HPE Labs

Beyond the Horizon:
Quantum, AI, and the tech shaping the next decade

Andrew Wheeler, HPE Fellow, SVP and Director, HPE Labs

MulticoreWorldXIII, February 2026

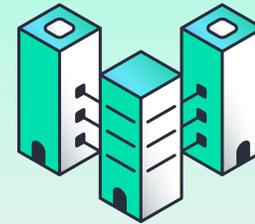# The Shape of the World in 2030

1

Agentic Enterprise

2

Beyond the Grid

3

Proof Before Trust

4

Quantum Utility

5

AI After LLMs

The Shape of the World in 2030
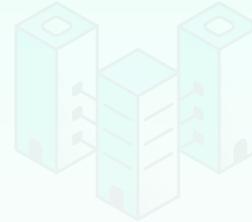
1

Agentic
Enterprise

2

Beyond
the Grid

3
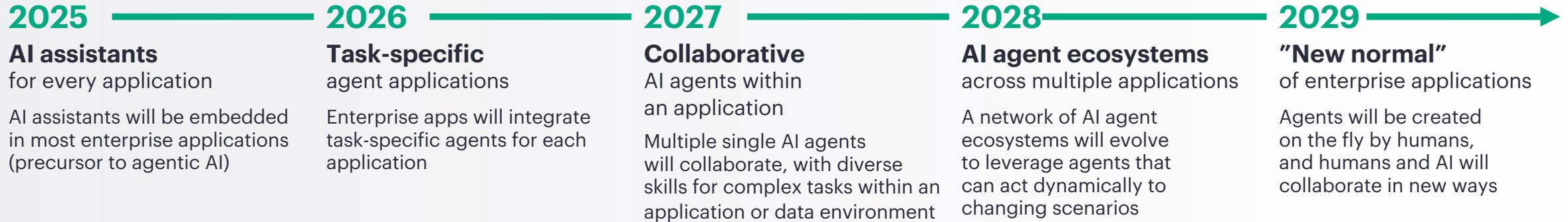
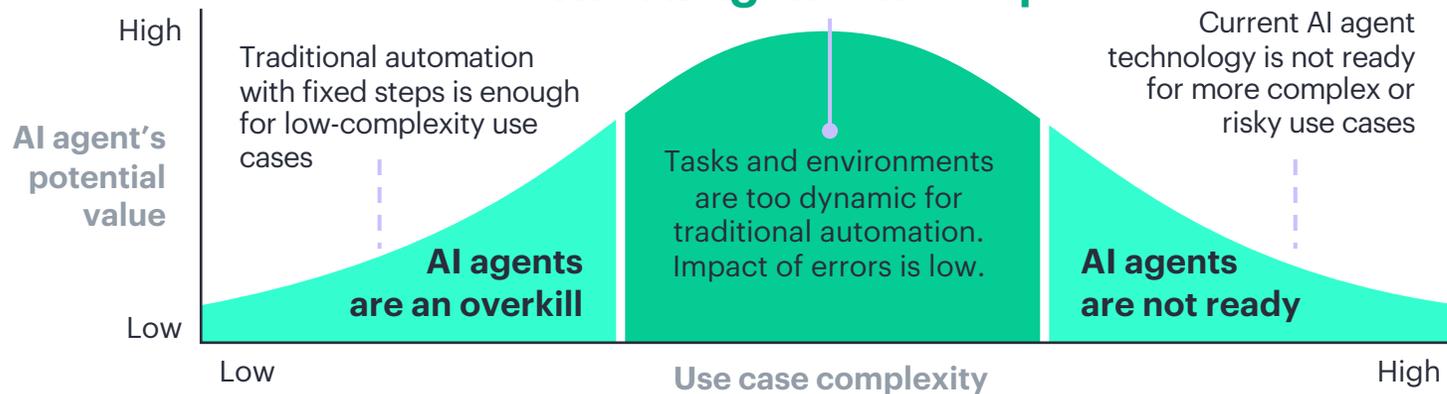Proof
Before Trust

4

Quantum
Utility

5

AI
After LLMs

Agentic AI and the hype cycle: Signal vs. Noise

The peak of inflated expectations

The future of agentic AI in enterprise applications

2025

AI assistants
for every application

AI assistants will be embedded in most enterprise applications (precursor to agentic AI)

2026

Task-specific
agent applications

Enterprise apps will integrate task-specific agents for each application

2027

Collaborative
AI agents within an application

Multiple single AI agents will collaborate, with diverse skills for complex tasks within an application or data environment

2028

AI agent ecosystems
across multiple applications

A network of AI agent ecosystems will evolve to leverage agents that can act dynamically to changing scenarios

2029

"New normal"
of enterprise applications

Agents will be created on the fly by humans, and humans and AI will collaborate in new ways

The AI agent sweet spot

High

AI agent's potential value

Traditional automation with fixed steps is enough for low-complexity use cases

Tasks and environments are too dynamic for traditional automation. Impact of errors is low.

Current AI agent technology is not ready for more complex or risky use cases

AI agents are an overkill

AI agents are not ready

Low

Low   **Use case complexity**   High

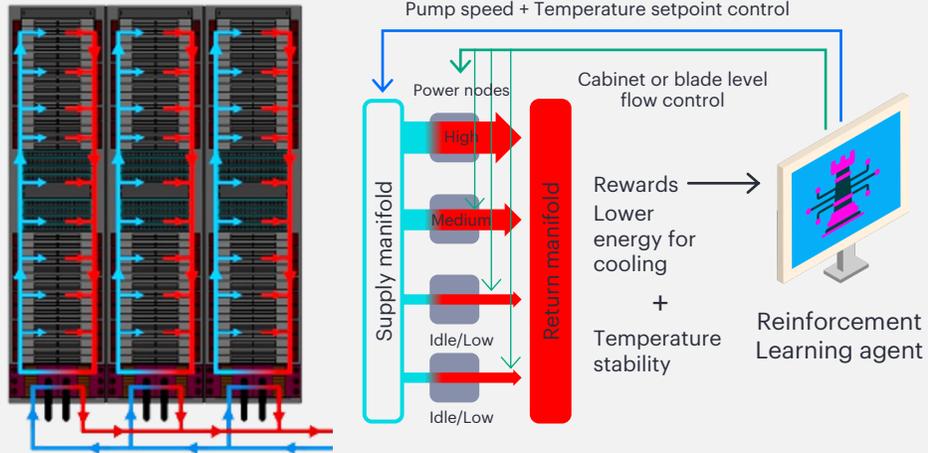By 2035, agentic AI will drive approximately

30% of enterprise application software revenue, surpassing **$450B**, up from 2% in 2025

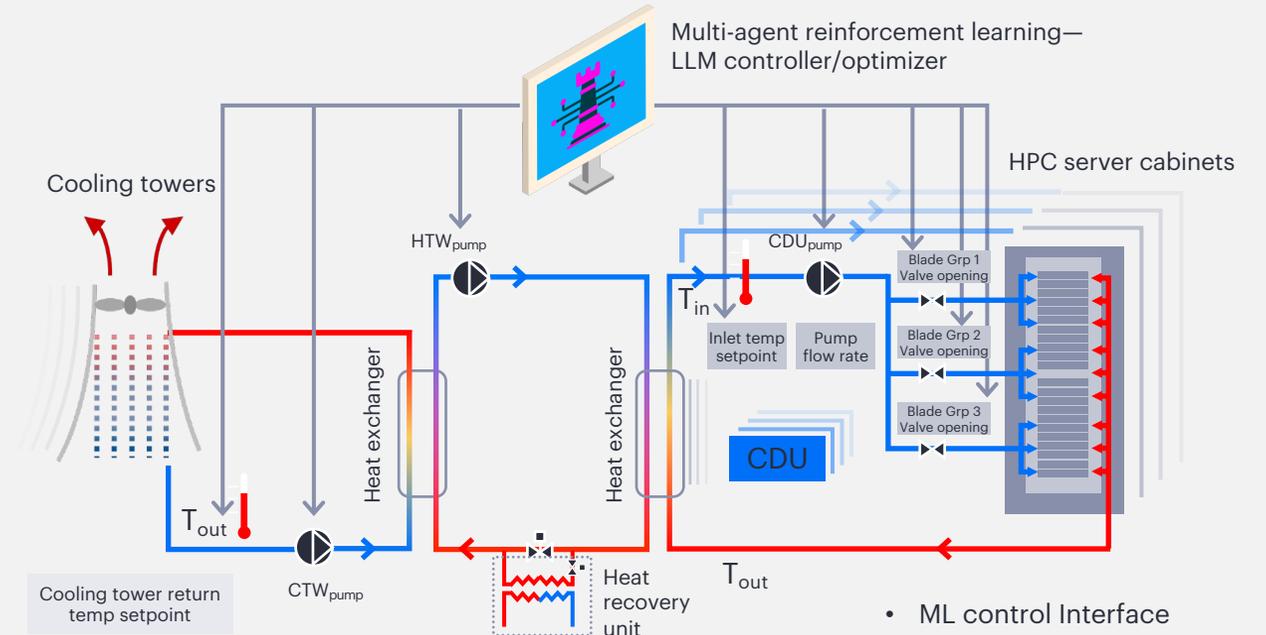Gartner predicts that over

40% of agentic AI projects will be canceled by the end of 2027

Source: Gartner (August 2025)

