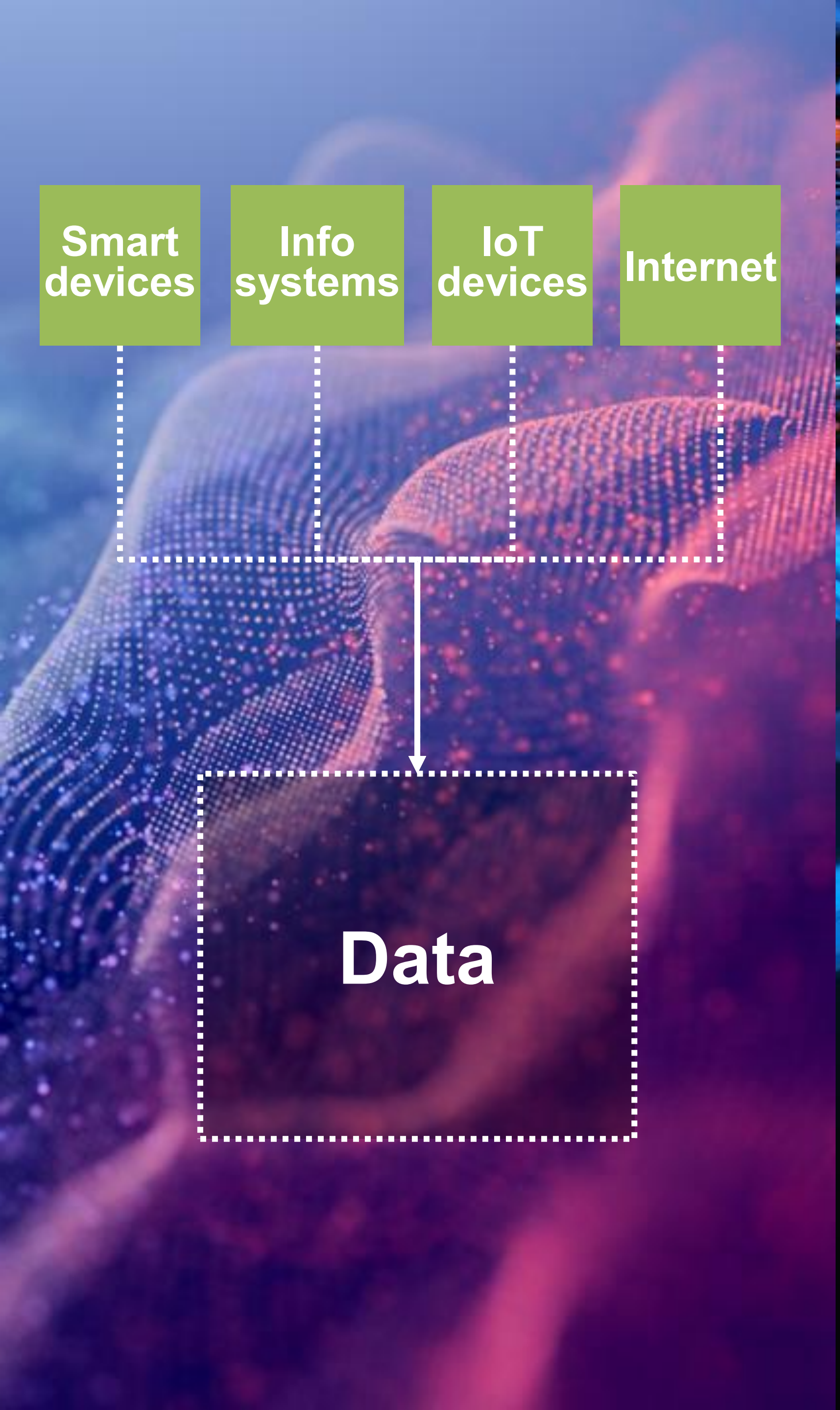
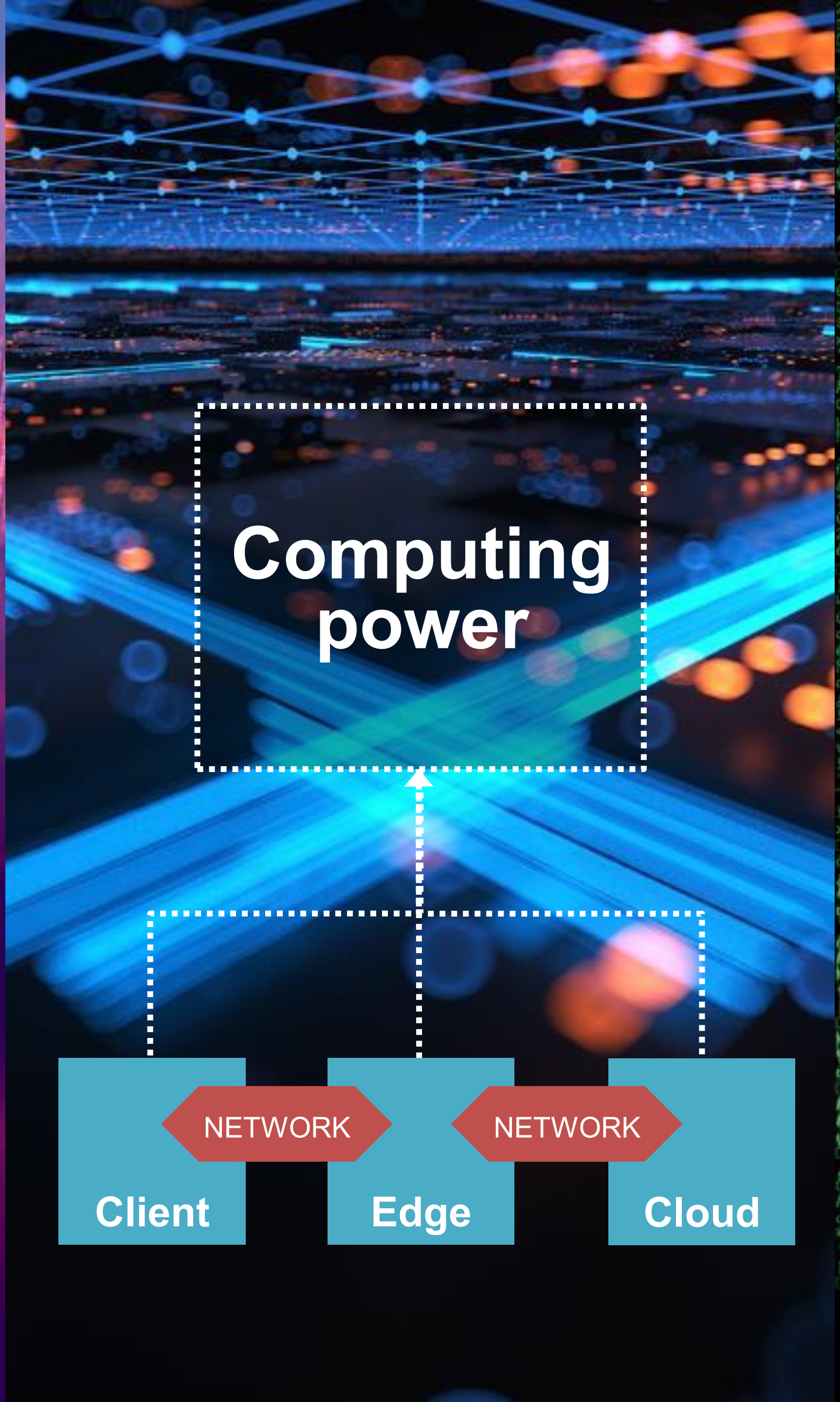
Il futuro dell’AI per la sostenibilità energetica

