VERTIV.

How to get AI ready? Evolution of Data Center

Tomislav Saric

Sales Director Indirect Business, CEE

NVIDIA Partner

VERTIV.

since 1965

Long history of pioneering data center infrastructure technologies for mission critical applications.

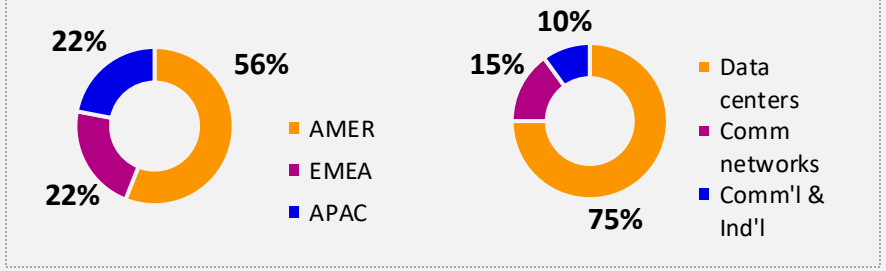
Today a global organization listed on the NY Stock Exchange with a team of 31,000+ people and revenues of $8bn (2024) from projects in 130+ countries.

Specialist in critical digital infrastructure with the most complete portfolio in power and cooling solutions wrapped in software platform and unmatched services offering.

Strong position in HPC/AI with technology partnerships and collaboration with industry leaders.

Sales ~$8.0B

Employees ~31,000



Customers include:
Alibaba, Alstom, America Movil, AT&T, China Mobile, Equinix, Ericsson, Reliance, Siemens, Telefonica, Tencent, Verizon, Vodafone

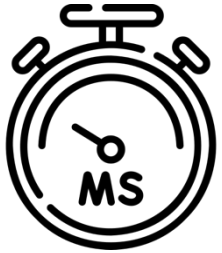
Manufacturing sites: 23

Service centers: 310+

Reach: 130+ countries

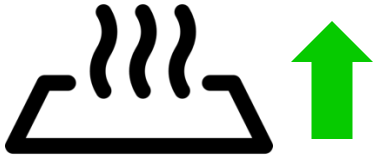
Note: All figures 2024FY.

IT Trends



Latency

Speed



Heat

Artificial Intelligence

Automation

HPC

Machine Learning

IoT(Internet of Things)

Media and Entertainment

Smart City

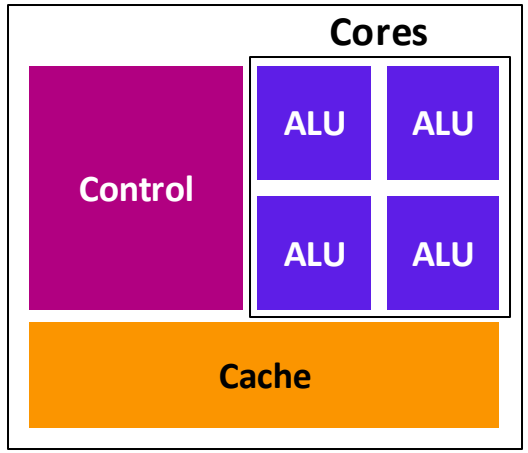
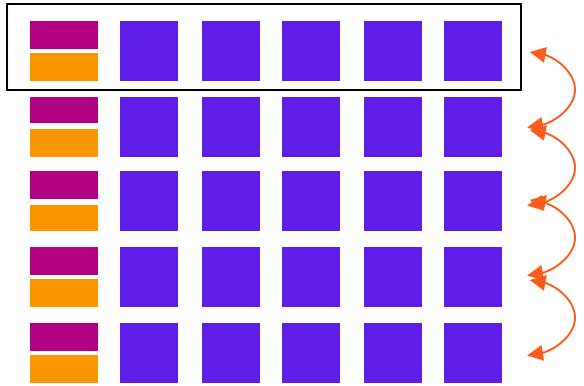
Healthcare

AI for Big Data

Finance-Trade automation Fraud detection

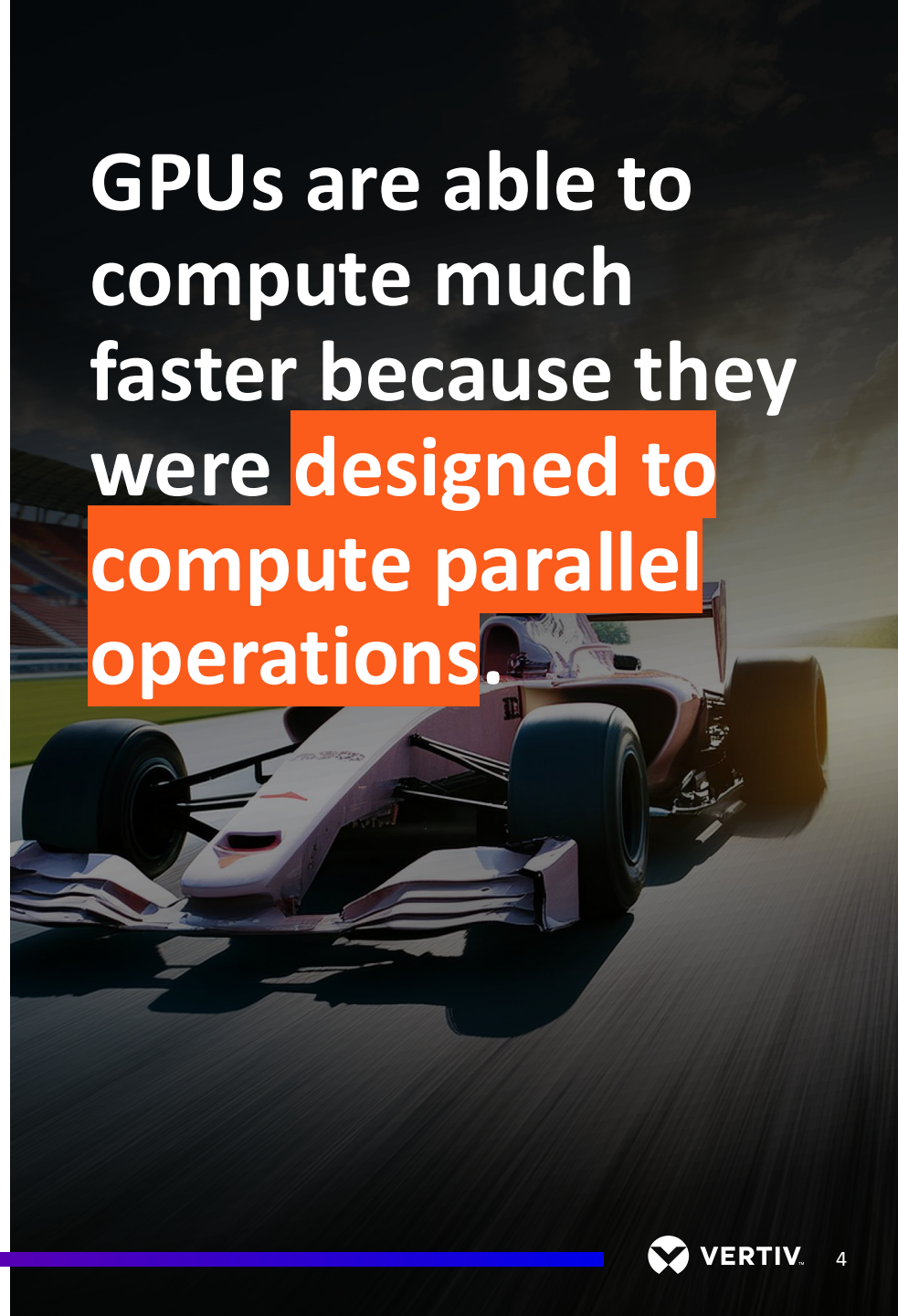
Manufacturing – simulation and testing

Data Infrastructure for Autonomous Vehicles

	CPUs	GPUs
Inside the chip		
Function	Generalized component that handles main processing functions of a server.	Specialized component that excels at parallel computing.
Processing	Designed for serial instruction processing.	Designed for parallel instruction processing.
Design	Fewer, more powerful cores.	100s × cores of CPUs, even if less powerful individually.
Best suited for	General purpose computing applications.	High-performance computing applications.