5

Precision liquid cooling for the AI era

Scalable and granular control of the Cooling Distribution Unit (CDU) for IT cabinets

Pump speed + Temperature setpoint control

Cabinet or blade level flow control

Power nodes

High

Medium

Idle/Low

Idle/Low

Supply manifold

Return manifold

Rewards
Lower energy for cooling
+
Temperature stability

Reinforcement Learning agent

Scalable cooling control of different components of a data center

Multi-agent reinforcement learning— LLM controller/optimizer

Cooling towers

HPC server cabinets

HTW_{pump}

CDU_{pump}

T_{in}

Blade Grp 1 Valve opening

Inlet temp setpoint

Pump flow rate

Blade Grp 2 Valve opening

Blade Grp 3 Valve opening

Heat exchanger

Heat exchanger

CDU

T_{out}

CTW_{pump}

Heat recovery unit

T_{out}

Cooling tower return temp setpoint

- High-fidelity cooling model of data center Digital Twin
- Model scales to HPC cluster

- End-to-end model spanning cooling Tower to IT cabinets and heat recovery unit

- ML control Interface
- Multi-agent Reinforcement Learning real-time control
- Optimizes energy and temperature stability

LLM-based Agentic control plane that can reason, monitor, and explain its actions in real-time
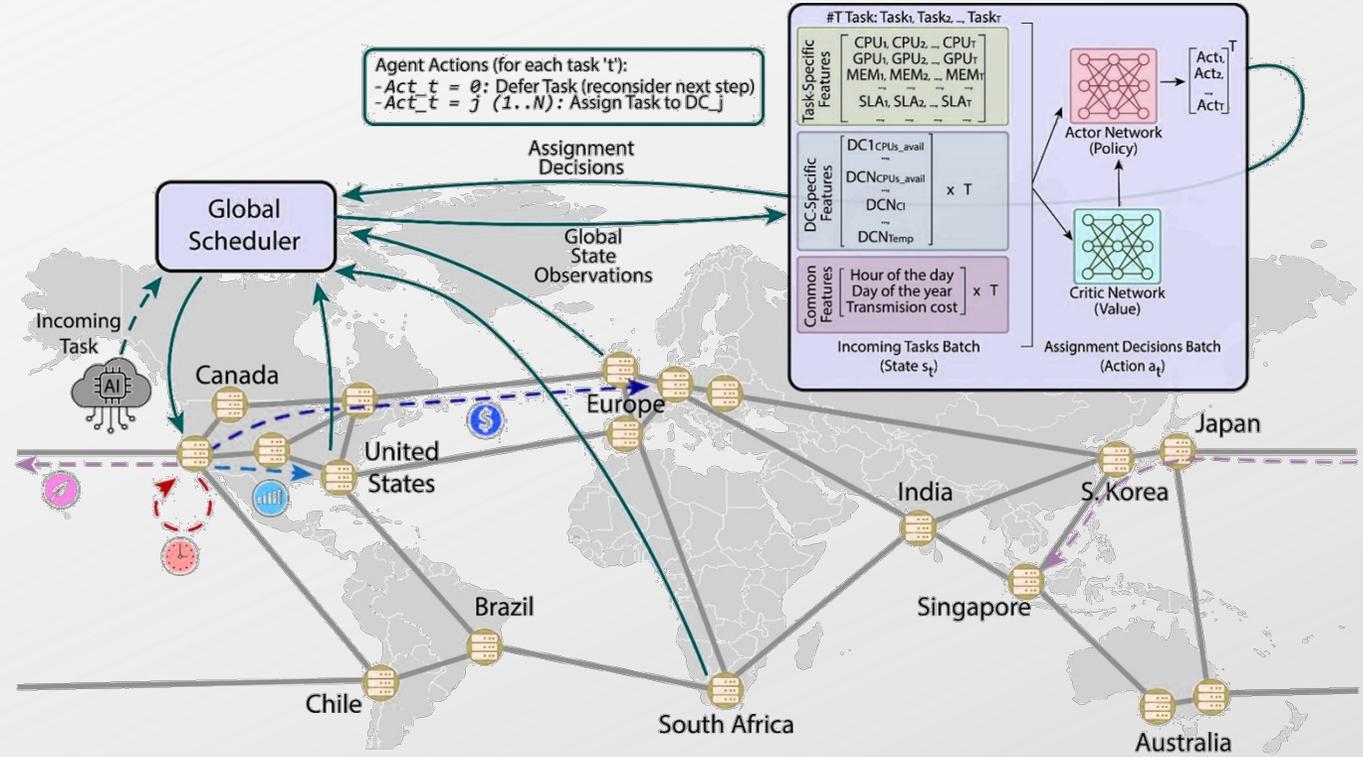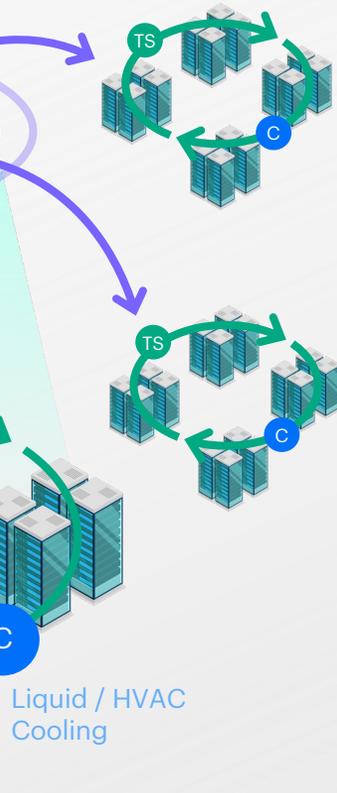
Global cloud orchestrator