Massimo Chiriatti
Chief Technical & Innovation Officer – Lenovo

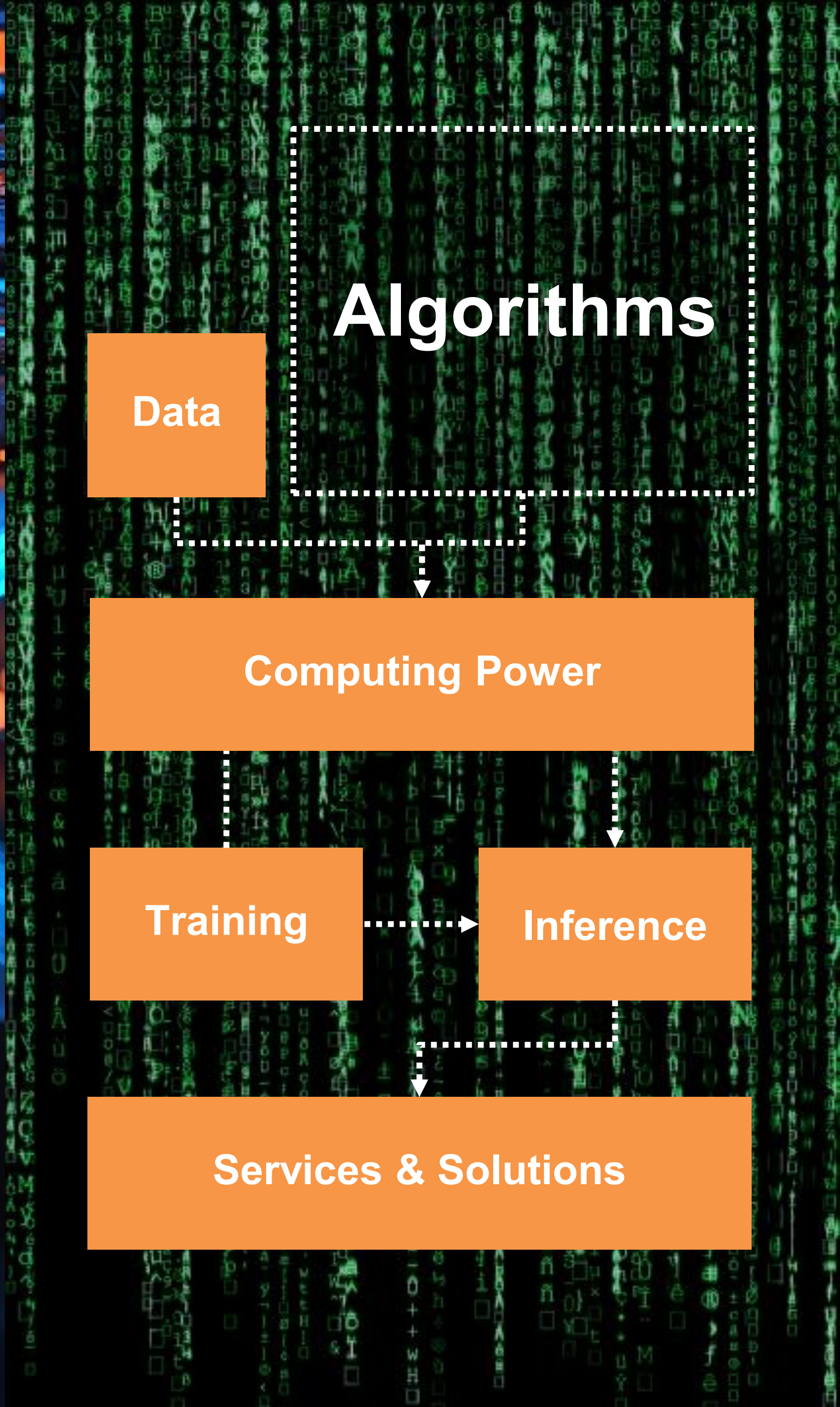
https://www.linkedin.com/in/massimochiriatti/



Smart devices
Info systems
IoT devices
Internet
Data



Computing power
NETWORK
NETWORK
Client
Edge
Cloud



Algorithms
Data
Computing Power
Training
Inference
Services & Solutions

Where is the data?