Control serves as the "coordinator" of the chip's hardware components.
Arithmetic Logic Unit (ALU) carries out the calculations (arithmetic and logical).
Cache holds frequently requested bits of data ready to be retrieved at high speed.

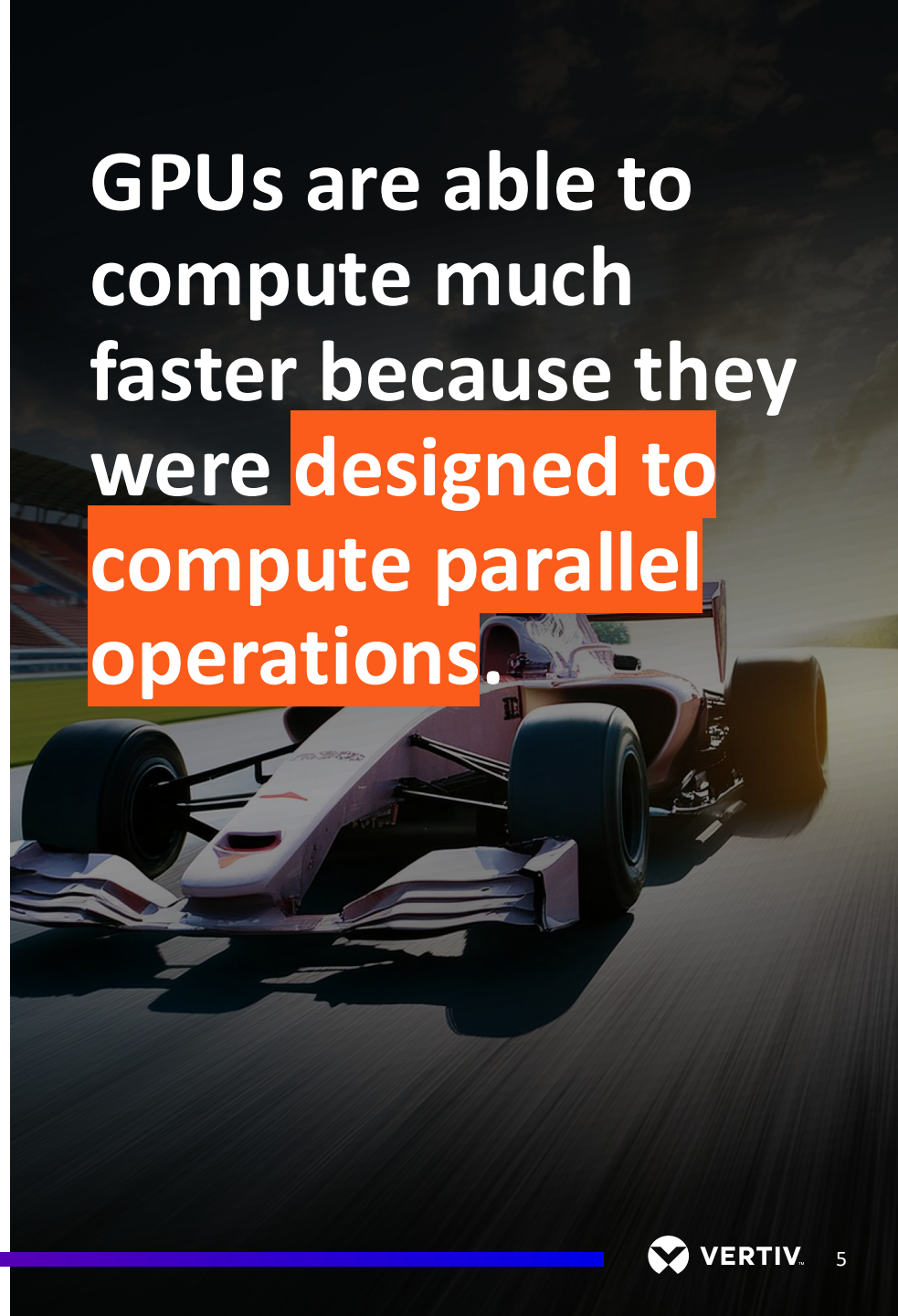
GPUs are able to compute much faster because they were designed to compute parallel operations.



Source: NVIDIA, AWS, IBM

	CPUs	GPUs
Scenario	Executing a single GPT-3 inference operation taking 350 TFLOP.	
Capability	*Today's processor core can typically execute 32 FP32 instructions per cycle.* *Processor clock rates have been stable around 3 GHz.*	*NVIDIA A100 GPU has 512 tensor cores that can perform a 4×4 matrix multiplication in a single cycle, with a nominal capacity of 312TFLOPs.*
Total processing time	$\frac{350\ \text{TFLOP}}{3\ \text{GHz} * 32\ \text{FLOP}} =$ ***~1 hour***	$\frac{350\ \text{TFLOP}}{312\ \text{TFLOPs}} =$ ***~1 second***

GPUs are able to compute much faster because they were designed to compute parallel operations.

