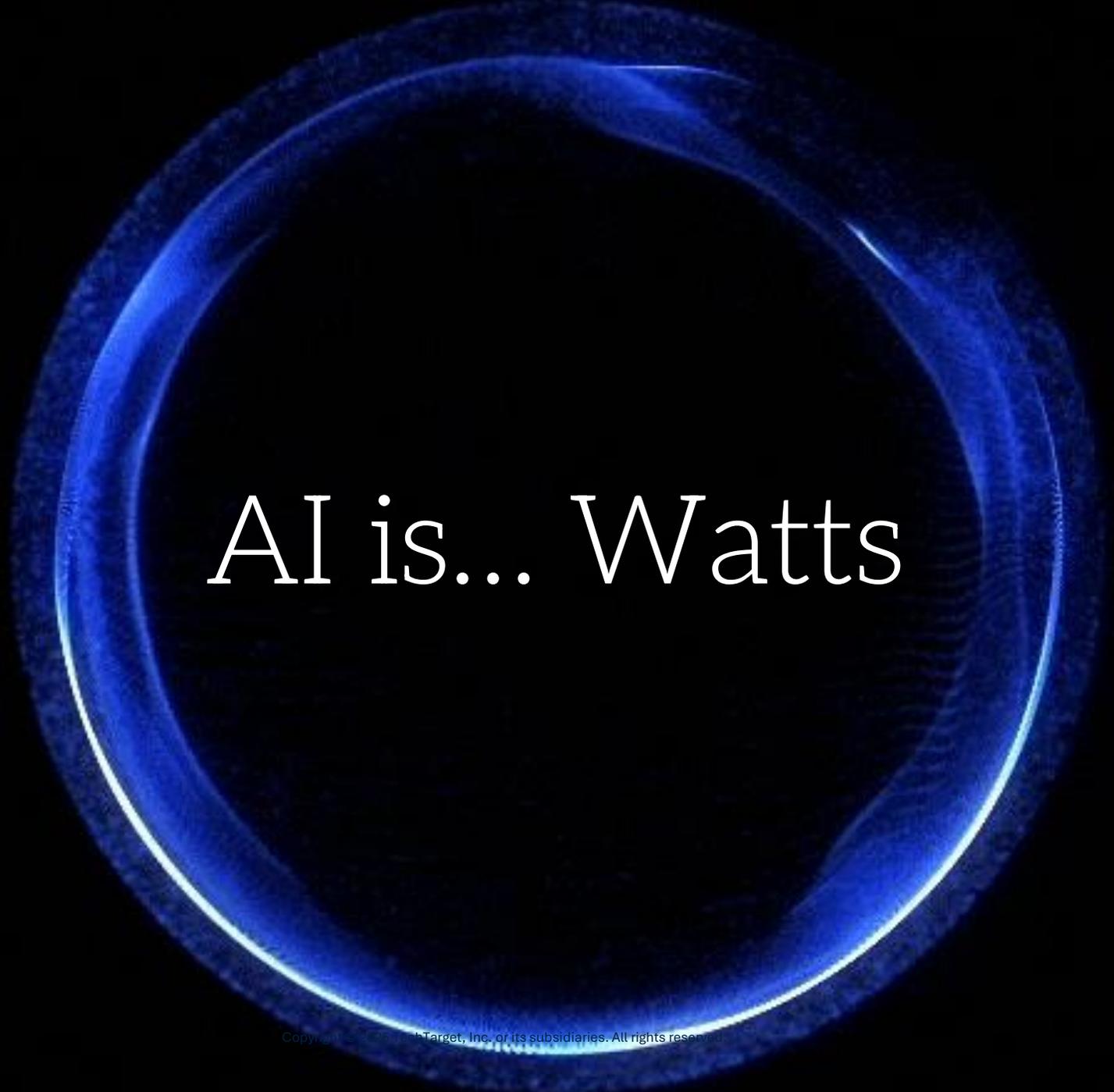
Vendor companies will/should not risk getting frozen out of the AI economy

Data center investment will not be constrained by cash

**Data center investment will not be constrained by power**

↑ Power access ↑

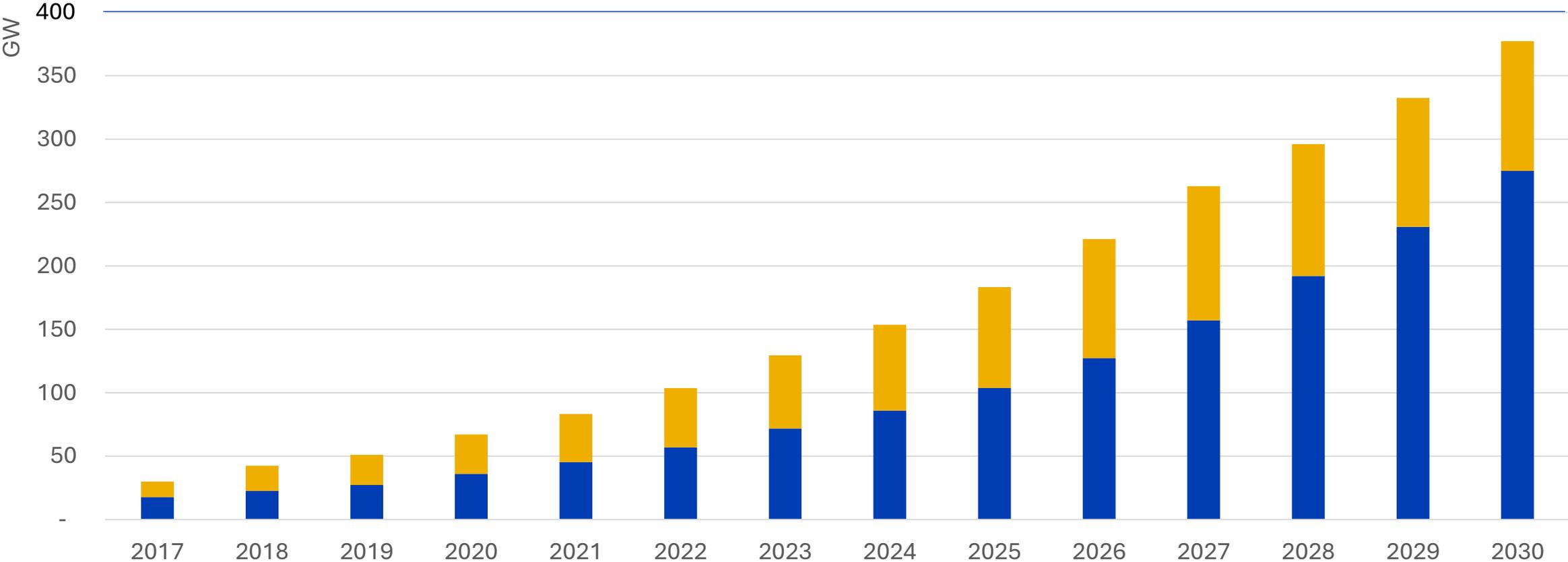




AI is... Watts

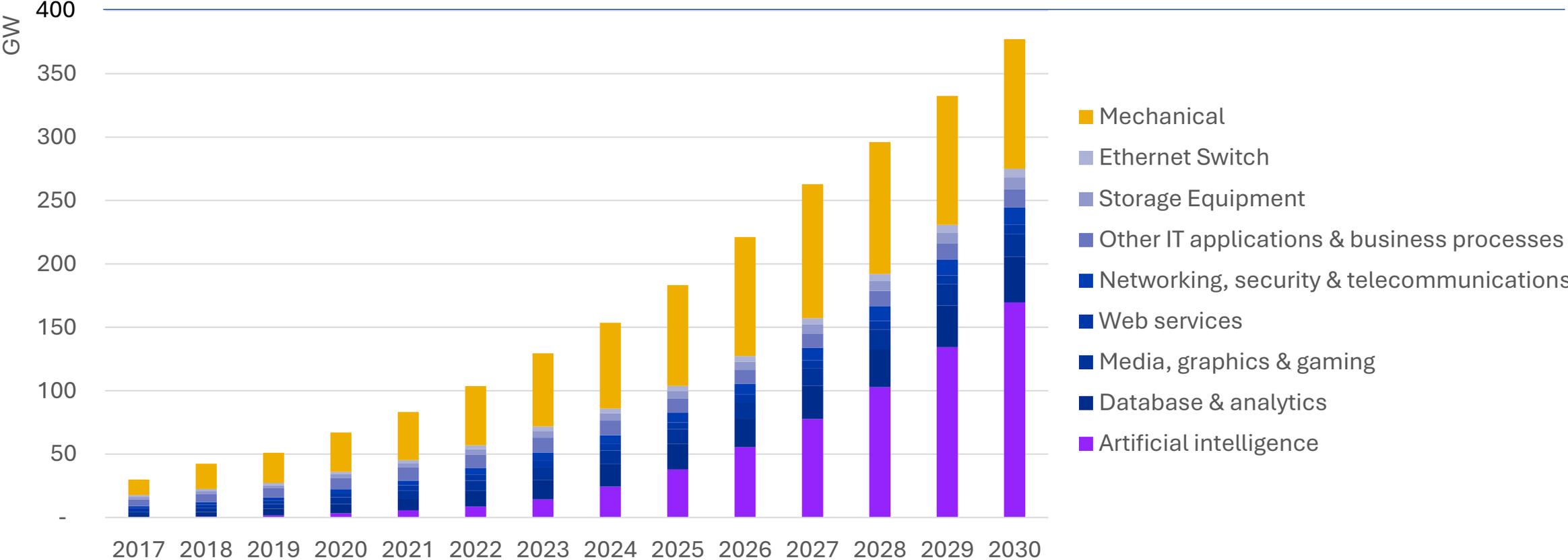
# Installed IT power capacity will double every three to four years

Cumulative data center power capacity



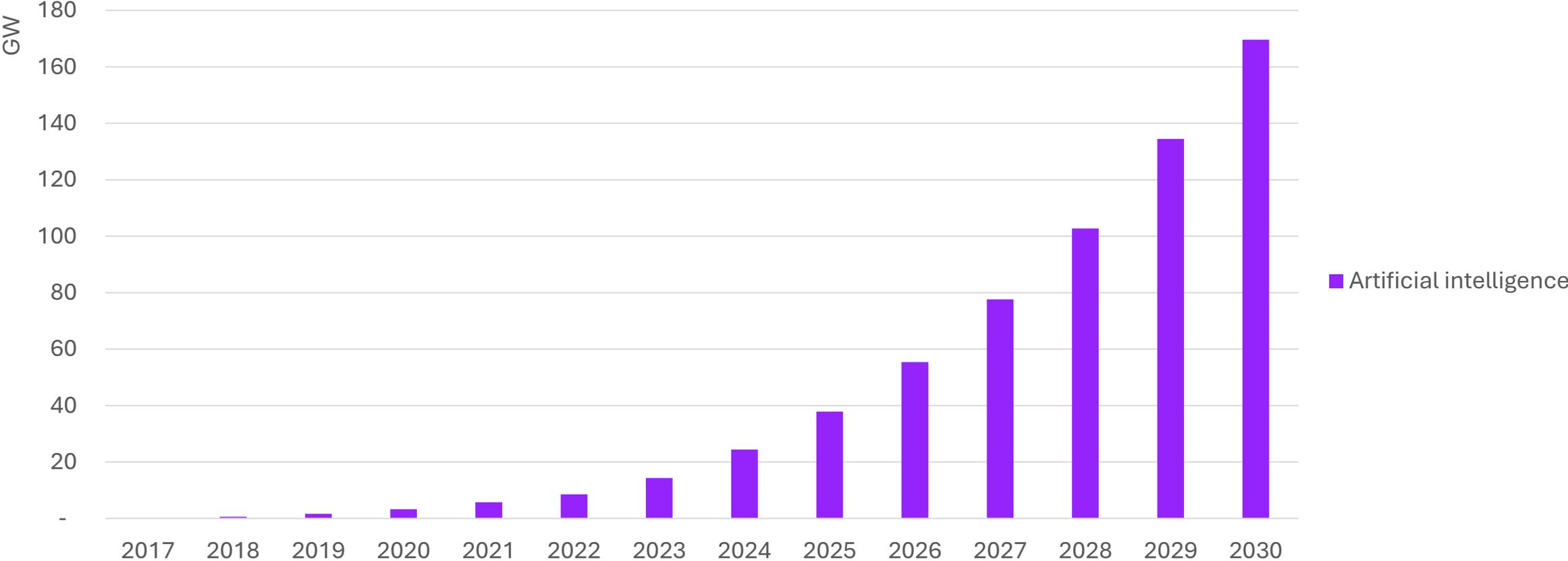
# By the end of the decade installed AI capacity will be half of the cumulative data center capacity

Cumulative data center power capacity



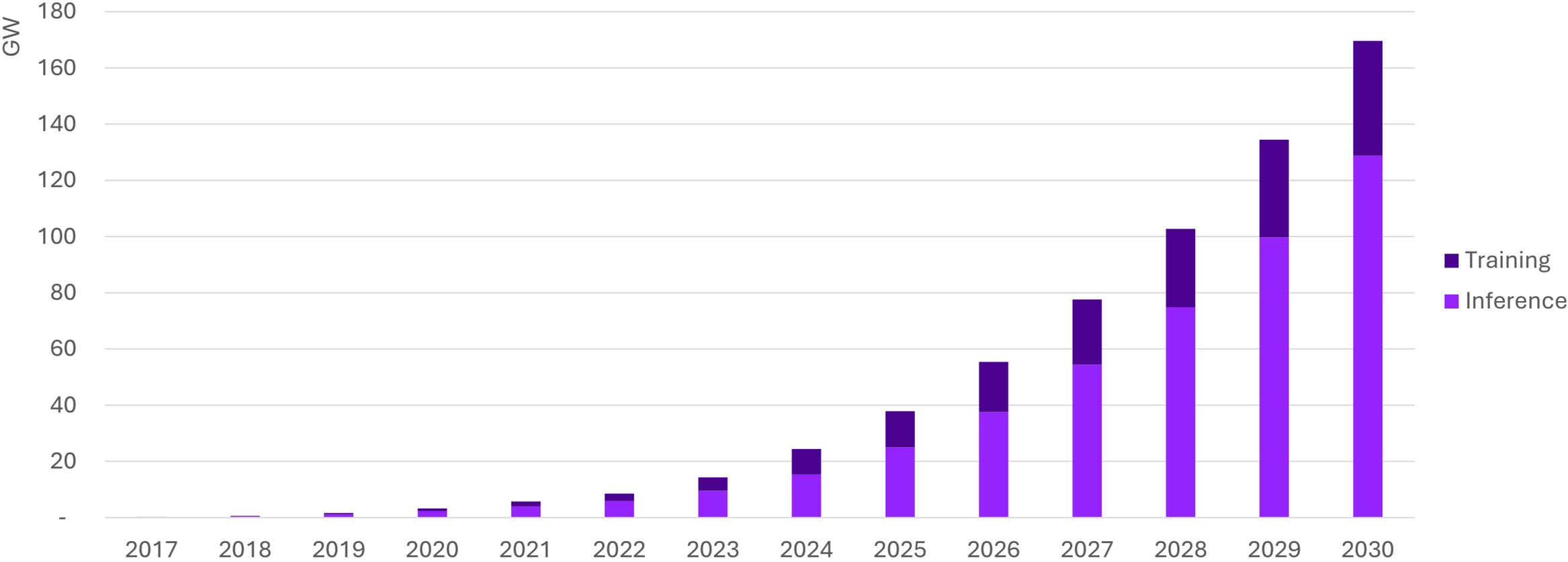
# We're still in the early stages of the AI computing capacity buildout

Cumulative AI computing power capacity (IT only)

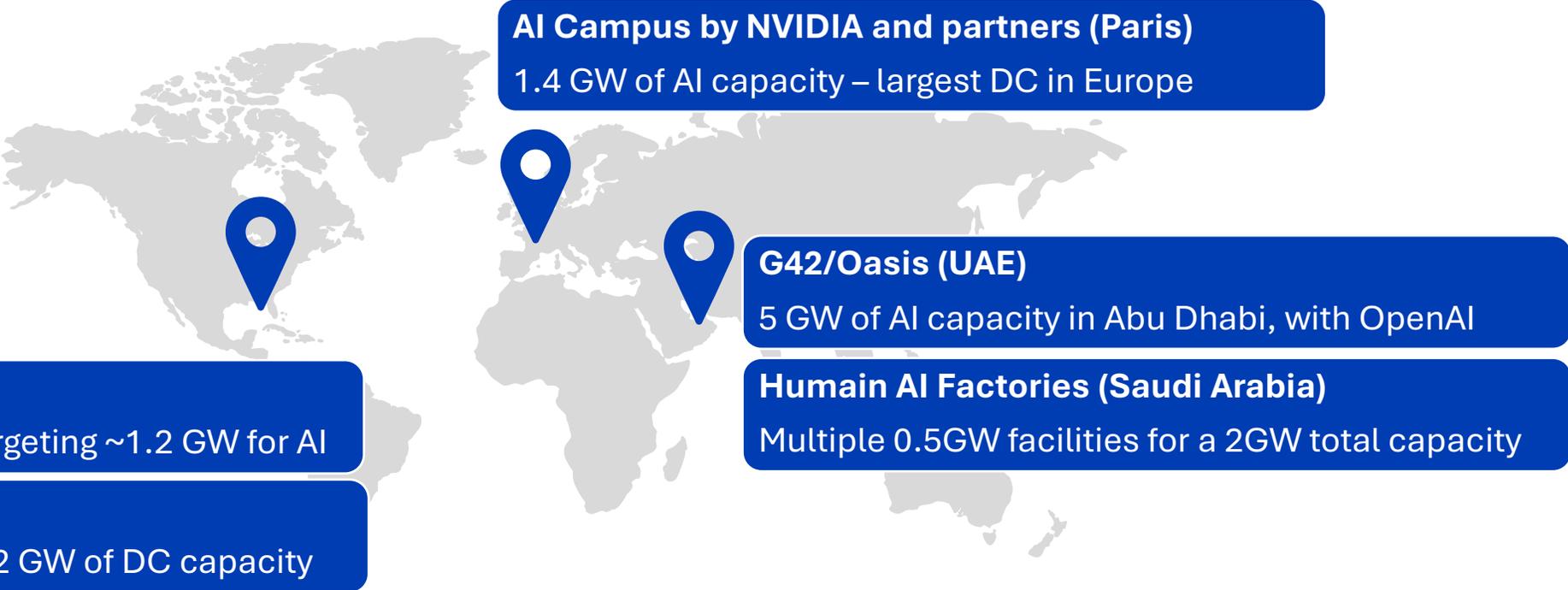


# Most of the installed AI computing capacity will be used for inference

Cumulative AI computing power capacity (IT only)



# The term hyperscale is being redefined and not by the companies we've been referring to as "hyperscale cloud SPs"



**Stargate (Texas, US)**

875-acre campus in Texas targeting ~1.2 GW for AI

**Tract (Texas, US)**

1,500-acre campus for over 2 GW of DC capacity

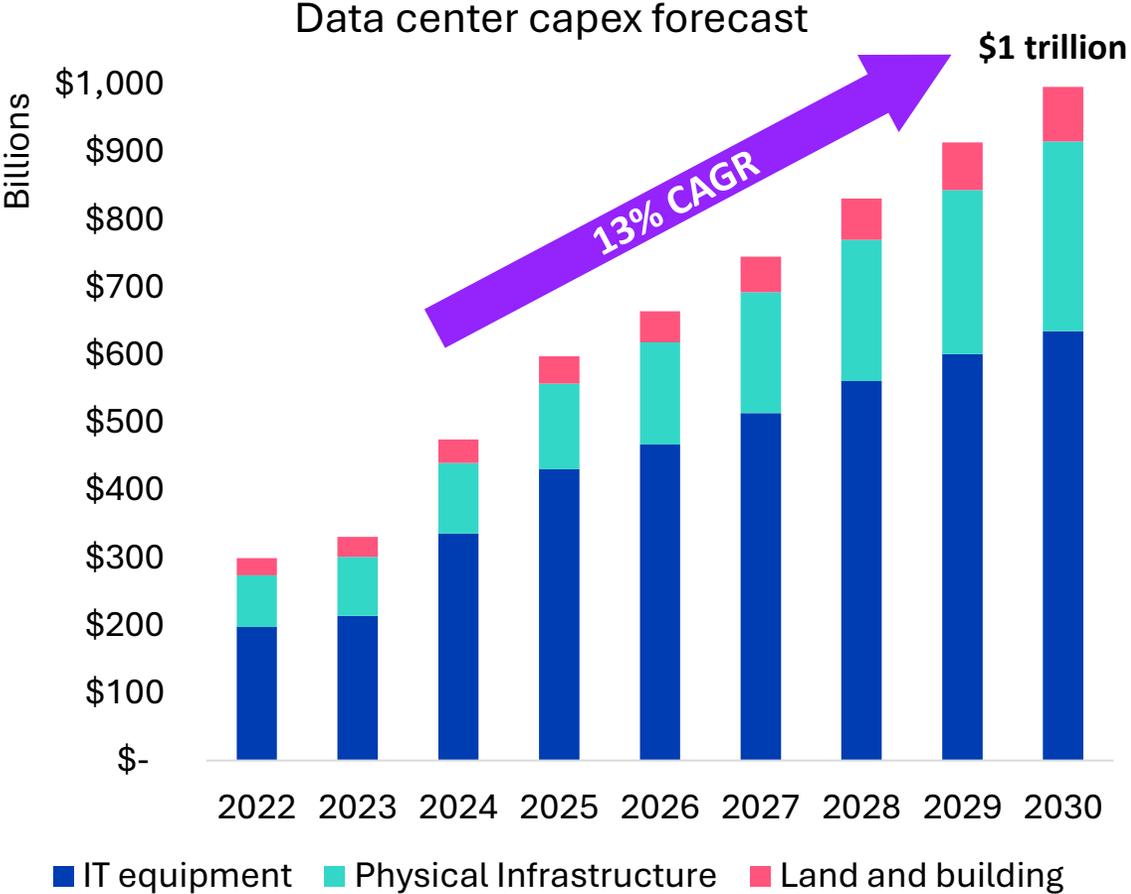
**AI Campus by NVIDIA and partners (Paris)**  
1.4 GW of AI capacity – largest DC in Europe

**G42/Oasis (UAE)**  
5 GW of AI capacity in Abu Dhabi, with OpenAI

**Humain AI Factories (Saudi Arabia)**  
Multiple 0.5GW facilities for a 2GW total capacity

AI is... \$

# Data center capex is on track to hit \$1 trillion by 2030



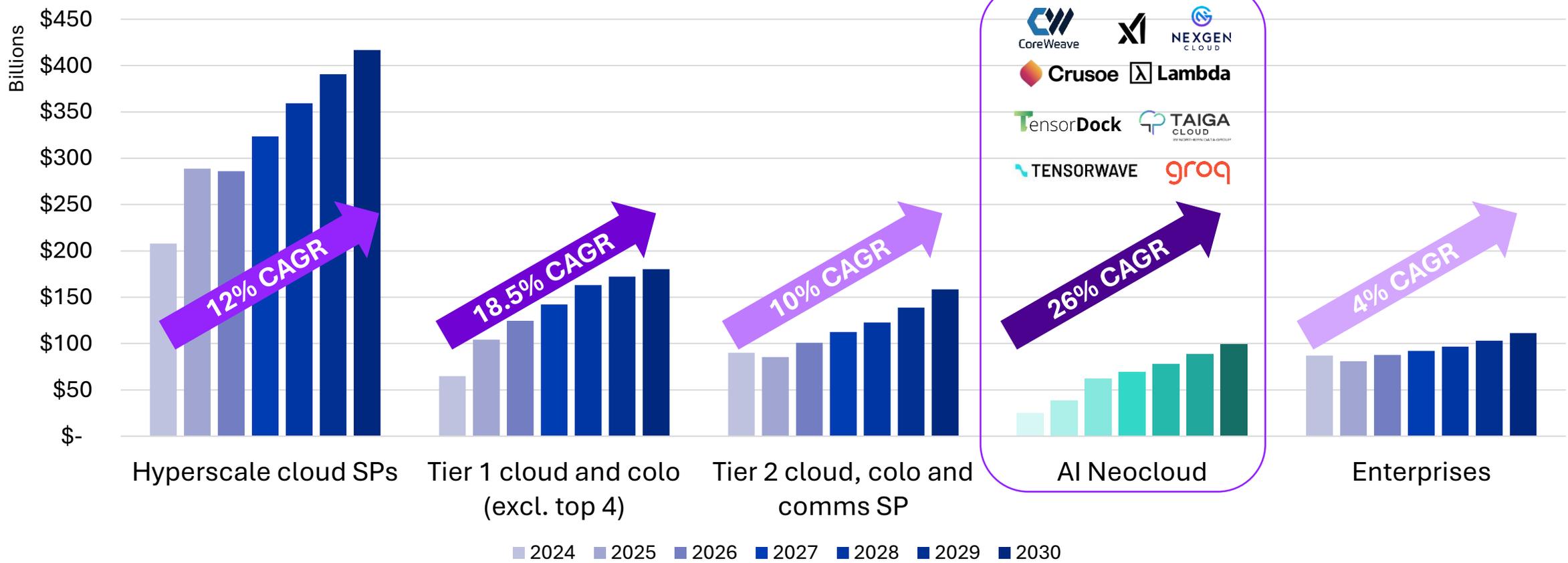
**Land and building: 15% CAGR**

**Physical infrastructure capex: 18% CAGR**  
 Compute densities and rack densities climb, the investment in physical infrastructure accelerates. The physical infrastructure aspect of 2024 computing clusters is rudimentary compared to that of 2027's. The industry will benefit from both new DC builds and retrofits.