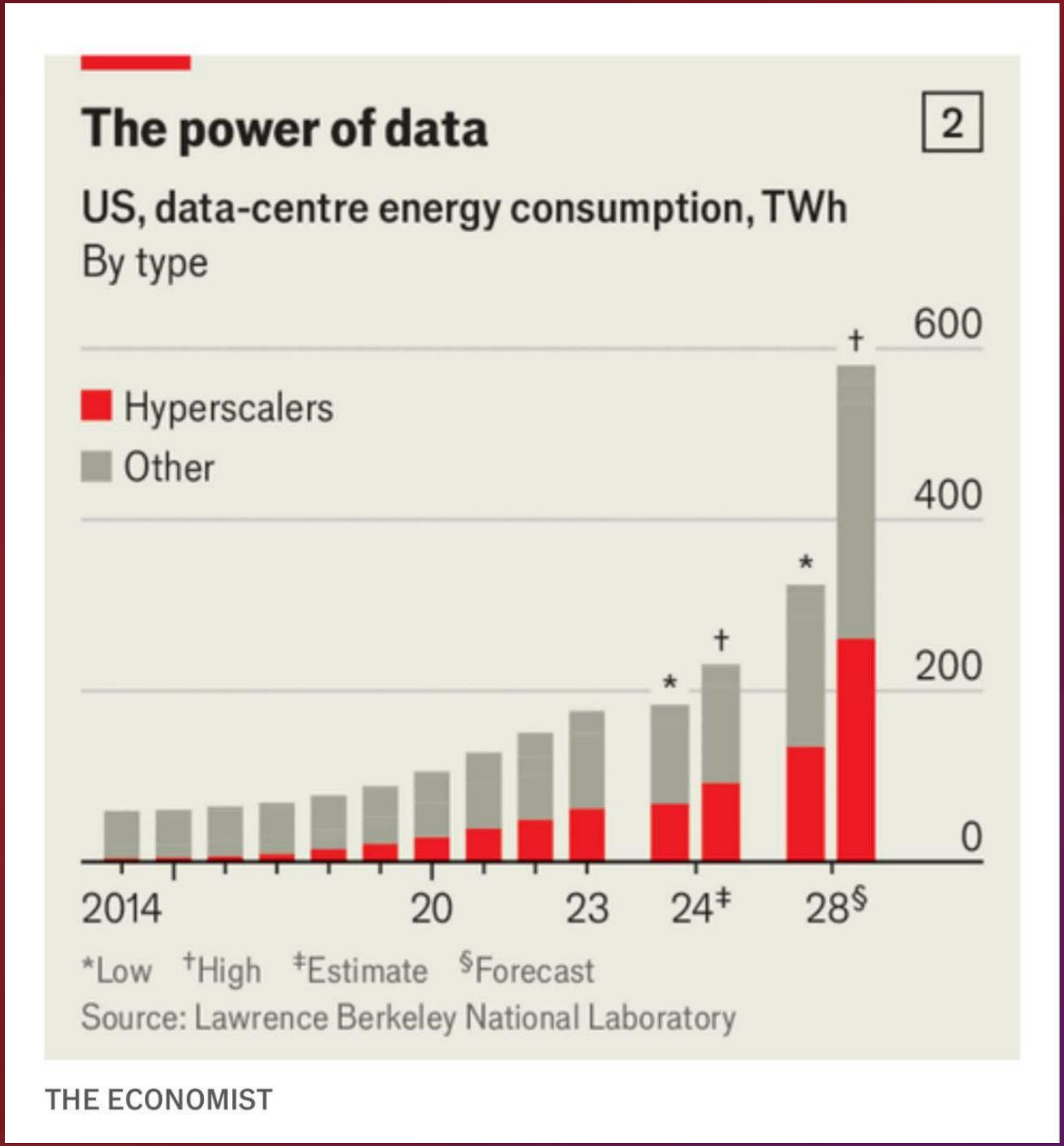
Cloud for Agility & Flexibility

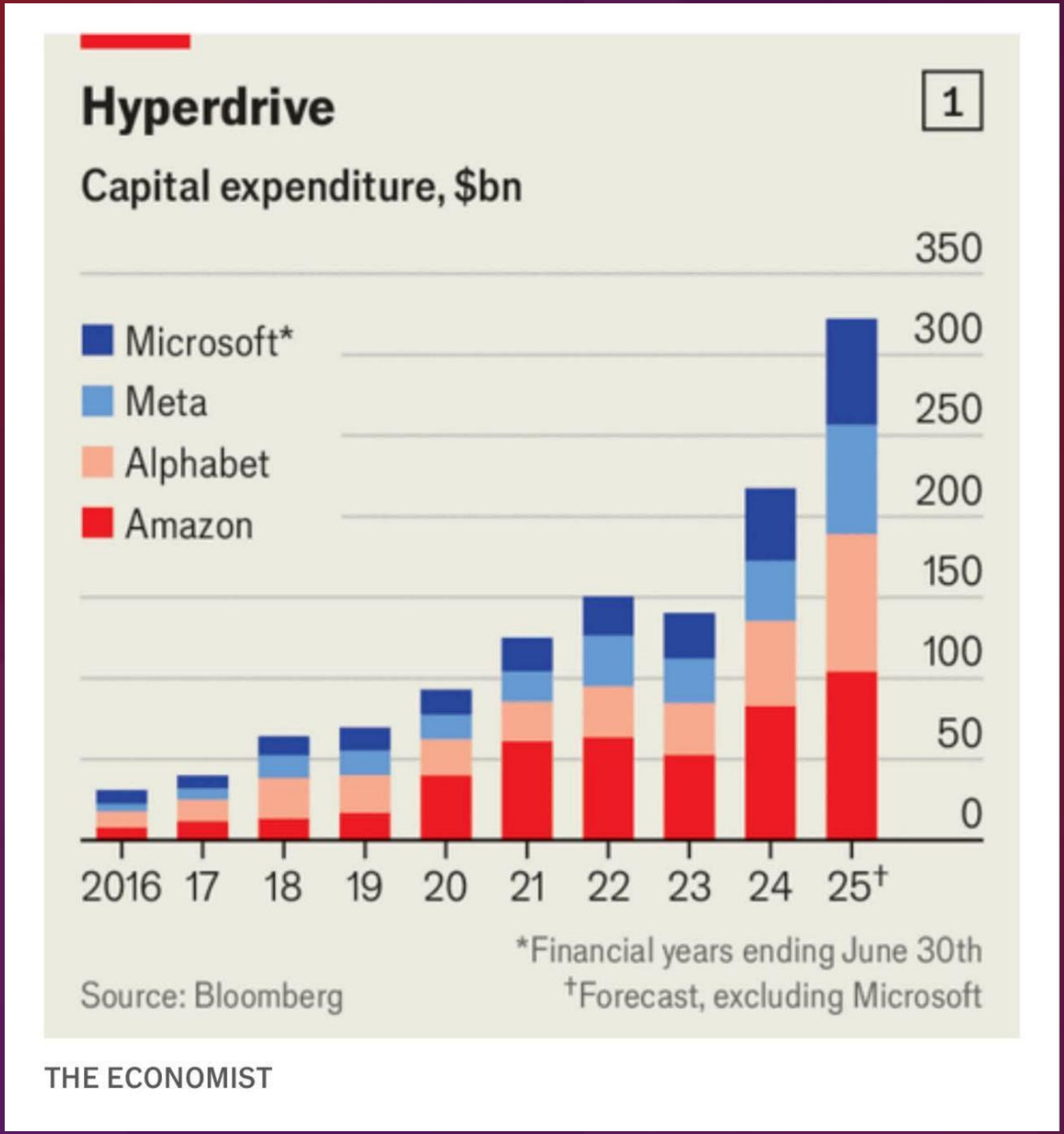
- Ideal for **short-term** needs: experimentation, fine-tuning, retraining.
- Offers rapid **scalability** and **on-demand** access.
- But: Costs scale quickly with sustained usage (compute, data transfer, storage).

