The Thermal Tsunami:
Riding the Wave of Liquid Cooling for Future Compute

Massimiliano Maistro

EMEA Application Engineer Manager, Thermal Management

Data Center Forum Helsinki, 11-April-2024

Massimiliano Maistro

1965 Long history of pioneering technologies for the data center critical digital infrastructure.

Today a global organization with a team strong of 27,000+ people generating revenues of $6.8bn serving 130+ countries.

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

Strong position in HPC/AI with deep partnerships and collaboration with leading players.

EMEA Application Engineer Manager with focus on High Density applications

Technical background and 6+ years of professional experience in HVAC and Data Center industry.

AI is here, and it is here to stay.

It will transform our lives.

It will transform our businesses.

It will transform our IT infrastructures.

> AI is one of the most important things humanity is working on. It is **more profound than electricity or fire**.
>
> — Sundar Pichai, CEO
> Google

> This is the **first time that a technology** developed in Silicon Valley **benefits the lives of everyday people so quickly and so tangibly**.
>
> — Satya Nadella, CEO
> Microsoft

> AI is on the fast track to **becoming ubiquitous** — at **home**, at **work**, and **every-where** in between.
>
> 60% of workers will use their own AI to perform their tasks.
>
> — Predictions 2024: AI
> FORRESTER®

> Generative AI could potentially raise annual labor productivity growth by around 1.5pp over a 10-year period, and eventually **raise global GDP by 7%**. Gen AI raises the potential for a boom in labor productivity that significantly increases global growth.
>
> — Joseph Briggs, Sr. Global Economist
> Goldman Sachs

> I am already seeing AI workloads leading to a broad proliferation of accelerated computing infra-structure. This will require investments in next-generation data center physical infra-structure to support new architectures with higher power and thermal management requirements.
>
> — Lucas Beran, Research Director
> DG

Source: Forrester, BCG, Goldman Sachs, Dell'Oro Group, IBM, News clippings

Technology

Higher coding productivity

Consumer

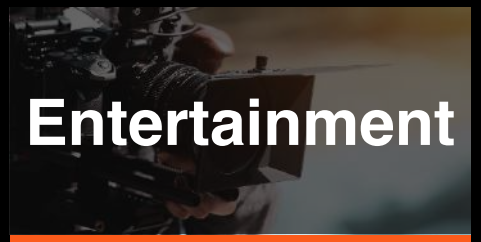
Conversion rate increase

Biopharma

Research timeline reduction

Financial Institutions

Higher fraud detection accuracy



Entertainment

Higher quality of animated images

Insurance

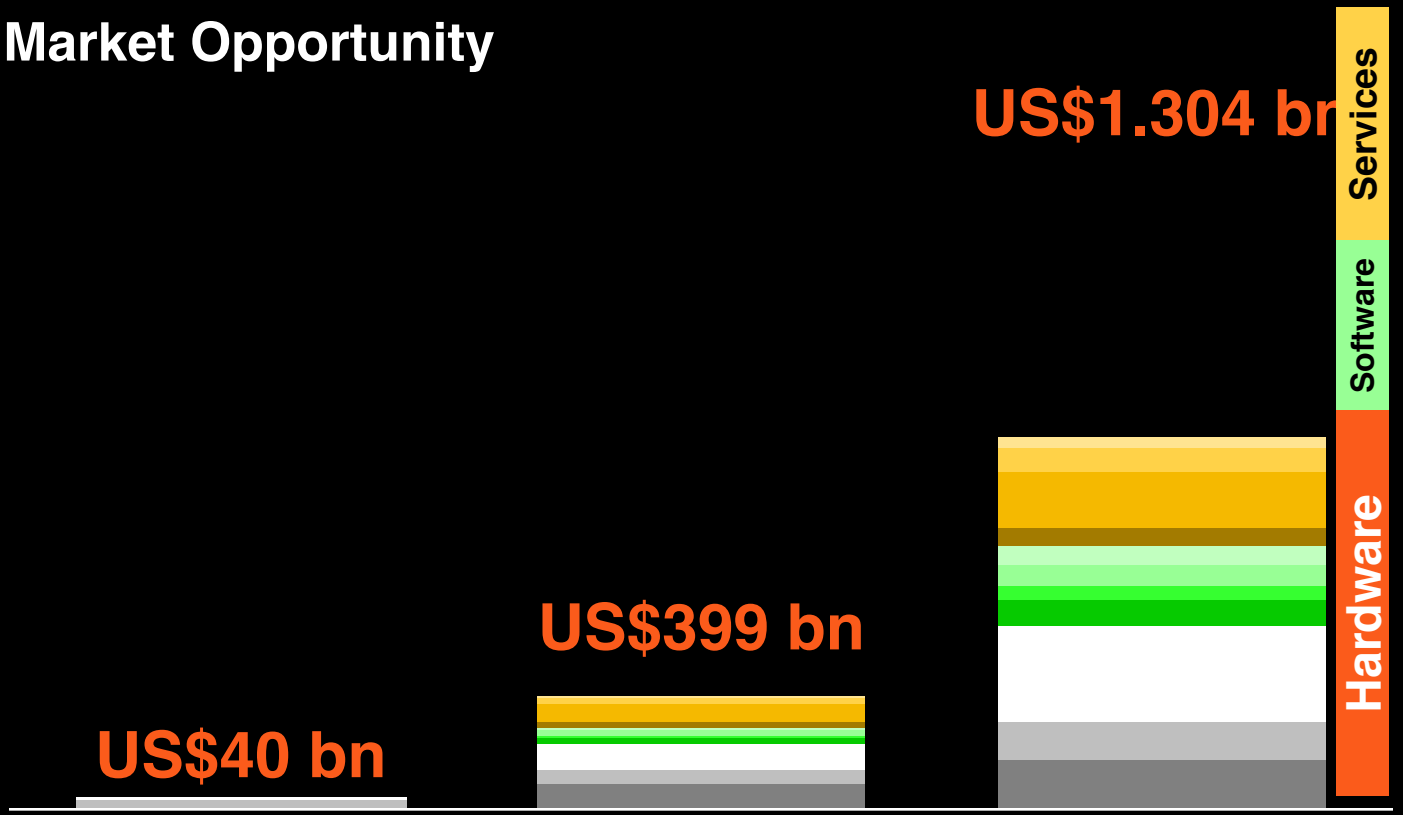
Customer service cost optimization

Encouraging AI use-cases are bringing in profit to businesses across all industry verticals.