**IT equipment capex: 11% CAGR**  
 Server spend grew dramatically in 2024 and will climb at a more modest rate going forward, despite continuous ongoing innovation and a potential refresh cycle of old servers. We expect a consolidation of server count where a small number of scaled up systems are preferred to a scaled-out strategy. The cost per byte/compute cycle is also decreasing.

# NVIDIA has fuelled the emergence of serious global contenders

Data center capex by market segment





AI is... still young

# Data center 2030

Every enterprise and consumer will use multiple AI applications to run their lives



AI applications will dominate every “app store”

AI will drive >50% of installed GW capacity and >70% of revenue opportunity



Improving productivity gains will accelerate end user adoption

Not everyone will survive in the data center wild wild west



Many startup DC campus developers and neoclods will fail to build a long-term business model

1MW rack may remain an ambition



Huge amount of coordinated innovation required to reach such an ambitious goal

Over 35GW of power will be self-generated



Data center dedicated off grid, behind the meter, power will be “self” generated

# Data center rack architecture and power density out to 2030

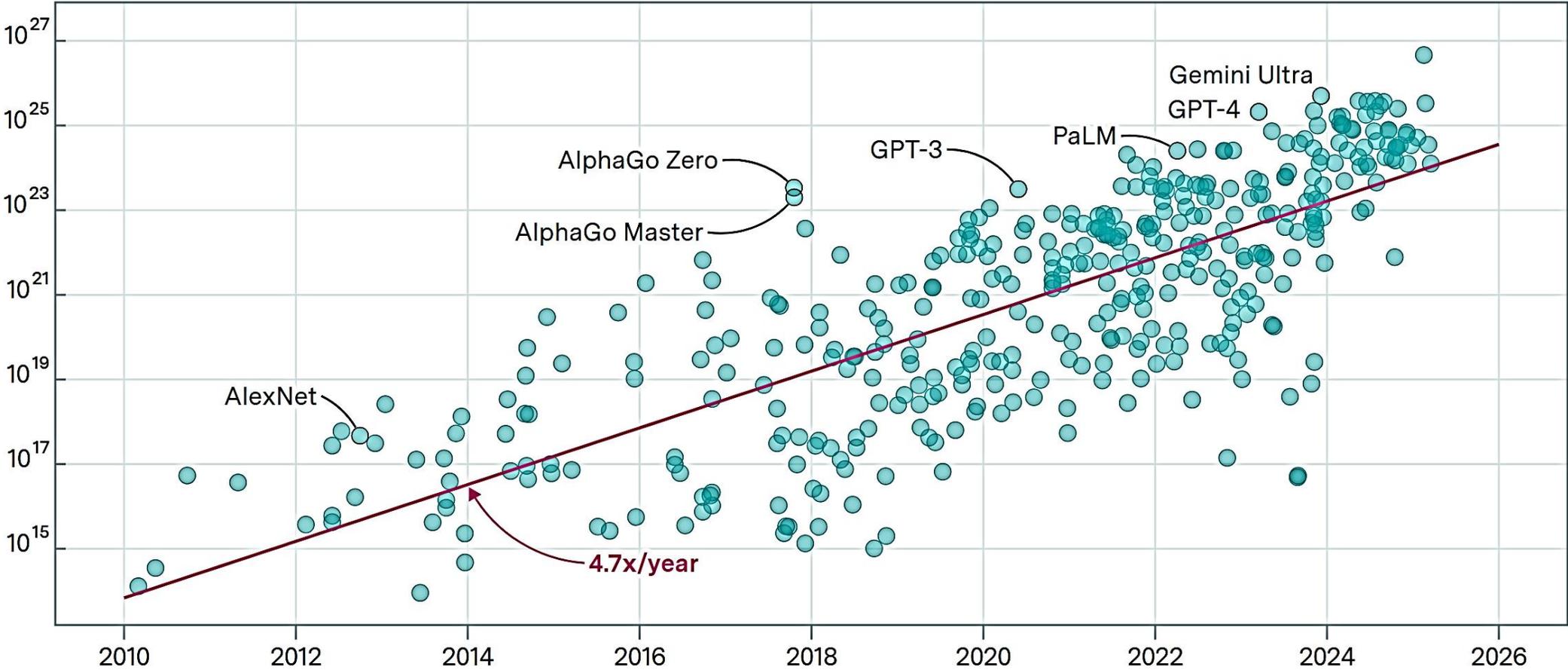
**Manoj Sukumaran**

Practice Lead,  
Data Center IT

# AI training compute requirement is doubling roughly every five months

Training compute (FLOP)

411 models

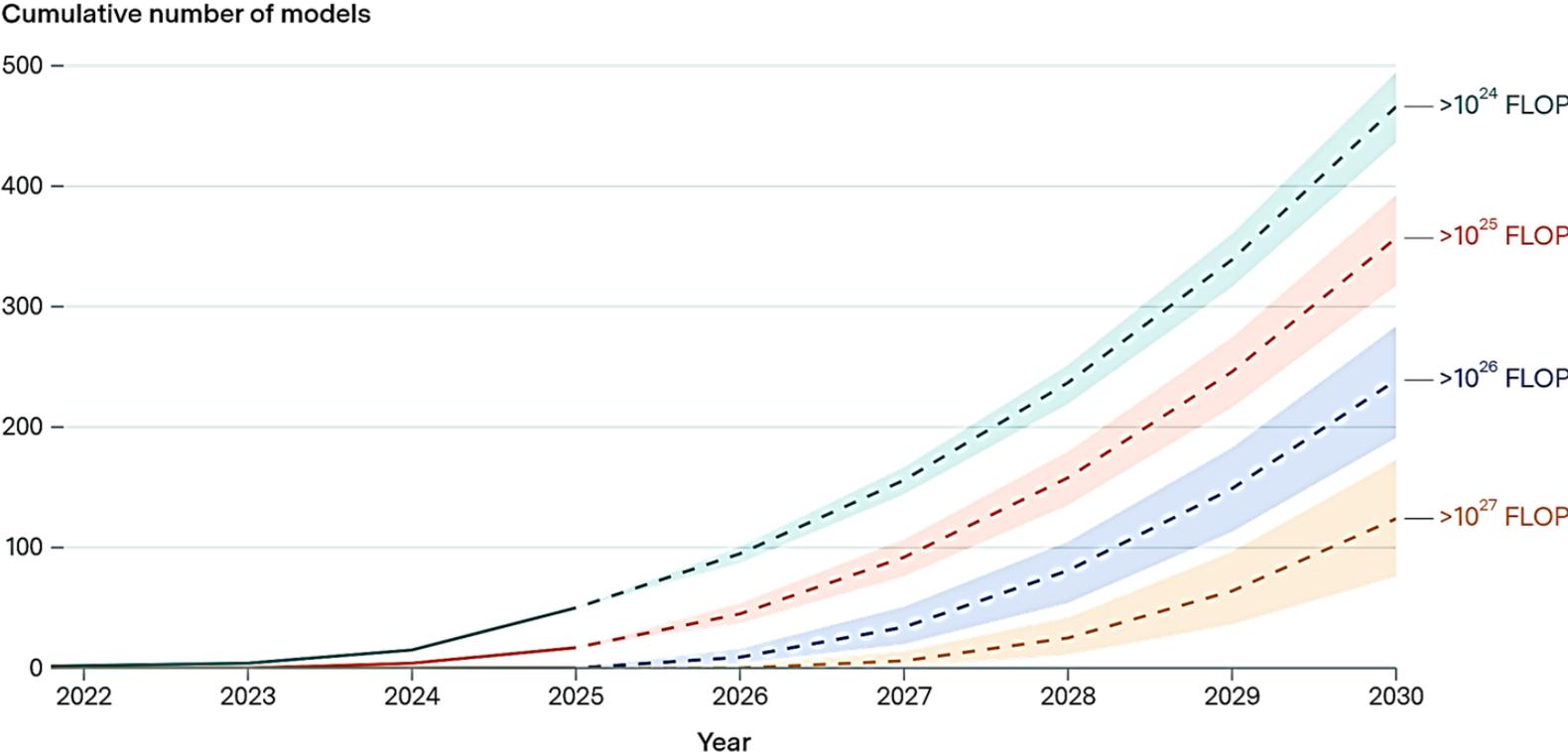


Robi Rahman and David Owen (2024), "The training compute of notable AI models is doubling roughly every five months".  
Published online at epoch.ai. Retrieved from: '<https://epoch.ai/data-insights/compute-trend-post-2010>'  
Copyright © 2025 TechTarget, Inc. or its subsidiaries. All rights reserved.

# More and more AI models coming up

## Cumulative number of notable AI models by year

Median projection and 80% confidence interval for different training compute thresholds.



As of mid-2025, **33+ models** from at least 11 organizations have been trained at or beyond GPT-4 scale ( $\geq 10^{25}$  FLOPs) despite the huge costs (tens of millions of dollars per training run)

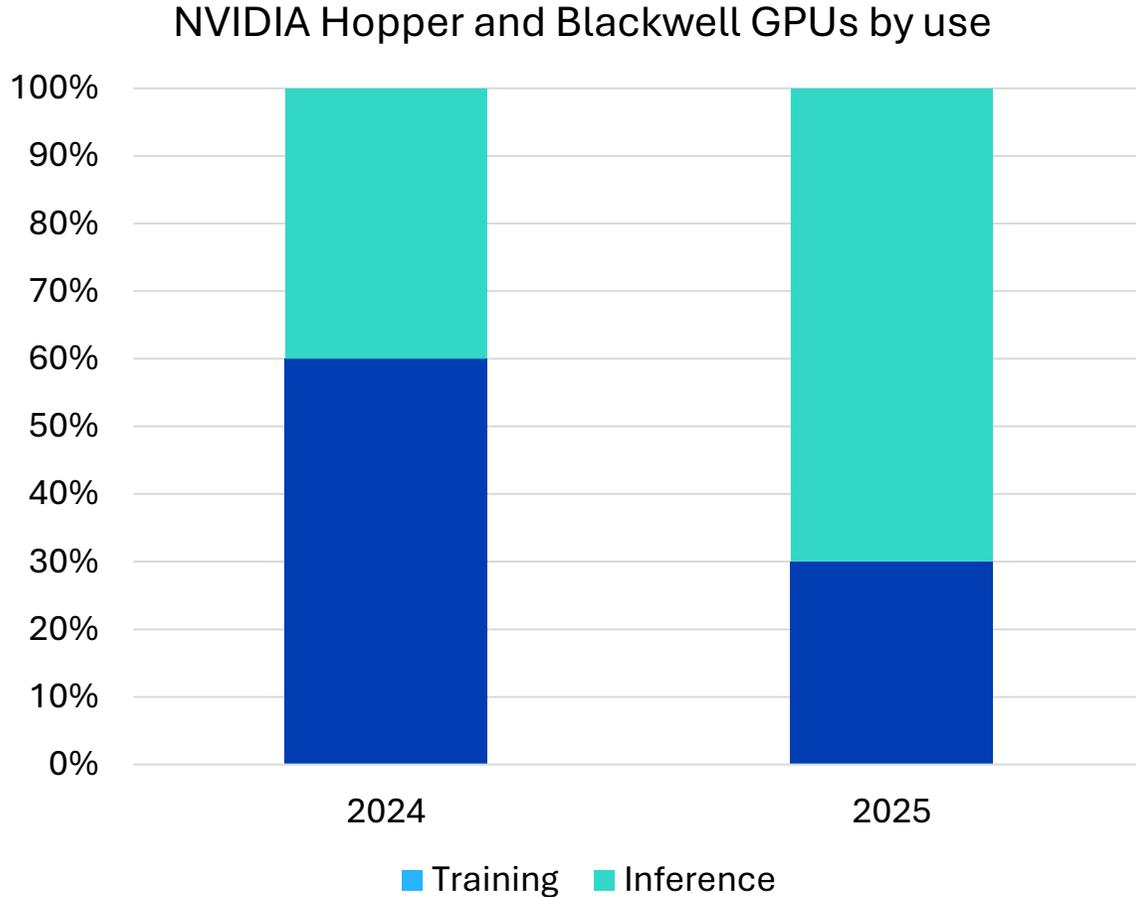
By 2030,

>10 <sup>24</sup> FLOP	466
>10 <sup>25</sup> FLOP	357
>10 <sup>26</sup> FLOP	239
>10 <sup>27</sup> FLOP	124

CC-BY

epoch.ai

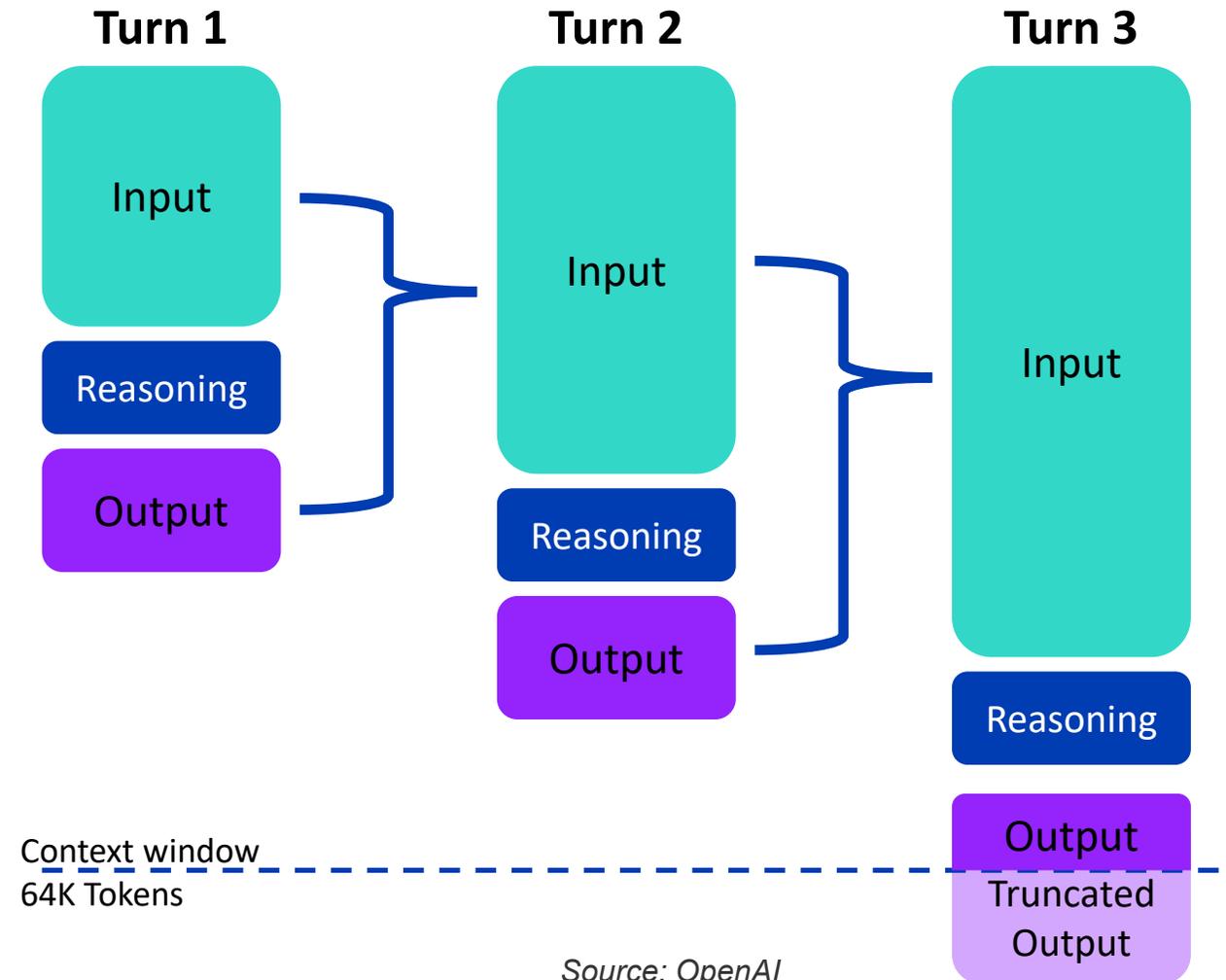
# Training vs. inference shift



- **Inference is now the dominant AI workload.** It will only continue to take share as AI goes mainstream.
- The generation/inference capability of AI models has transitioned **from static responses to complex reasoning**:
  - Early LLMs (e.g. GPT-3) mostly completed sentences
  - Newer models (GPT-4, Claude 3, Gemini 1.5) handle multi-turn reasoning, multi-modal inputs, tool use, and dynamic planning
- **Reasoning requires more compute cycles per generation request.**
  - Reasoning models must track context across longer conversations, maintain state, run structured computations, or call APIs and tools. This increases latency, token count, and memory usage per inference.
- **Key-Value Cache demands are exploding.**
  - As context windows grow (Gemini 1.5 and Claude 3 exceed 100K tokens), the KV cache size per query increases massively – this consumes more high-bandwidth memory and increases per-query GPU utilization.

# Reasoning capability is making AI inference more compute intensive

- o1 by OpenAI
- Gemini 2.0 Flash Thinking by Google
- Grok 3 by xAI
- R1 by DeepSeek
- Claude 3.7 Sonnet by Anthropic
- Thinking QwQ by Alibaba



# Agentic AI Case Study Call Center

## Customer Query

My Smart Home thermostat isn't connecting to Wi-Fi, and I keep seeing an error code E-22. How can I fix this?

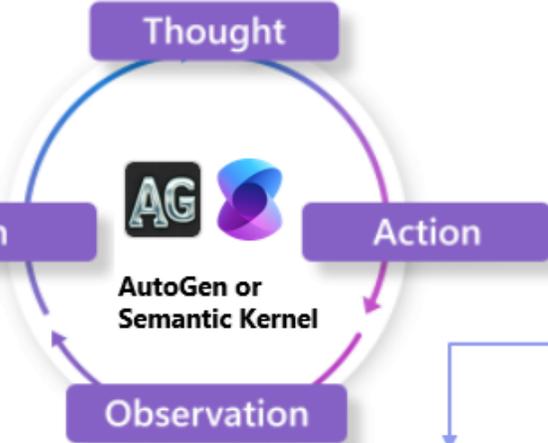
Query Parsing  
LLM

- { Wi-Fi }
- { Error: E-22 }
- { Thermostat X10 }

## Dynamic Plan

- Doc Agent** → Search internal support documentation for error code "E-22" and Wi-Fi troubleshooting.
- Video Agent** → Locate and timestamp troubleshooting video content related to Wi-Fi connection and error code E-22.
- Web Search Agent** → Search external sources for recent discussions, tips, or product-specific issues regarding Wi-Fi and the error code.

Planner Agent  
Azure AI Agent Service

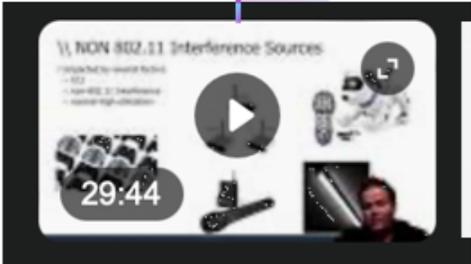


\* Abstraction

Video Agent  
Azure AI Agent Service

Doc Agent  
Azure AI Agent Service

Web Agent  
Azure AI Agent Service



Gathers internal troubleshooting steps.



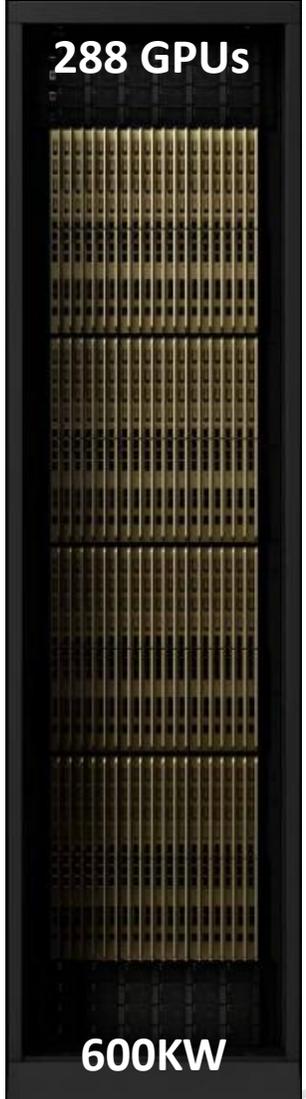
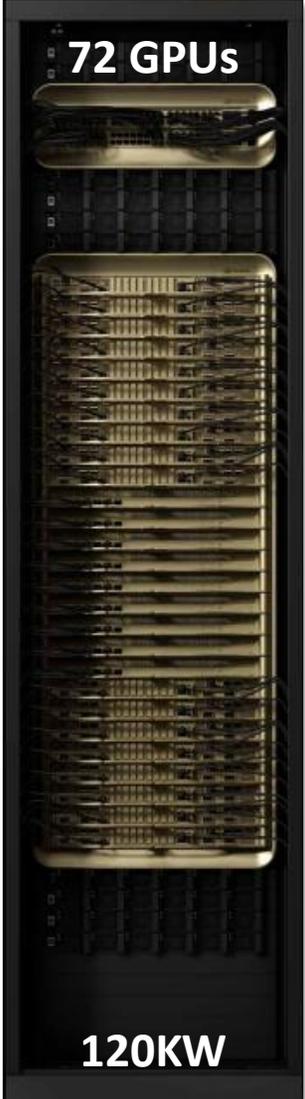
Summarizes video clips related to Wi-Fi and error code E-22.



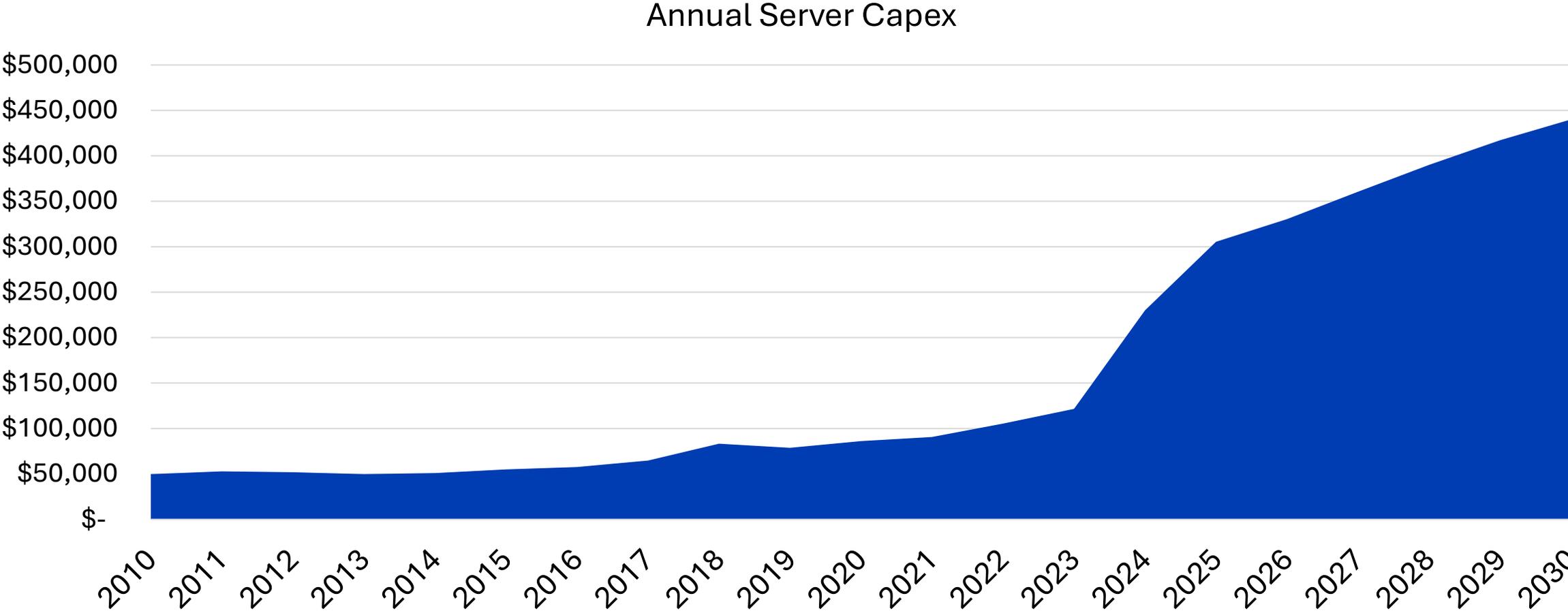
Finds the latest external resources related to the issue

Source: Microsoft

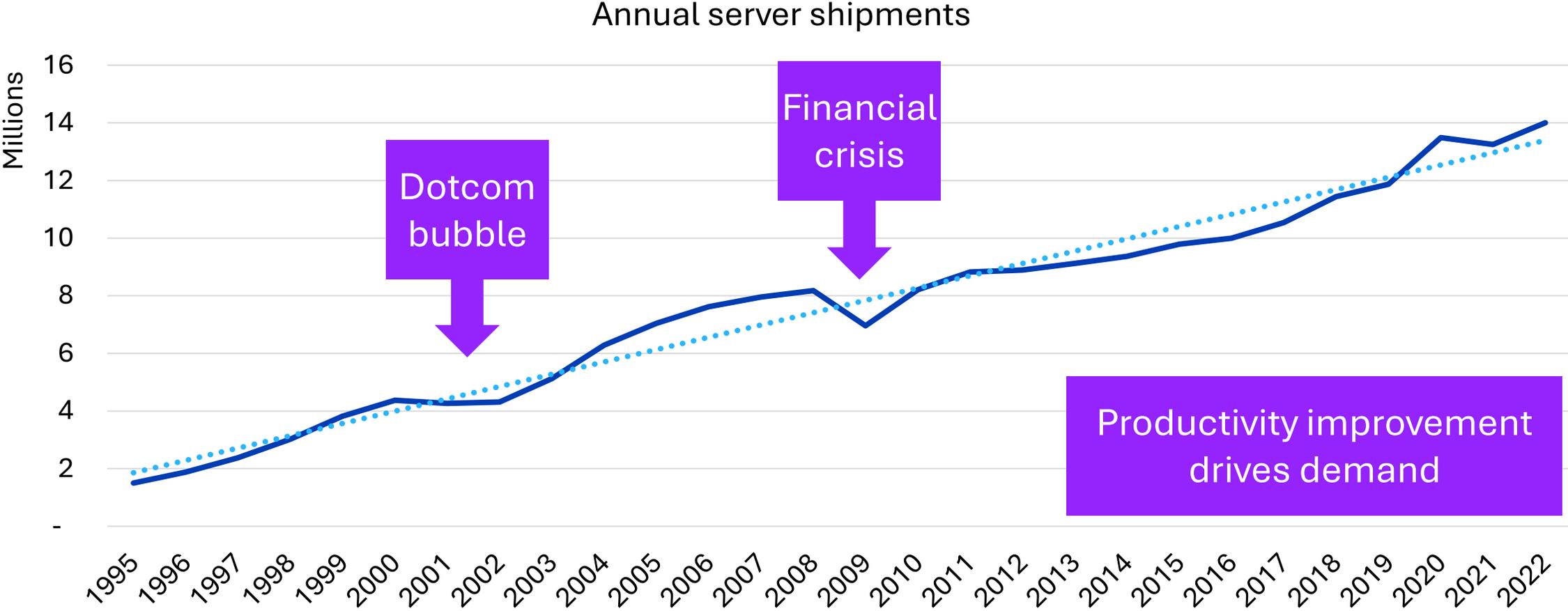
# Increasing compute and memory demands are driving rack scale system designs



# And the numbers speak for themselves



# “Bubbles” are a stock market phenomenon not an adoption curve



# AI services continue to be capacity constrained!

 **Sam Altman**   @sama · Apr 1  ...

we are getting things under control, but you should expect new releases from openai to be delayed, stuff to break, and for service to sometimes be slow as we deal with capacity challenges.