On-Prem for Predictability & Savings

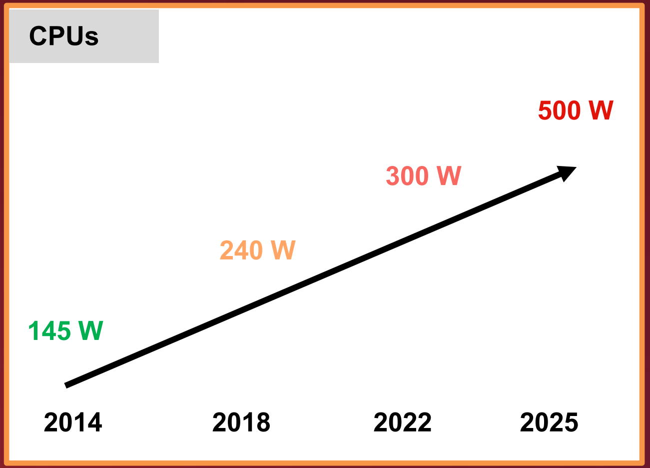
- Best for **long-term**, **predictable** workloads like inference/model serving.
- Higher upfront investment but lower TCO beyond the breakeven point.
- Avoids recurring cloud charges and offers **greater cost control**.

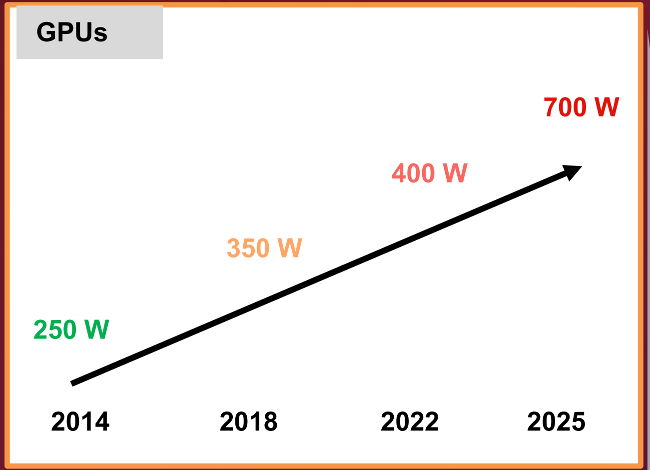
Energy & Capital

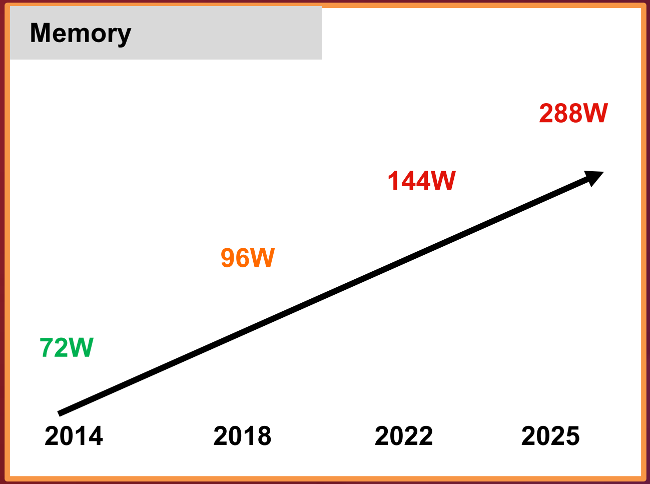


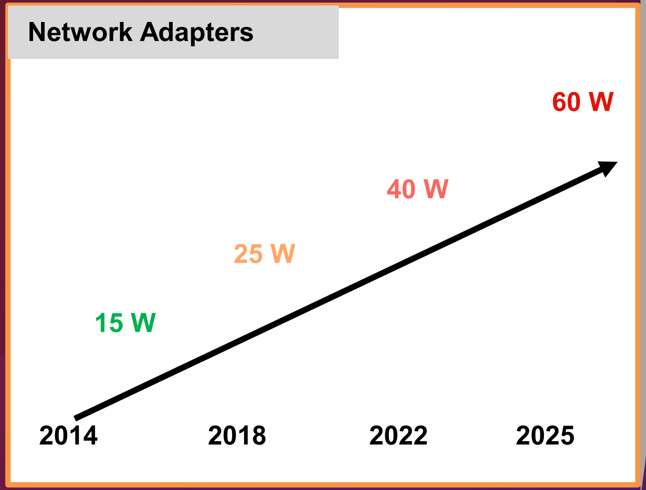


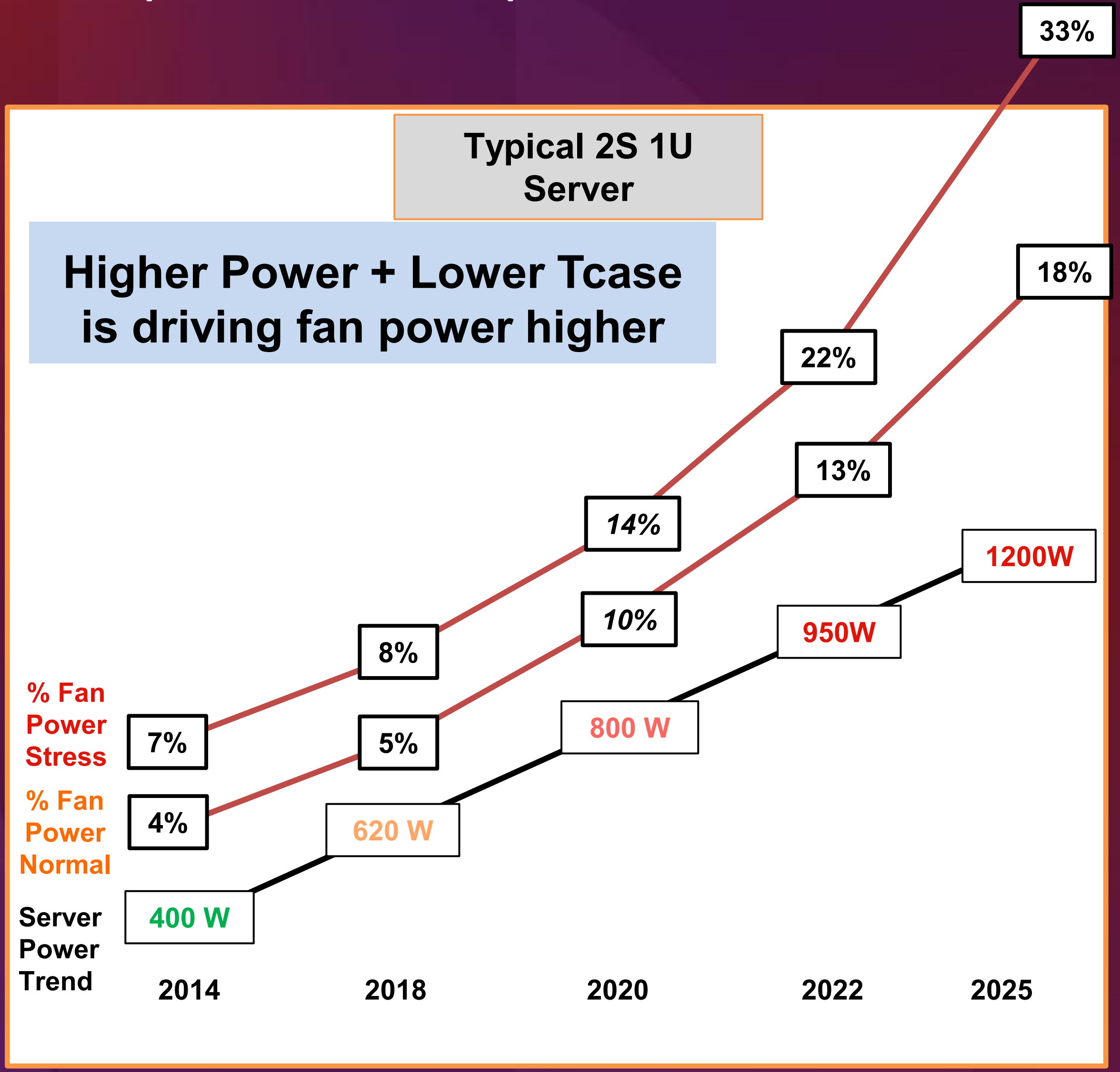
Power for cooling is growing faster than component level power