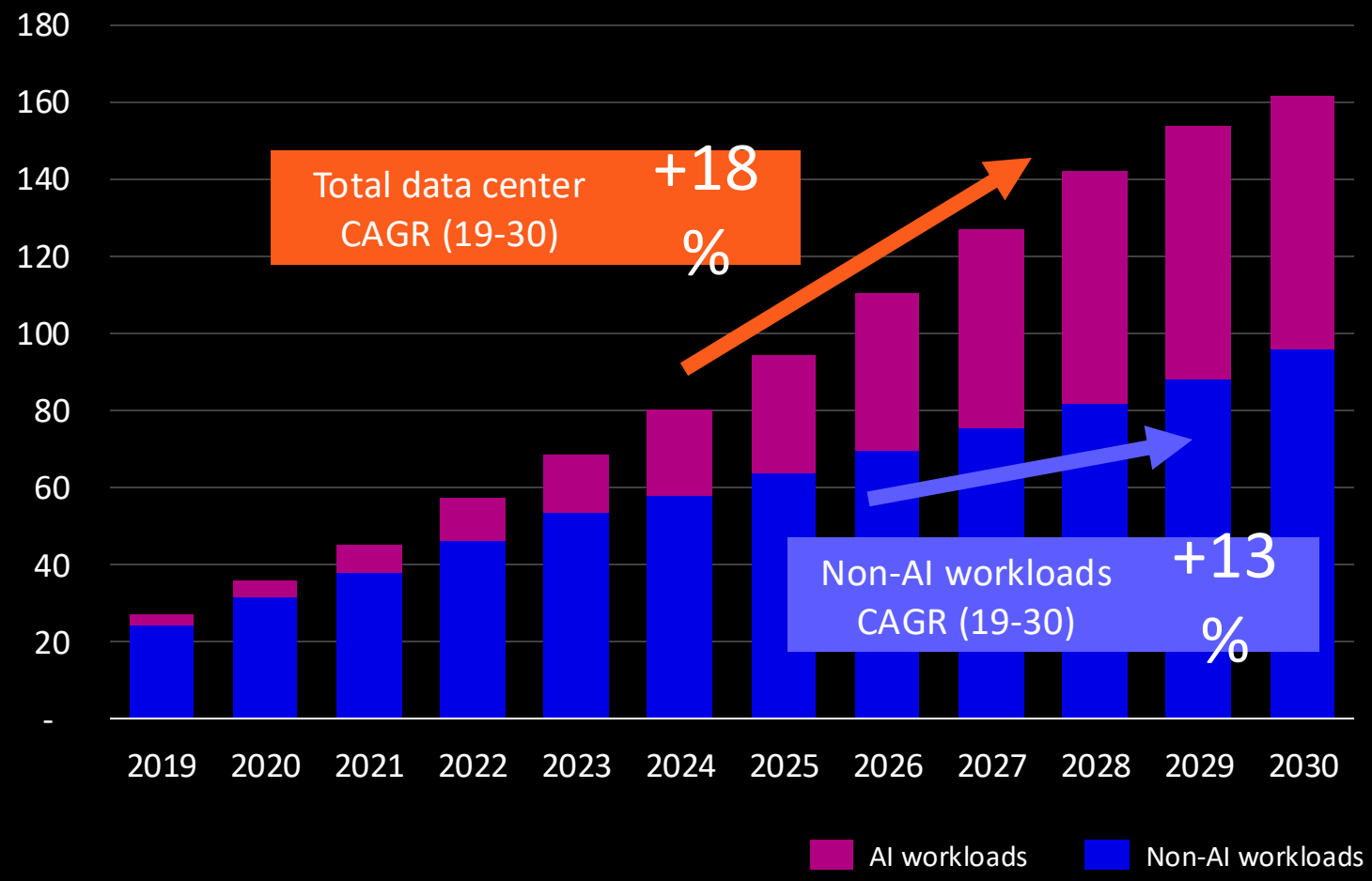


Source: Andreessen Horowitz

In the data center business, AI will accelerate capacity expansion.

AI workloads are incremental to conventional IT loads, not substitutive.

Total data center installed capacity (GW)

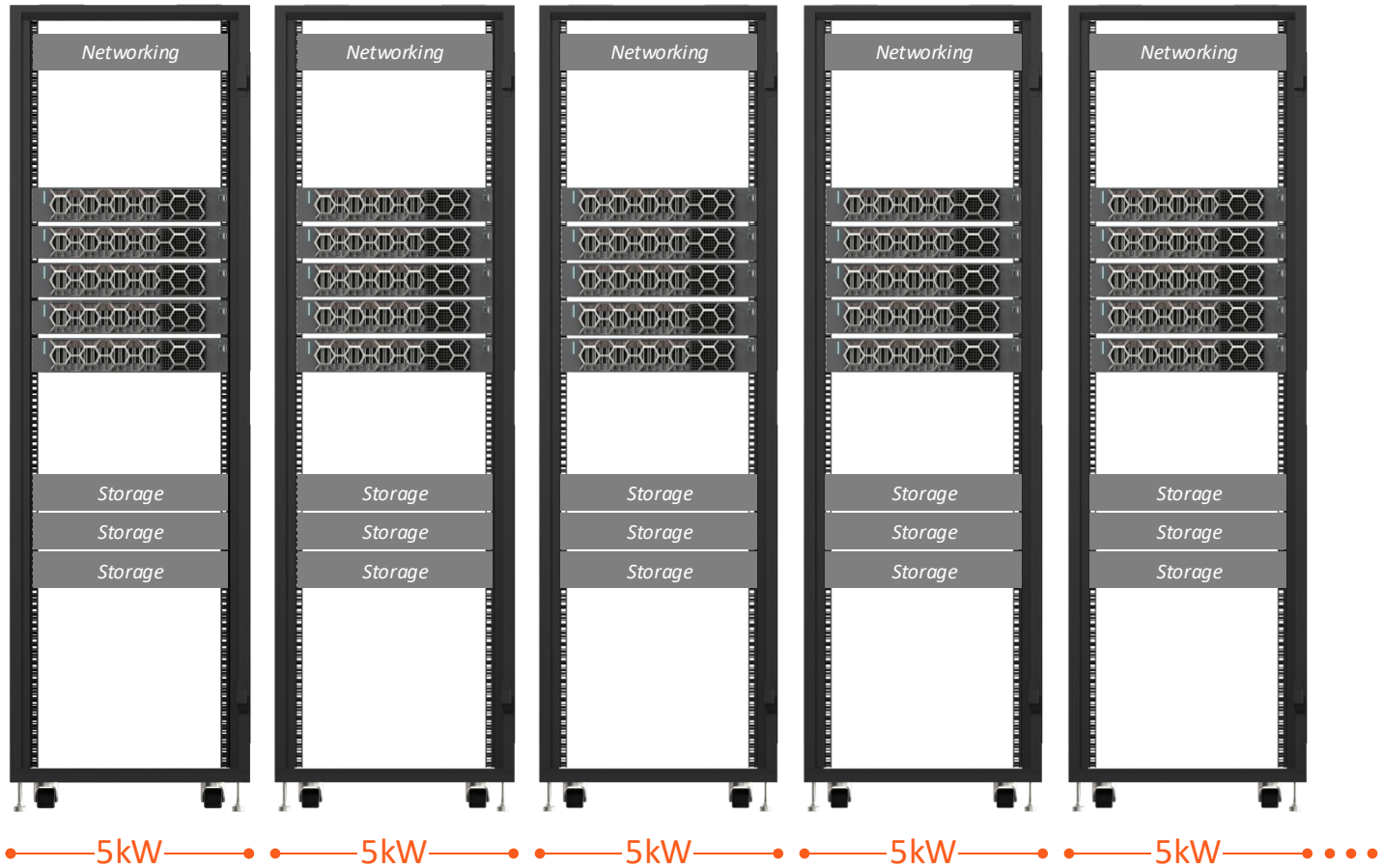


Source: Omdia

50kW

workload?

Conventional IT solution.



Traditional workloads were able to keep densities low by splitting loads across multiple racks.

50kW

workload?

AI solution.

AI workloads are driving rack density up: servers are drawing more power per rack unit, and they need to be closer together to minimize network latency and costs.

Traditional workloads were able to keep densities low by splitting loads across multiple racks.

50kW

workload?

Air-cooled.

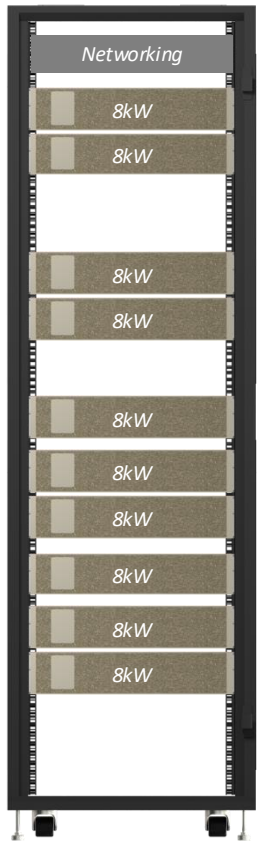
AI workloads are driving rack density up: servers are drawing more power per rack unit, and they need to be closer together to minimize network latency and costs.

Traditional workloads were able to keep densities low by splitting loads across multiple racks.

80kW

workload?

Liquid-cooled.