Source: BCG, GitHub, Vue.AI, Nature, ScienceDirect, BusinessWire

Bloomberg

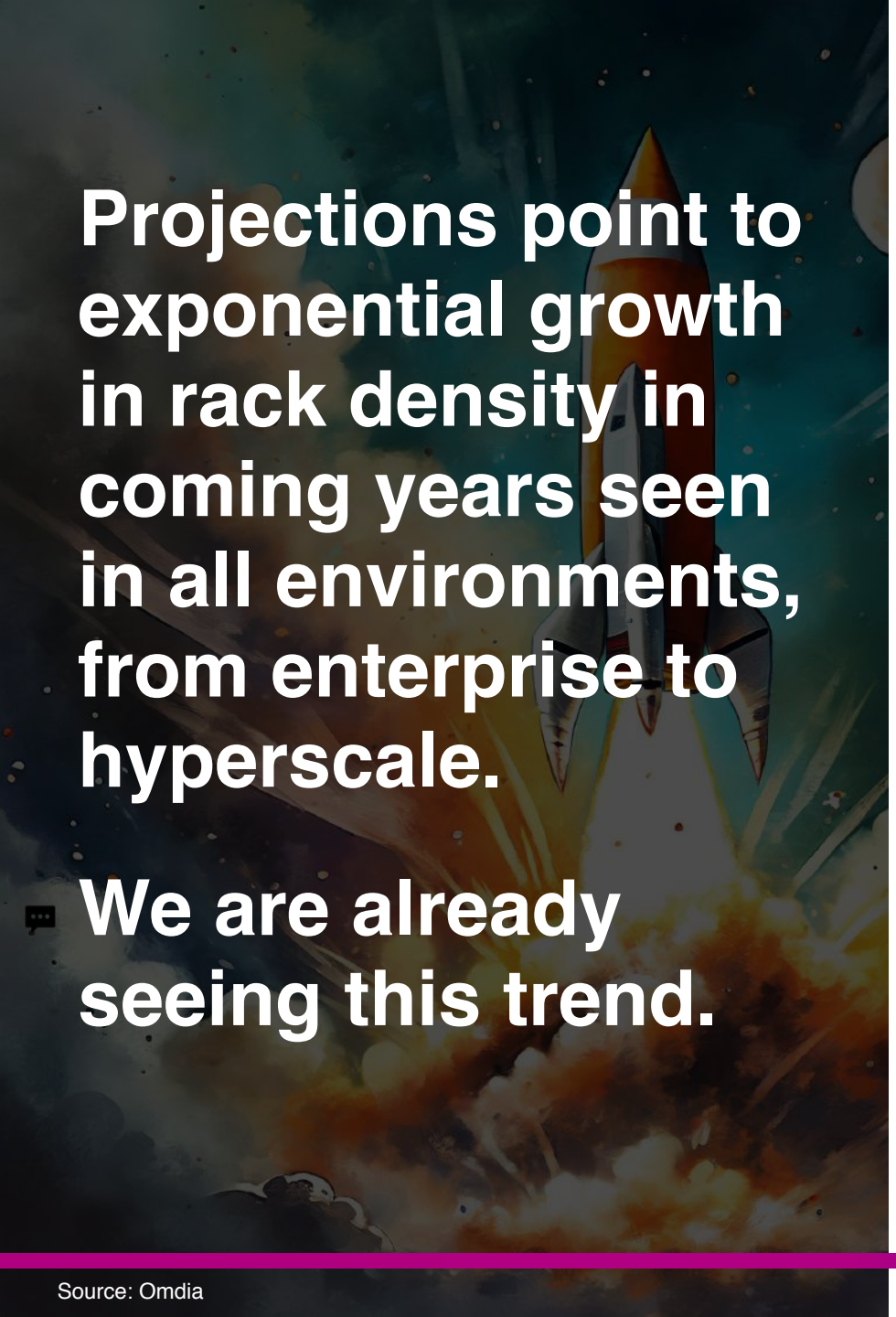
Generative AI Market Opportunity



1%	5%	12%
of total technology spend		

Gen AI is expected to be a $1.3-trillion business by 2032.