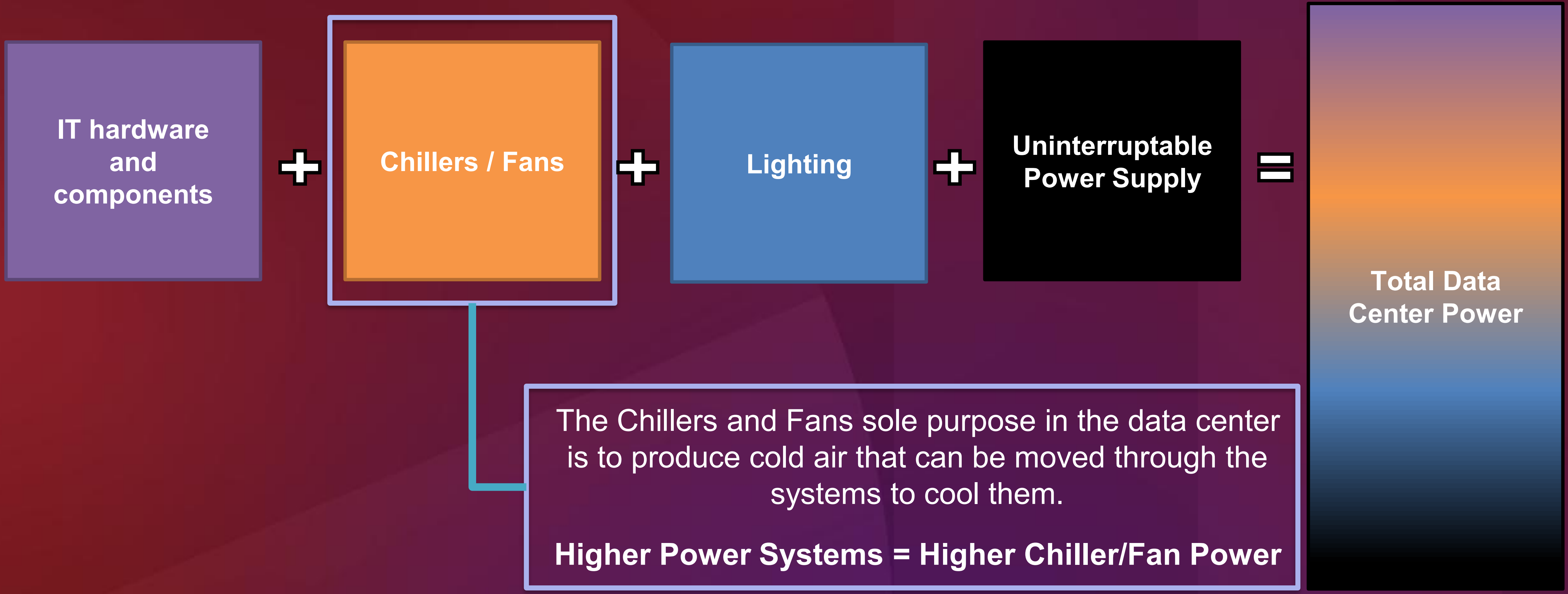






Data center power equation

Power is consumed from other sources in the data center beyond the IT hardware and components. But in the data center power equation, traditional air-cooling methods such as chillers and fans can be replaced with more efficient methods.



Power Use Efficiency (PUE)