Agentic AI pipeline

GreenLake Intelligence

HPE Agentic AI Mesh

Active Learning AI Agents

Agentic Liquid Cooling

Data Center Cloud Control

HPE
Evolutionary

HPE Labs
Disruptive

Learn More

NeurIPS
Liquid Cooling

NeurIPS
Data Center Cloud

9

The Shape of the World in 2030

1 Agentic Enterprise
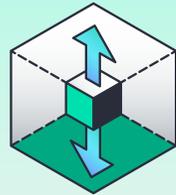
2 Beyond the Grid

3 Proof Before Trust
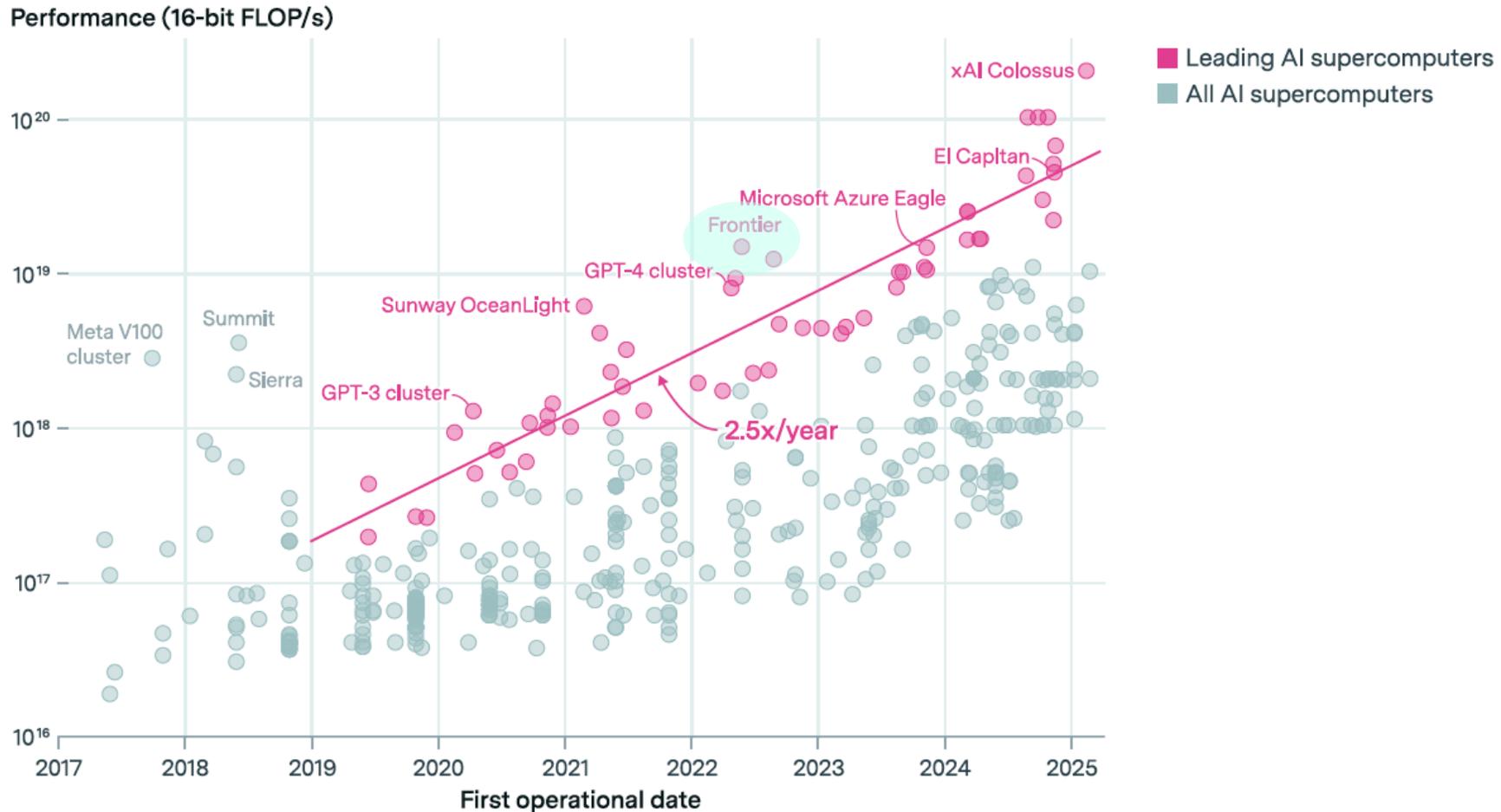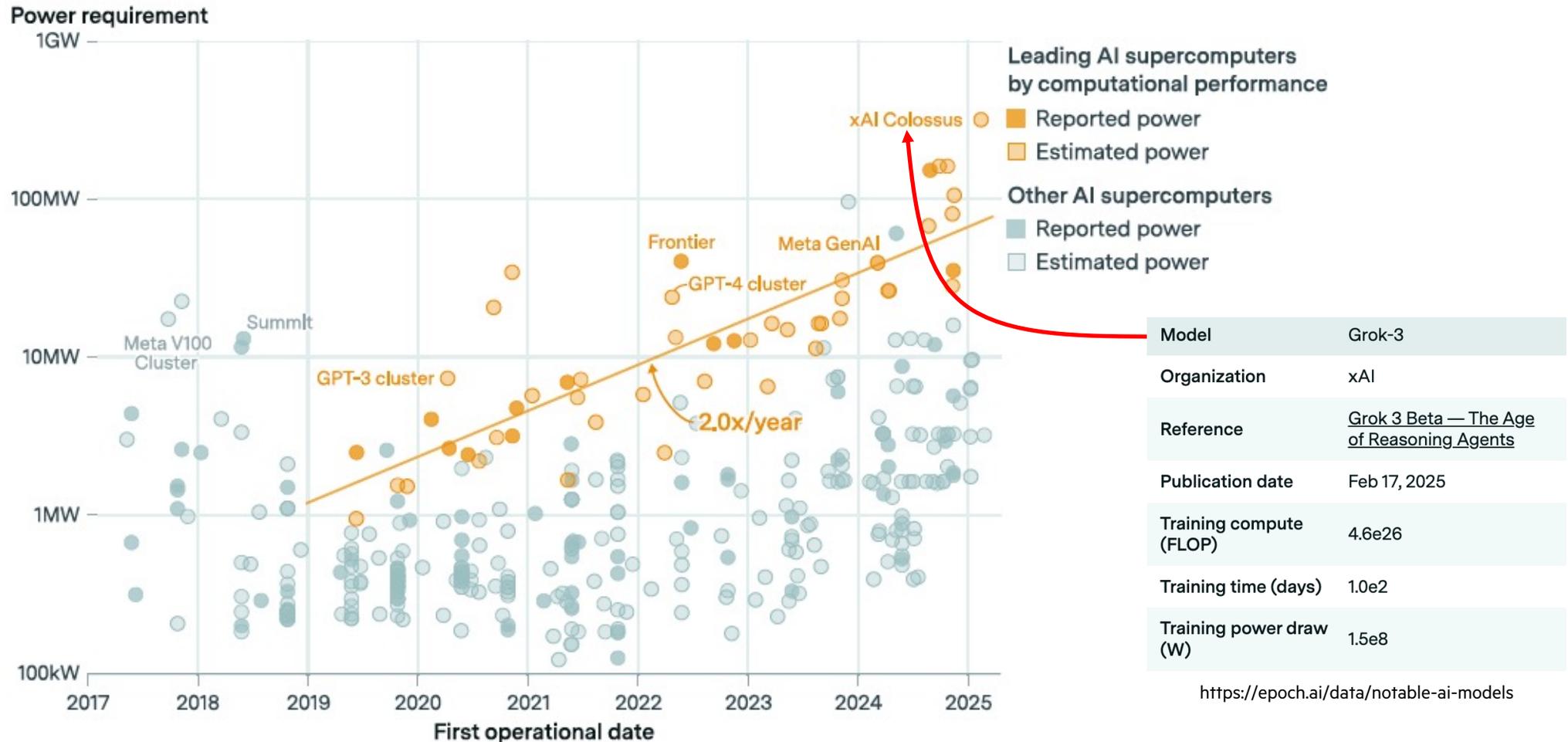
4 Quantum Utility

5 AI After LLMs

Post-Exascale Trends - Performance

Computational performance of AI supercomputers

Pilz, K., Sanders, J., Rahman, R., and Lennert H. Trends in AI Supercomputers, 2025

Post Exascale Trends - Power

Power requirement of AI supercomputers

Model	Grok-3
Organization	xAI
Reference	Grok 3 Beta — The Age of Reasoning Agents
Publication date	Feb 17, 2025
Training compute (FLOP)	4.6e26
Training time (days)	1.0e2
Training power draw (W)	1.5e8

https://epoch.ai/data/notable-ai-models

Pilz, K., Sanders, J., Rahman, R., and Lennert H. Trends in AI Supercomputers, 2025