 719  583  10K  1.2M  

 **Sam Altman**   @sama  ...

working as fast we can to really get stuff humming; if anyone has GPU capacity in 100k chunks we can get asap please call!

9:20 PM · Apr 1, 2025 · **628.5K** Views

 Seeking Alpha

## Google said to be closing in on deal to rent Nvidia servers from CoreWeave: report (GOOG:NASDAQ)

Google (GOOG) (GOOGL) is close to agreeing to a deal with CoreWeave (CRWV) to rent out data center servers using Nvidia's (NVDA) Blackwell...

2 weeks ago

 **Mark Zuckerberg**  16m ·  ...

For our superintelligence effort, I'm focused on building the most elite and talent-dense team in the industry. We're also going to invest hundreds of billions of dollars into compute to build superintelligence. We have the capital from our business to do this.

SemiAnalysis just reported that Meta is on track to be the first lab to bring a 1GW+ supercluster online. 🙌

We're actually building several multi-GW clusters. We're calling the first one Prometheus and it's coming online in '26. We're also building Hyperion, which will be able to scale up to 5GW over several years. We're building multiple more titan clusters as well. Just one of these covers a significant part of the footprint of Manhattan.

Meta Superintelligence Labs will have industry-leading levels of compute and by far the greatest compute per researcher. I'm looking forward to working with the top researchers to advance the frontier!



受衣餘國大員人習會

受衣餘國大員人習會

MAR-A-LAGO

# Omdia Analyst Summit: Where is AI really headed?

Time	Session	Speakers
11:00 – 11:45	Where is AI really headed? AI的发展方向在哪里?	Vlad Galabov
11:45 – 12:30	Data center rack architecture and power density out to 2030 2030年前的数据中心机柜架构与功率密度	Manoj Sukumaran
12:30 – 14:00	Break 茶歇	
14:00 – 15:00	Powering the AI data center now and in 2030 为当下和2030年的数据中心供电	Shen Wang
15:00 – 15:30	Break 茶歇	
15:30 – 16:30	Cooling the AI data center now and in 2030 为当下和2030年的数据中心散热	Jessica Nian
16:30 – 17:00	Wrap up and audience Q&A 总结与问答环节	Vlad Galabov

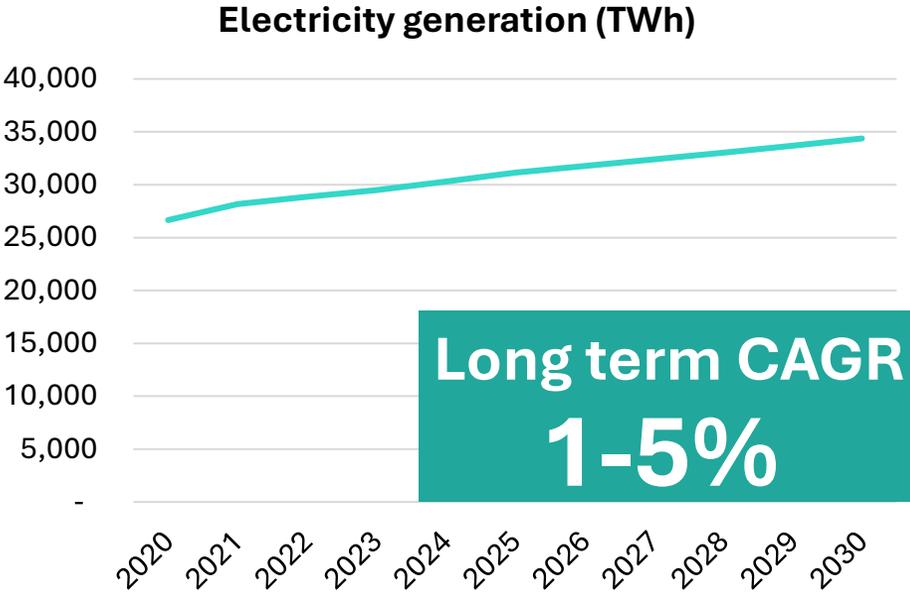
# Powering the AI data center now and in 2030

**Shen Wang**

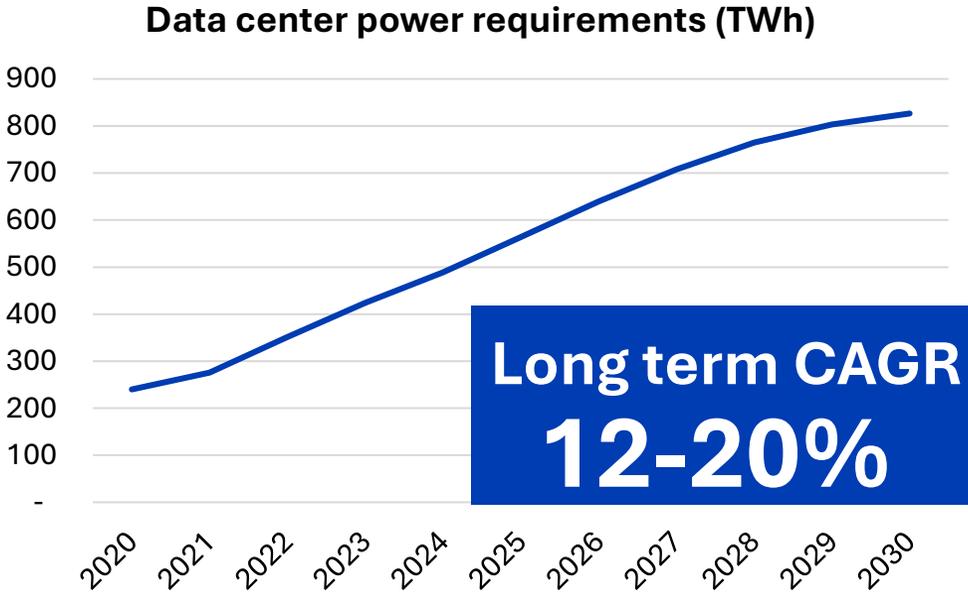
Practice Lead,  
Data center capacity and infrastructure

© 2025 TechTarget, Inc. or its subsidiaries. All rights reserved.

# Global power generation capacity growth is slower than data center demand growth



Source: Omdia consensus based on IEA and Rethink Energy



Source: Omdia

Source: IEA

### Surge in demand

Global electrification is accelerating, with electric vehicles increasing sixfold in six years

The growth rate of power generation lags behind, and the competition for electricity is intensifying

### Supply bottleneck

Traditional power grids cannot meet the expansion needs of data centers

Operators are forced to turn to off-grid power generation (self-built microgrids)

### Industry Transformation

Renewable energy + energy storage system becomes the key solution

Power independence will determine data center expansion capabilities

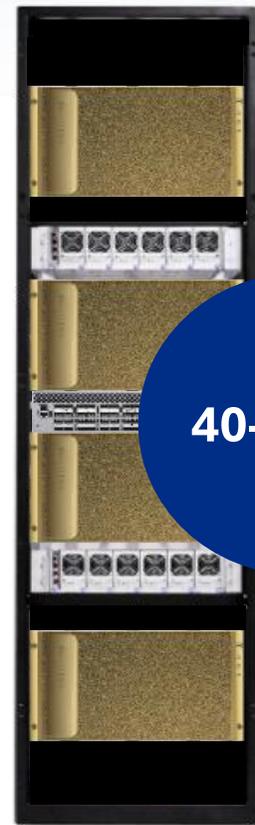
# What people used to expect to happen:

2020 - 2023



20-30KW

2022 - 2025



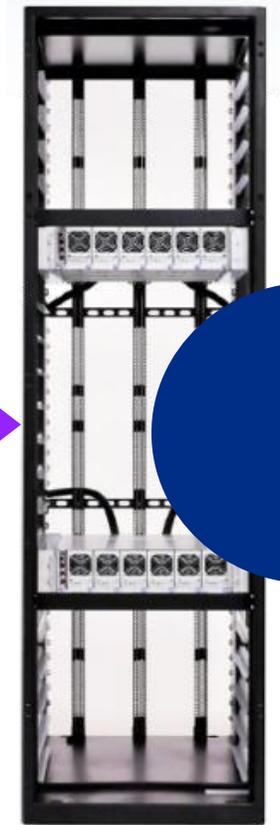
40-50KW

2024 - 2027



70-100KW

2028 & Beyond



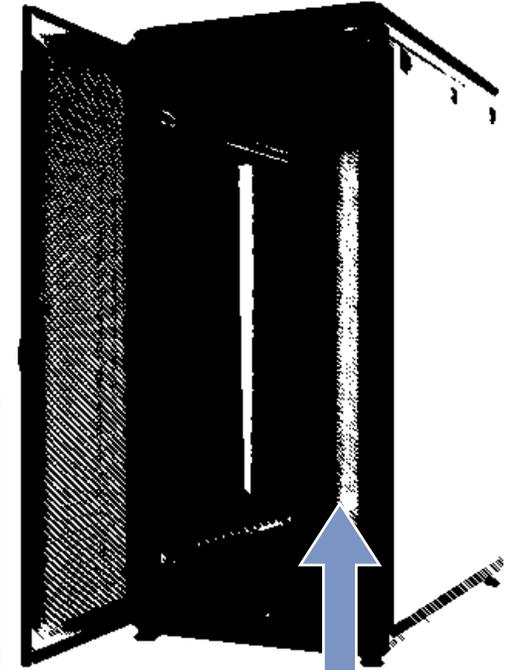
?

# What truly going to happen is...

AI Rack density double almost every year

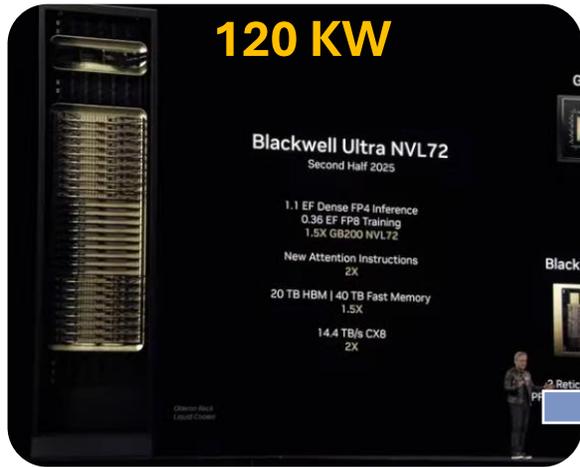
1MW

?



2H 2025

120 KW



**Blackwell Ultra NVL72**  
Second Half 2025

- 1.1 EF Dense FP4 Inference
- 0.36 EF FP8 Training
- 1.5X GB200 NVL72
- New Attention Instructions 2X
- 20 TB HBM | 40 TB Fast Memory 1.5X
- 14.4 TB/s CX8 2X

2H 2026

240 KW



**Vera Rubin NVL144**  
Second Half 2026

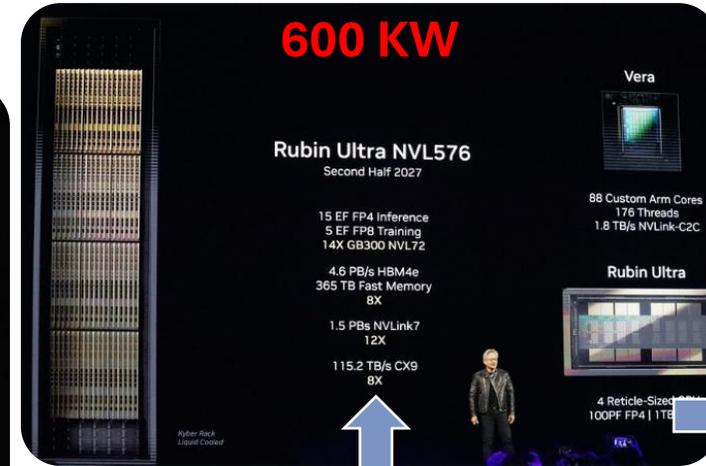
- 3.6 EF FP4 Inference
- 1.2 EF FP8 Training
- 3.3X GB300 NVL72
- 13 TB/s HBM4
- 75 TB Fast Memory 1.6X
- 260 TB/s NVLink6 2X
- 28.8 TB/s CX9 2X

**Rubin**

- 2 Reticle-Sized GPUs
- 50PF FP4 | 288CB HBM4

2H 2027

600 KW



**Rubin Ultra NVL576**  
Second Half 2027

- 15 EF FP4 Inference
- 5 EF FP8 Training
- 14X GB300 NVL72
- 4.6 PB/s HBM4e
- 365 TB Fast Memory 8X
- 1.5 PBs NVLink7 12X
- 115.2 TB/s CX9 8X

**Vera**

- 88 Custom Arm Cores
- 176 Threads
- 1.8 TB/s NVLink-C2C

**Rubin Ultra**

- 4 Reticle-Sized GPUs
- 100PF FP4 | 1TB

Soon enough we will be dealing with 1MW/rack

# Why? Cluster networking is driving IT to be denser

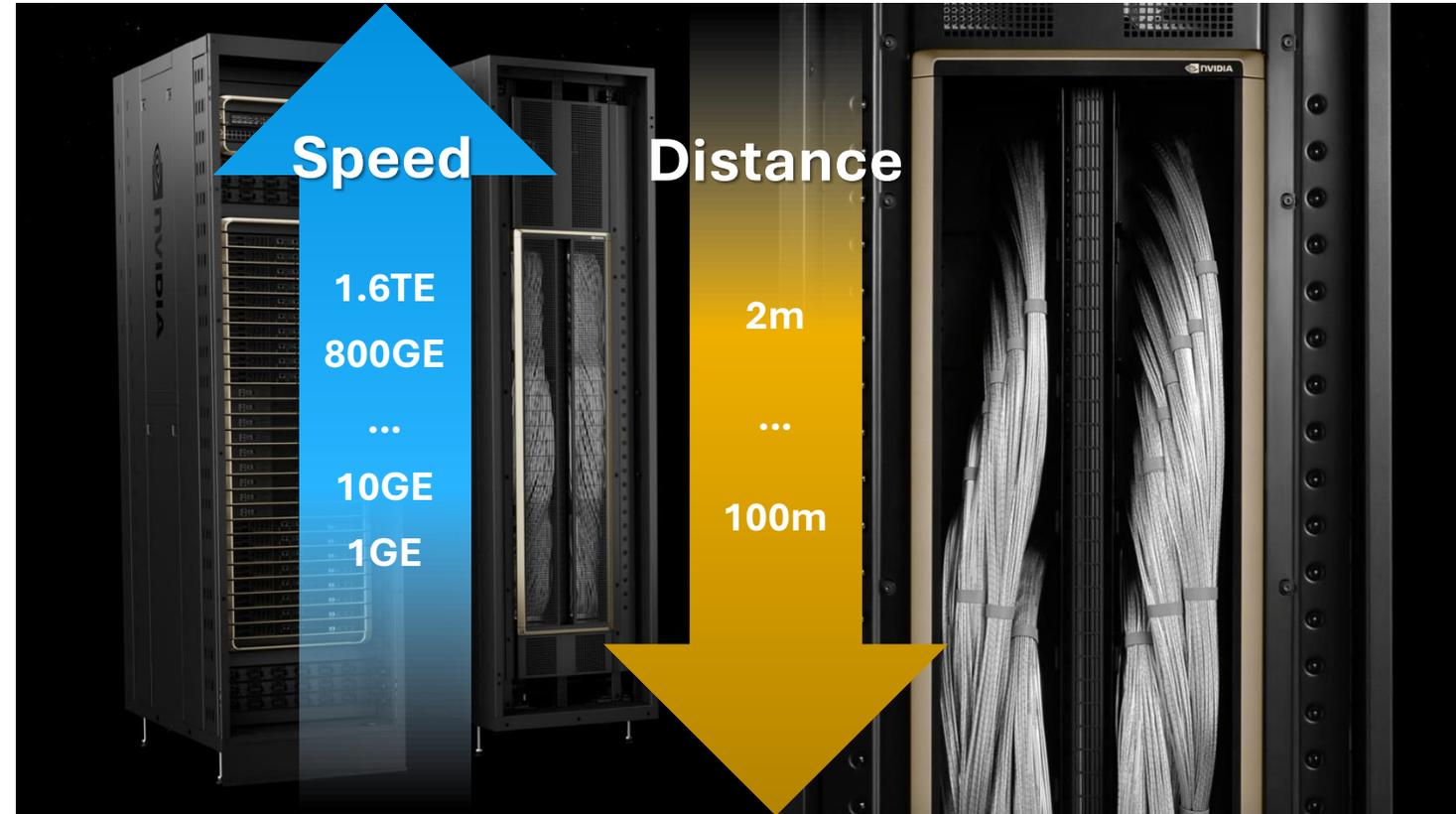
*Faster processing*

*Closer compute*

*Closer distance*

*Higher density*

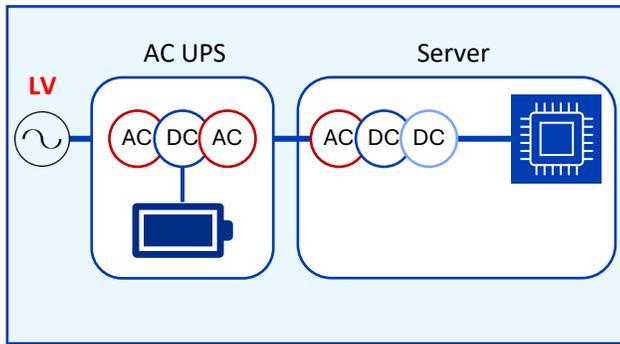
- Unless.. optical network could overcome a series of problems in a single rack
- But.. It will still take a while to achieve



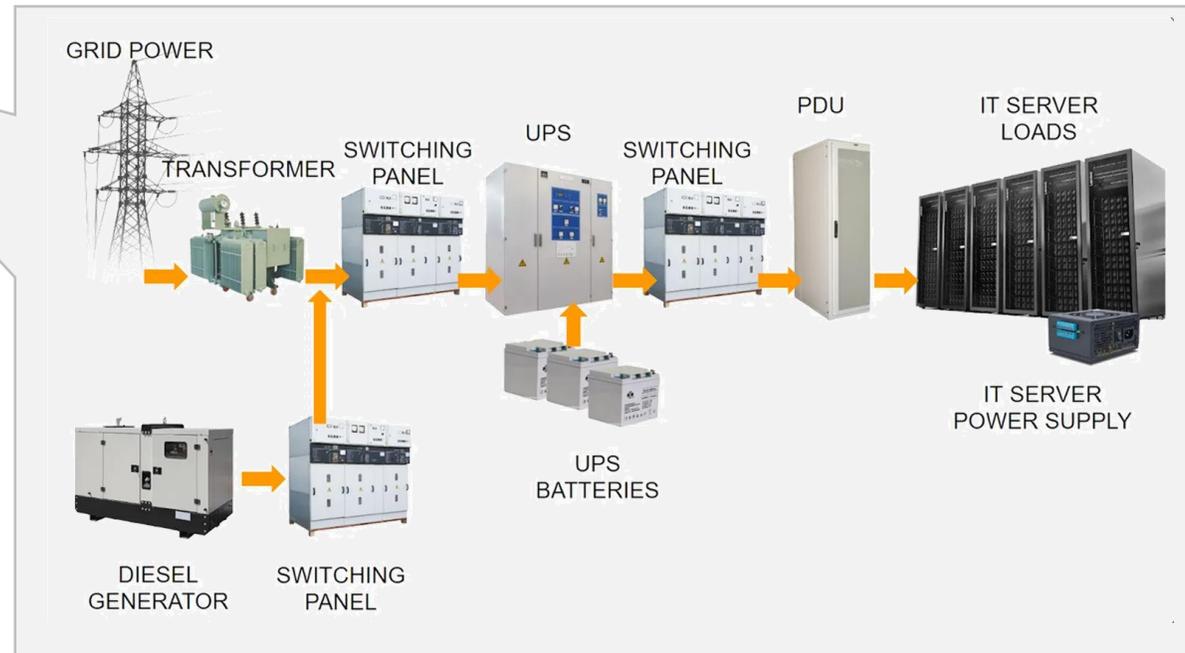
# And direct current distribution is needed in those NVL racks