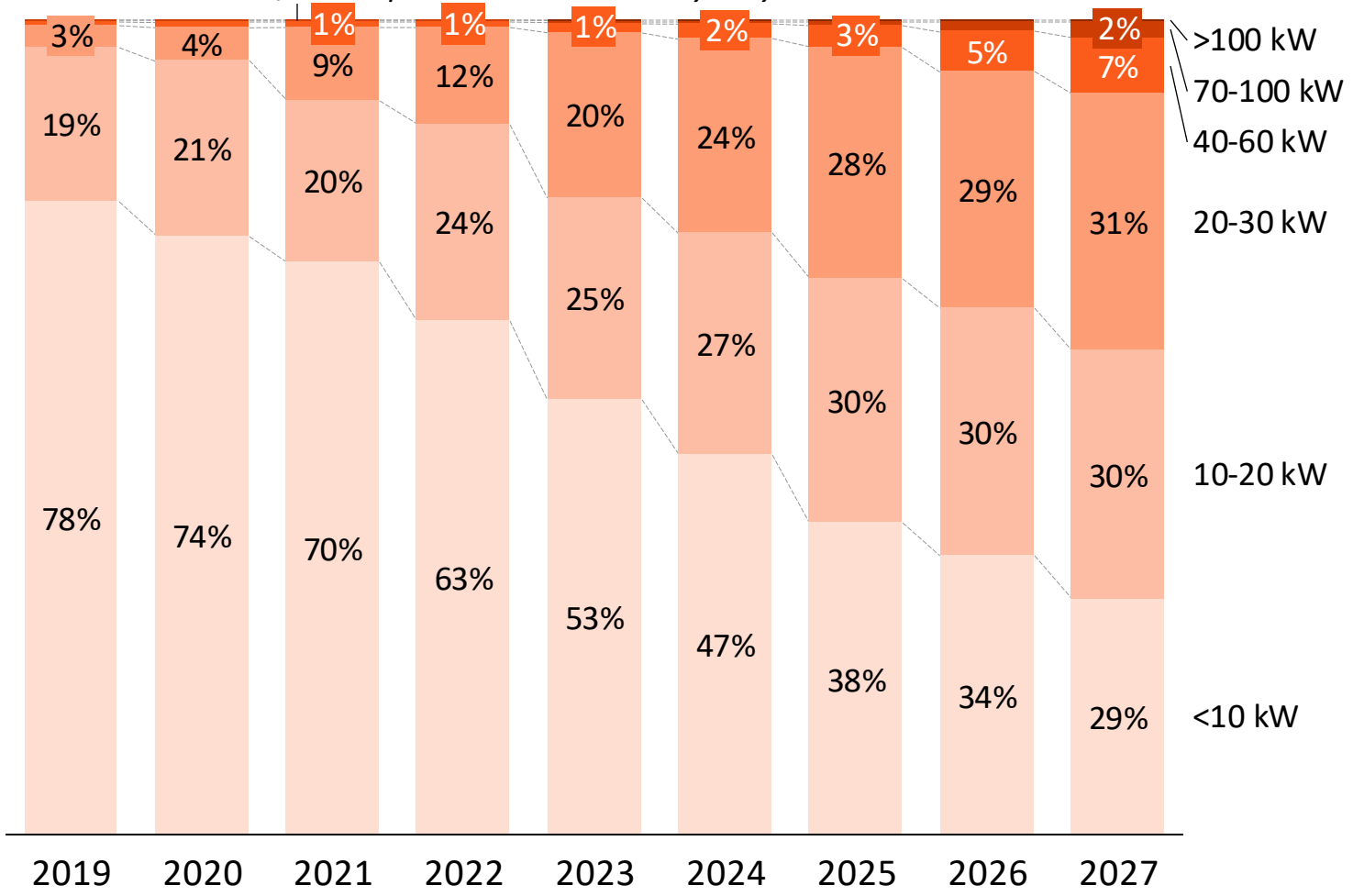
Liquid cooling is enabling further densification with high-powered servers taking less U-space.

AI workloads are driving rack density up: servers are drawing more power per rack unit, and they need to be closer together to minimize network latency and costs.

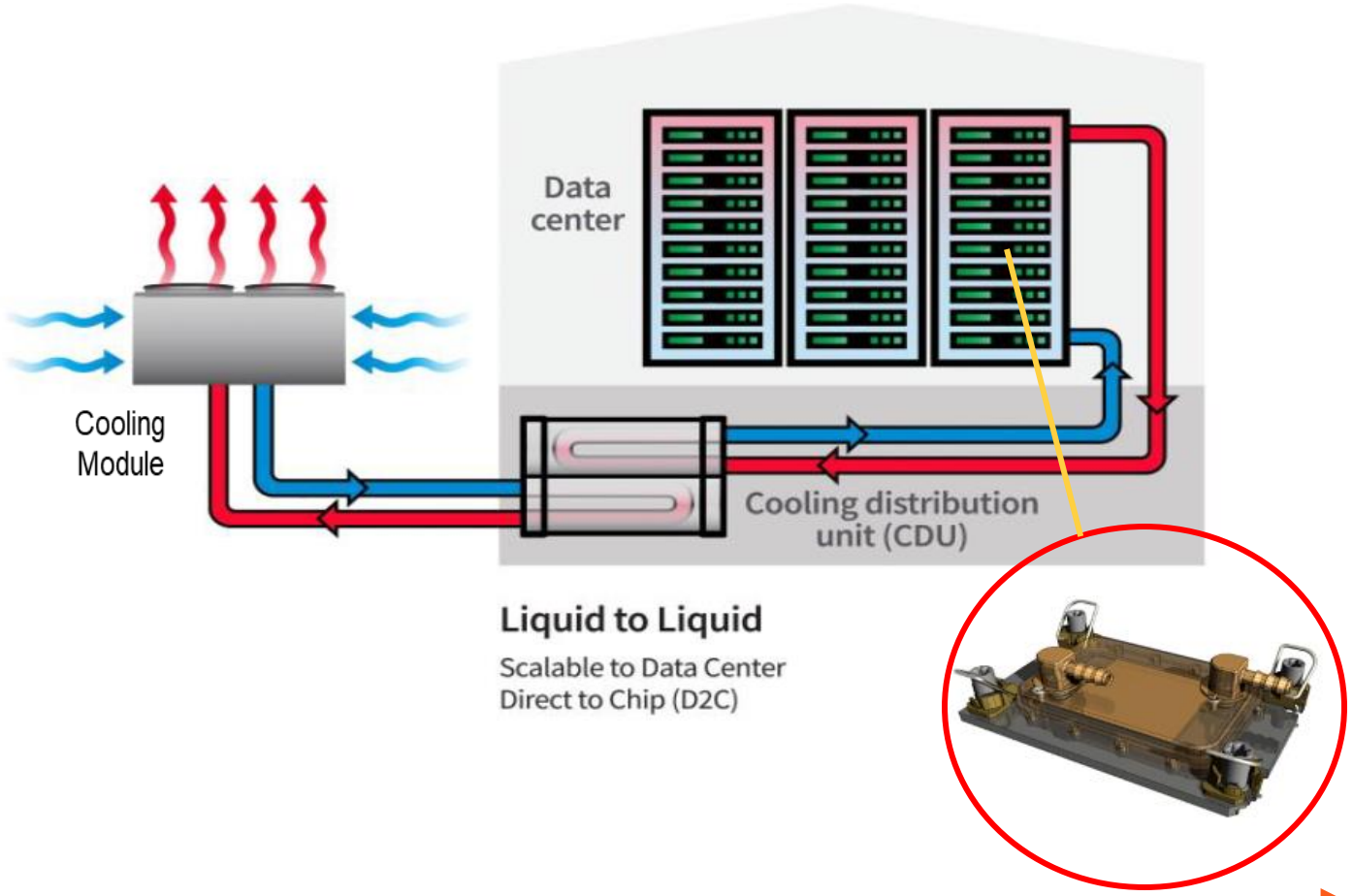
As AI expands, rack density is on the rise with low density racks becoming the exception rather than the rule.