Source: Bloomberg Intelligence

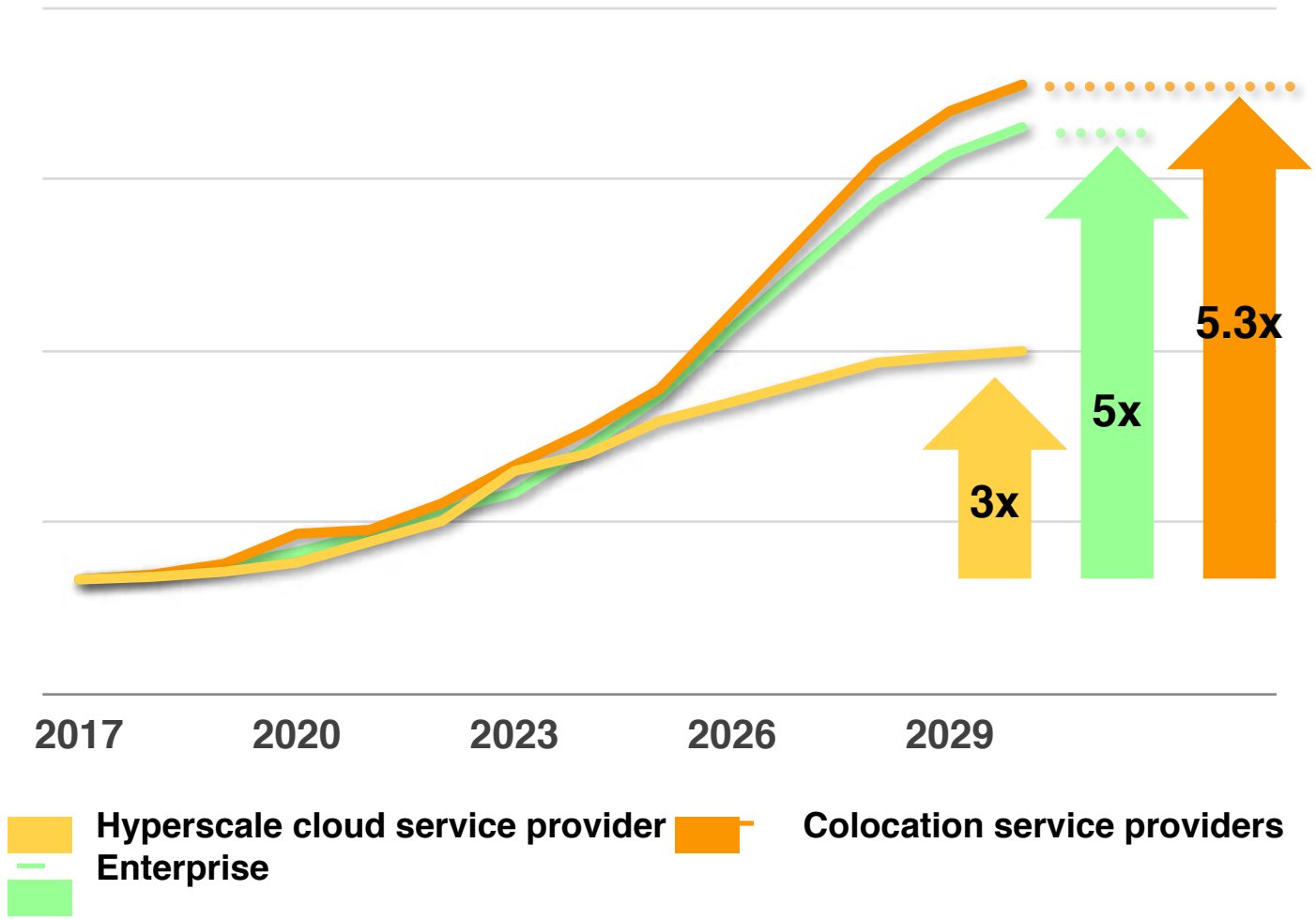


Projections point to exponential growth in rack density in coming years seen in all environments, from enterprise to hyperscale.

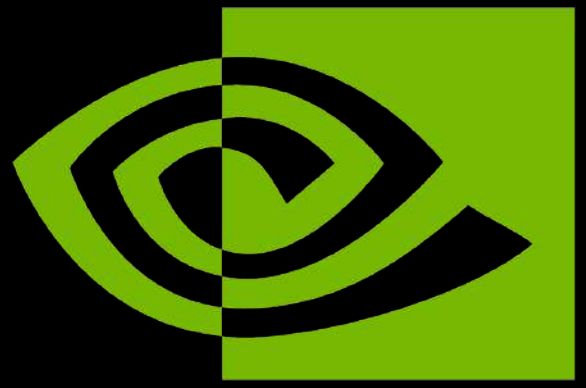
- We are already seeing this trend.

Rack density growth is expected to take off.

Avg kW per rack trend



Source: Omdia



NVIDIA®

habana®

An Intel Company

Solution advisor consultants provide consultation services and expert advice to customers looking to implement NVIDIA based solutions or technology.

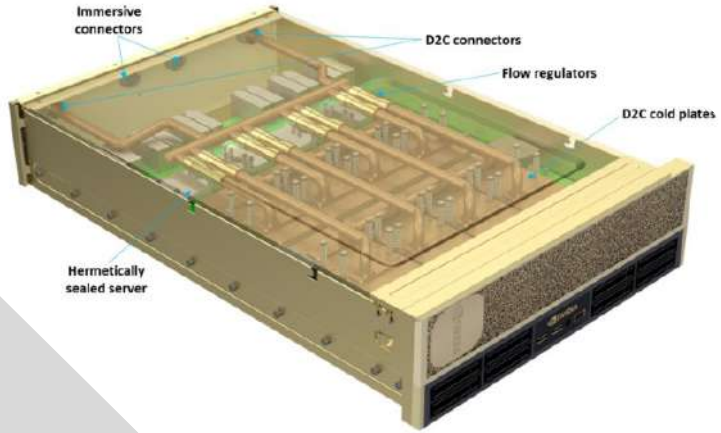
- VERTIV has met product and technology competency attainments required by NVIDIA.
- VERTIV Sales and Technical associates have undergone training of NVIDIA products.

Vertiv is a Solution Advisor Consultant partner in the NVIDIA Partner Network.

NVIDIA Partner

U.S. Department of Energy awarded a $5m to an NVIDIA-led team to the development of advanced liquid-cooling technology as part of its Coolerchips program.

The team came with a novel approach combining two liquid-cooling technologies as never seen before.



NVIDIA partnered with Vertiv, other leading manufacturers, and research institutions.

Vertiv is partnering with NVIDIA to develop cutting edge cooling technology.