## In-server direct current

Typical power setup



Source: Omdia



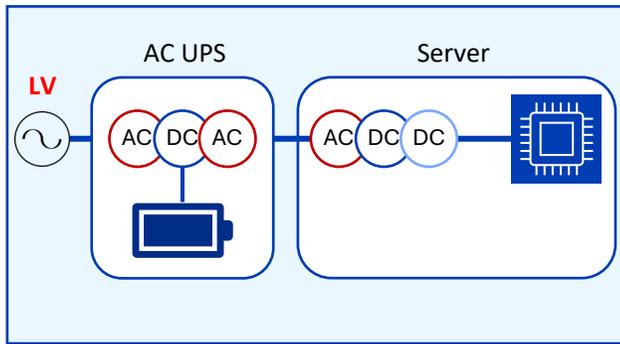
© 2025 Omdia

# And direct current distribution is needed in those NVL racks

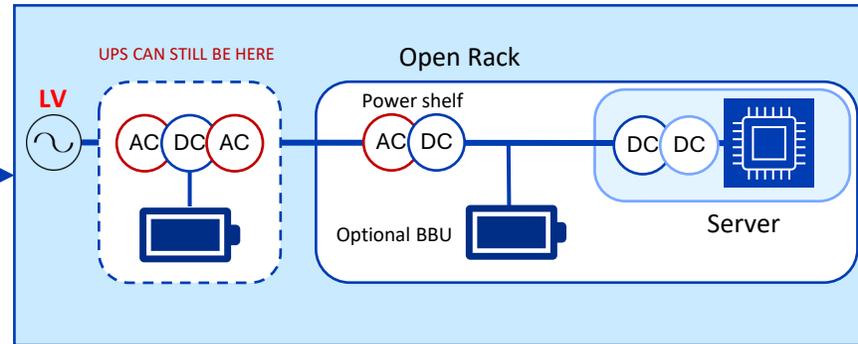
In-server direct current

In-rack direct current

Typical power setup



OCP architecture with/without BBU

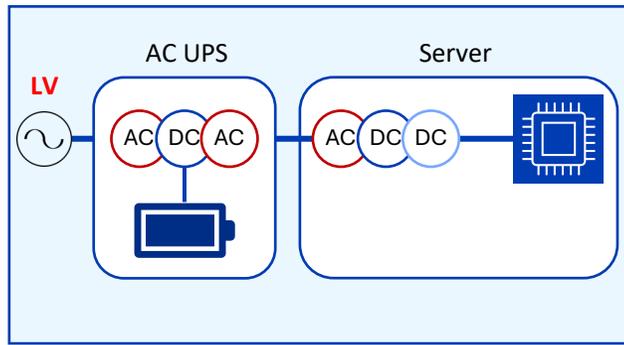


# And direct current distribution is needed in those NVL racks

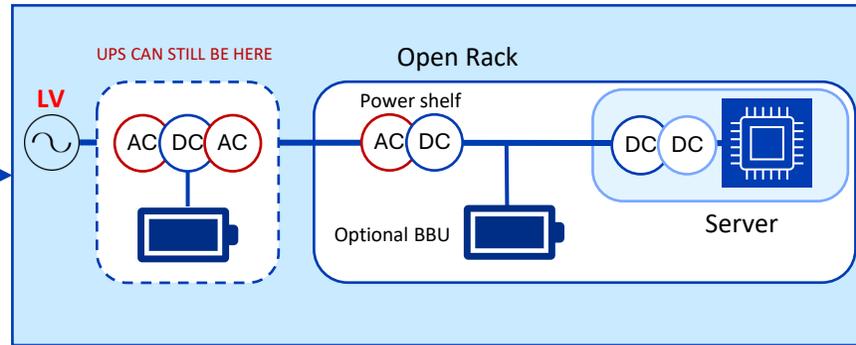
In-server direct current

In-rack direct current

Typical power setup

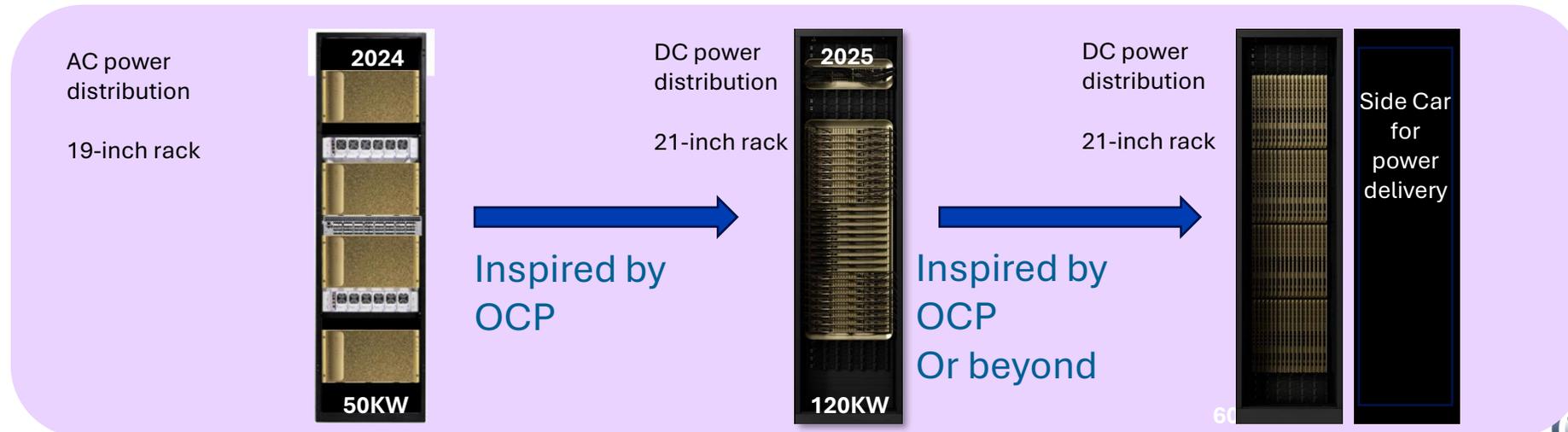


OCP architecture with/without BBU



Source: Omdia

© 2025 Omdia

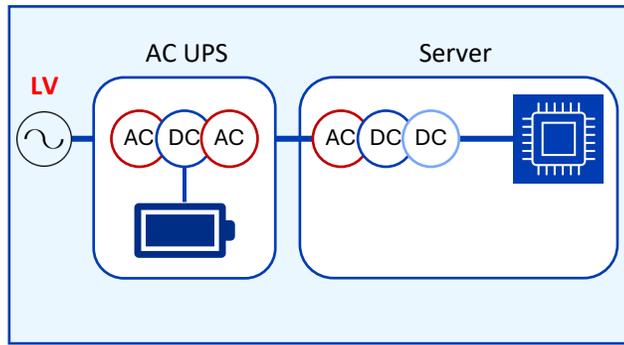


# And direct current distribution is needed in those NVL racks

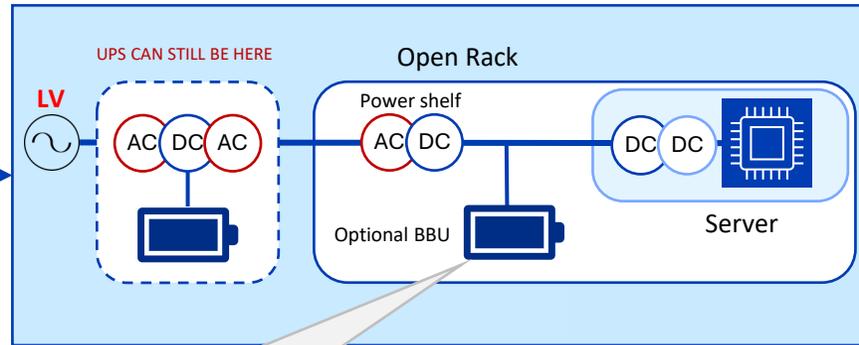
In-server direct current

In-rack direct current

Typical power setup

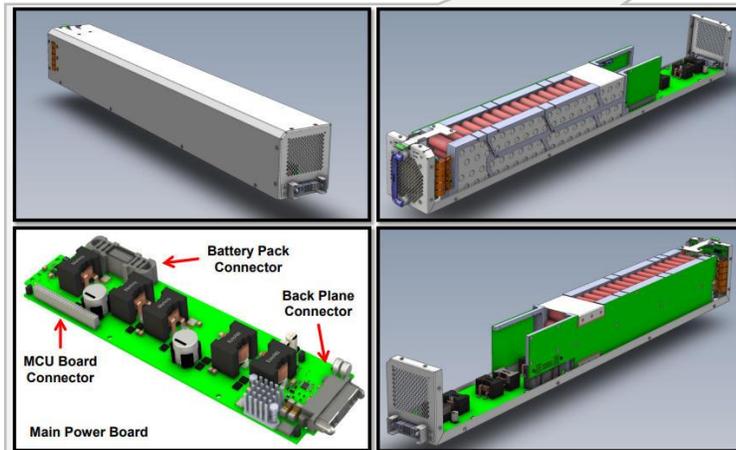


OCP architecture with/without BBU

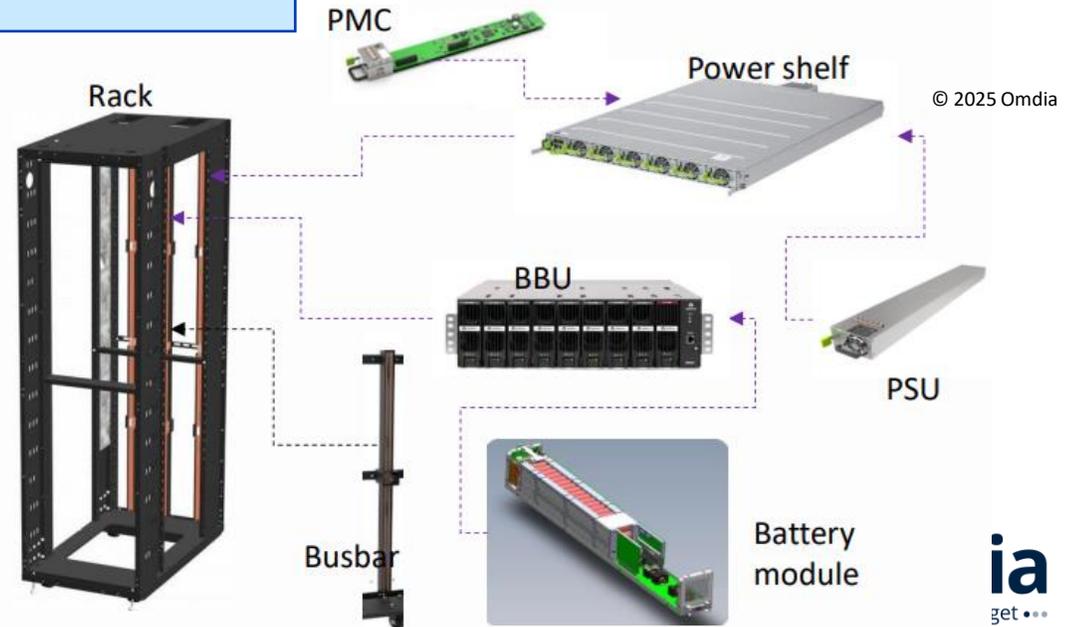


Source: Omdia

ORV3 continues iterative updates, enabling in-rack DC distribution and BBU as an option



st, Inc. or its subsidiaries. All right



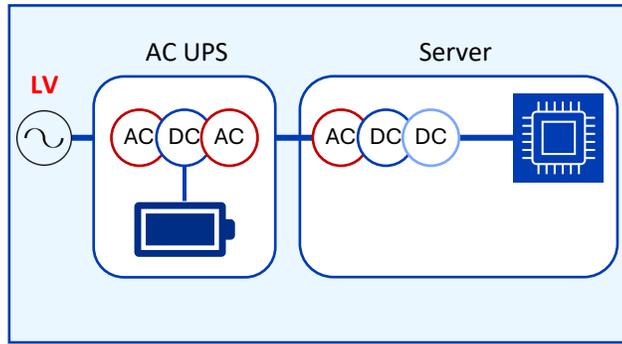
© 2025 Omdia

# And direct current distribution is needed in those NVL racks

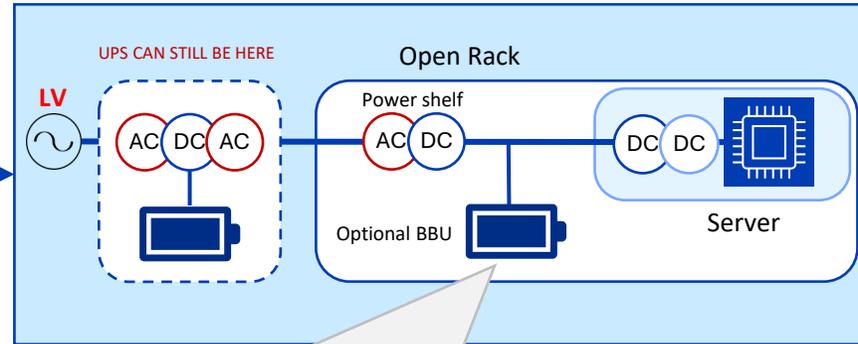
In-server direct current

In-rack direct current

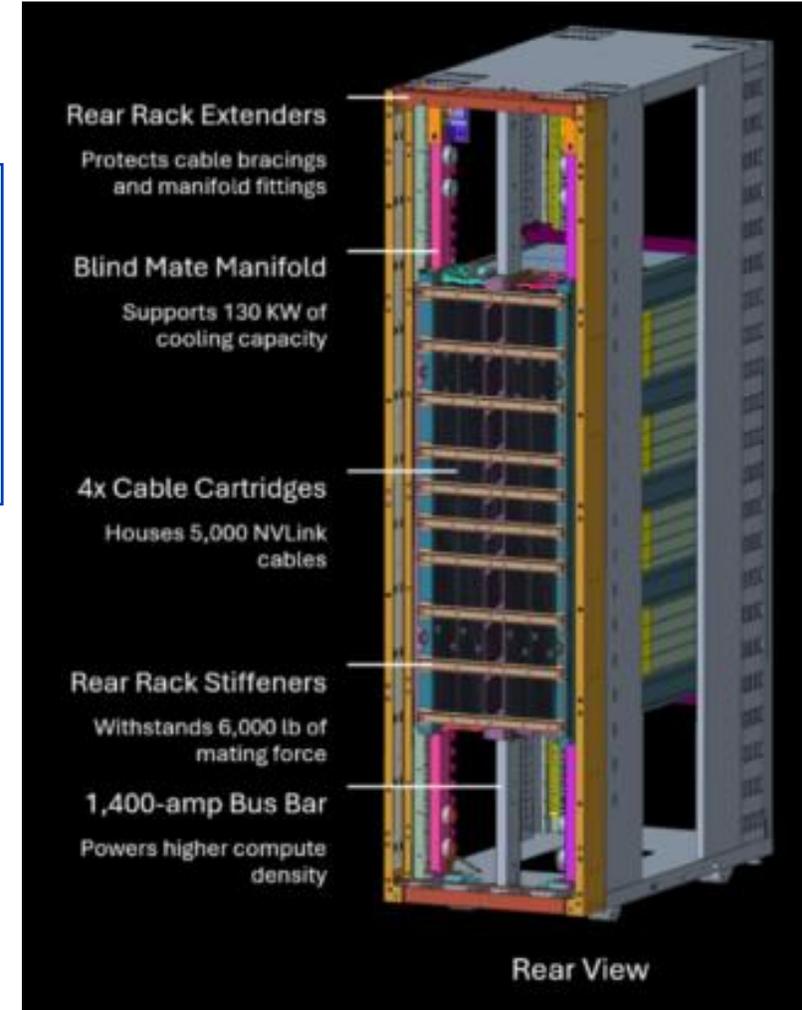
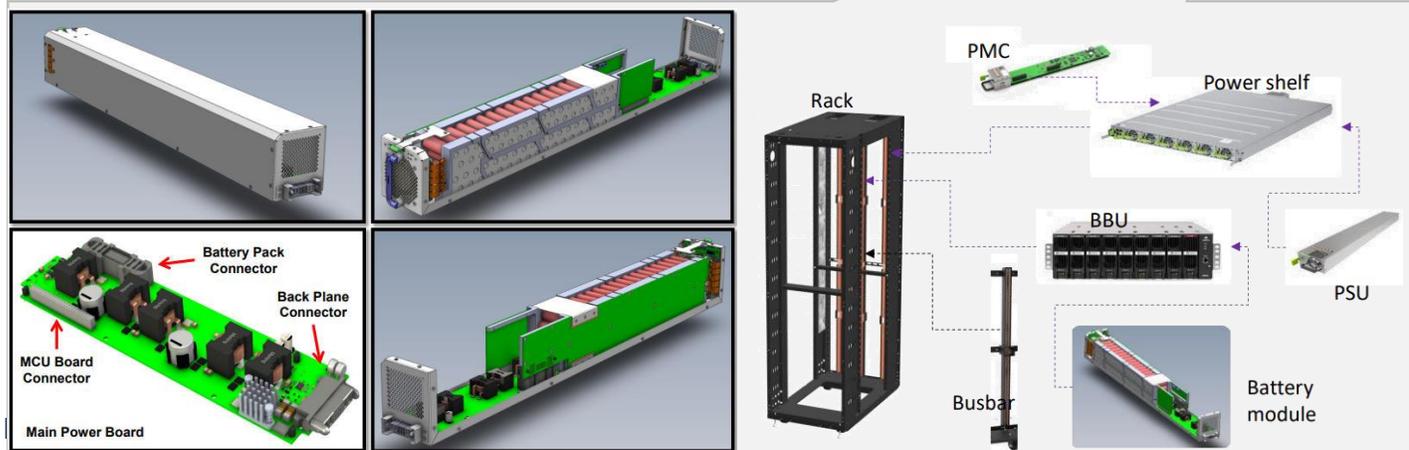
Typical power setup



OCP architecture with/without BBU

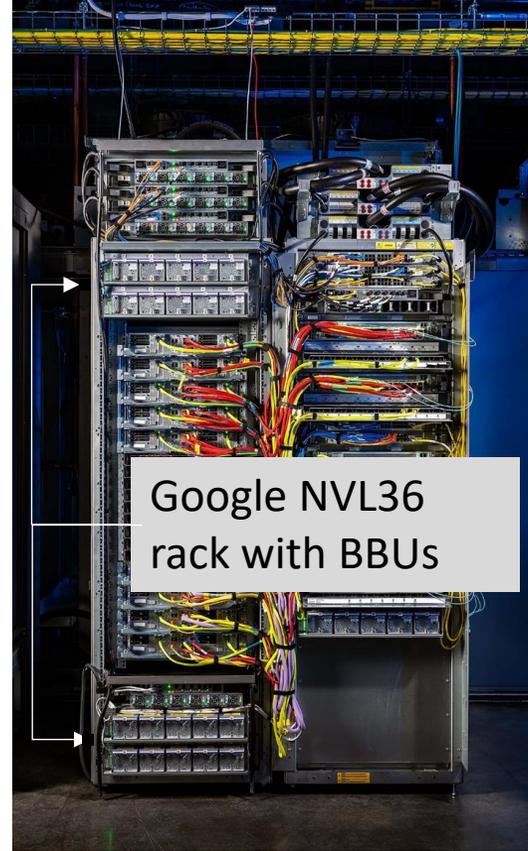


Source: Omdia



# BBU: already in some DCs

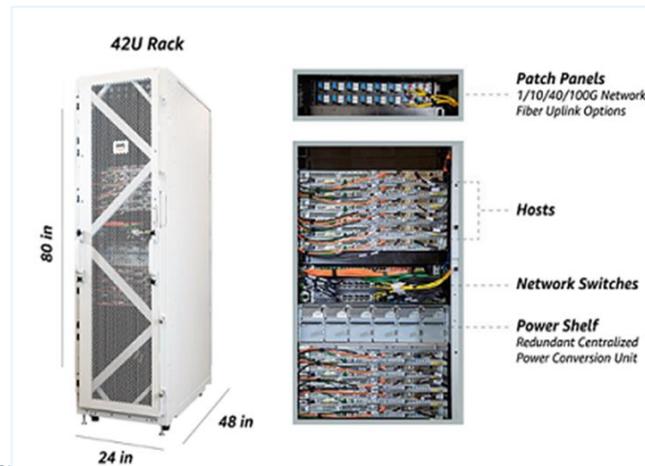
- Not every rack is equipped with BBU
  - Meta: 12V BBU in ORV2, 48V BBUs in ORV3 racks.
  - Google used BBU since ORV3, and continues with NVL36/72
  - Microsoft’s NVL32/72 rack will have a BBU
  - Amazon will also not use BBUs in their current and next gen racks
  - xAI will have no BBU and no UPS, but only genset and
  - Alibaba and Baidu also uses BBUs at small scale since 2016, starting with a 12V and now uses 48V 3kW BBU.



Microsoft NVL32 rack



Alibaba's first generation of BBU is 12V and occupies one PSU slot in the server



Amazon's Outpost rack does not have BBU, and there is no plan to deploy BBU in the short term.

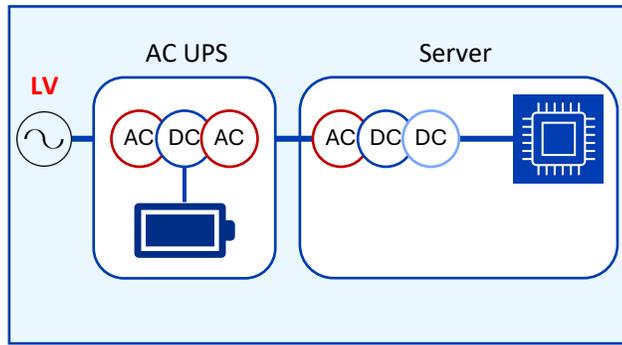
# And direct current distribution needs higher voltage

In-server direct current

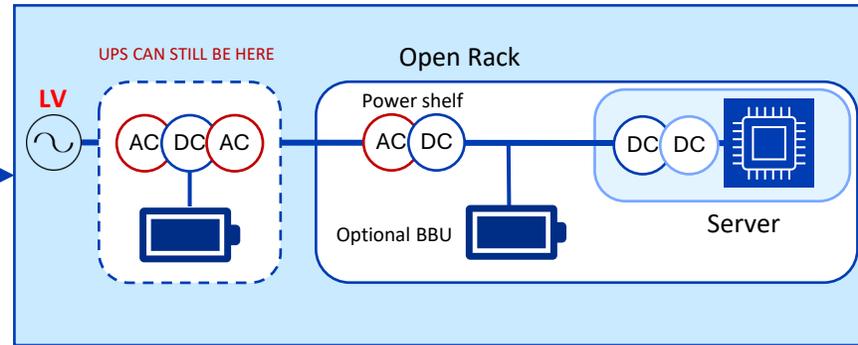
In-rack direct current

In-row direct current

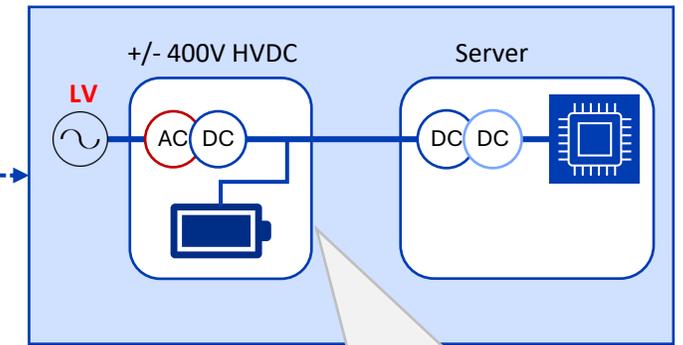
Typical power setup



OCP architecture with/without BBU



HVDC architecture setup



Source: Omdia

© 2025 Omdia

**Current HVDC solution run on lower voltage**



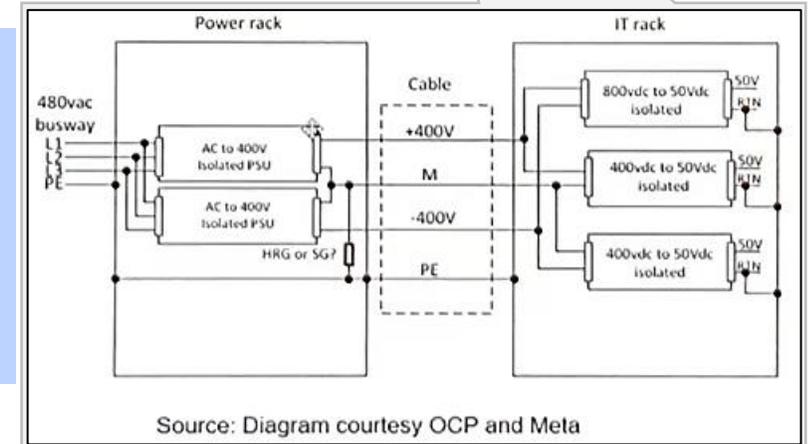
**Next ±400VDC will have two sub-stages**



- China Telecom, Alibaba HVDC: 240VDC
- China Mobile: 336VDC
- Baidu: 270VDC & 750VDC

- HVDC between power rack and IT rack
- HVDC infra-level to all racks

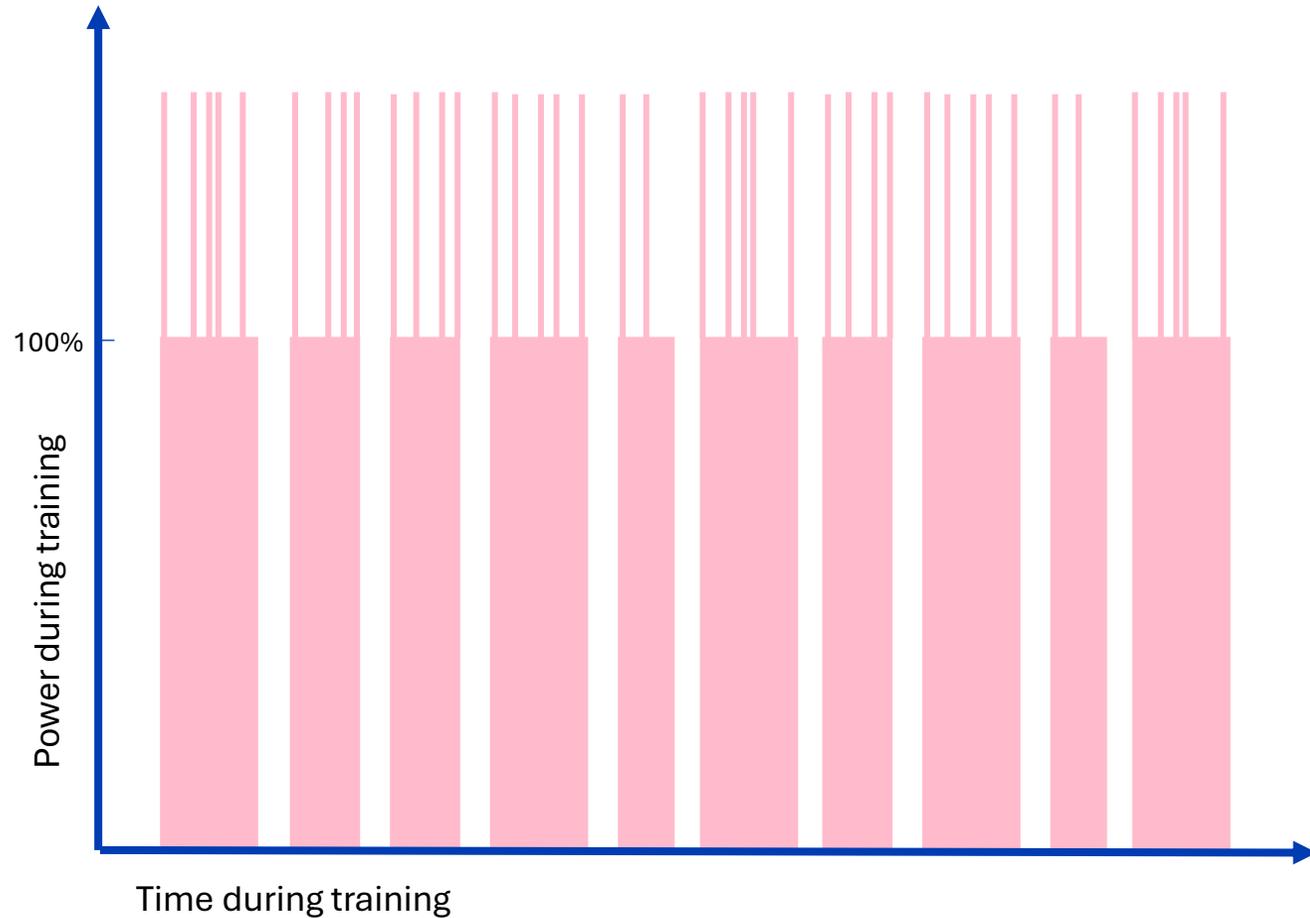
The ±400VDC solution builds on existing power infrastructure, enabling two DC voltage configuration via power racks.