Number of racks installed by rack density.

From 2021, more than a thousand racks over 100kW/rack expected to be installed yearly.



Source: Omdia



Coolant gets closer to the IT hardware

Liquid Cooling Direct To Chip Technologies

Single phase **or**

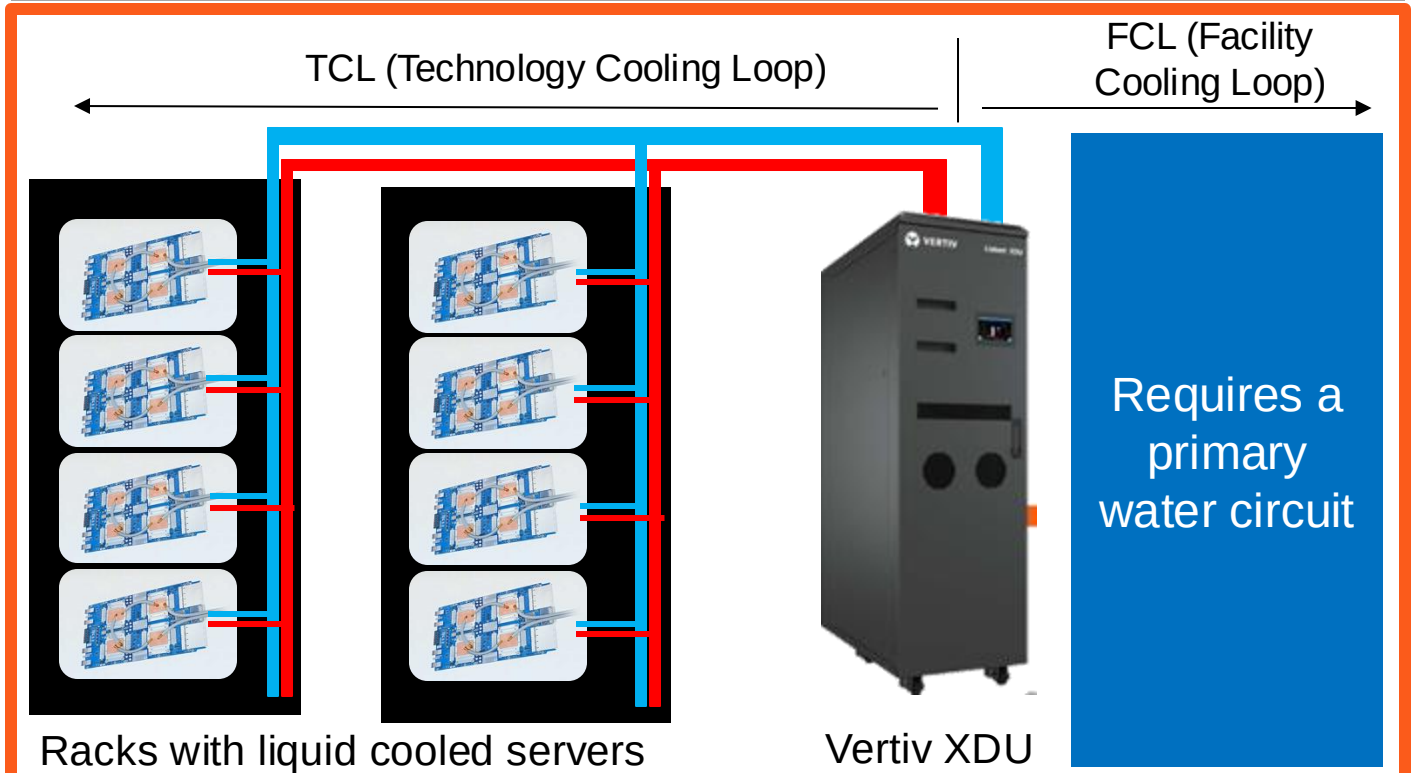
Dual phase

Liquid to Liquid CDU

Liquid to Air CDU

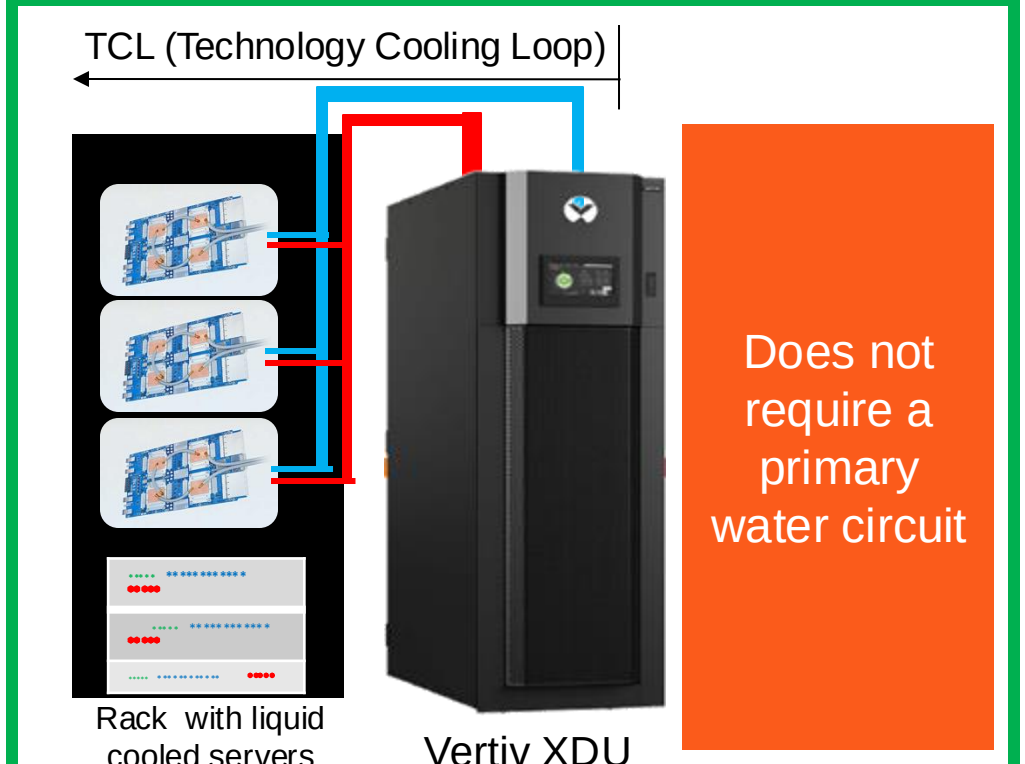
Liquid Cooling, Direct to Chip. How does it work?