Measures how efficiently a data center uses power

Total Power Used

÷

IT Equipment Power

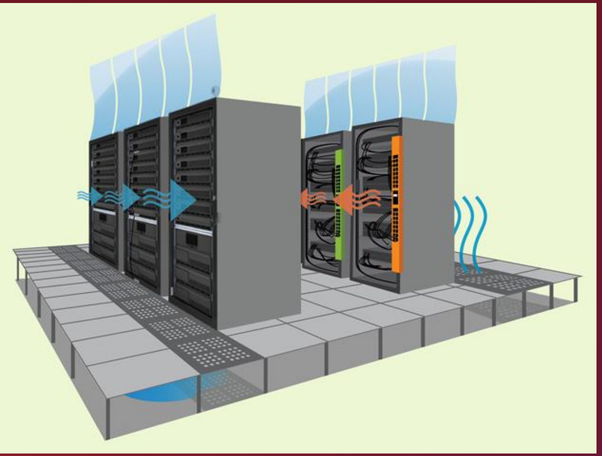
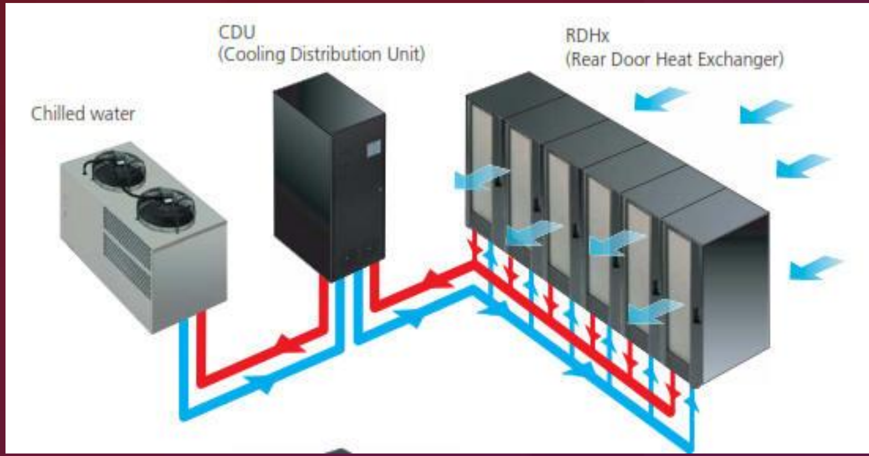
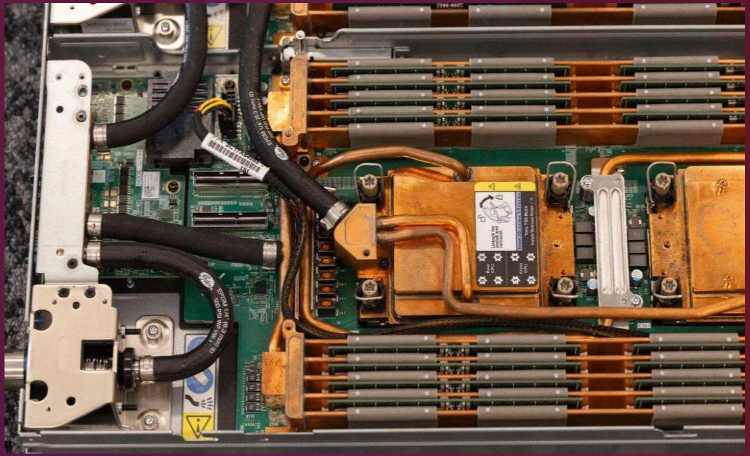
Low PUE value*