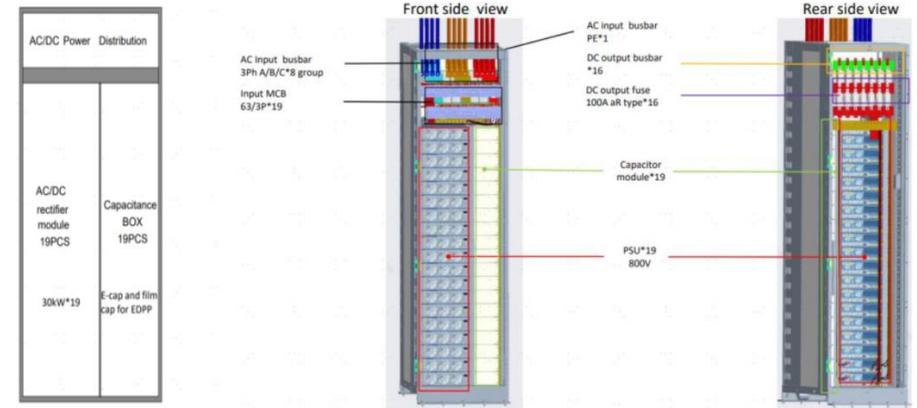
Source: Diagram courtesy OCP and Meta

Copyright © 2025 TechTarget, Inc. or its subsidiaries. All rights reserved.

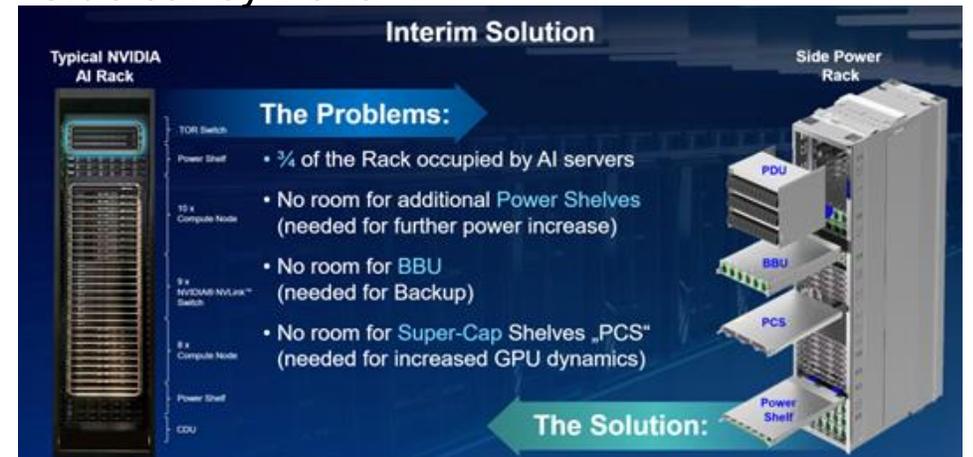
# Power sidecar will become a must, just in different architecture



## Side car by Magmeet



## Side car by Delta



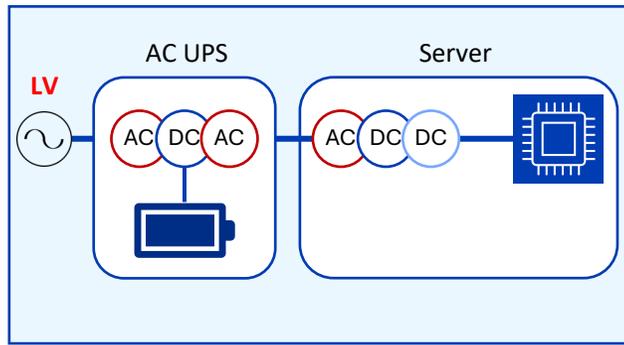
# And direct current distribution needs higher voltage

In-server direct current

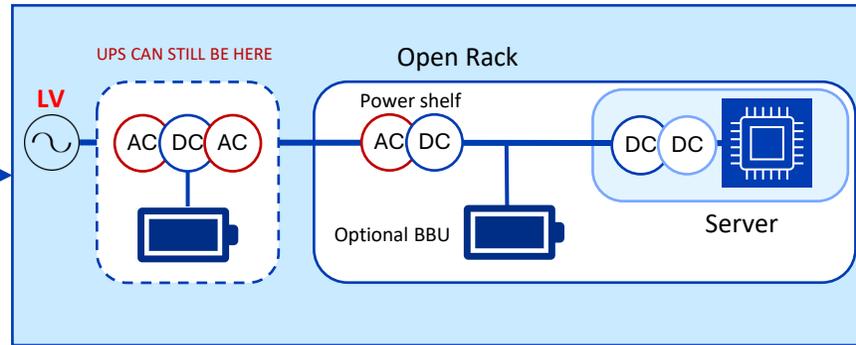
In-rack direct current

In-row direct current

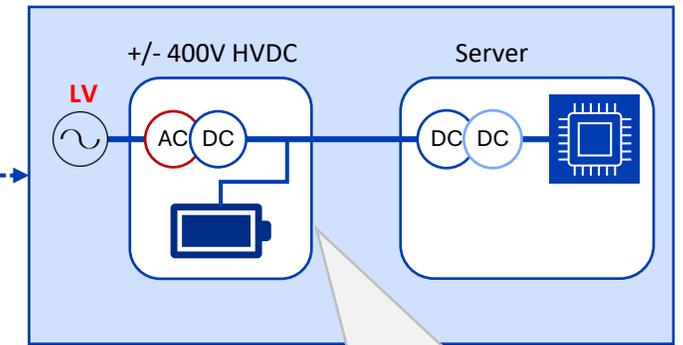
Typical power setup



OCP architecture with/without BBU

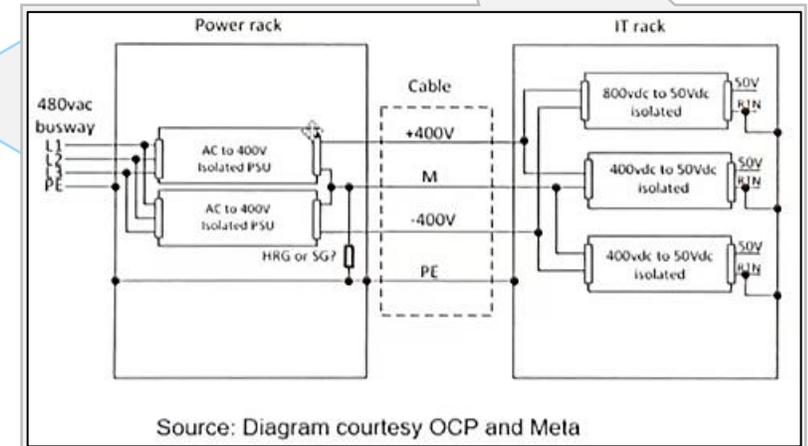
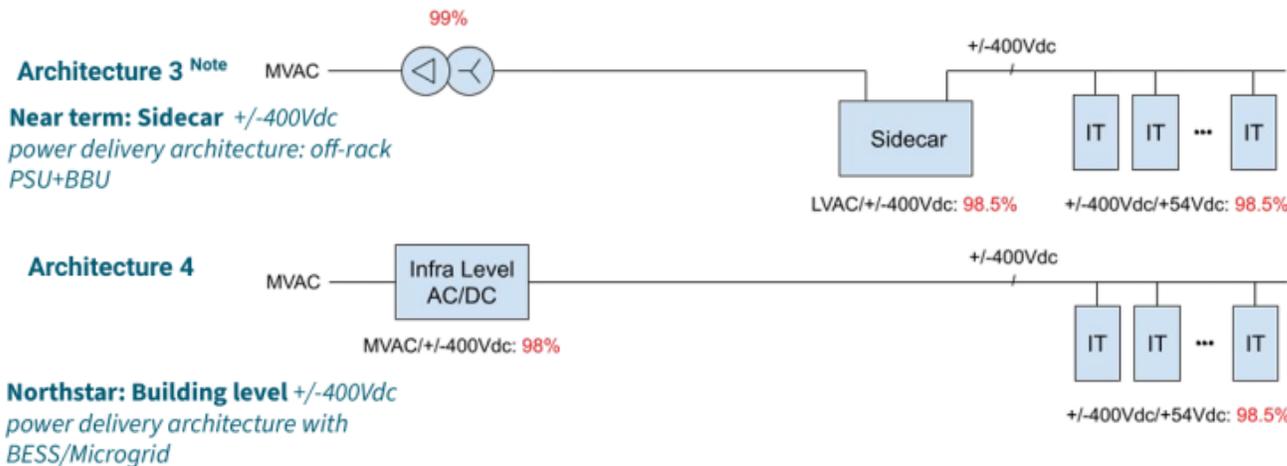


HVDC architecture setup



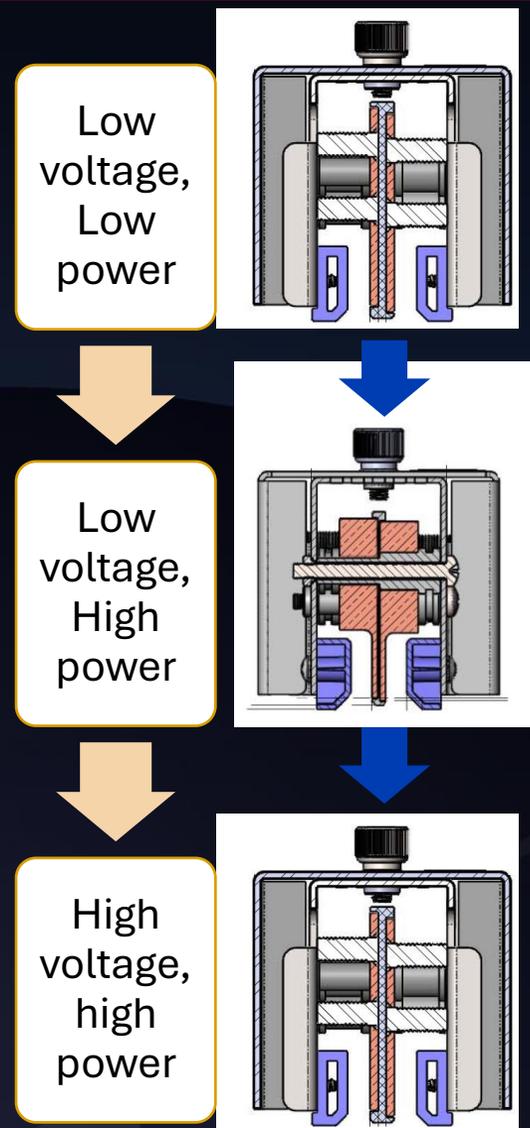
Source: Omdia

© 2025 Omdia

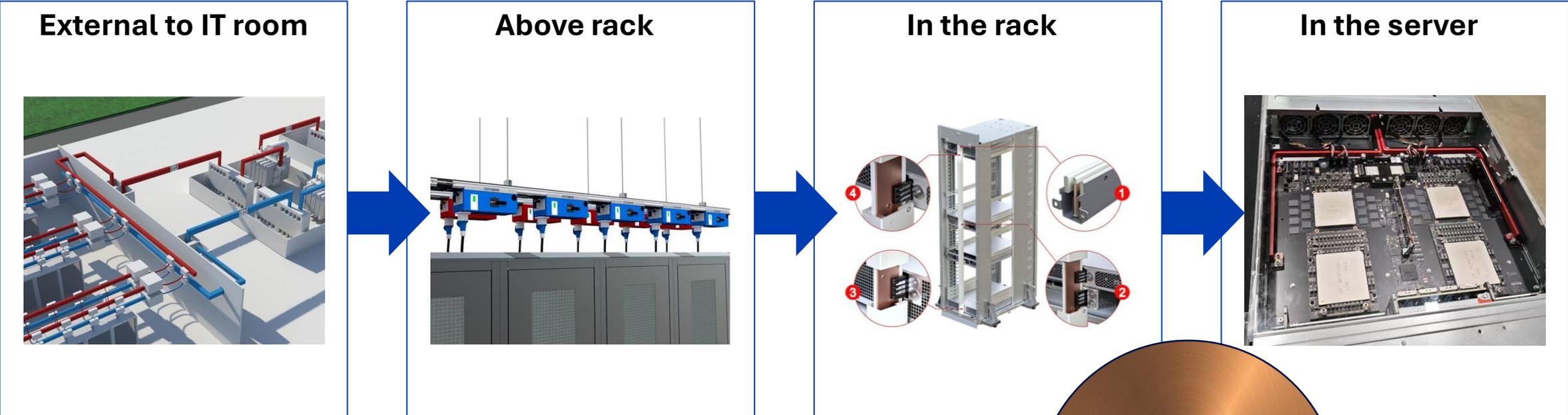


Source: Google

# Higher copper price will further accelerate adoption of higher voltage



# Busway/busbar in server is also often seen in server



**~3 tons  
of copper  
per MW**

**~4% of total copper demand are in data centers**

# The 2027 target is native 800VDC and Nvidia is accelerating it

## Facility-level native 800VDC

### Why?

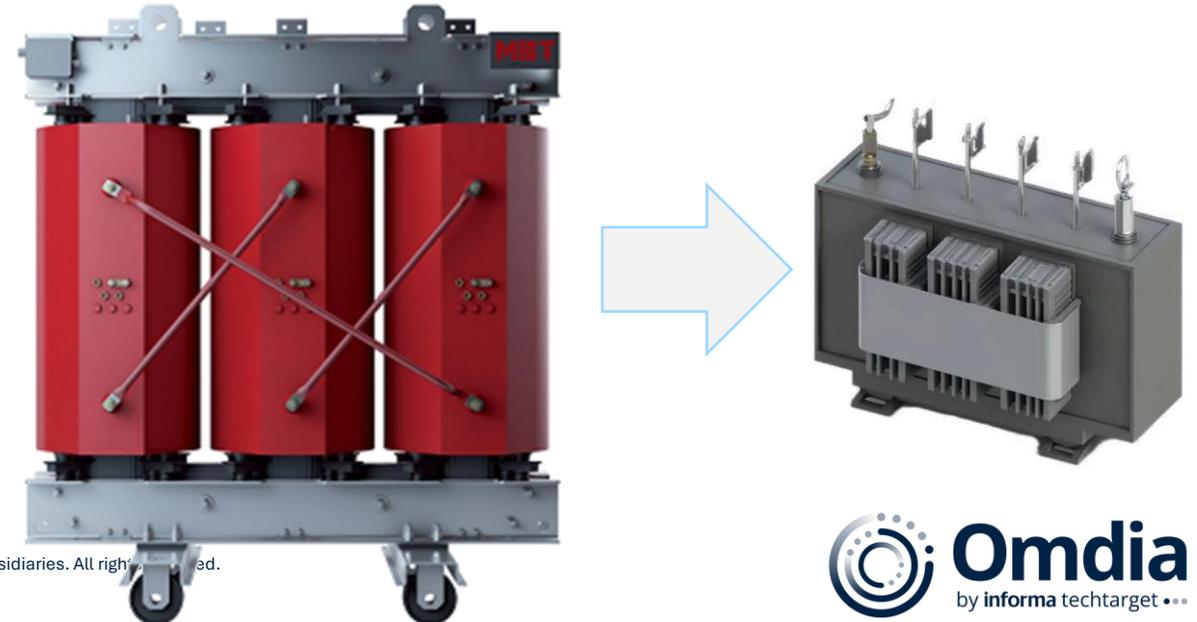
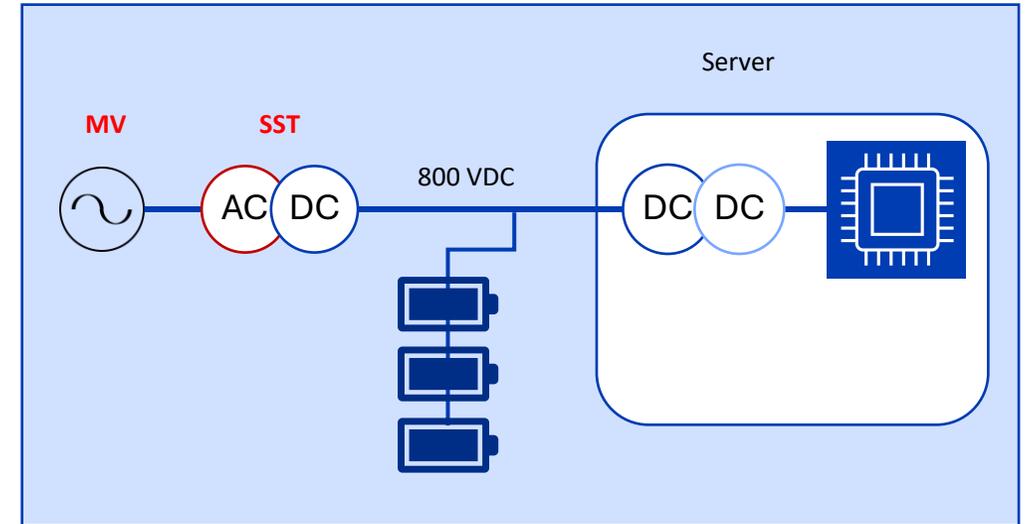
- Saves Copper – 85% more power transmission vs. 415V AC (critical under US copper tariffs)
- Higher Efficiency – Fewer conversions, lower cooling/cable losses, and reduced failure points
- Compact Design – Fewer components free up space for IT equipment

### How?

- Solid-state transformer – One-time power transition only, MV AC input, 800VDC output.
- Existing powersemi solutions – from 800VDC EV industry.
- Relies on SiC/GaN semiconductors and specialty magnetics

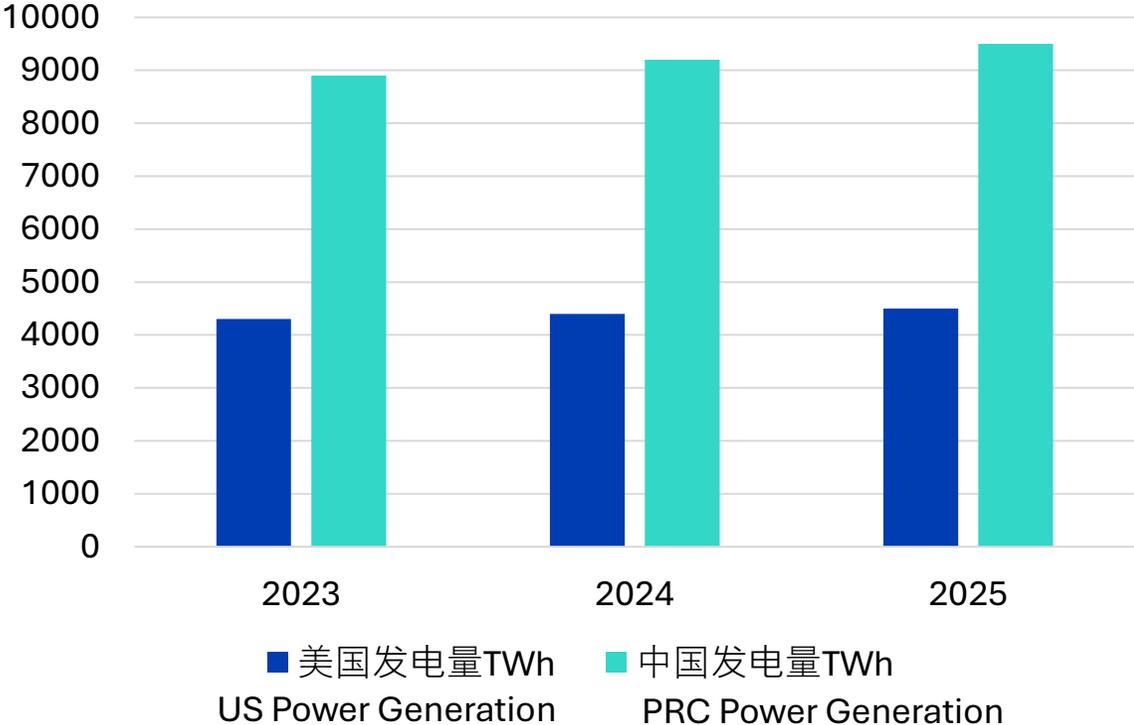
### Target

- 1MW rack
- Market tipping point expected by 2026

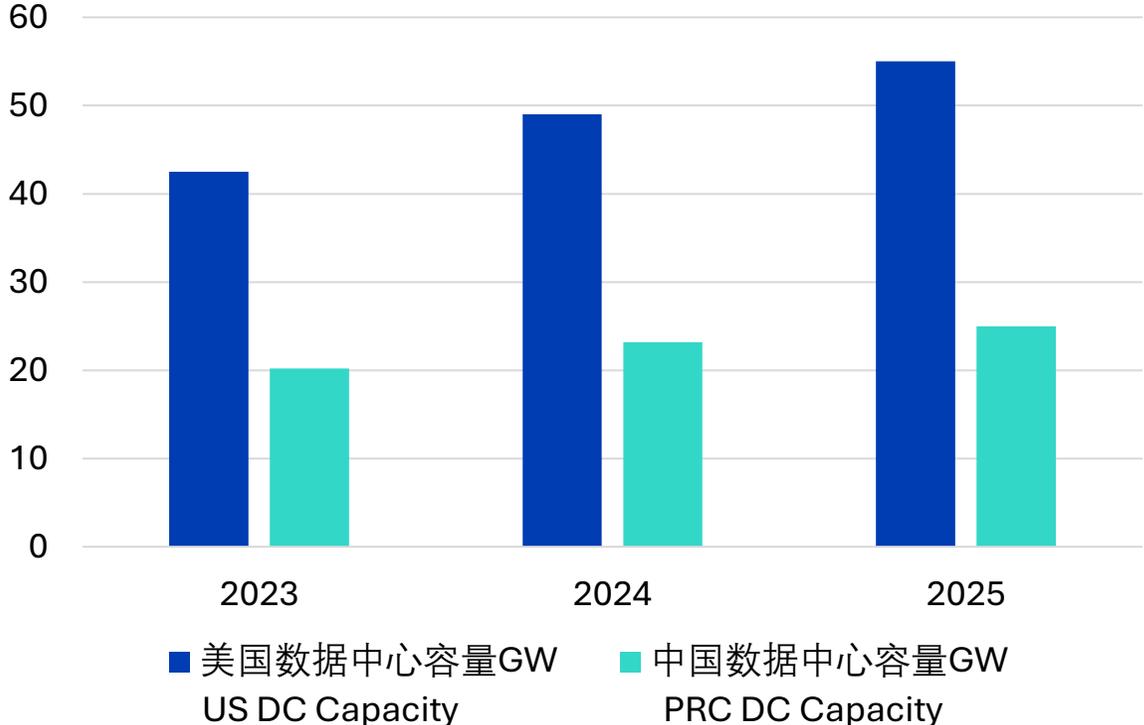


# Infra-level DC power in China and the United States

China's electricity generation is about twice that of the United States



The U.S. data center capacity still far exceeds that of China and is growing at a higher rate



# Case: xAI switches to full natural gas power generation in the absence of grid supply



xAI Memphis data center site



Source: Google Maps



Source: VNZoom

# Power architecture within China and US data centers



## Power Architecture in US Data Centers:

- The power density of single cabinets from major cloud vendors is higher, with scale up as the main development direction
- The backup time of UPS is relatively short, about 5 minutes
- Backup batteries have long adopted lithium batteries, which have a high penetration rate
- The four major cloud vendors mostly use OCP architecture, among which Meta and Google use BBU in the cabinet for backup power, while AWS and Microsoft also have some involvement.



## Power Architecture in China Data Centers:

- Due to supply chain limitations, the power density of single cabinets in China is lower, with a focus on scale out and a development direction of scale up.
- UPS backup time is relatively long, about 15 minutes
- The backup solution is still mainly lead-acid, while the lithium solution started relatively late
- Cloud vendors have relatively few in cabinet DC power supply solutions such as OCP or ODCC, and there is a lack of unified open design solutions.
- The application of high-voltage direct current (HVDC) in data centers was the earliest and has a larger scale, which is conducive to subsequent upgrades.