LIQUID TO LIQUID CDU



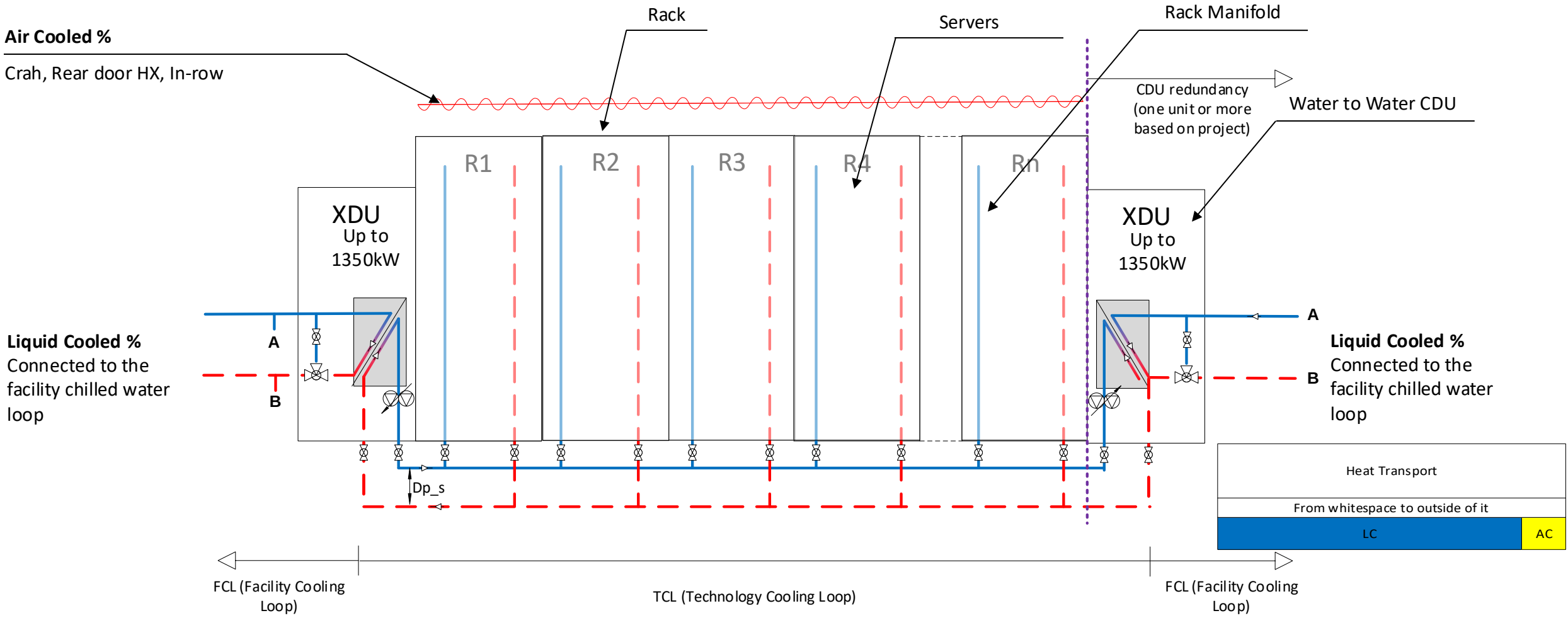
Vertiv Models **XDU450, XDU1350**

LIQUID TO AIR CDU

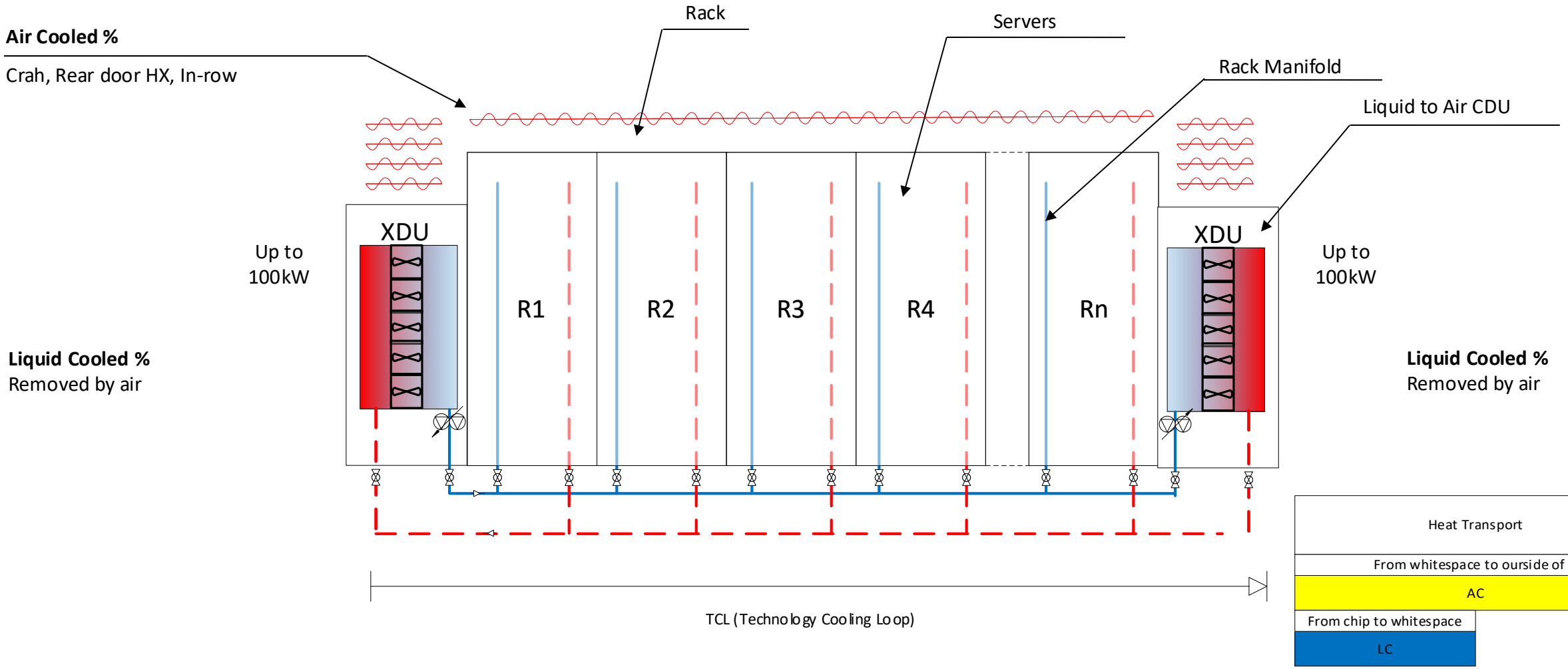


Vertiv Models **XDU070**

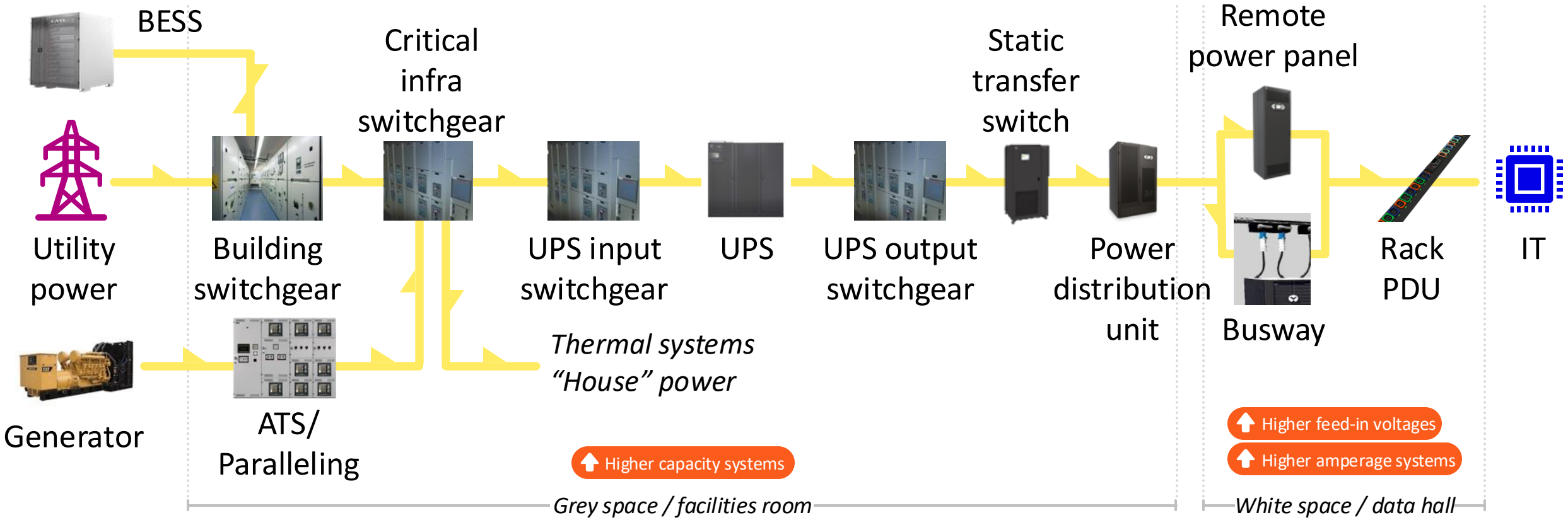
Liquid to Liquid CDU. How does it work?