NVIDIA | Partner

Source: NVIDIA, US Department of Energy

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.**

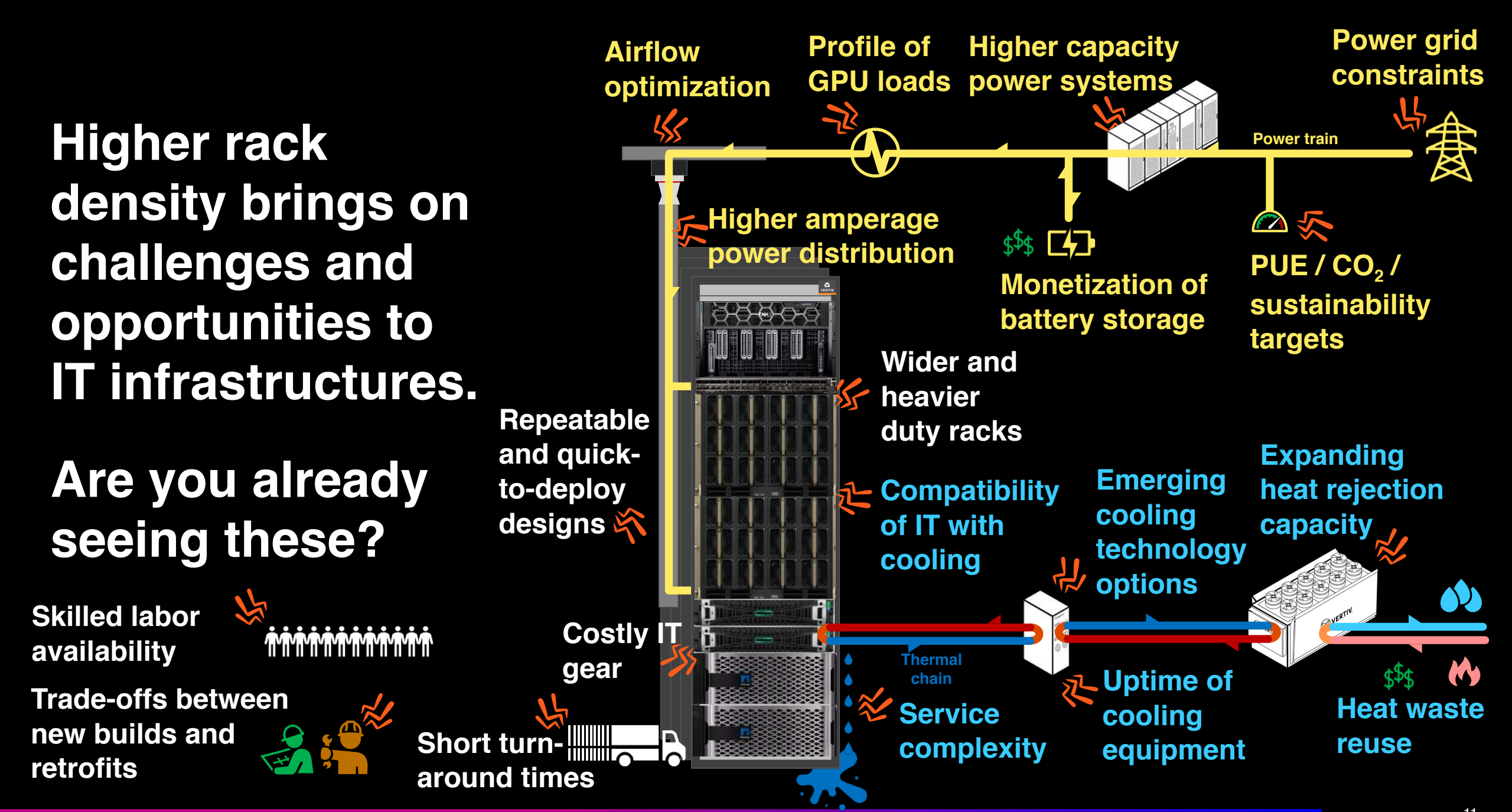
- **Two different liquid cooling design options:**
  - **Refrigerant-to-air-cooled solutions handling one rack up to 40kW.**
  - —or—
  - **Refrigerant-to-liquid cooling system able to remove up to 160kW of heat load.**

> 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

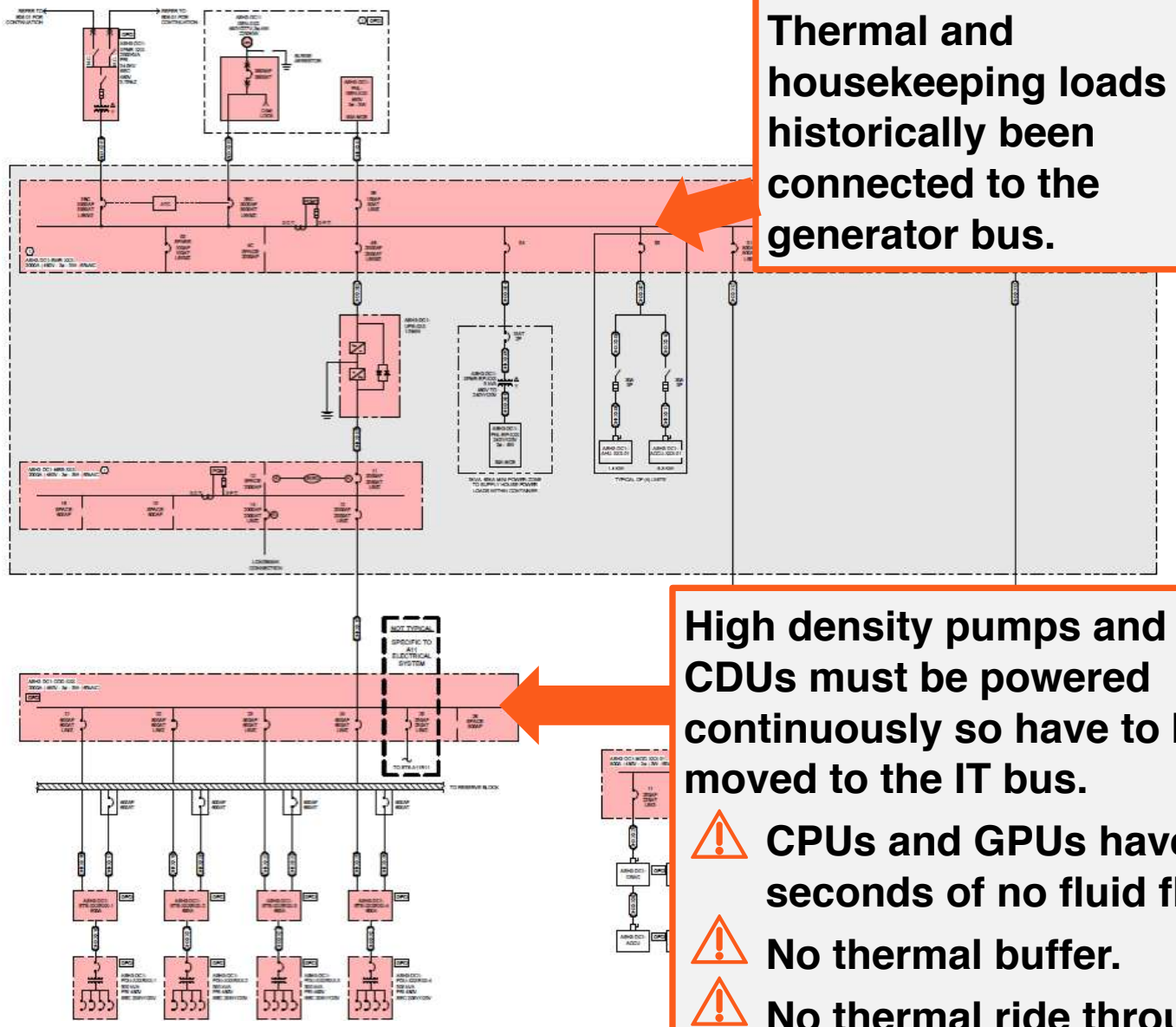
Scan me

Higher rack density brings on challenges and opportunities to IT infrastructures.