# Trade network and electricity for computing power - China's solution

**Scale Out最大集群**  
The largest scale-out cluster  
单集群提供超 10万PFLOPS 算力  
100,000 PFLOPS compute per cluster

**Scale Up最强超节点**  
The strongest scale-up supernode  
算力规模达 300PFLOPS  
Up to 300 PFLOPS  
单卡推理吞吐 2300Tokens/s  
2,300 tokens/s throughput per card

**资源使用 最优配比**  
The smartest resource ratios  
训推一体 灵活调度  
Integrated training and inference with flexible scheduling

**CloudMatrix384 超节点**  
CloudMatrix384 supernode

- DC power supply design
- Divided into multiple cabinets, higher power consumption and floor space
- Lower density per cabinet, fully compatible with cold plate liquid cooling
- 2-layer switch network, fully interconnected lossless network based on optical fiber
- Scale-up + Scale-out
- Software and hardware collaboratively optimize large model training efficiency

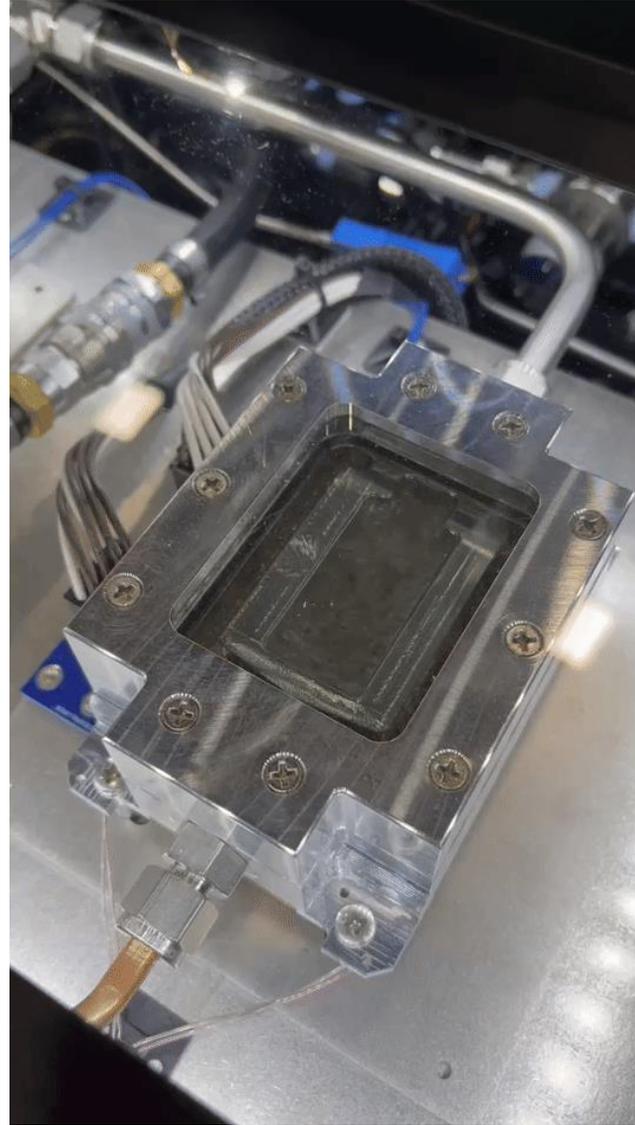
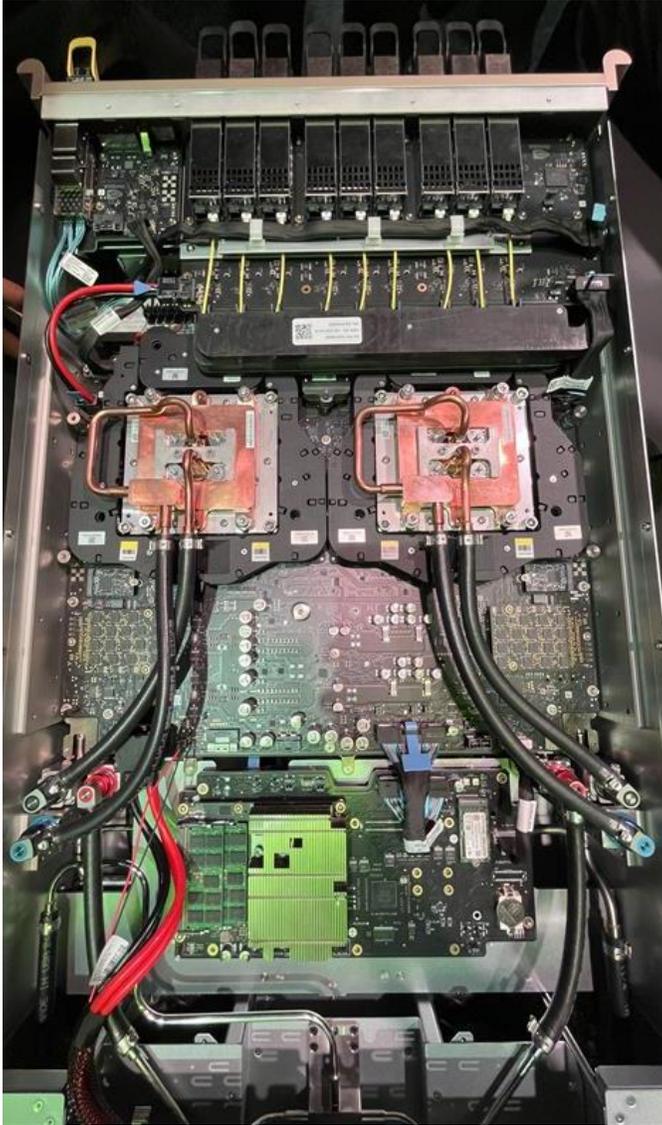
# Cooling the AI data center now and in 2030

**Jessica Nian**

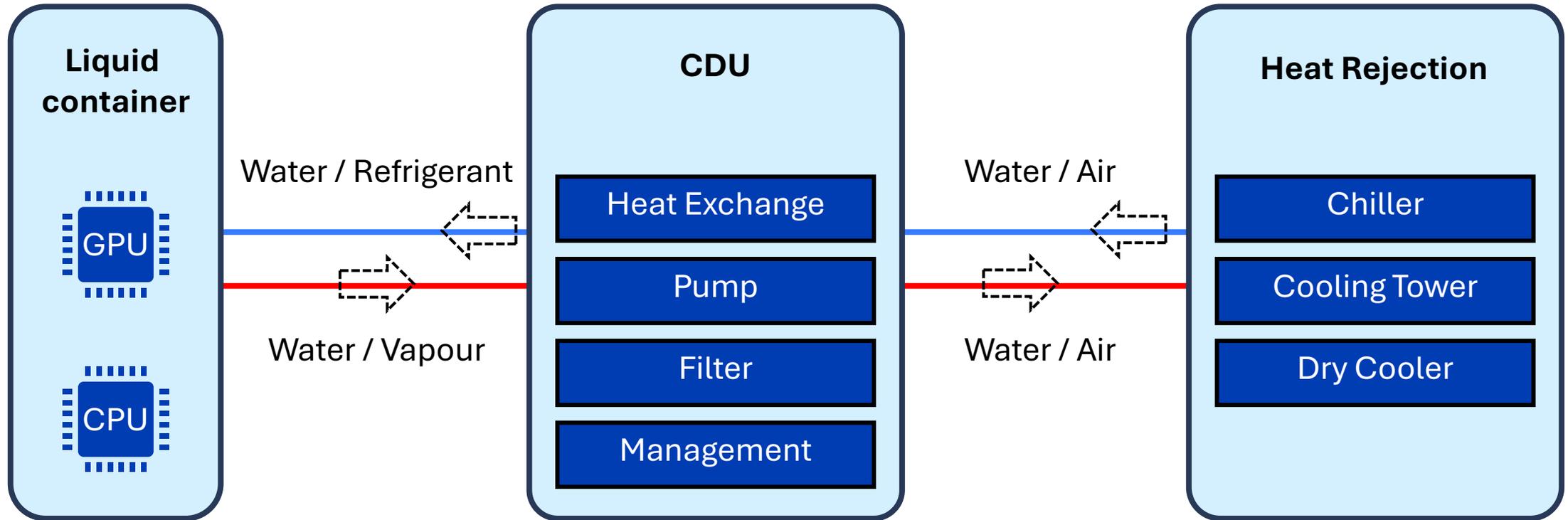
Senior analyst,

Data center physical infrastructure

Copyright © 2025 TechTarget, Inc. or its subsidiaries. All rights reserved.



# Liquid cooling 101



# Content

Market drivers

Regional trends

Technical trends

System trends

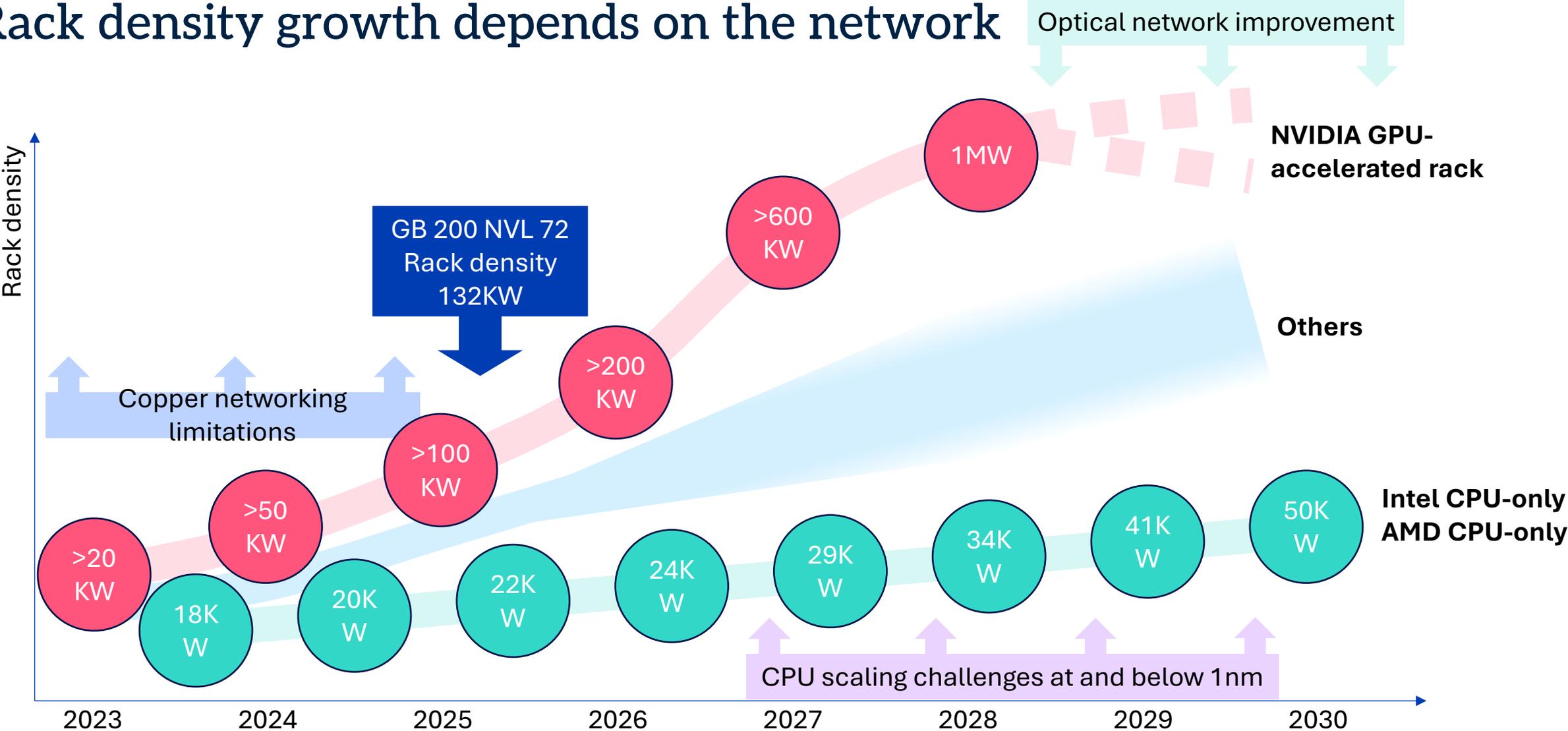
Case studies

The competition is heating up

Standards and ecosystems

# Market drivers

# Rack density growth depends on the network



# China and Southeast Asia PUE and WUE regulatory policies



# European PUE and WUE regulatory policies

California: Global Leader in Stringency:

**EU's Energy Efficiency Directive:** mandatory energy efficiency reporting system for data centers above 1MW;  
**Germany's Energy Efficiency Promotion and Energy Services Act Amendment** builds a phased PUE control system ( $\leq 1.5$  in 2027,  $\leq 1.3$  in 2030).

**China:** "East Data West Compute" PUE  $\leq 1.25$

**Singapore:** "Green Data Center Roadmap, PUE  $\leq 1.3$ ; WUE as a core part of its certification system

**Malaysia, Indonesia, Thailand, and Vietnam:** Lack mandatory PUE/WUE regulations.

**Middle East:** No mandatory regulations but extreme climate and corporate demands drive PUE  $\leq 1.5$  and near-zero WUE

# U.S. PUE and WUE regulatory policies

**California:** Global Leader in Stringency: PUE Requirement:  $\leq 1.2$  (Cold Zones) to  $\leq 1.3$  (Warm Zones) (Effective 2026)

**EU's Energy Efficiency Directive:** mandatory energy efficiency reporting system for data centers above 500kW;

**Germany's Energy Efficiency Promotion and Energy Services Act Amendment** builds a phased PUE control system ( $\leq 1.5$  in 2027,  $\leq 1.3$  in 2030).

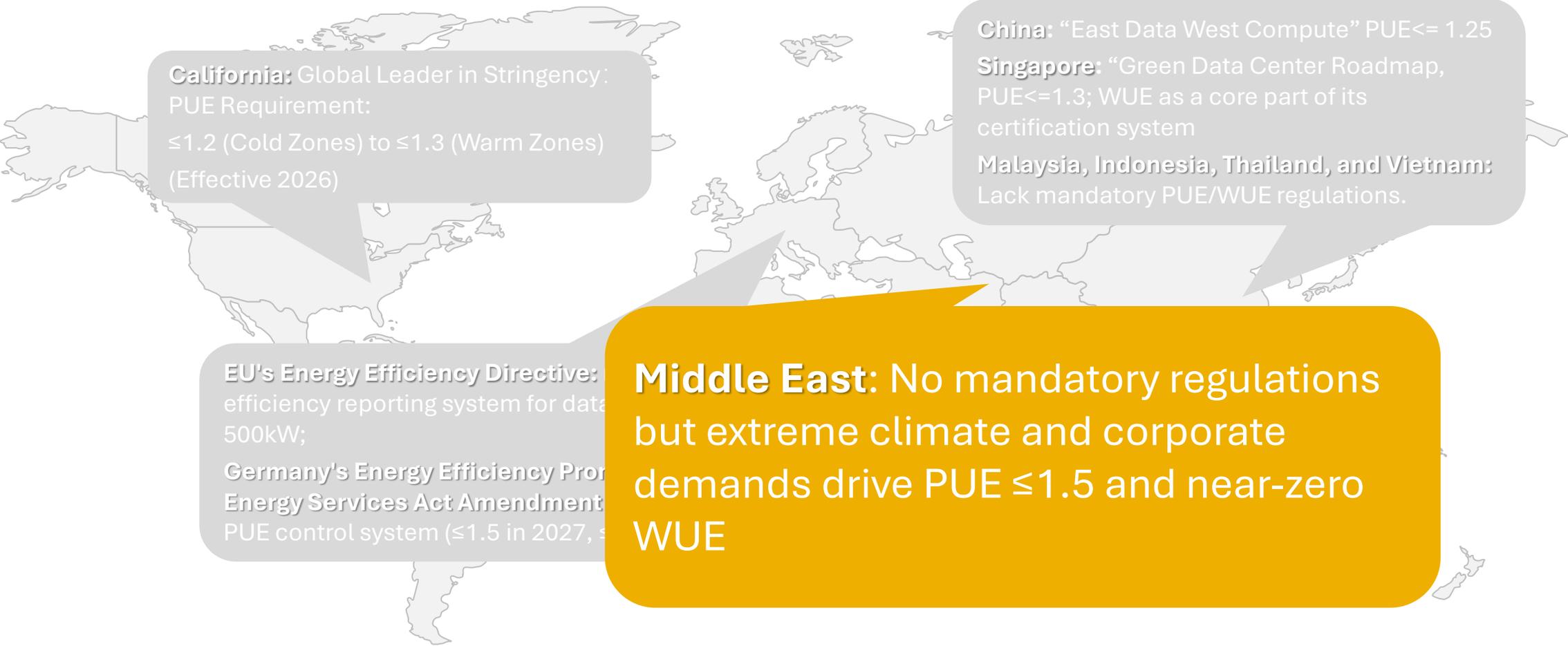
**China:** "East Data West Compute" PUE  $\leq 1.25$

**Singapore:** "Green Data Center Roadmap, PUE  $\leq 1.3$ ; WUE as a core part of its certification system

**Malaysia, Indonesia, Thailand, and Vietnam:** Lack mandatory PUE/WUE regulations.

**Middle East:** No mandatory regulations but extreme climate and corporate demands drive PUE  $\leq 1.5$  and near-zero WUE

# Middle East PUE and WUE regulatory policies



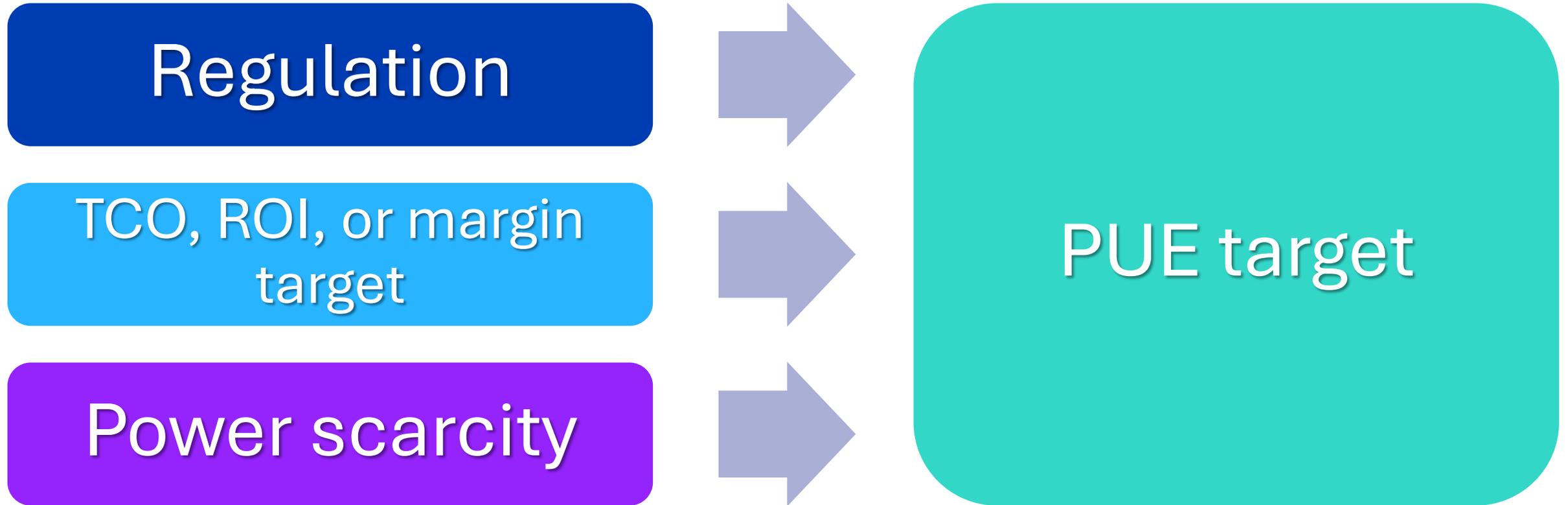
**California:** Global Leader in Stringency: PUE Requirement:  $\leq 1.2$  (Cold Zones) to  $\leq 1.3$  (Warm Zones) (Effective 2026)

**EU's Energy Efficiency Directive:** efficiency reporting system for data 500kW;  
**Germany's Energy Efficiency Promotion Act and Energy Services Act Amendment:** PUE control system ( $\leq 1.5$  in 2027,  $\leq 1.3$  in 2030)

**Middle East:** No mandatory regulations but extreme climate and corporate demands drive PUE  $\leq 1.5$  and near-zero WUE

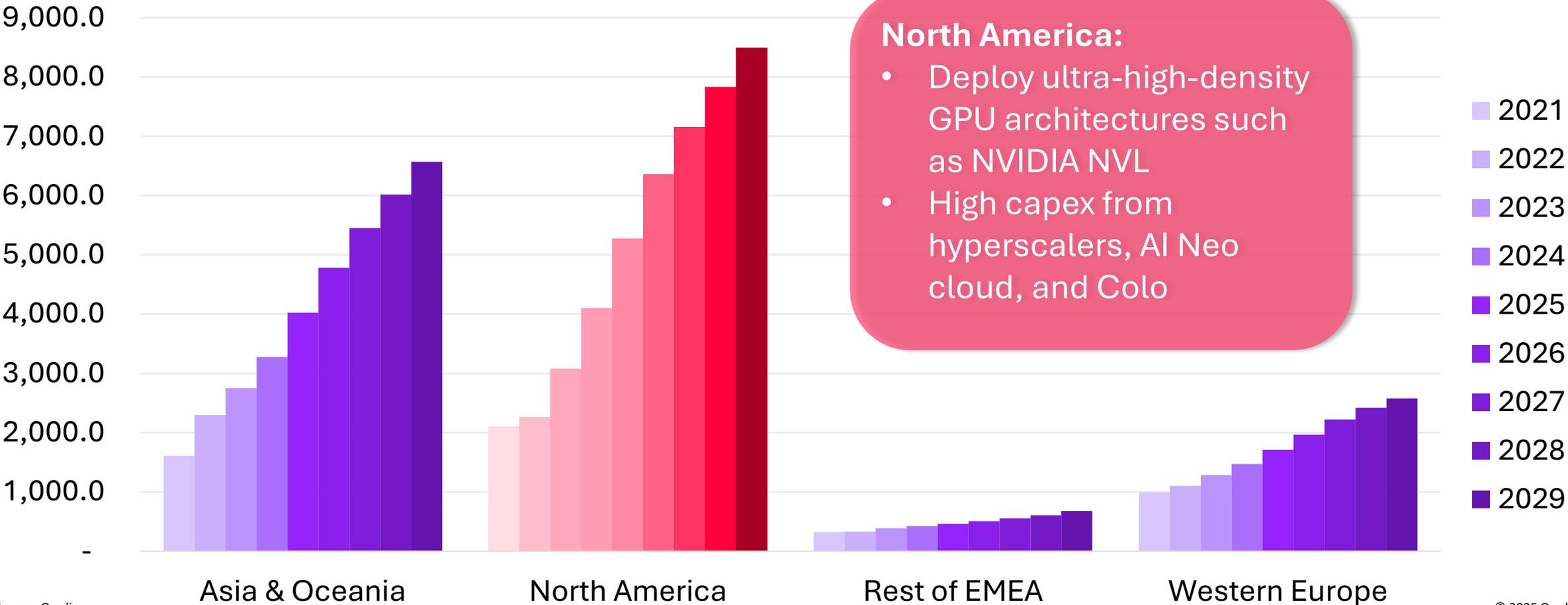
**China:** "East Data West Compute" PUE  $\leq 1.25$   
**Singapore:** "Green Data Center Roadmap, PUE  $\leq 1.3$ ; WUE as a core part of its certification system  
**Malaysia, Indonesia, Thailand, and Vietnam:** Lack mandatory PUE/WUE regulations.