United States Data Center Energy Usage Projections

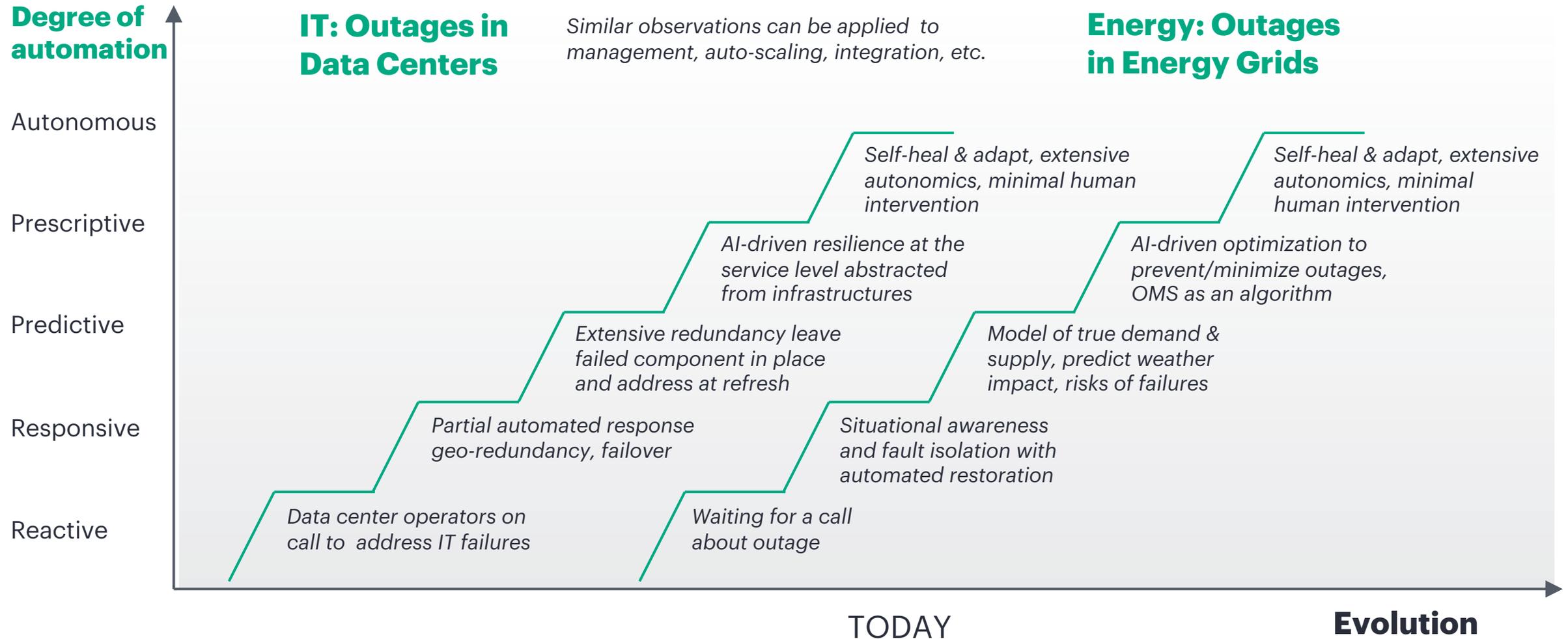Total U.S. data center electricity use from 2014 through 2028
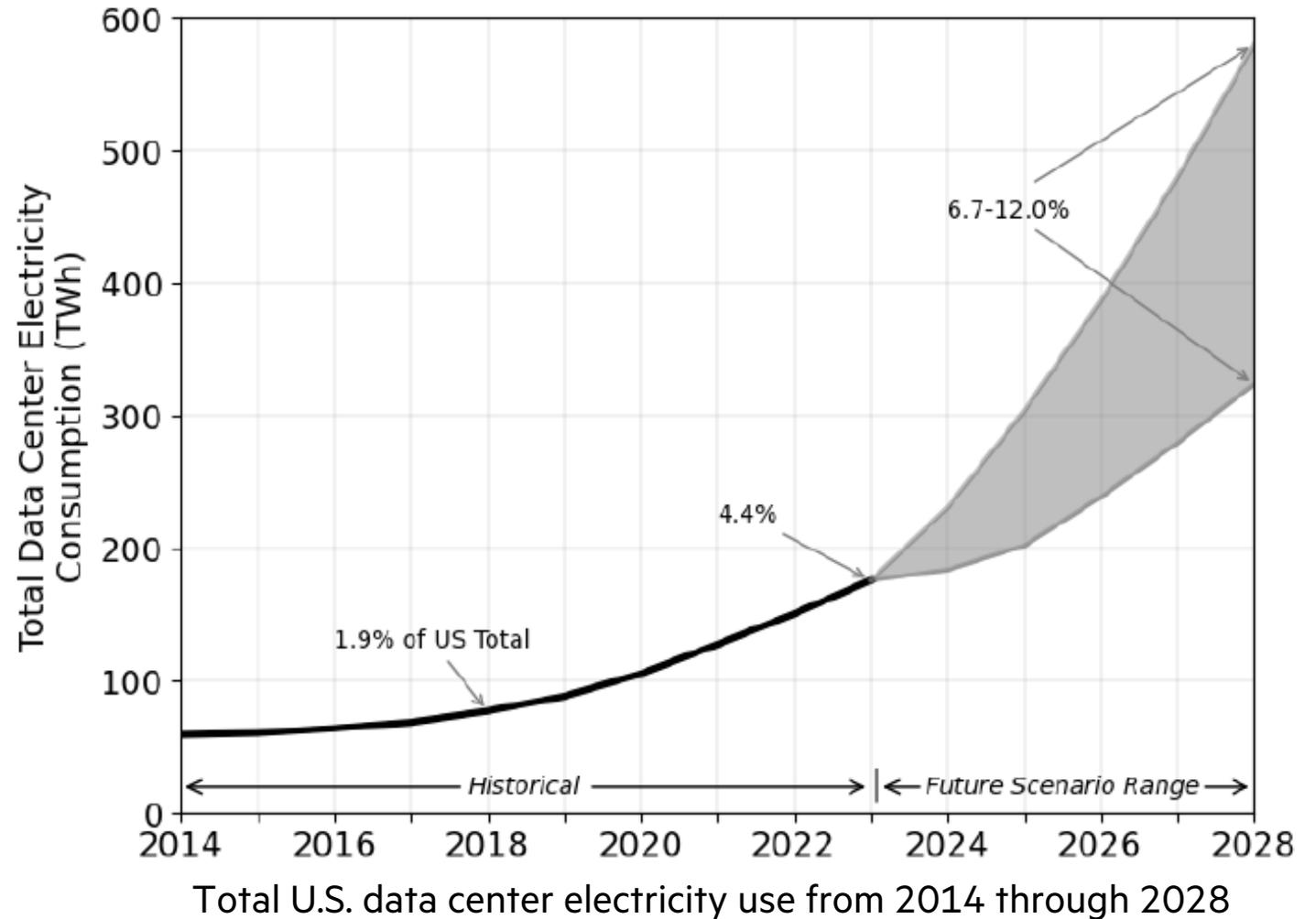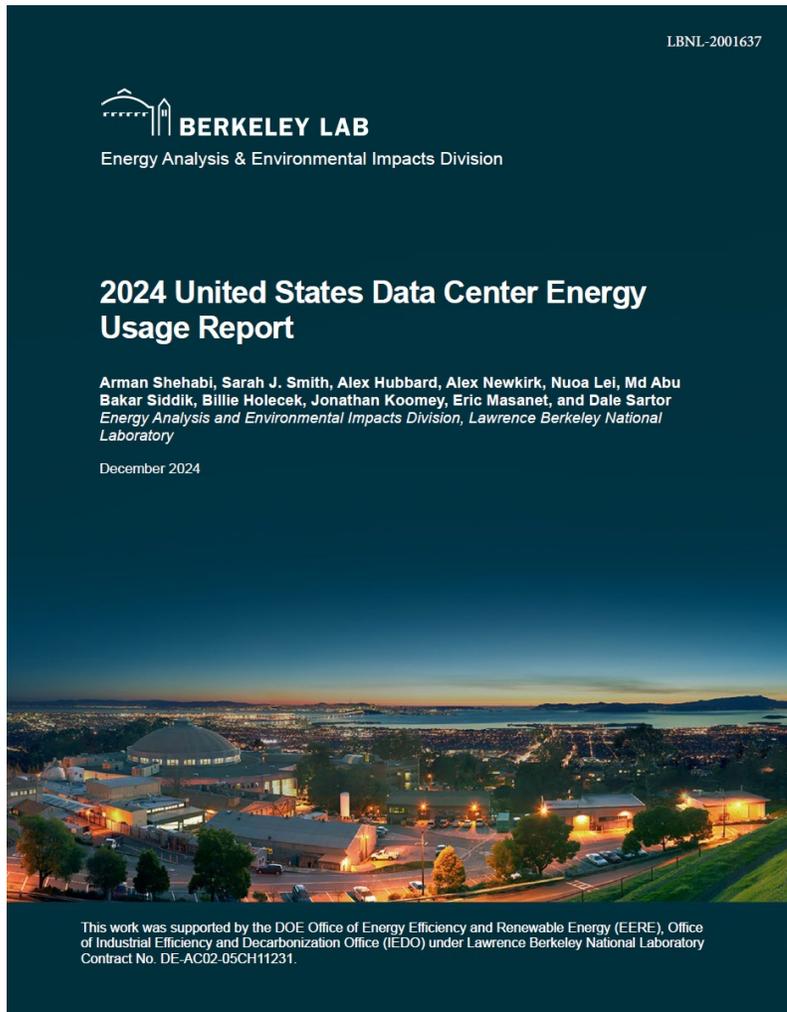
Shehabi, A., Smith, S.J., Hubbard, A., Newkirk, A., Lei, N., Siddik, M.A.B., Holecek, B., Koomey, J., Masanet, E., Sartor, D. 2024. 2024 United States Data Center Energy Usage Report. Lawrence Berkeley National Laboratory, Berkeley, California. LBNL-2001637