Are you already seeing these?



Cooling equipment is added to the critical load as GPUs generate so much heat that thermal loads cannot go down.

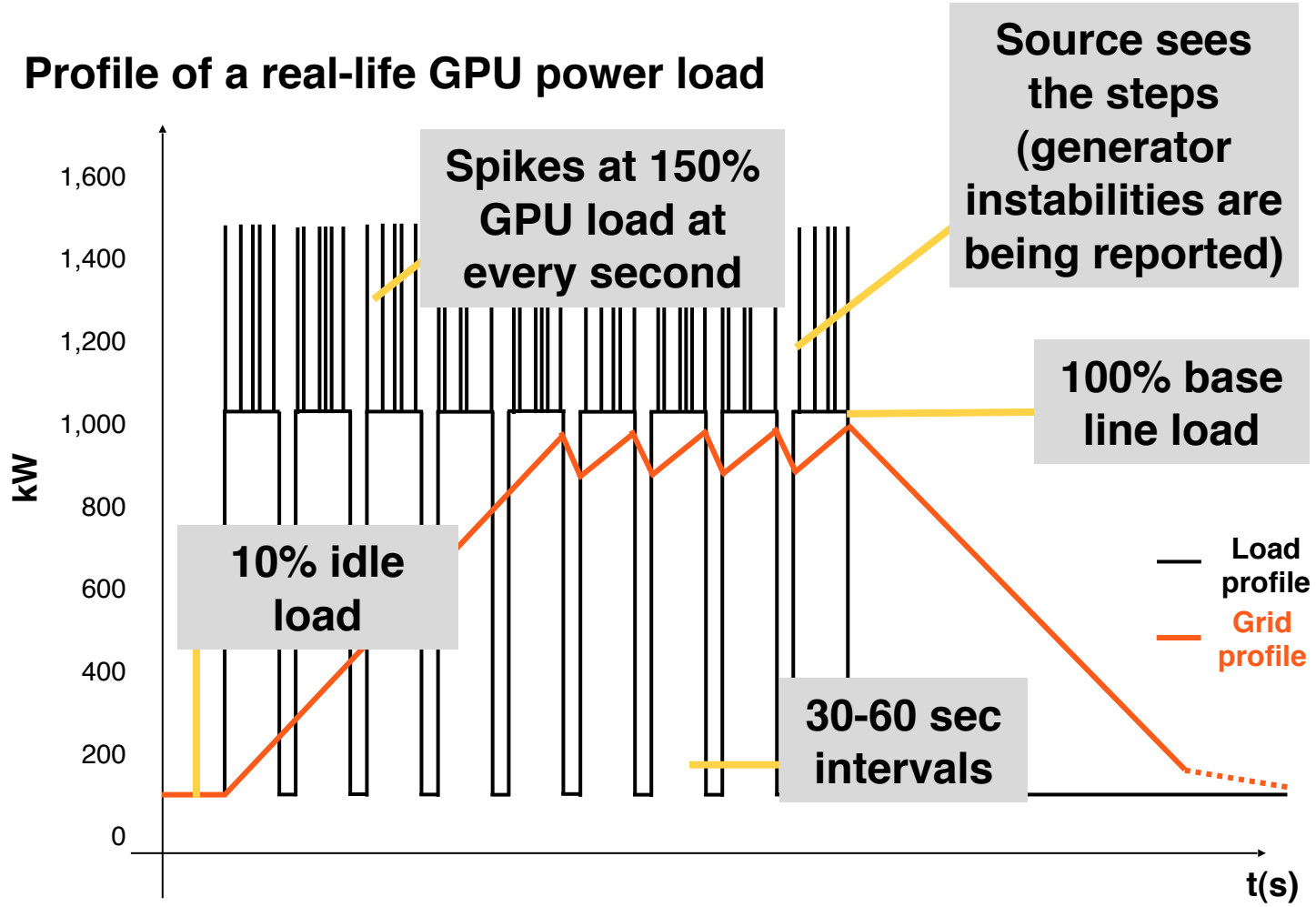


Thermal and housekeeping loads have historically been connected to the generator bus.

High density pumps and CDUs must be powered continuously so have to be moved to the IT bus.

- ⚠ **CPUs and GPUs have <5 seconds of no fluid flow.**
- ⚠ **No thermal buffer.**
- ⚠ **No thermal ride through.**

AI workloads don't have uniform patterns and that add complexity.