=

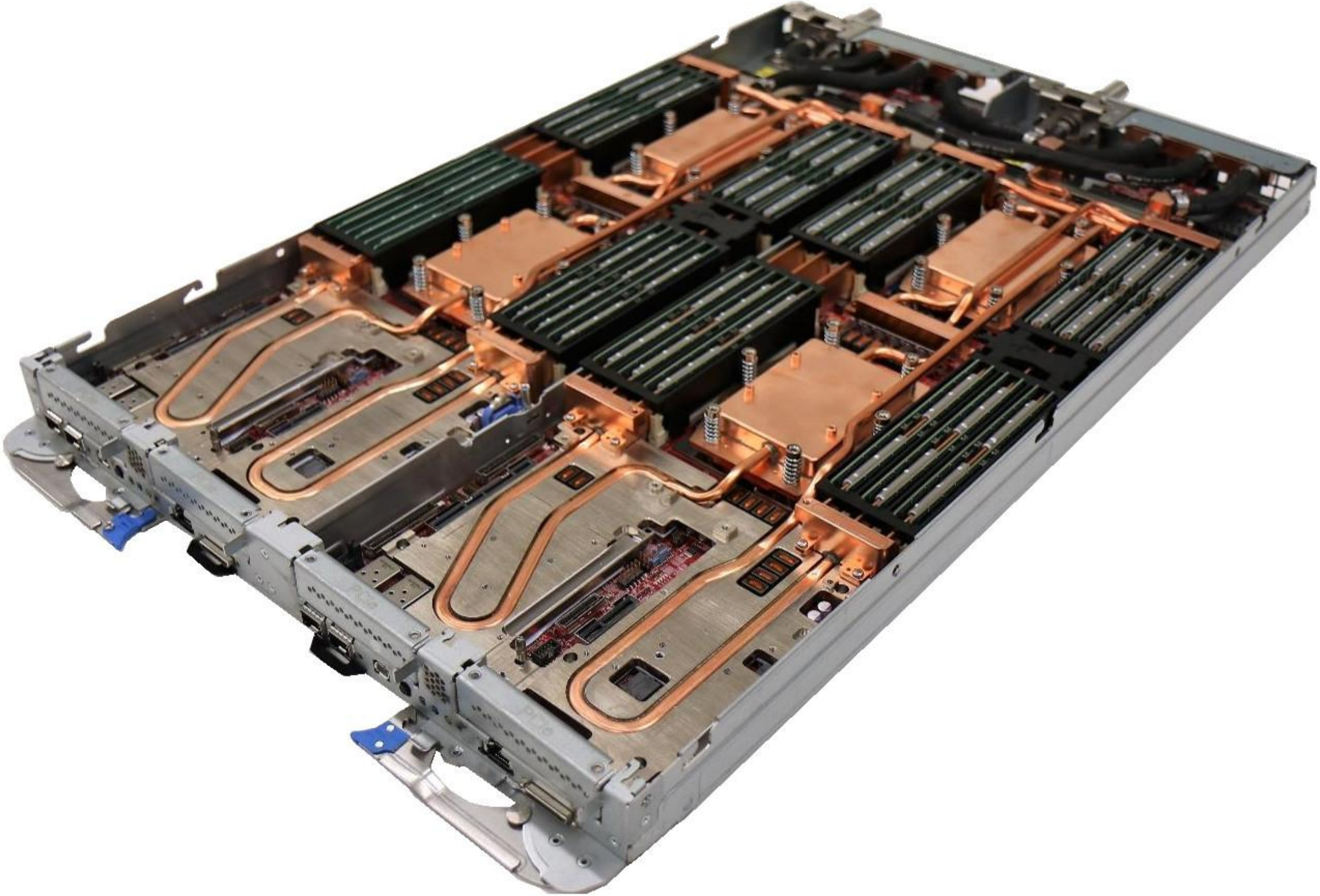
Better energy efficiency

*** PUE value of 1 means all the power going into a data center is consumed by the IT equipment.**

Data Center Cooling Topologies

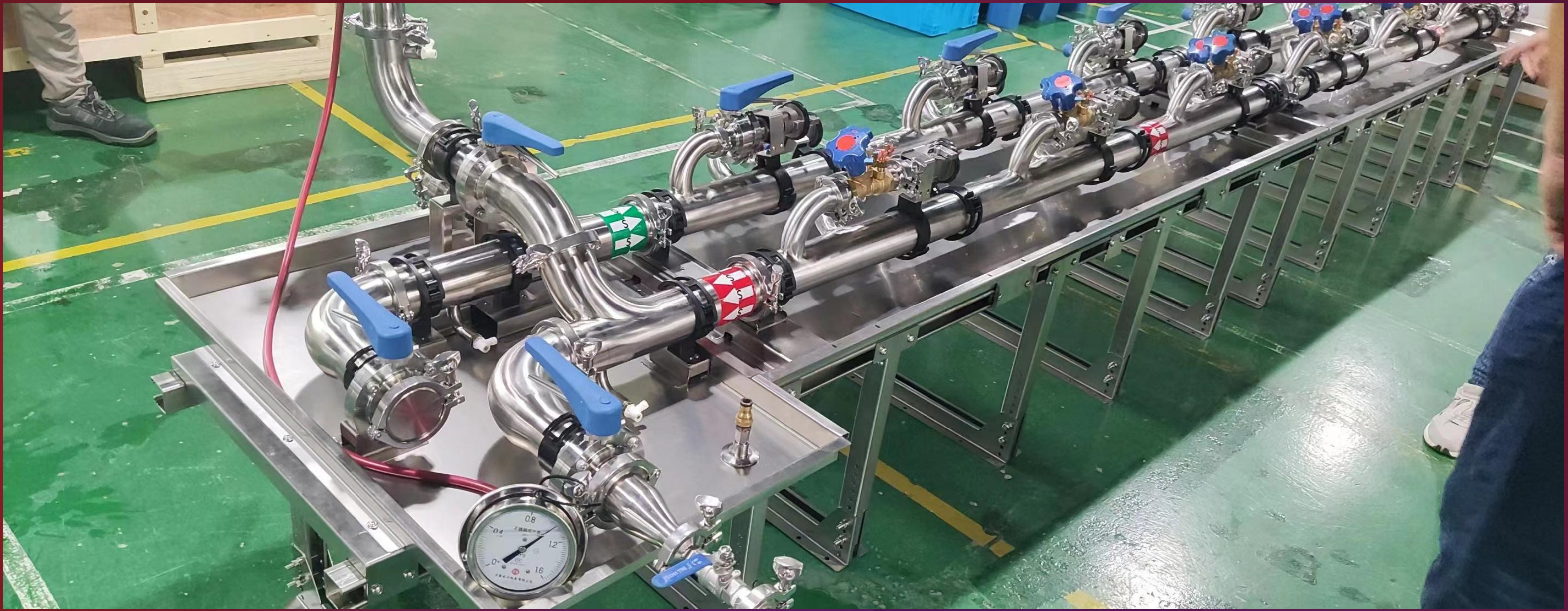
Air Cooled	Rack Level Heat Exchangers	Direct Water Cooled
		
PUE = ~ 1.5	PUE = ~1.2	PUE <= 1.1
Conventional	**Flexibility & Efficiency**	**Performance & Efficiency**

$$PUE = \frac{\text{Total Facility Energy}}{\text{IT Equipment Energy}}$$




Neptune

Data Center Cooling Topologies



AI Factory – LLM Supercomputer

Gen AI Training Server:

	Air	DWC
Total Server Power	**9.7 kW**	**8.1 kW**
GPU/CPU	**6.5 kW**	**6.5 kW**
Other IT	**1.2 kW**	**1.2 kW**
Fan Power	**2.0 kW**	**0.4 kW**
Cooling Air (PUE 1.6) *	**5.8 kW**	**1.0 kW**
Cooling Water (PUE 1.06)	**0 kW**	**0.4 kW**
Total Non-IT Power	**~ 8 kW**	**~2 kW**

* 35°C ambient, allocated fan power

- Real Case example
  - 5,000 Server w/ 40,000 GPUs / 10,000 CPUs

→ Aircooled: ~78 MW w/ 39 MW "wasted"

→ Watercooled: ~47 MW w/ 9 MW "wasted"