The Problem (and Solution space)
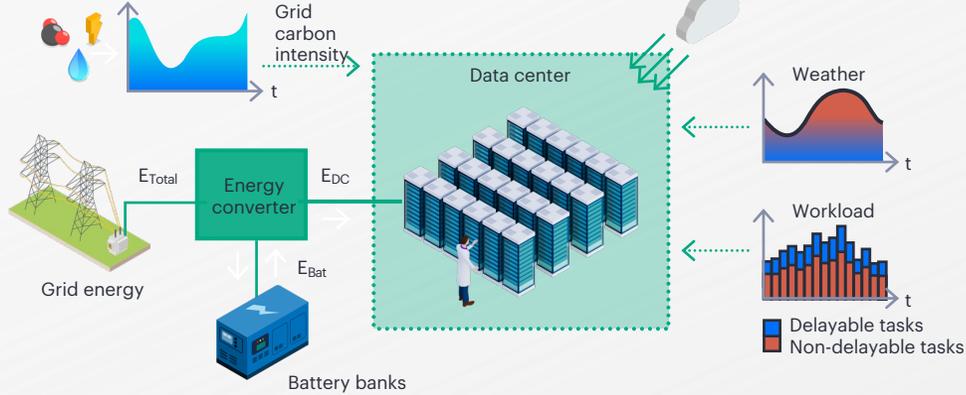
Data center energy flow

4 **Minimize water consumption**

Customer or HPE data center

PUE_{Power} $ITUE_{Power}$ **Compute performance** $ITUE_{Cooling}$ $PUE_{Cooling}$

HPE IT equipment

From building primary AC feed → To DC logic level on silicon

Maximize productive capability of hardware

From heat generated on silicon → To heat leaving the building

6 Enable responsible energy sourcing

2 Minimize losses in the supply of electrical energy

1 Maximize productivity of system

3 Minimize energy burden for heat removal

5 Maximize energy reuse

Evolving outage management: IT and energy grids

Degree of automation

- Autonomous
- Prescriptive
- Predictive
- Responsive
- Reactive

IT: Outages in Data Centers

Similar observations can be applied to management, auto-scaling, integration, etc.

Energy: Outages in Energy Grids

Self-heal & adapt, extensive autonomics, minimal human intervention

AI-driven resilience at the service level abstracted from infrastructures

Extensive redundancy leave failed component in place and address at refresh

Partial automated response geo-redundancy, failover

Data center operators on call to address IT failures

Waiting for a call about outage

Situational awareness and fault isolation with automated restoration

Model of true demand & supply, predict weather impact, risks of failures

AI-driven optimization to prevent/minimize outages, OMS as an algorithm

Self-heal & adapt, extensive autonomics, minimal human intervention
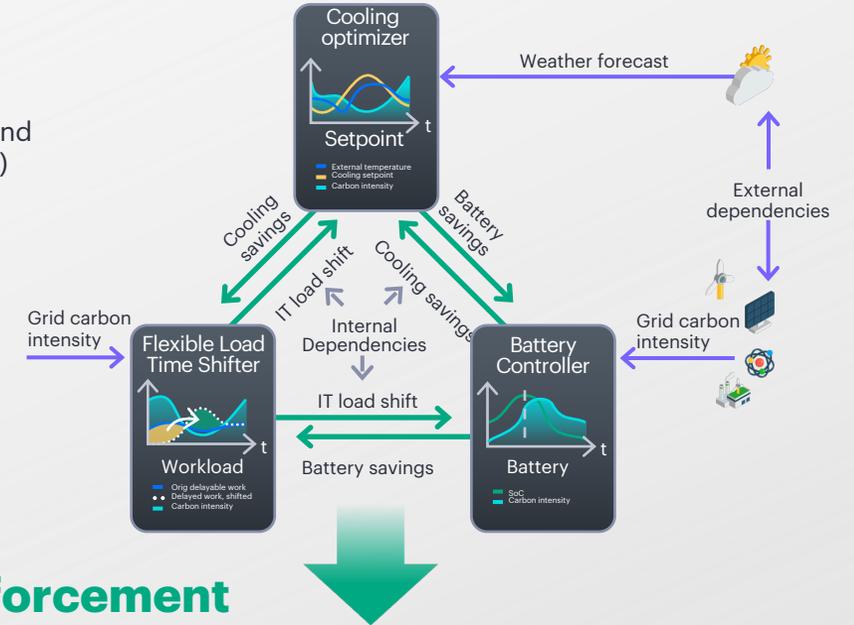
TODAY

Evolution

15

Holistic data center intelligence
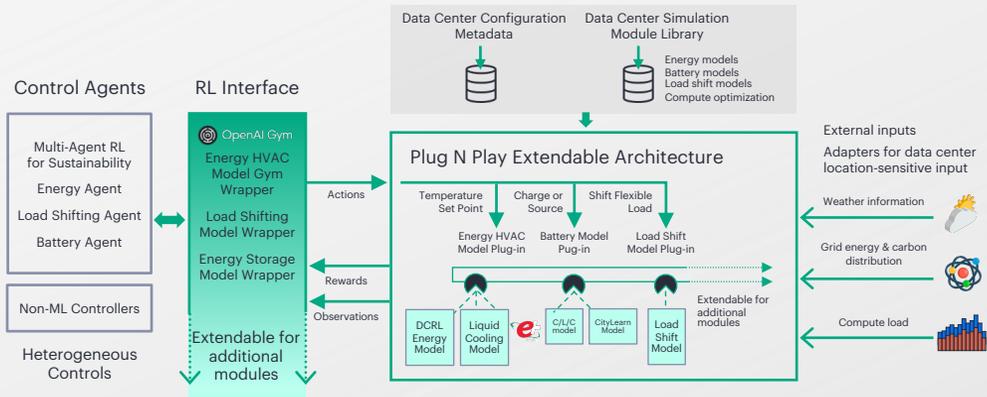

Physical system
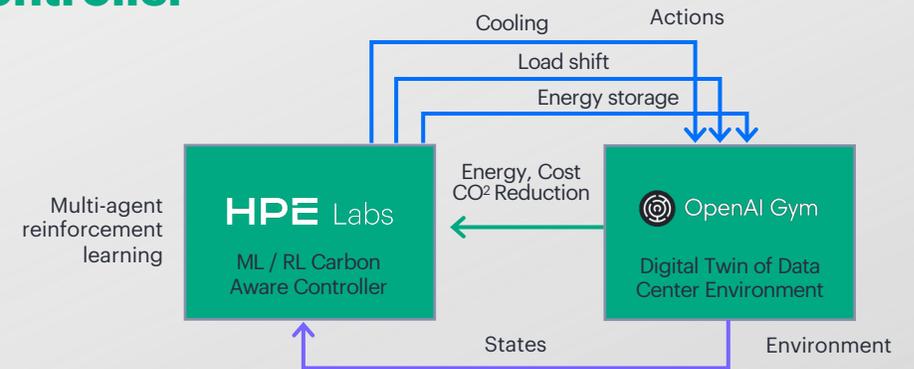
Digital system
Resolving internal and external dependencies for the cooling, and scheduling (and energy storage)

SustainDC RL Environment Framework

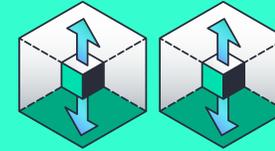**Multi-agent reinforcement learning controller**

Beyond the Grid pipeline

| Direct liquid cooling | Sustainability Insight Center | Holistic Power Management | Digital Twins | Next-Gen Grid |

HPE
Evolutionary

HPE Labs
Disruptive

Learn More

Convergence of Energy and IT Sectors

The Shape of the World in 2030
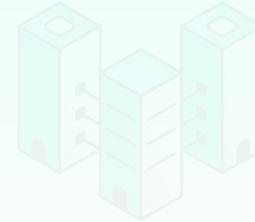
1	2	3	4	5
Agentic Enterprise	Beyond the Grid	**Proof Before Trust**	Quantum Utility	AI After LLMs