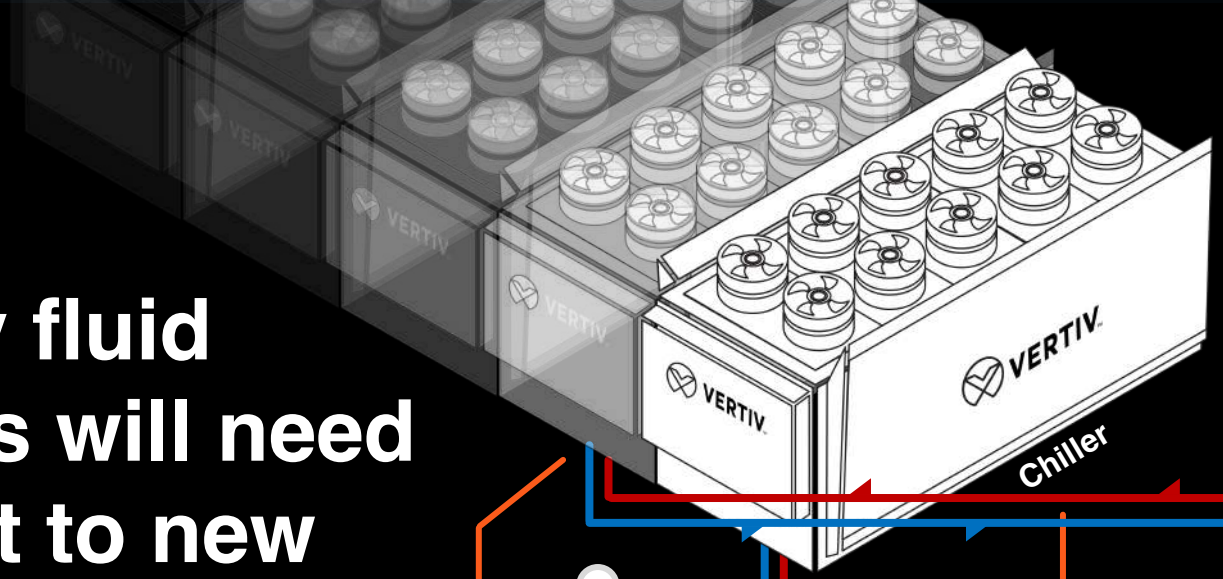


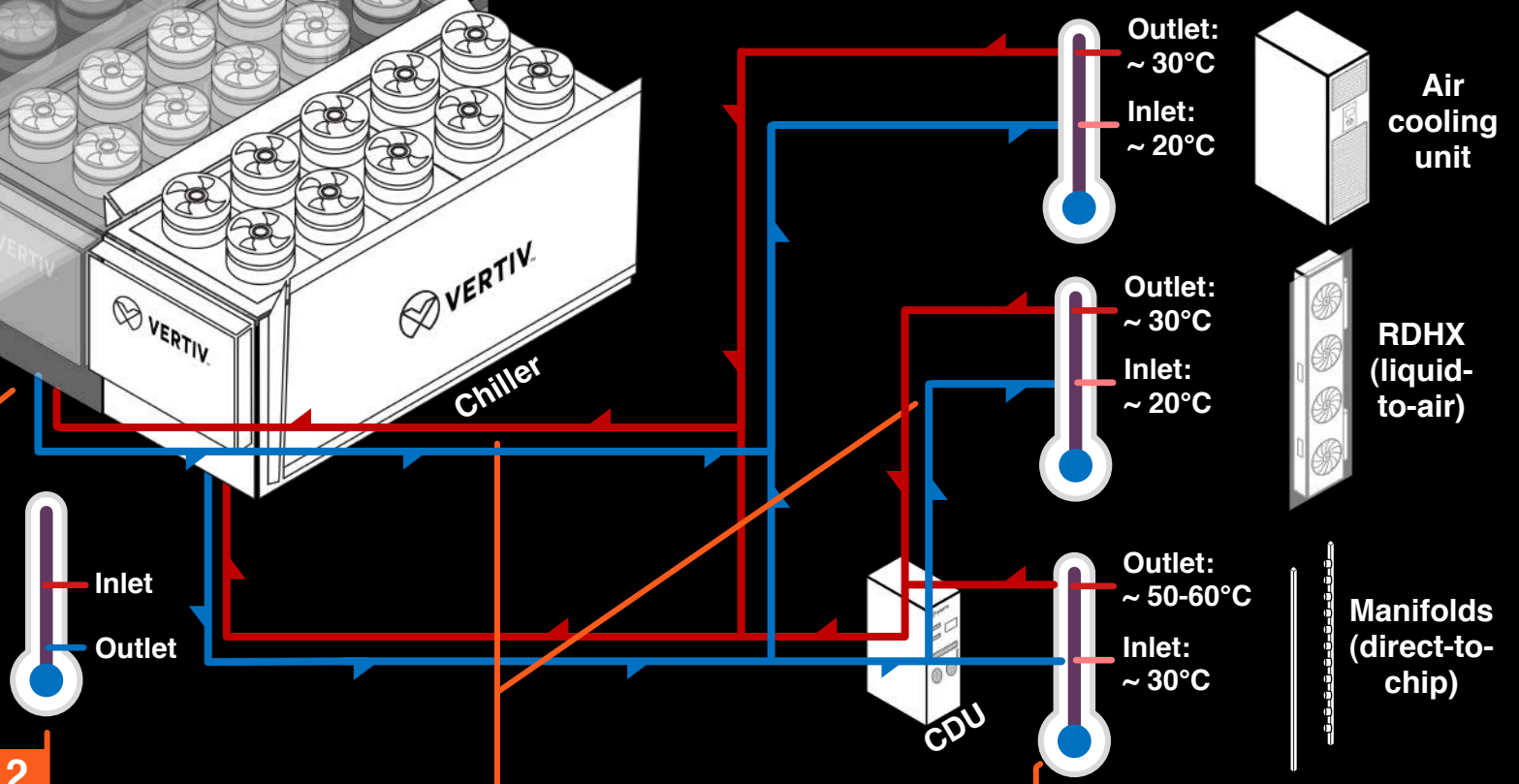
Vertiv is developing behavioural models with GPU manufacturers to model these loads in live data centres.

Vertiv is supporting live testing with leading MTDCs.

Primary fluid systems will need to adapt to new technologies.

Approach temperatures key to performance and efficiency.



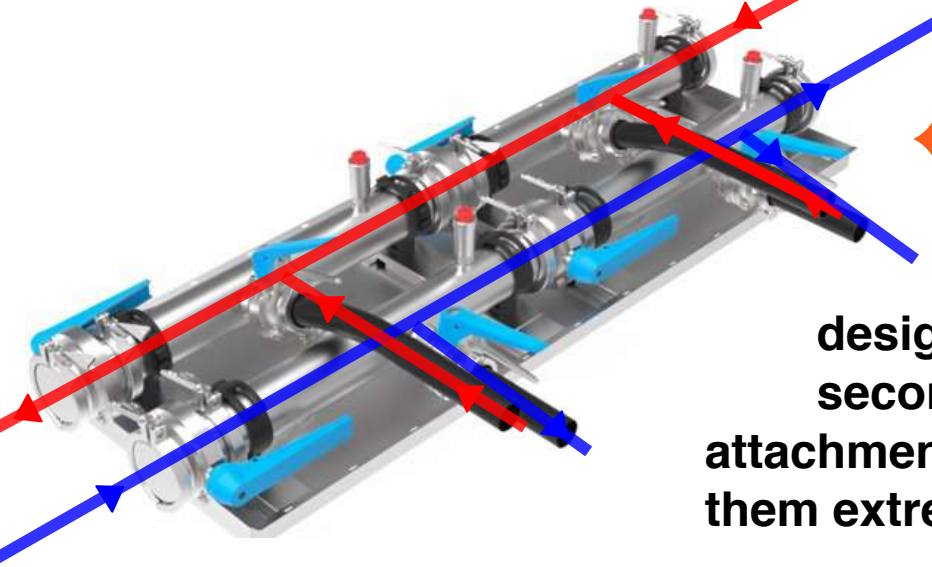


1 If space is available, chillers can easily scale up capacity at the pace of your workloads.

2 Inlet and outlet temperature flexibility support both air and liquid cooling solutions out-of-the-box.

3 Different requirements might push chiller to lowest temperature across multiple cooling systems, limiting energy saving benefits from direct-to-chip.