Liquid to Air CDU. How does it work?

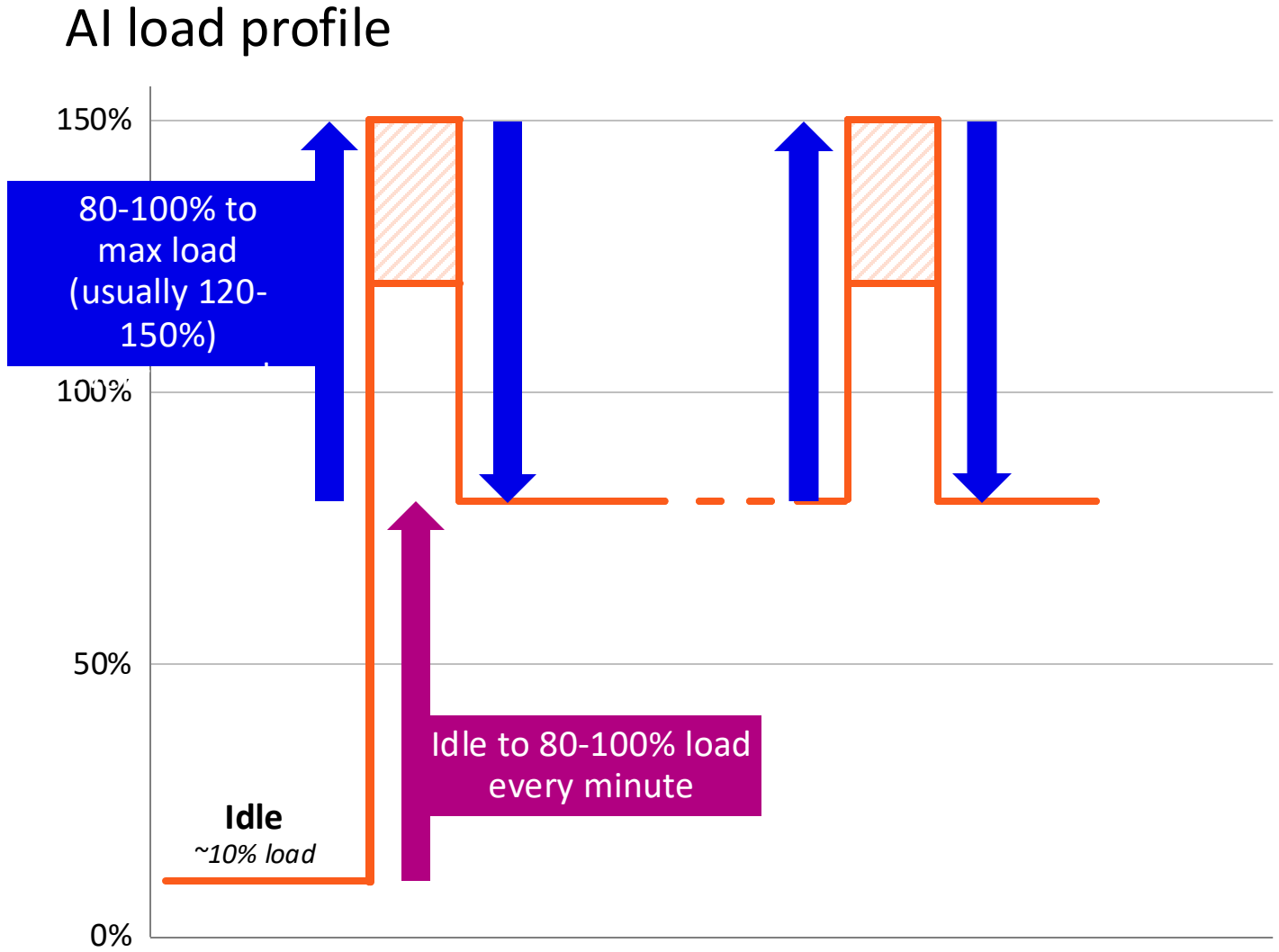


Amount of energy feeding high-density AI chips impacts forces redesign of the entire power train of the data center.



Power is critical to the success of AI deployments. IT and facilities teams must collaborate and partner with specialists who understand the power train end-to-end.

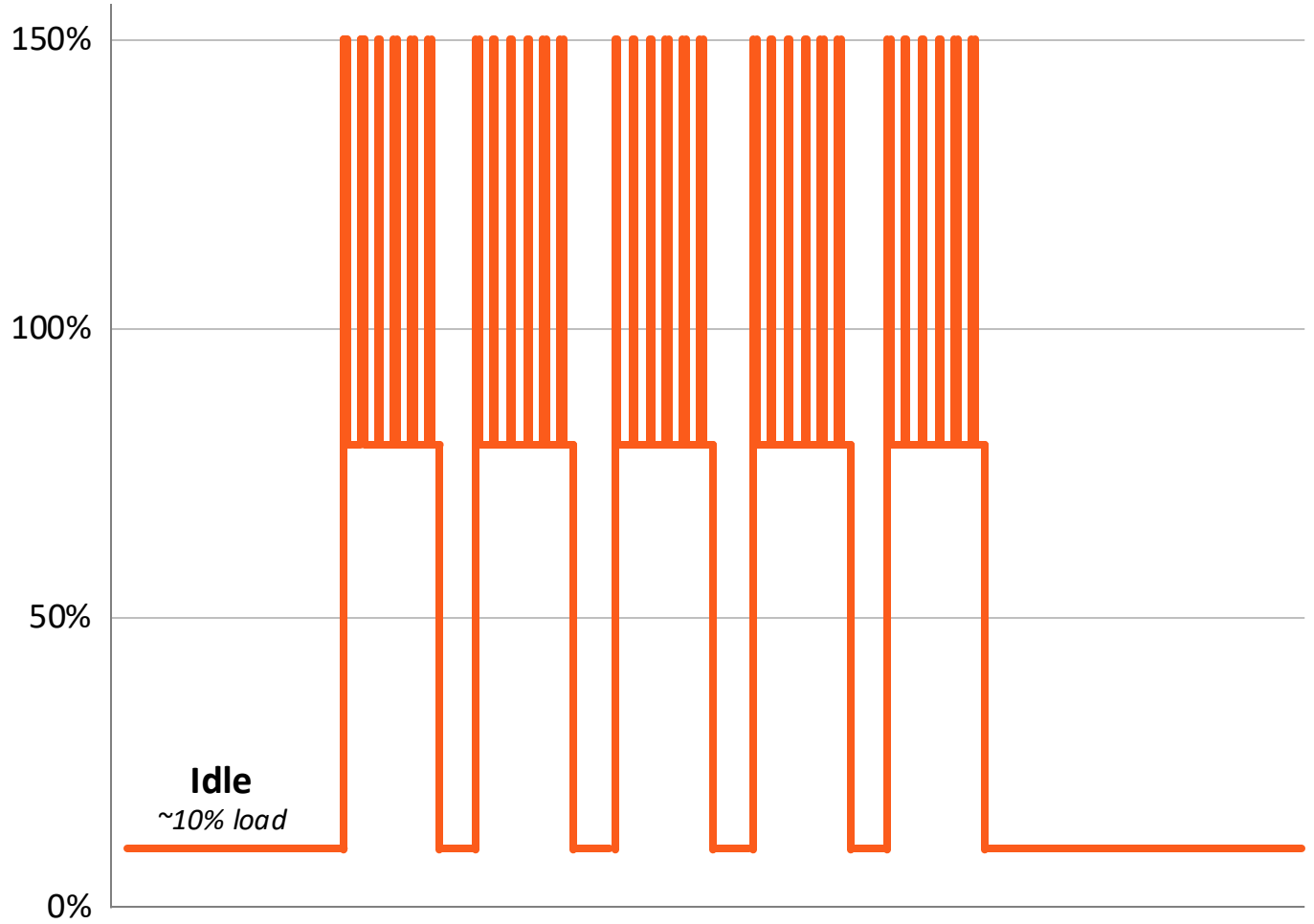
As GPUs become ubiquitous in data halls, they bring in challenging load profiles for supporting power infrastructure.