The Problem

The challenge of end-to-end data protection and verification

Remote exploit

Physical access

Storage

Network

At rest

Compute

In transit

Encryption

In use

Malicious or compromised insider
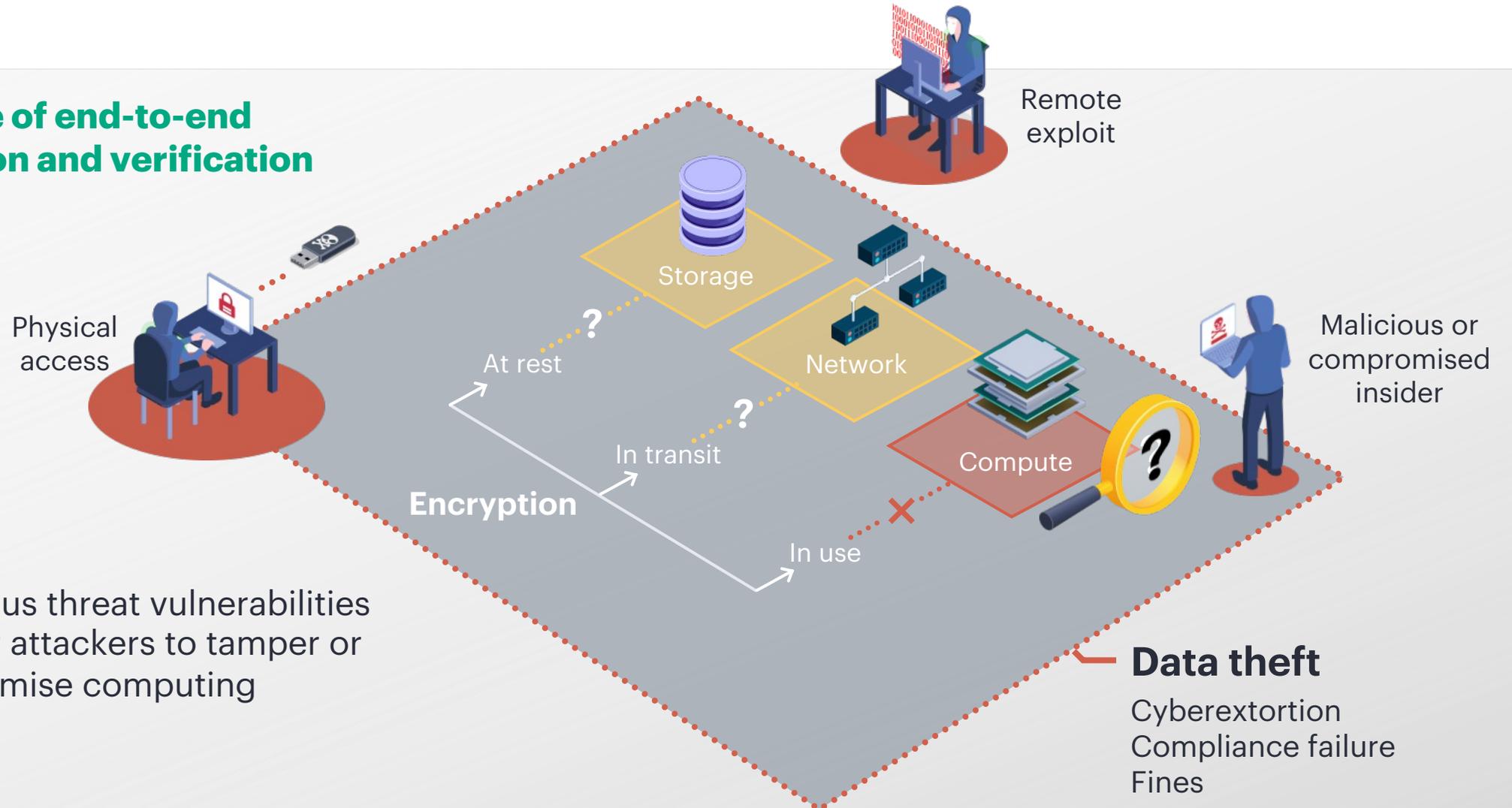
Numerous threat vulnerabilities exist for attackers to tamper or compromise computing

Data theft
Cyberextortion
Compliance failure
Fines

What we're solving with Confidential Computing

The challenge of end-to-end data protection and verification
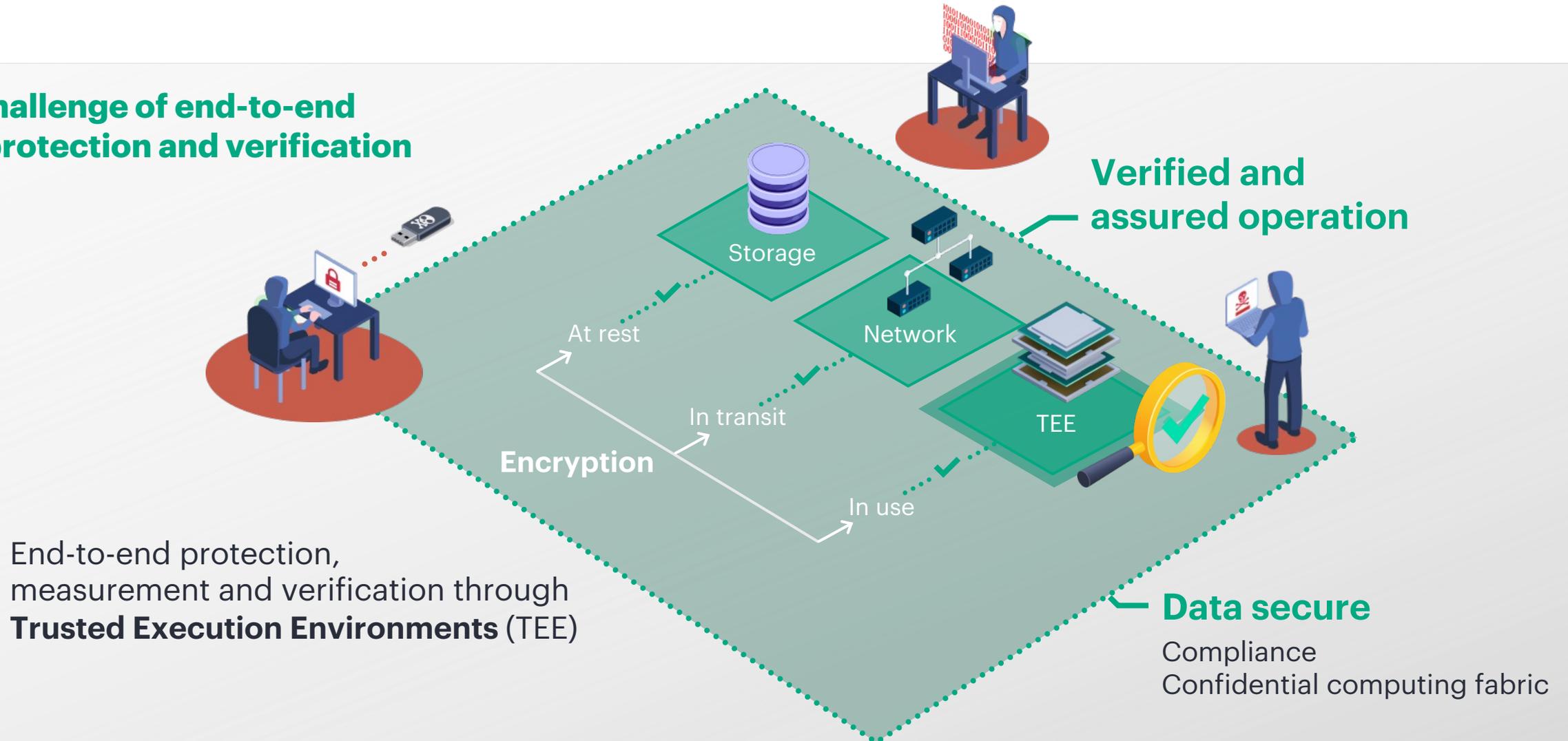
Storage

Network

TEE

At rest

In transit

Encryption

In use

Verified and assured operation

End-to-end protection, measurement and verification through **Trusted Execution Environments** (TEE)

Data secure

Compliance
Confidential computing fabric

The Shape of the World in 2030

1	2	3	4	5
Agentic Enterprise	Beyond the Grid	Proof Before Trust	**Quantum Utility**	AI After LLMs

The journey to Quantum Computing

**Multi-platform simulation engine** with adaptive circuit knitting

Hybrid HPC/ Quantum scaling

Large-scale quantum computing

Simulation engines

Small scale simulation engine

Large scale simulation engine

Quantum computing today

New materials research

Fundamental physics

Accelerators

FPGAs

GPUs

CPUs

Fundamental physics

New materials research

Advanced cryptography

Drug discovery

Today

+5 years

+10 years

Heterogenous computing development

Unlocking the promise of Quantum requires a change in the pace of development

Commentary

- Reaching utility-scale by 2033 requires **increasingly unrealistic scaling trajectories**

- **Many orders of magnitude** separate the qubit counts of today and the counts necessary for tomorrow's utility-scale machines

- **Millions of phys qubits will be required** Standalone quantum systems will continue struggling

- **Heterogeneous solutions are the answer** Connect multiple quantum processors with traditional HPC software and hardware
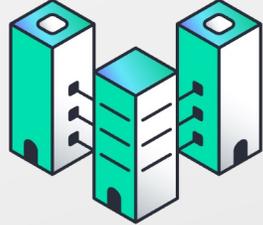