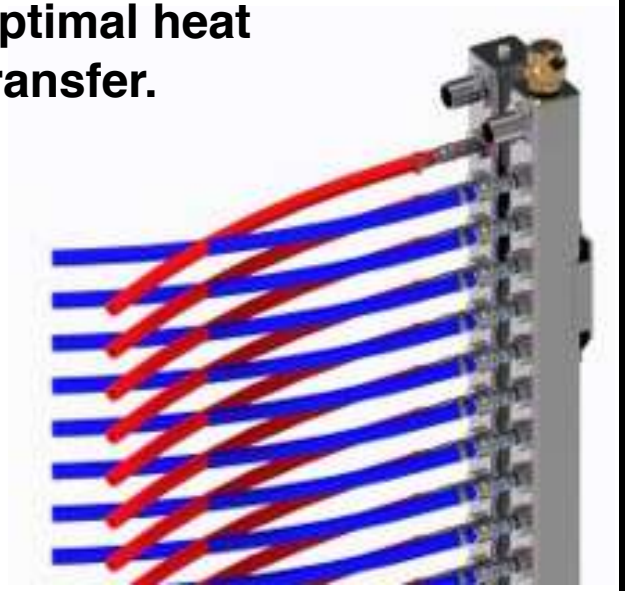
4 Or different cooling methods can be isolated in individual closed loops if energy saving can pay back additional piping.



Cleanliness is paramount. Unlike pumps and CDUs, current server designs do not offer multiple secondary fluid network attachment for redundancy, making them extremely critical.

Demanding **material quality** often food-grade or medical-grade to avoid corrosion and leaks.

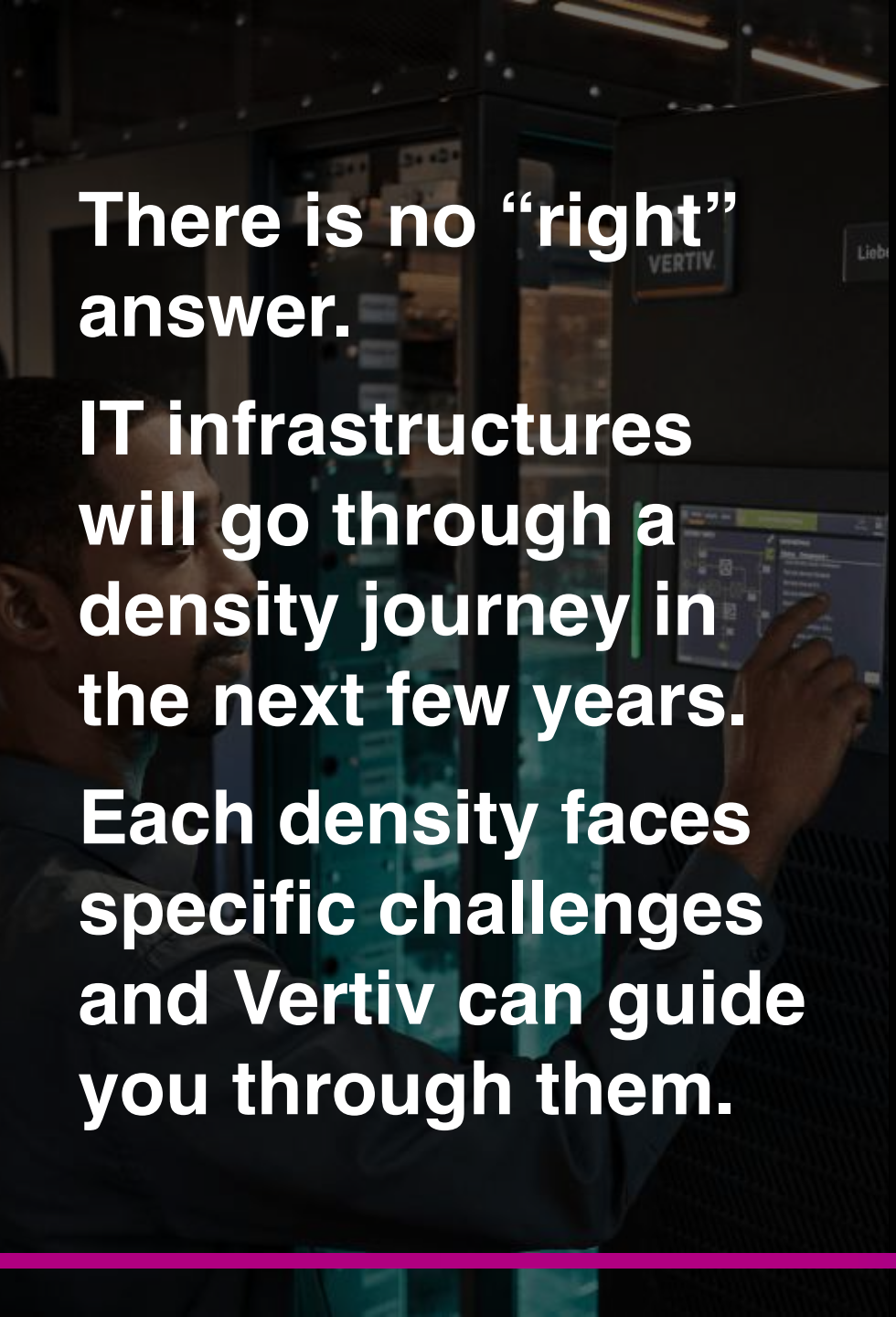
Cold plates require **27µm filtration** to ensure optimal heat transfer.



Whether PG25 or DI water, fluid running in secondary networks require **specialised attention**:

- Flushing
- Regular testing
- Topping up
- Avoiding air in system

Secondary fluid networks are in direct contact with GPU cold plats and have become as critical as power.



There is no “right” answer.

IT infrastructures will go through a density journey in the next few years.

Each density faces specific challenges and Vertiv can guide you through them.