Designing a more sustainable data center

No/low emission power
Now more affordable
than ever before

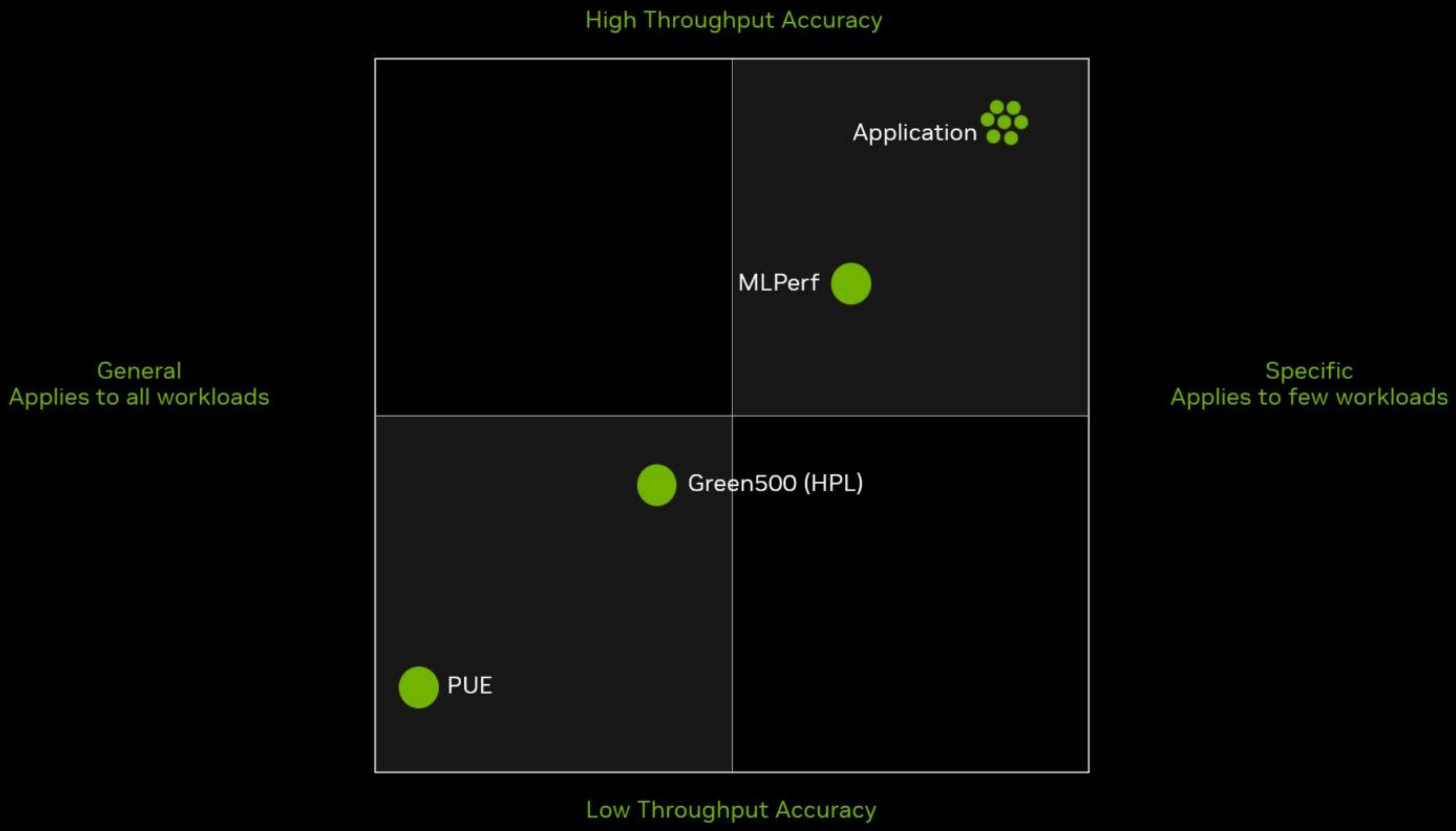
Direct liquid cooling

Liquid to liquid exchanger
Transfer heat
from IT to facilities

Facility reuse
Hot water, heat, pool

Energy Efficiency

PUE is the only the first step



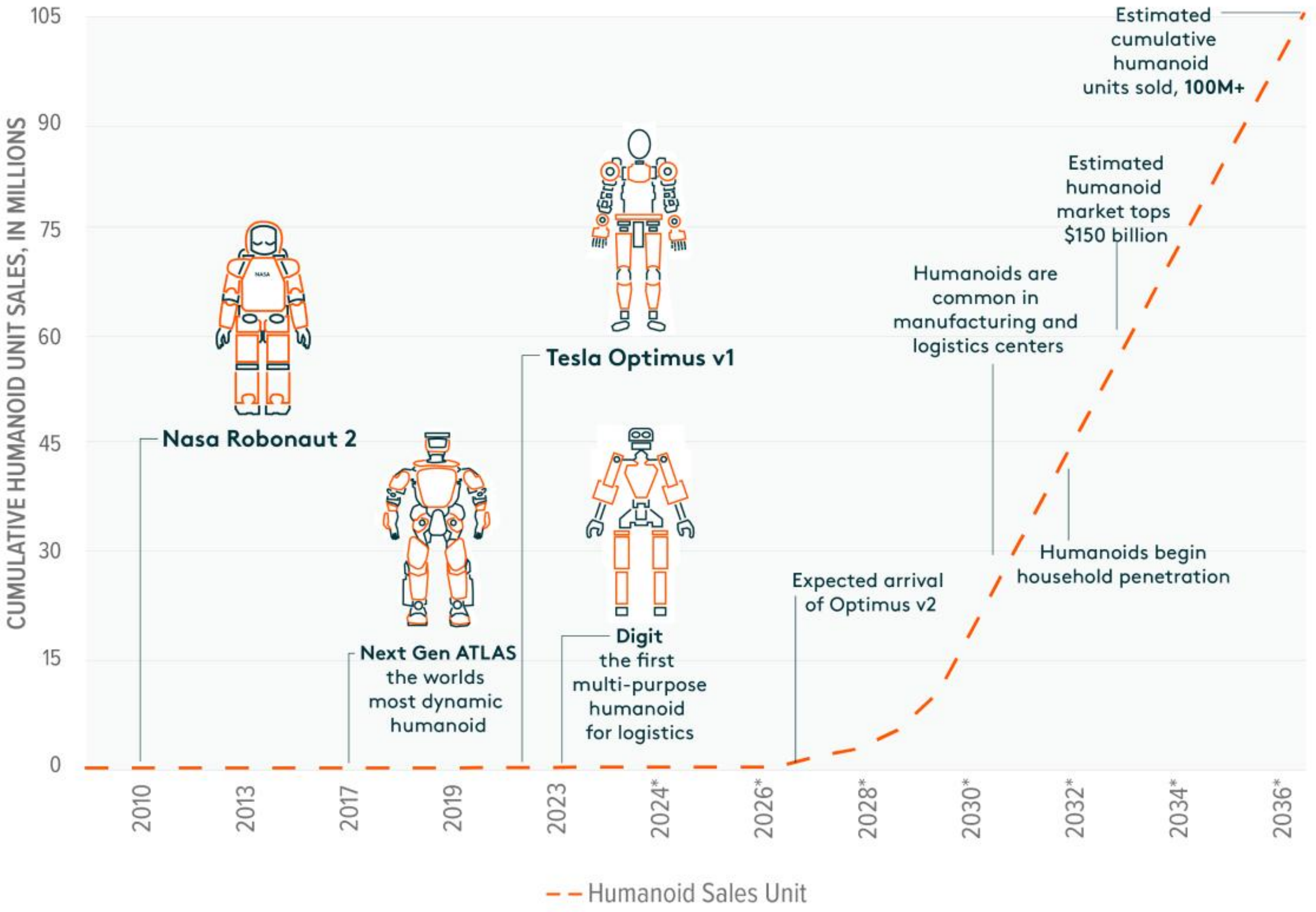
NVIDIA

Key risks associated with datacenters

Risk	Description
Cybersecurity	DDoS attacks, data theft, supply chain
Resiliency and downtime	Economic and reputational impact due to disruptions
Climate and water impact	High energy consumption, use of water for cooling
Regulation and compliance	Regulatory constraints on data and energy
Supply chain HW	Dependence on chips and rare materials

Opportunities and use cases

Opportunity	Examples and benefits
AI and HPC	Increased demand for infrastructure for training and inference
Edge computing	Low latency for IoT and AR/VR applications
Sustainability	Use of renewable energy, heat recovery
Managed and Cloud Services	Adoption of OPEX models and managed services
Operational optimization	Automation, PUE efficiency



https://www.globalxetfs.com/articles/the-rise-of-humanoids-explained

Policy recommendations

Integrate sustainability policies with renewable energy suppliers.
Design for resilience: redundancy, DR plans, continuous testing.
Invest in efficiency (PUE optimization, heat recovery).
Predict edge/hybrid strategies for latency and local compliance.
Assess supply-chain risk and diversify HW suppliers.

Strategic vision

The operator no longer provides just computing, but trust, sustainability, and operational continuity, positioning themselves as enablers of digital ecosystems.
Datacenters must be considered digital sovereignty assets, with long-term geopolitical and economic implications.

Thanks

Massimo Chiriatti
Chief Technical & Innovation Officer - Lenovo
https://www.linkedin.com/in/massimochiriatti/