# PUE is more than a regulatory requirement



# Regional trends

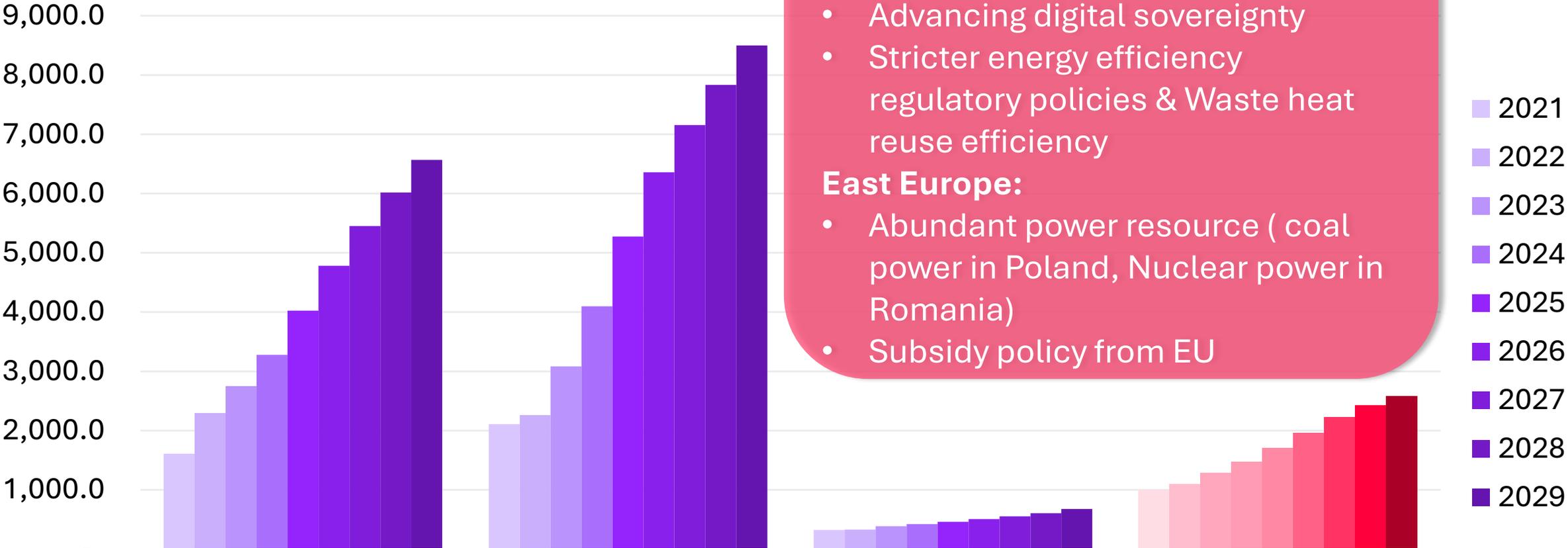
# The liquid cooling market in North America



Source: Omdia

© 2025 Omdia

# The liquid cooling market in Europe



**Western Europe:**

- Significant investments in AI
- Advancing digital sovereignty
- Stricter energy efficiency regulatory policies & Waste heat reuse efficiency

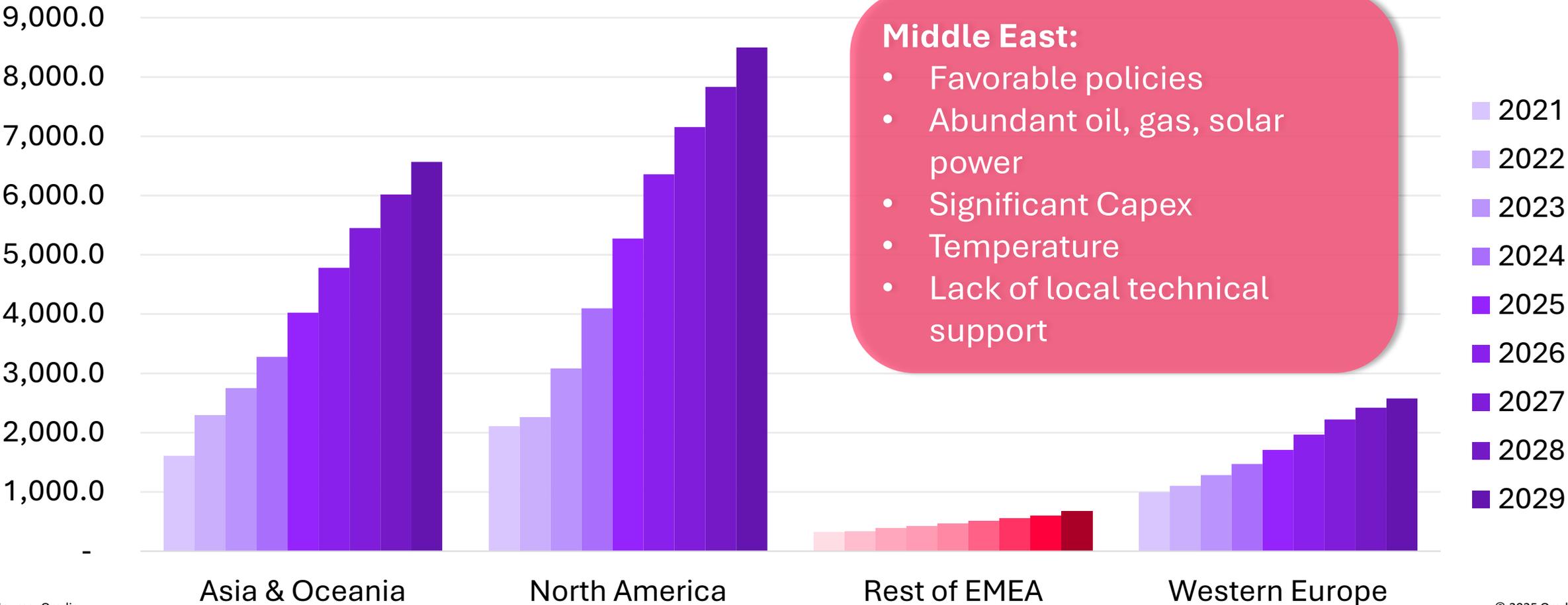
**East Europe:**

- Abundant power resource ( coal power in Poland, Nuclear power in Romania)
- Subsidy policy from EU

Source: Omdia

© 2025 Omdia

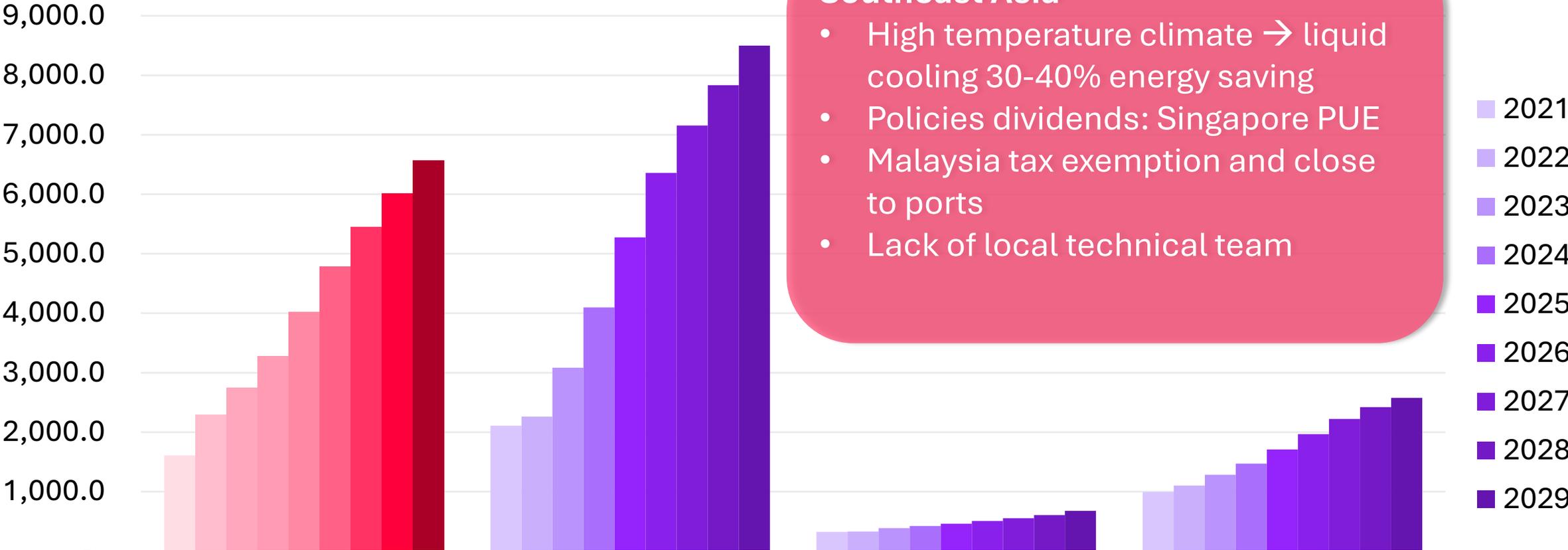
# The liquid cooling market in Middle East



Source: Omdia

© 2025 Omdia

# The liquid cooling market in Asia & Oceania



**Southeast Asia**

- High temperature climate → liquid cooling 30-40% energy saving
- Policies dividends: Singapore PUE
- Malaysia tax exemption and close to ports
- Lack of local technical team

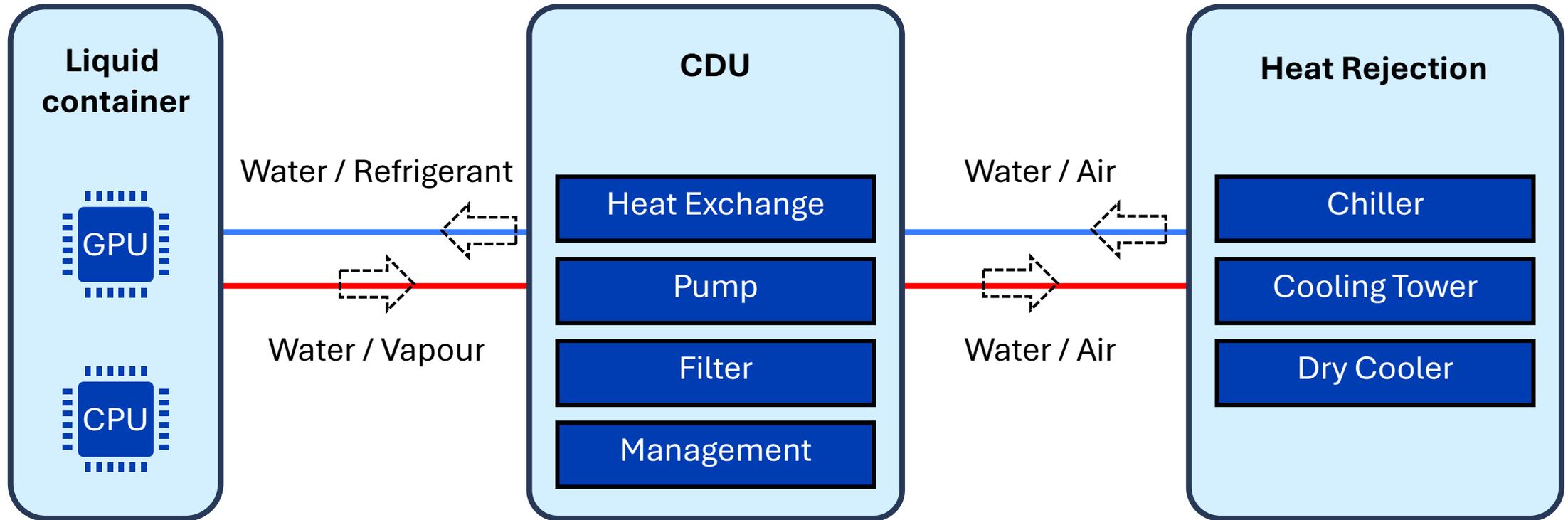
Source: Omdia

© 2025 Omdia

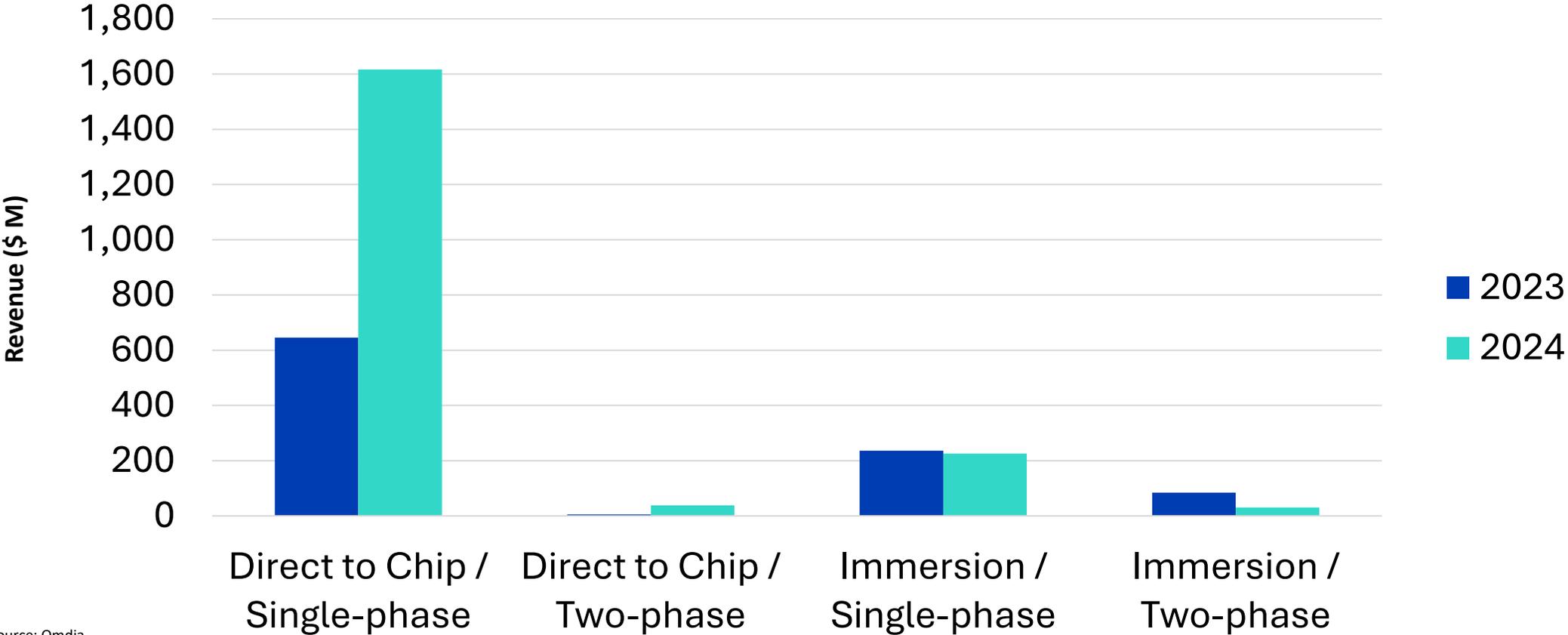


# Technical trends

# Liquid cooling 101



# DTC/1P dominates the liquid cooling market in next five years



Source: Omdia

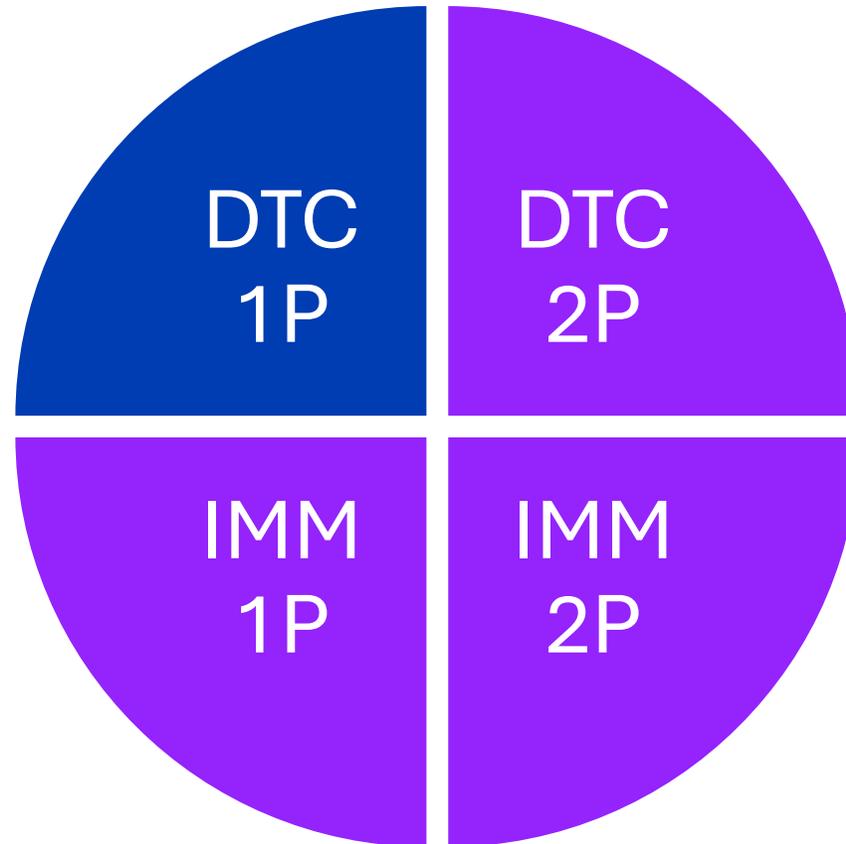
© 2025 Omdia

# DTC 1P

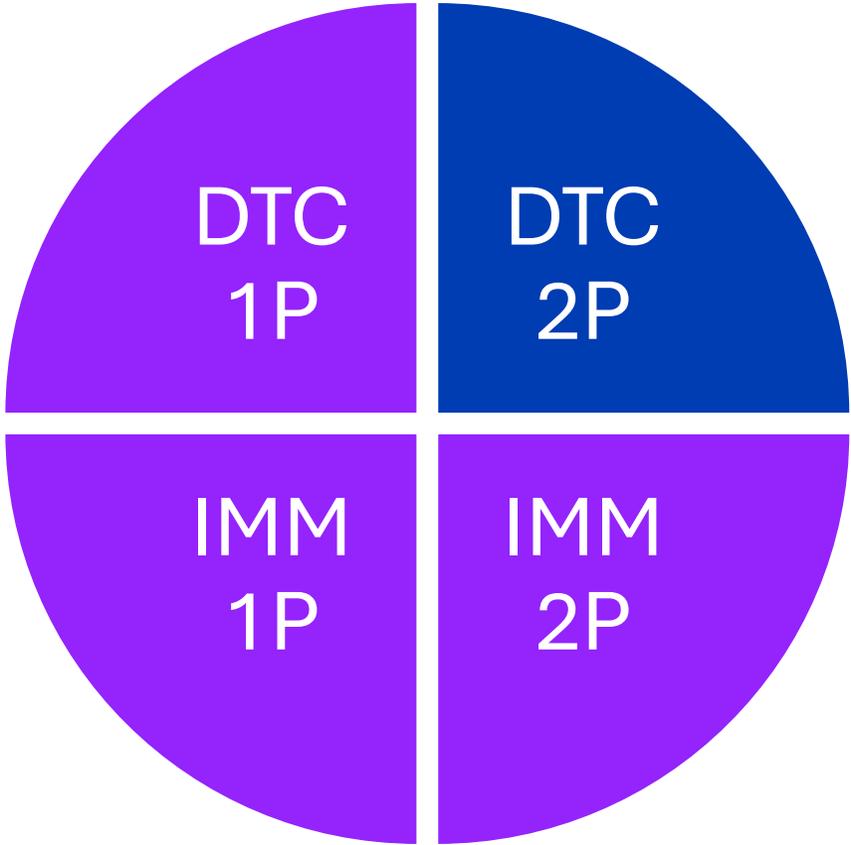
**Advantages:**

- Tech Maturity
- Low Renovation Cost

**Challenges:** TDP increase →  
Flow speed → Leaking Risk



# DTC 2P



**Advantages:**

- Heat Dissipation
- Efficiency

**Challenges:**

- Stability of Coolant
- Press Control Technology
- Maintenance System

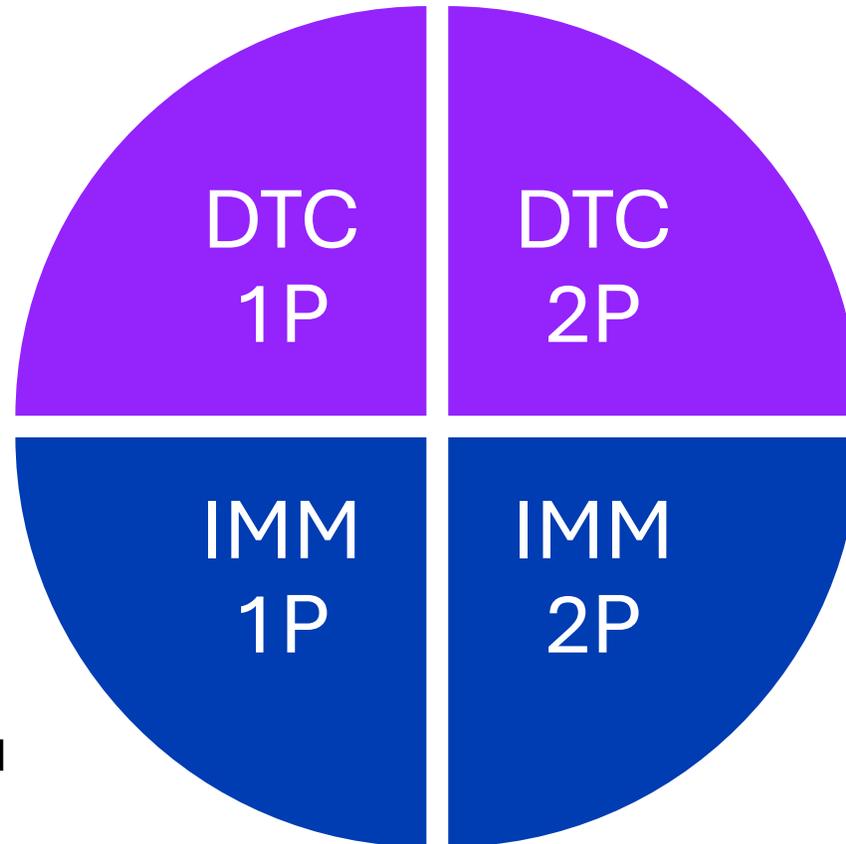
# Immersion

## Advantages:

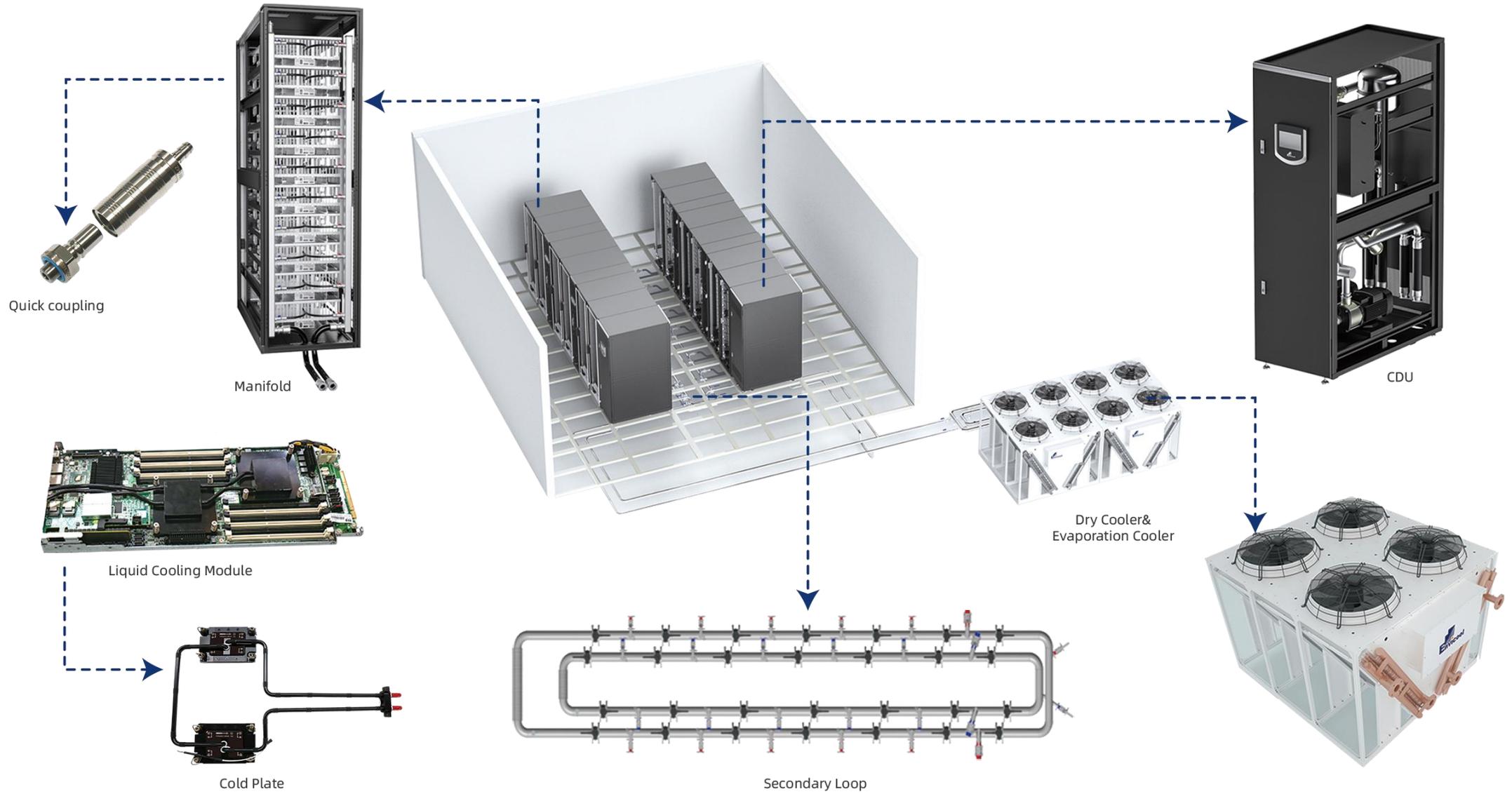
- Heat Dissipation Performance
- WUE

## Challenges:

- Renovation of Servers
- Maintenance & Operation System
- Long – term Stability
- NVIDIA no Warranty
- Compatibility of Coolant and Hardware