Low density
Up to 10kW

Conventional compute workloads

Some accelerated compute, AI use case testing or inferencing workloads

Medium density
10 to 25kW

High density
25 to 80kW

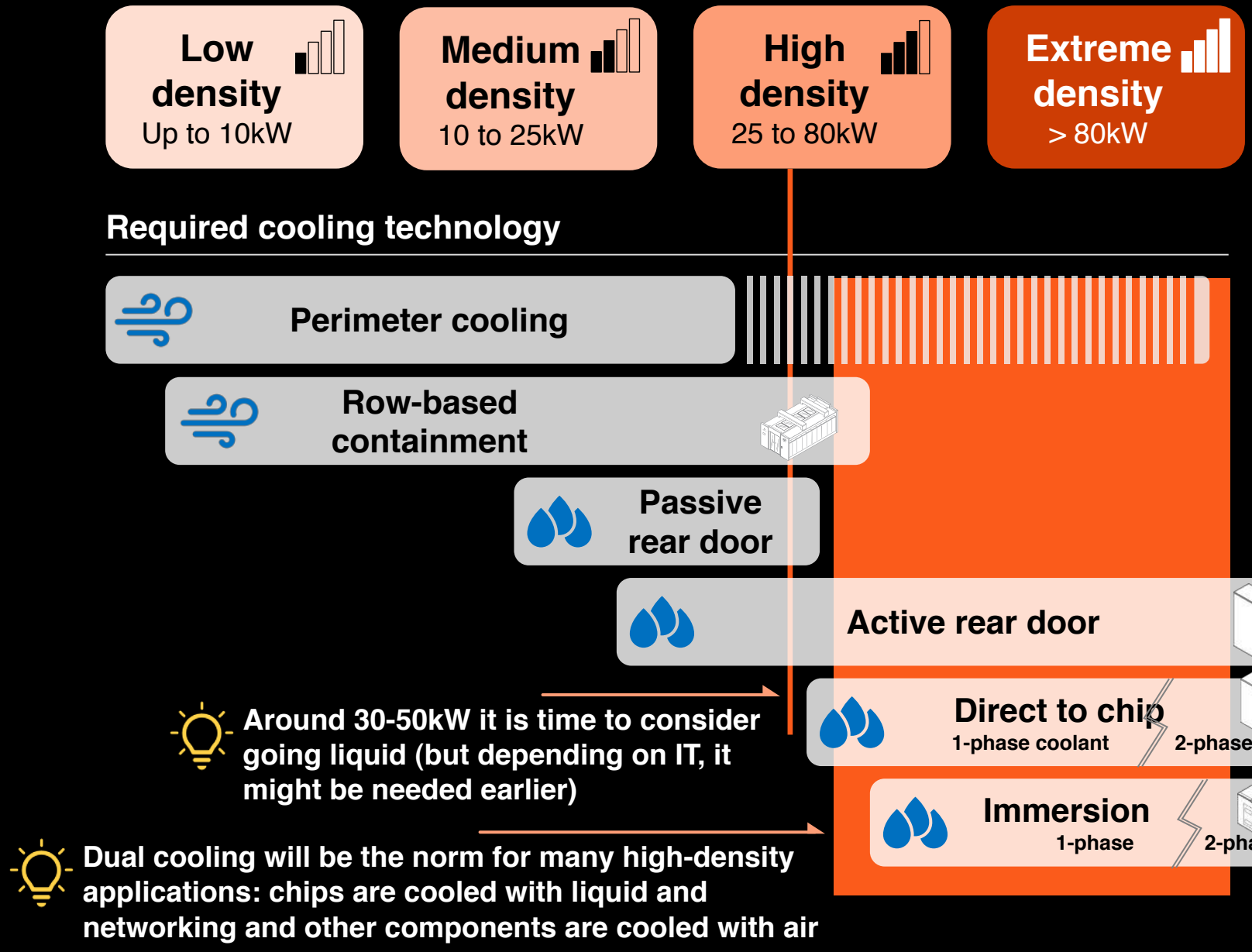
Most accelerated compute, dedicated to both AI training and inferencing workloads

All accelerated compute, AI factory

Extreme density
> 80kW

Beefing up existing cooling may not be enough to meet needs of AI work-loads.

Vertiv can help choosing the right cooling technology for each operation.



Liquid cooling infrastructure from chip to atmosphere is composed of 7 key elements.

1 Heat Rejection

Raised temperature Fluid Cooler & Efficient Chillers to reject heat outdoors.

2 DTC Cold Plates

Single- and Two-Phase Cold Plates to remove heat from the processor into the liquid.

3 Row & Rack Manifolds

Row and Rack Manifolds distribute liquid to IT or other Liquid Cooling products.

Liquid Cooled Facility with Dual Air and Liquid Cooling

4 Chillers, Condensers & Fluid Coolers

Free cooling chillers, fluid coolers and condensers to reject heat for single and 2 phase air and liquid cooling.

5 Coolant Distribution Units

1 & 2 Phase Row & Rack mounted CDUs to manage temperature liquid quality to server cold plates.

6 Indoor Chilled Water

Chilled Water room cooling units to address residual heat from liquid cooled racks.

7 Immersion Tanks

Immersion Racks to support server compaction and High -Density IT.

Our offer of coolant distribution units (CDUs) for direct-to-chip liquid cooling cover a wide range of requirements.

Liquid to Liquid CDU

- Transfer station for rows
- Heat transfer to chilled water network
- Cabinet design
- Cooling for many server racks
- Large applications

450kW 1350kW

- Transfer unit for rack installation
- Heat transfer to chilled water network
- Compact design 19" only 4U
- Cooling per rack
- Small applications (few racks)

19” 4HE **100kW**

Liquid to air CDU

- Heat dissipation into the room
- Almost no heat recovery
- Bridging technology
- Mostly for testing of few racks
- Hot/cold aisle arrangement

600mm Rack **70kW**

Vertiv™ Liebert® VIC immersion cooling solutions portfolio is designed with different applications in mind.

Self contained option

- Up to 50kW total power.
- Ideal for pilot applications or small isolated high-density workloads.

Multi-tank + CDU

CDU 240kW

Immersion tank

- Up to 240kW total power.
- Flexible and scalable design allowing to add more immersion tanks to same shared CDU.
- Ideal to future-proof infrastructure to accommodate expansion.

Thank you.

Albér™ | Avocent® | Cybex™

Liebert® | NetSure™