Load profile of GPUs running AI workloads is unique and very different from traditional IT equipment

As GPUs become ubiquitous in data halls, they bring in challenging load profiles for supporting power infrastructure.

AI load profile

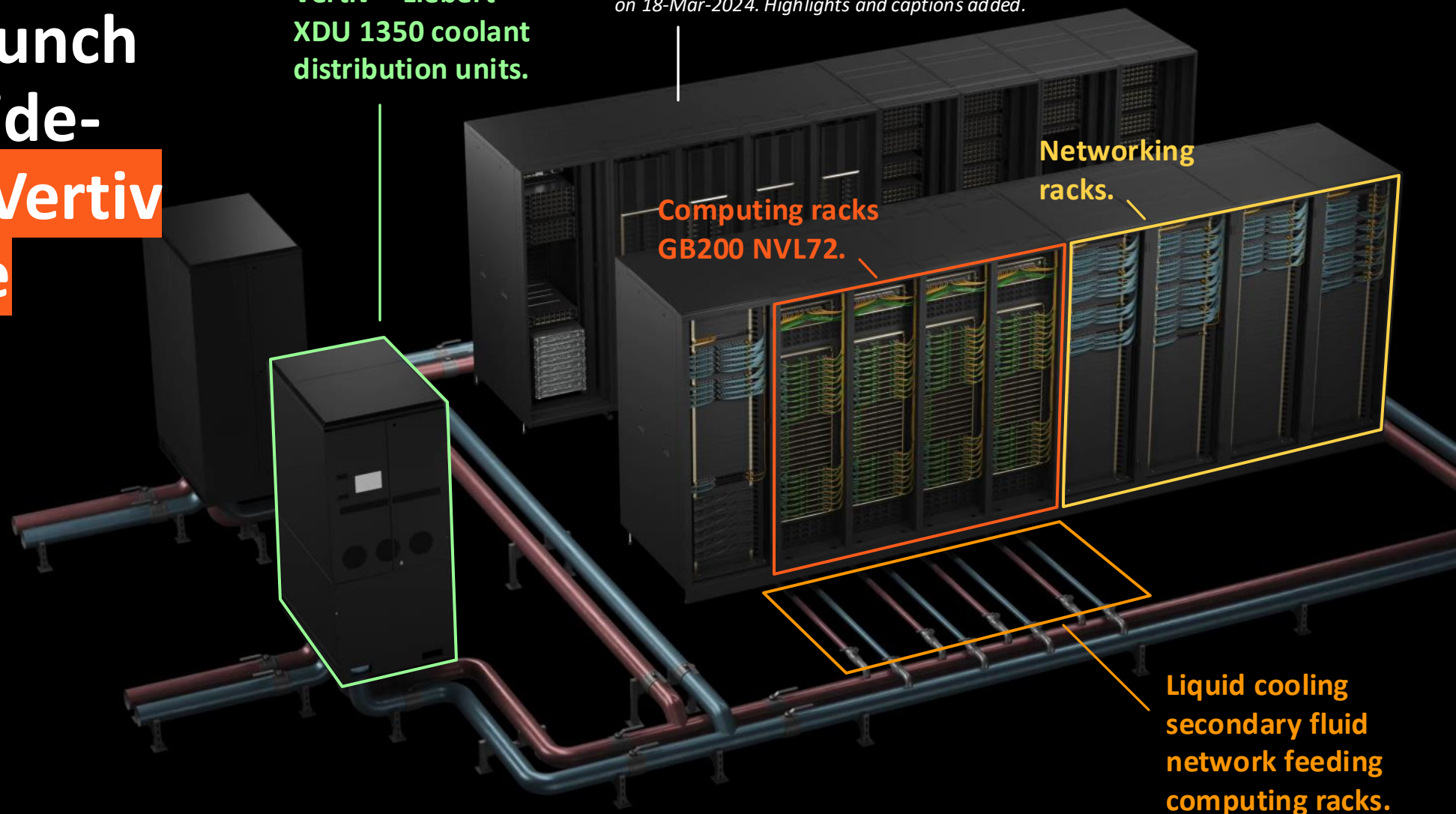


Pulses result in system instabilities. UPS must be able to avoid triggering battery and generator.

NVIDIA introduced its flagship launch at GTC2024 side-by-side with Vertiv infrastructure solutions.

NVIDIA DGX SuperPOD™ with GB200 systems

Rendering as presented by Jensen Huang at GTC2024 on 18-Mar-2024. Highlights and captions added.



NVIDIA Partner

Source: NVIDIA GTC2024 Keynote Speech

NVIDIA's leadership sees Vertiv as a key partner to create effective solutions in tandem with NVIDIA's growth.

NVIDIA Partner

BARRON'S

Nvidia CEO Says Partnership with Vertiv Will Help with Power Issue

By Tae Kim

Nvidia CEO Jensen Huang isn't too worried about the higher power draw from the company's latest Blackwell GPUs.

When asked about increasing power demands from Nvidia's latest AI GPUs during a financial analyst Q&A session at GTC Tuesday, the executive said rising requirements for absolute power, cooling, or power delivery will not be an issue.

He cited Nvidia's partnership with Vertiv, a leading provider of power and cooling infrastructure equipment, in creating effective cooling solutions for customers.

Source: Barron's

Vertiv is working side-by-side with Intel to develop power and cooling infrastructure and support its highest-grade AI solutions.

- Joint development of a product strategic for Intel's growth in the AI space.

- Two different liquid cooling design options:
  - Refrigerant-to-air-cooled solutions handling one rack up to 40kW.
  - *or*
  - Refrigerant-to-liquid cooling system able to remove up to 160kW of heat load.

> *To support increasing thermal design power and heat flux for next-generation accelerators, Intel has worked with Vertiv and other ecosystem partners to enable an innovative cooling solution that will be critical in helping customers meet critical sustainability goals.*
>
> — Devdatta Kulkarni, Principal Engineer, intel

Scan me

- Close partnership and collaboration with companies leading the AI revolution.
- Leader with most complete portfolio in power from grid to chip and cooling from chip to reuse, wrapped In monitoring and control software and AI-ready services capabilities, covering product lifecycle.
- Trailblazer Innovator from the days of Ralph Liebert to a NYSE-listed global organization.
- Unique capabilities to meet deployment needs from single rack to entire data center.
- Global scale with resilient operational footprint, in-market expertise and field service coverage.

Your AI critical infrastructure needs are in safe hands with Vertiv. Our offering is unparalleled in the industry.


VERTIV™


NVIDIA
Partner