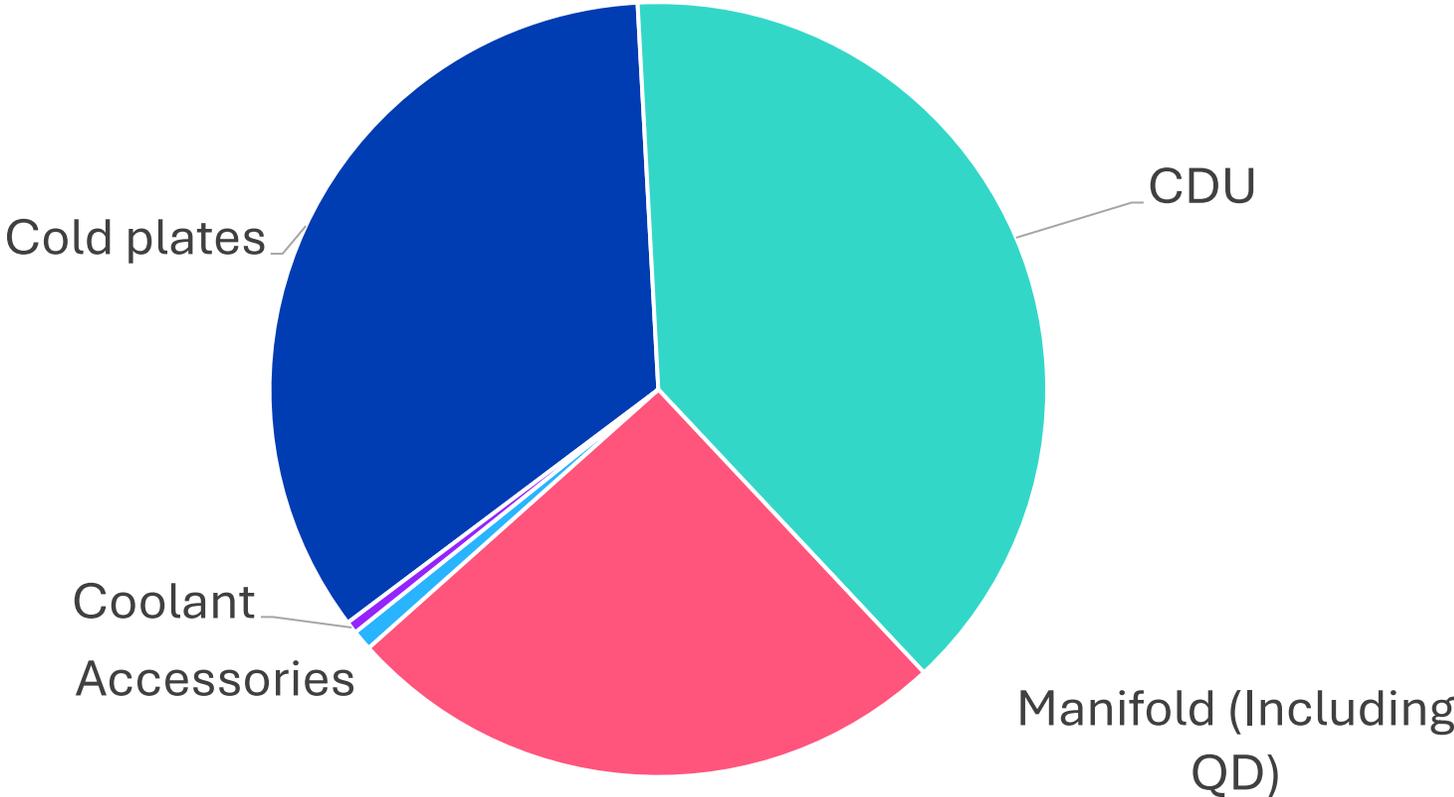


# System trends



# Cost dollar per watt of cold plates is estimated to surpass that of CDU

The DTC liquid cooling market by product



Source: Omdia

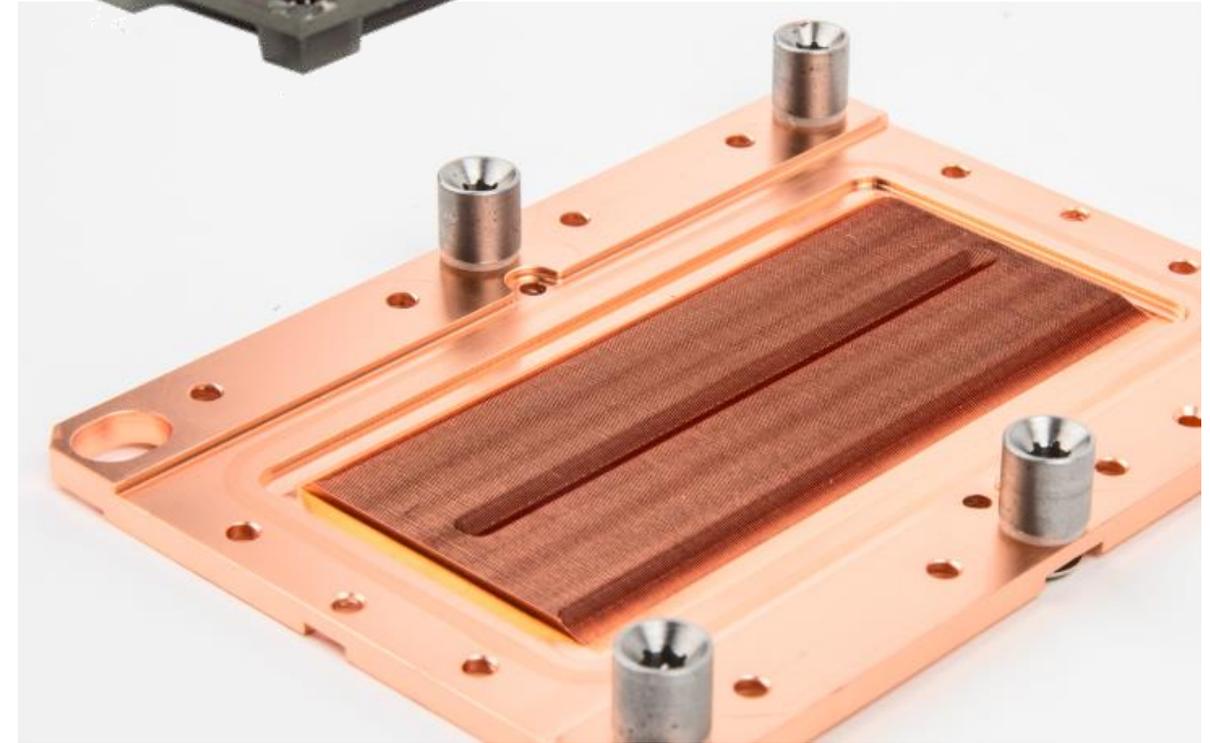
© 2025 Omdia

# Cold Plates Trend

- **Cold plates:**

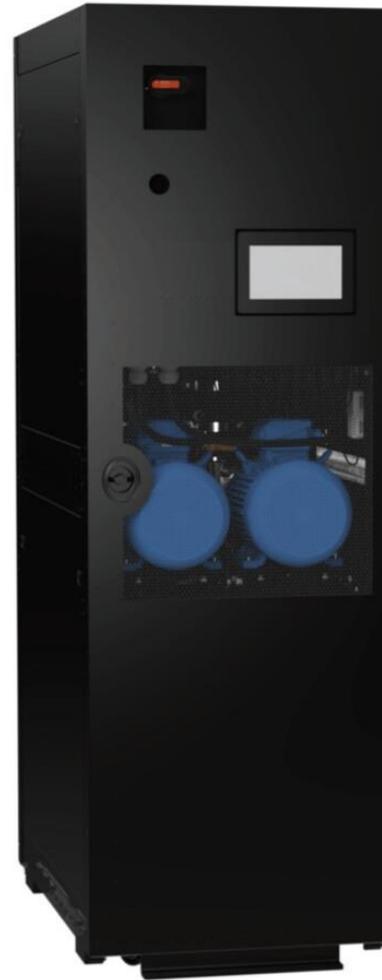
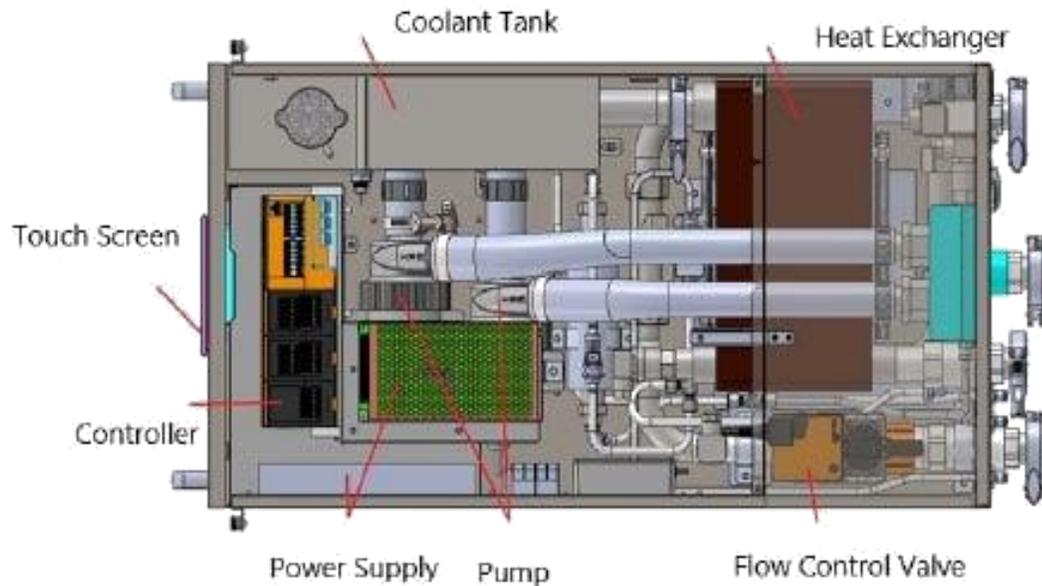
- Based on all-copper cold plate and exploring composite material
- Heat dissipation coverage
- Micro-channel design optimization

Examples: Boyd 4000W, CoolIT 4000w



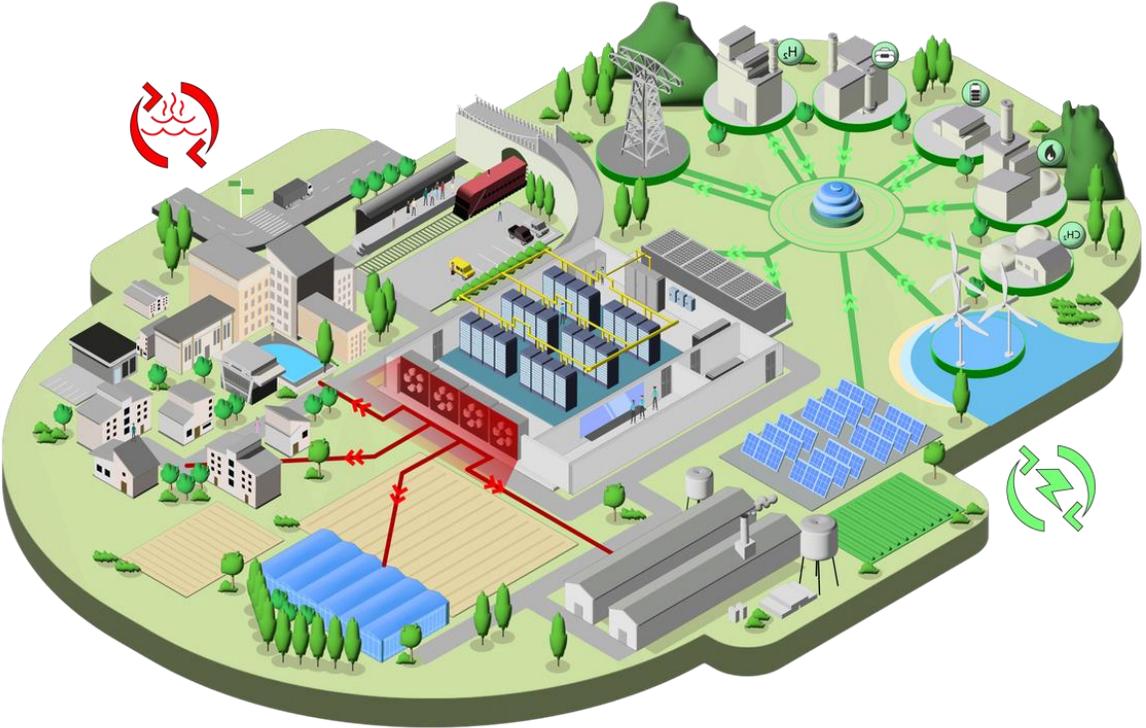
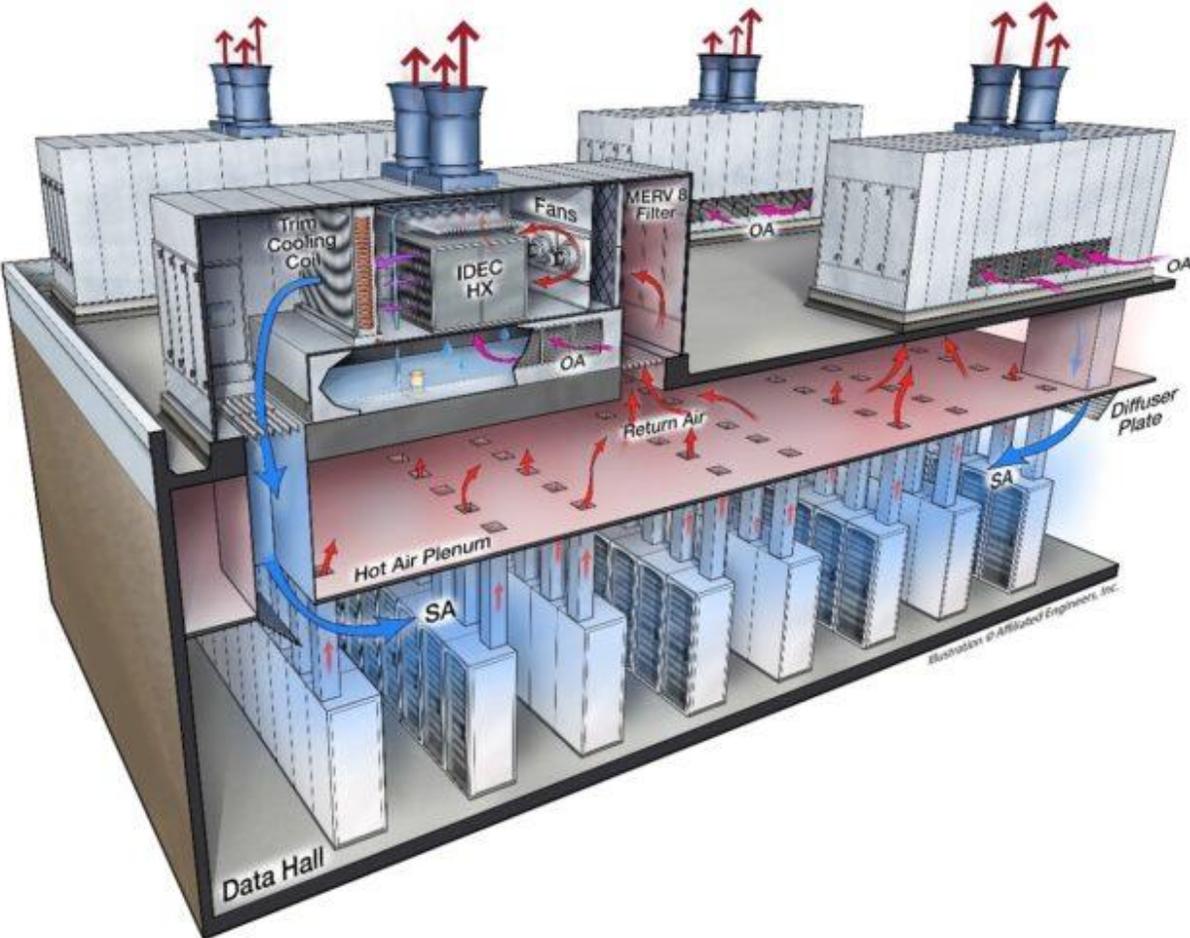
# The Trends of CDU

- **CDU:**
    - In-rack CDU to in-row CDU to gray space facility CDU
    - Liquid to air CDU to Liquid-to-liquid CDU
- Examples: Vertiv 2MW CDU, CoolIT 2 MW



Liquid cooling is also about air...

...and heat reuse



# Case studies

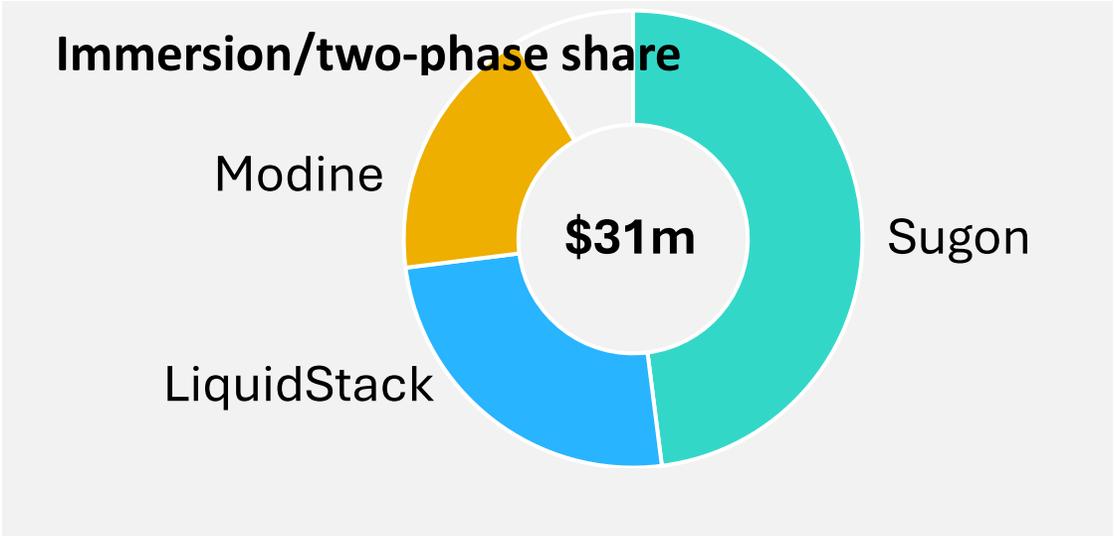
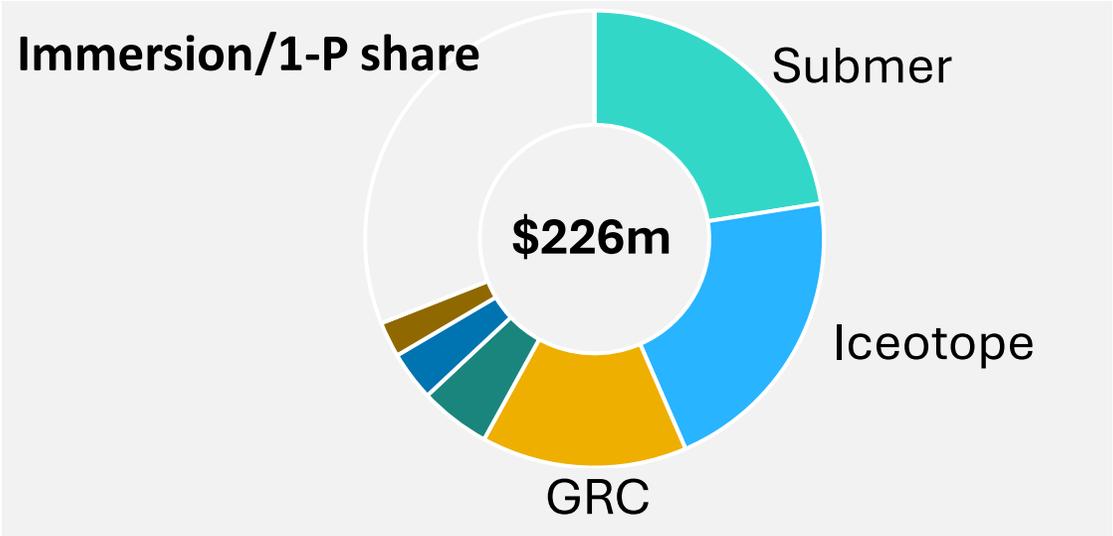
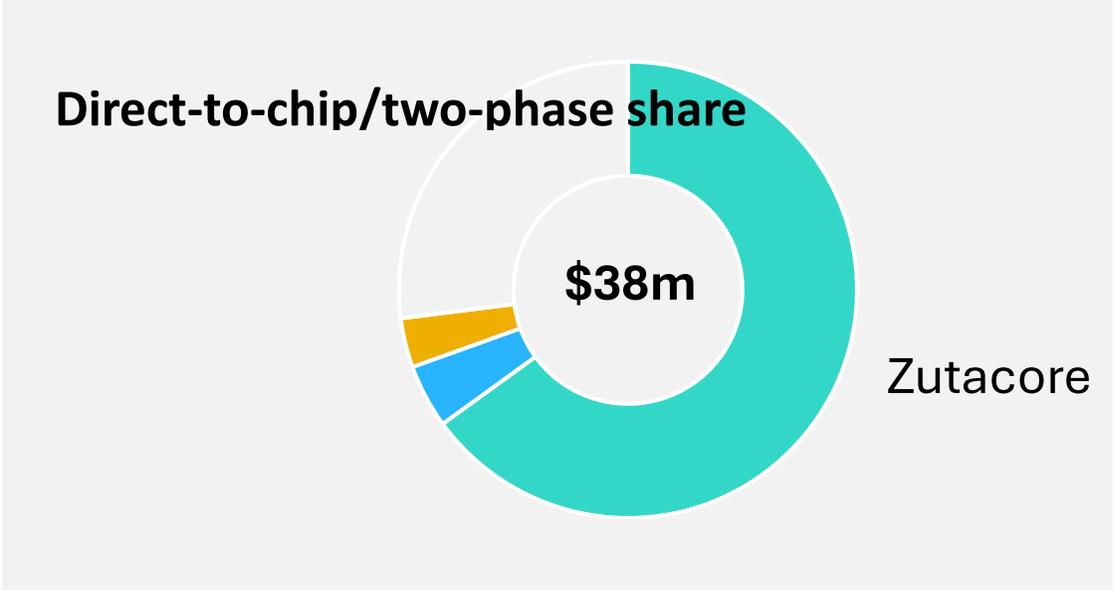
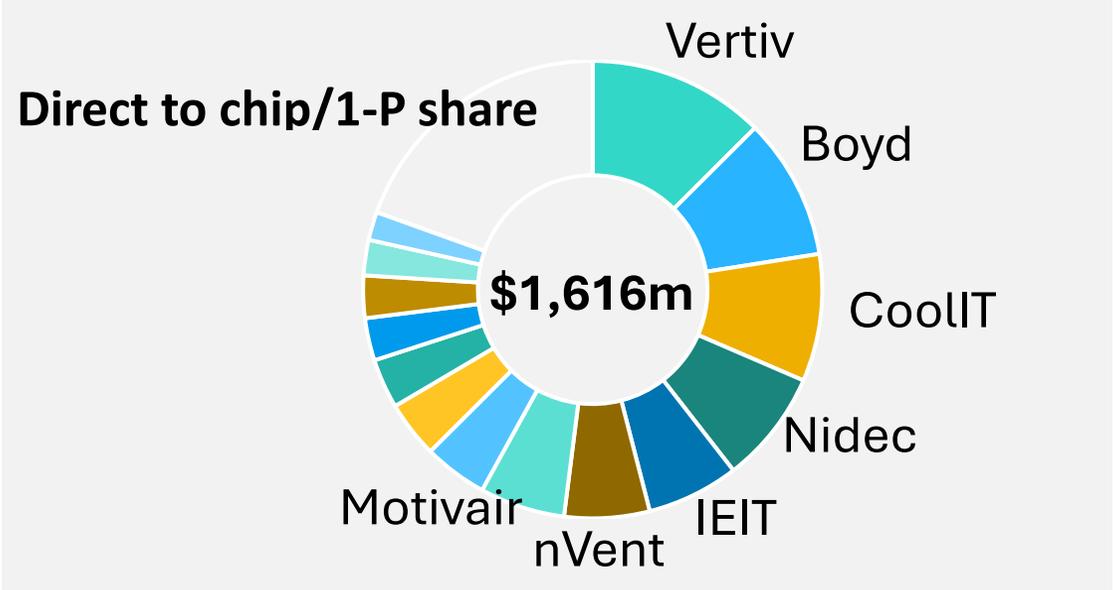
# Cooling solution of Datacenters





# The competition is heating up

# The market share for liquid cooling technology by products

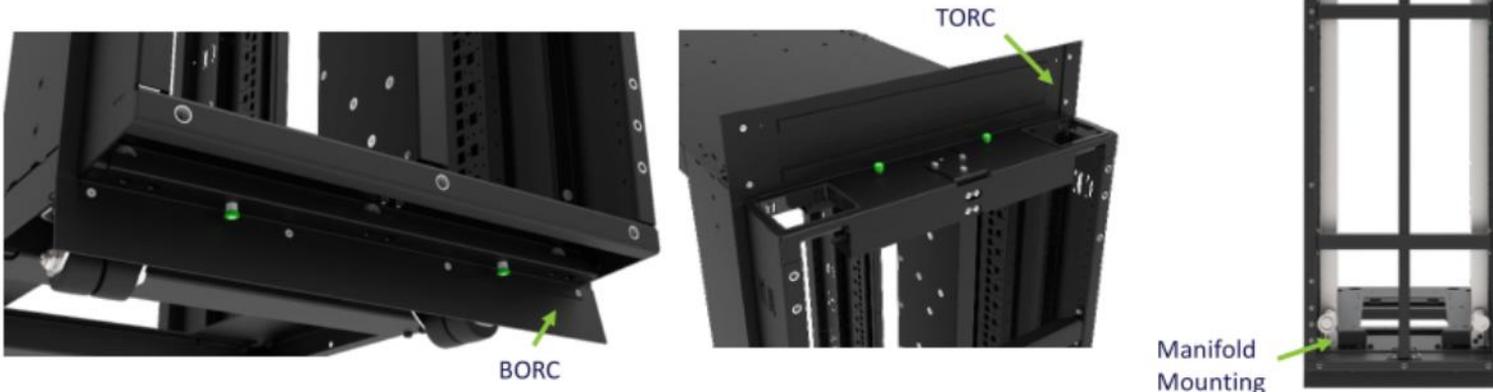


# Standards and ecosystem

# Now and Future: OCP standard ecosystem

## ORv3 HPR Rack

- New ORv3 HPR-specific TORC/BORC (Top/Bottom of Rack Containment) kits to seal off airflow between hot and cold aisle in data center.
- Same blind mate manifold mounting locations as ORv3, but extension frame now provides protection for potential deeper manifolds



- The OCP is becoming standard leader in the liquid cooling field
- The reduce of adaptation costs of all links (cold plate interfaces, rack-level integration and system-level collaboration)
- The cost of liquid-cooled component adaptation was reduced by 29%
- "Cooling Environment" project

# Contact us if you have any questions

## **Vlad Galabov**

Senior Director, Enterprise Infrastructure Research  
vgalabov@omdia.com

WeChat



LinkedIn



## **Shen Wang**

Practice Lead, Data center capacity and infrastructure  
shen.wang@informa.com

## **Manoj Sukumaran**

Practice Lead, Data Center IT  
manoj.sukumaran@omdia.com

## **Jessica Nian**

Senior analyst, Data center physical infrastructure  
jessica.nian@omdia.com