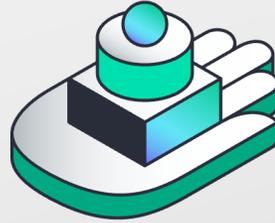Expectations for Scaling Standalone Quantum Computers
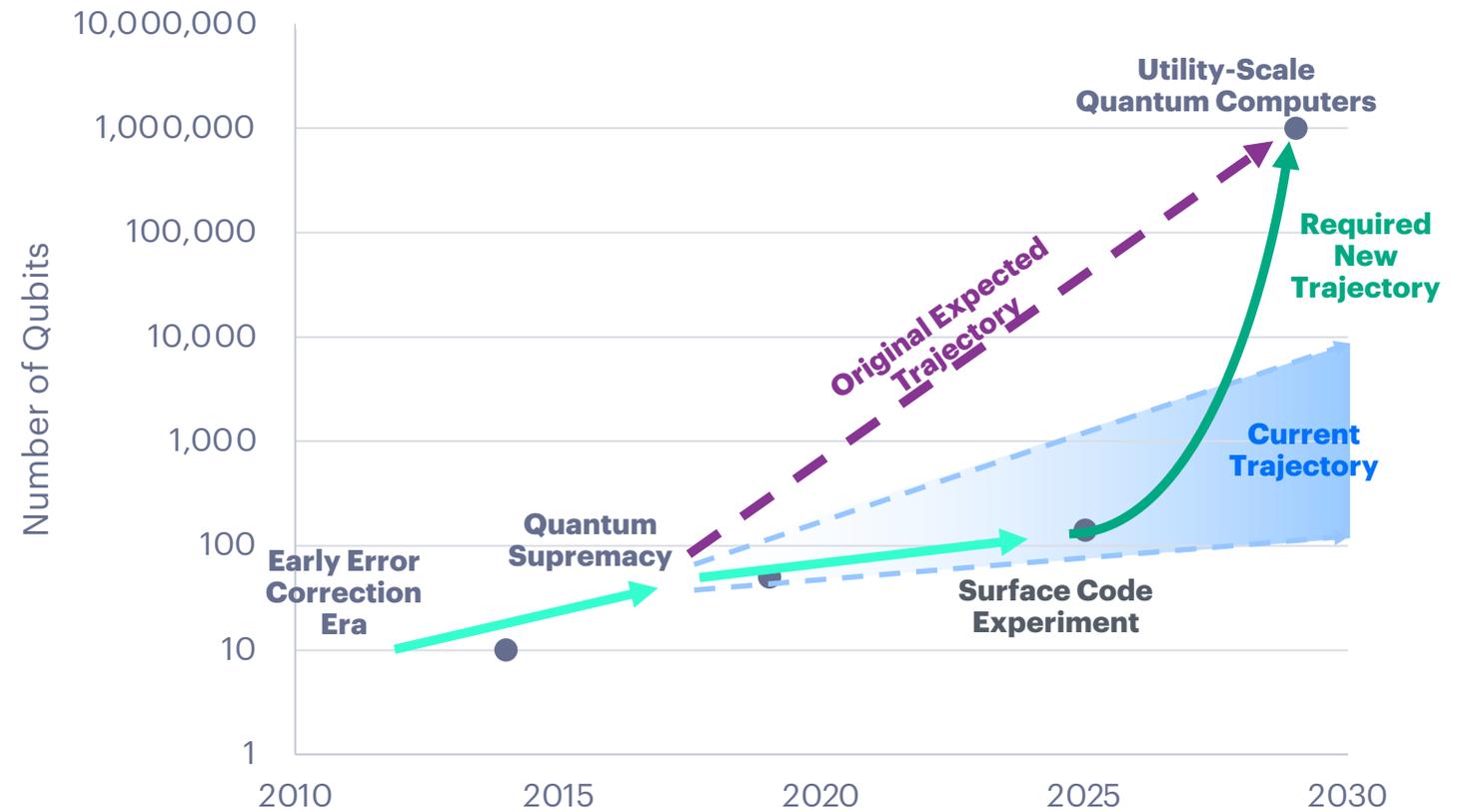
Unlocking Quantum Computing at scale
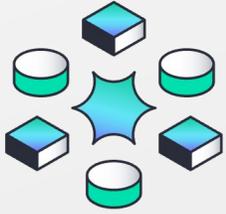
NETWORKING + **HPC/AI** + **QUANTUM DEV ENVIRONMENT**
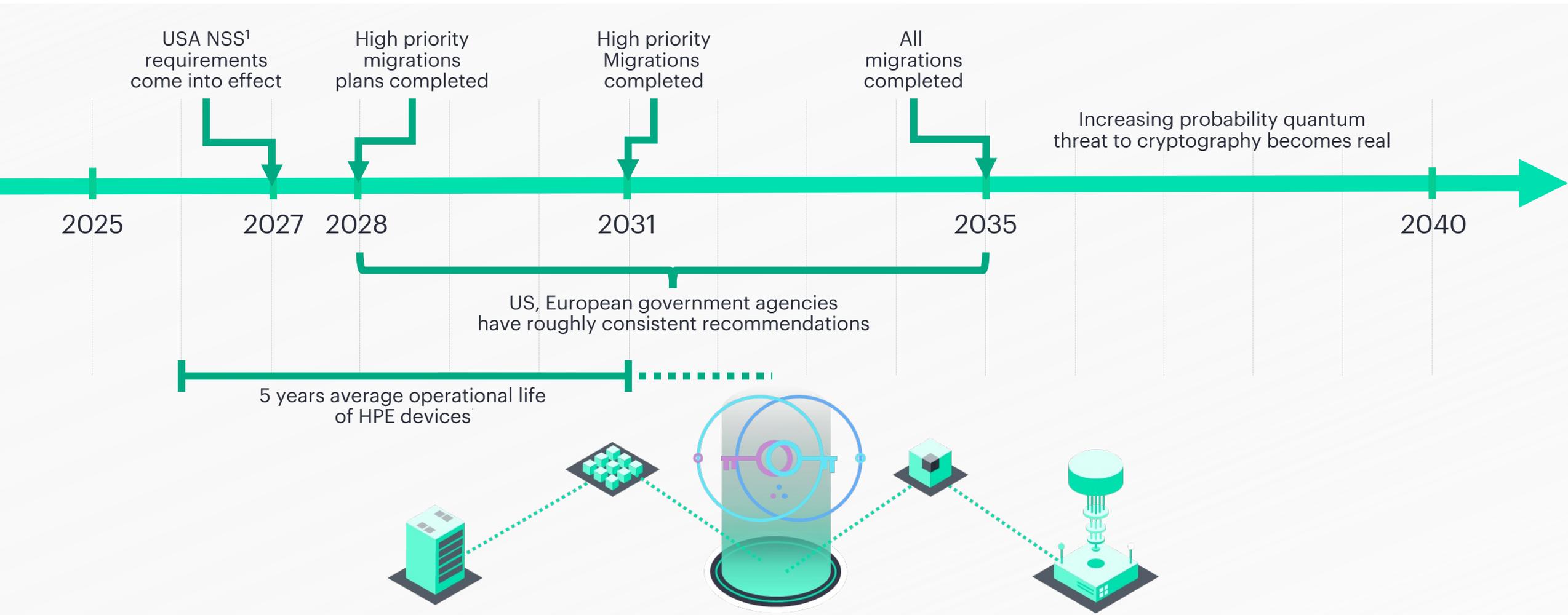
HPE Labs

Learn More

Quantum Scaling Alliance

How to Build a Quantum Supercomputer

Post Quantum Cryptography

Customer PQC migration timeline

USA NSS[1] requirements come into effect

High priority migrations plans completed

High priority Migrations completed

All migrations completed

Increasing probability quantum threat to cryptography becomes real

2025　　　2027　2028　　　　2031　　　　　　2035　　　　　　2040

US, European government agencies have roughly consistent recommendations

5 years average operational life of HPE devices

The Shape of the World in 2030

1
Agentic
Enterprise

2
Beyond
the Grid

3
Proof
Before Trust

4
Quantum
Utility

5

AI
After LLMs

Current limitations of LLMs

Practical limitations	Computational limitations	Fundamental limitations
• **Limited access to high-quality data** • **Cost of pre-training, fine-tuning** (RLHF) , and inference • Limited access to GPU	• **Quadratic scaling of compute** with number of model parameters in pre-training • **Combinatorial hardness** and latency in inference (for multimodal data)	• **Trustworthiness** • Alignment with human/customer values • **Lack of ability for complex reasoning** (compositional gap persists with scaling) • Lack of understanding emergence skills

- **Lack of complex reasoning** significantly limits LLM applications to decision making, planning, analytical thinking, scientific reasoning
- **Black-box nature of LLM** limits applications that require guarantees, confidence interval, transparency, security

# Why are LLM's limited for complex reasoning problems?

Insights from "The Illusion of Thinking" by Apple (2025)

Three regimes as we increase the complexity of the problem (tower of Hanoi in this case)

For simple problems, LLM (Large Language Model) match or event beats the LRM (Large Reasoning Model based on RL)

In the intermediate, LRM opens a gap in performance and outperforms the LLM

For sufficiently complex tasks both completely fail

Claude 3.7 Sonnet (+thinking)

Claude 3.7 Sonnet

Accuracy (%)

Complexity (number of disks)
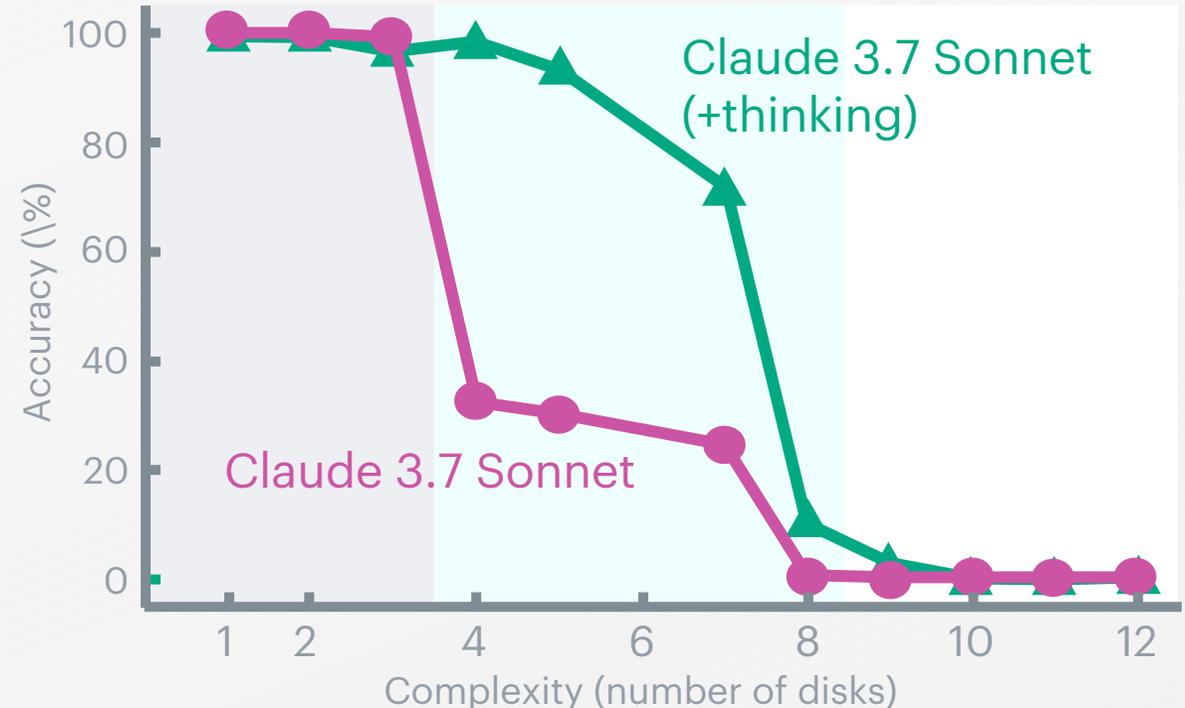
Modern LRMs often present an **illusion of reasoning** — fluent explanations without scalable, consistent problem-solving

How are we addressing the limitations of LLMs?

Daniel Kahneman (Nobel laureate in Economic Sciences 2002):

"System 1 operates automatically and quickly, with little or no effort and no sense of voluntary control."

"System 2 allocates attention to the effortful mental activities that demand it, including **complex computations.**"

How are we addressing the limitations of LLMs?

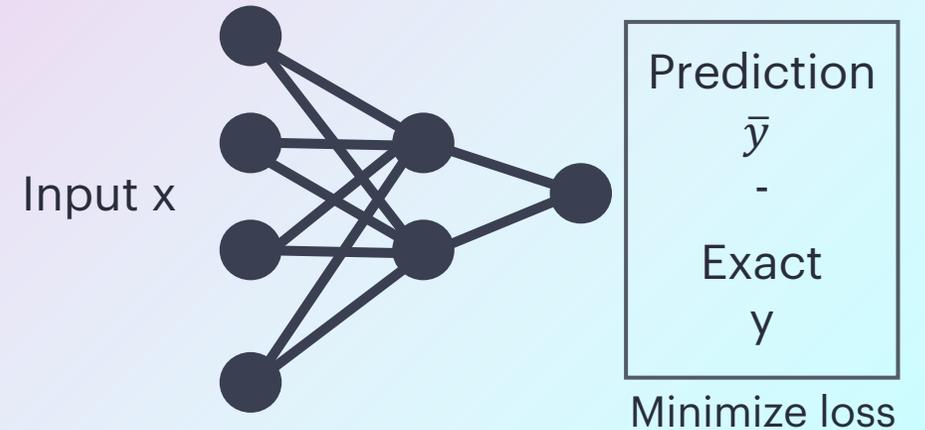## Augmenting LLMs with system 2 thinking

System 1 =
LLM
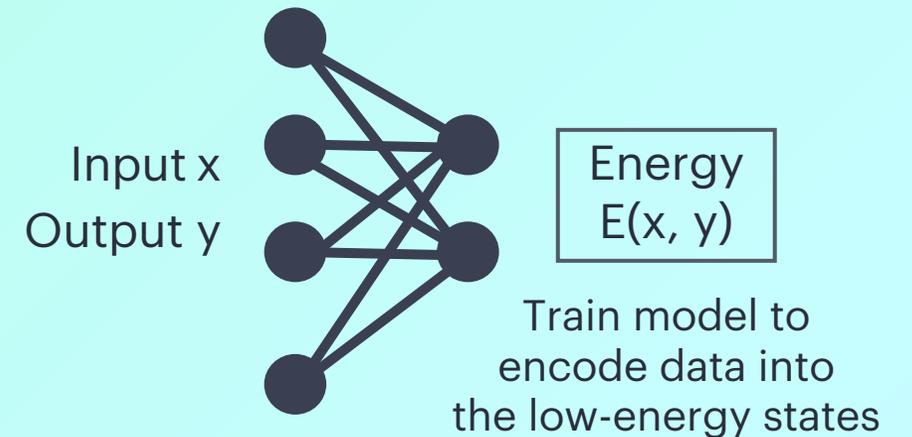
System 2 =
Energy Based Model (EBM)

Energy based models Benefits

- **Unified framework** for unsupervised, supervised, or self-supervised learning

- More compact models (high sample quality)

- **Quadratically less compute** than existing LLM as input size increases (linear scaling with tokens)

- **Confidence interval** for generative models (reliability and alignment, similar to Bayesian leaning)

- Interpretable/explainable emergence properties

- Directly can be trained or fine-tuned for reasoning and planning tasks

- Applicable to causal learning

- Generally hard to train (focused work at Labs)

Feed-forward model

Input x

Prediction $\bar{y}$ - Exact y

Minimize loss

Energy-based model

Input x
Output y

Energy $E(x, y)$

Train model to encode data into the low-energy states

Emergent Machine Intelligence

Physics-inspired heterogeneous HPC for scientific computing, complex reasoning, and building world models

Integration of quantum accelerators

Scaling by integrating classical and quantum systems
to harness diverse accelerators that maximize run-time, efficiency, sustainability

Full-stack quantum supercomputer
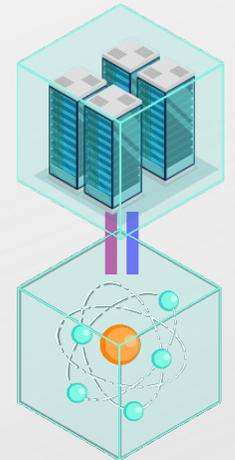For efficient quantum simulations and quantum machine learning

Distributed quantum algorithms and applications
Toward utility scale quantum computing

Hybrid Quantum-Classical Sampling
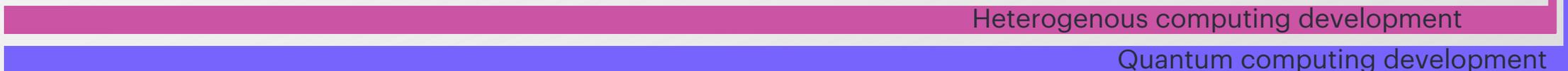Solve combinatorial optimization and faster inference

Physics-based generative AI
Augmenting LLMs for complex reasoning

HPE Labs ✛ **HPE HPC & AI Business Group** + **Partners** (academic, industrial, government)

Accelerators
FPGAs
GPUs
CPUs

Heterogenous computing development

Quantum computing development

Thank You

andrew.wheeler@hpe.com
hpe.com/labs