



Data Center Industry Overview and Insights

Day 2 – Data center (and IT) technology

Prepared for the Lloyd's Market Association (LMA)

Confidential: For internal use by LMA and Uptime Institute only.

Course overview

Day 1 Data centers in the modern economy



Day 2 Data center (and IT) technology



Day 3 Resiliency, availability, and performance

The Uptime Team



Doug Donnellan
(Moderator)

Research Analyst,
Uptime Institute



Andy Lawrence

Executive Director of
Research, Uptime Institute



Dr. Owen Rogers

Senior Research Director of
Cloud Computing,
Uptime Institute



Dr. Tomas Rahkonen

Research Director
Sustainability (Europe),
Uptime Institute

About today's session:

- Run time: 90 minutes with a break at 45 minutes
- Ask questions at any time
- There will be dedicated pauses to answer questions throughout

1

IT building blocks

Servers, storage and networking make up the IT side of data centers, but each layer has risks.

2

Why AI is different

GPU clusters intensify density, power demand, cooling needs and supply chain dependencies.

3

Power and cooling keep IT running

Facility systems convert utility power and heat rejection into continuous IT availability.

4

Costs, forecasts and future pressure

Growth creates capital intensity, grid pressure, technology choices and operational trade offs.

Recap of Day 1

- Data centers house the servers, storage and networking that run digital services.
- They provide the power, cooling, connectivity and protection needed to keep IT available.
- Data centers capacity is commonly measured by the amount of IT power a facility can support, usually in MW.



Day 1 recap - Different types of data centers

Enterprise: Built and operated mainly for one organization's own needs



Colocation: Many tenants share the building and utilities, but each controls what is inside its own space.



Hyperscale: Very large, standardized infrastructure designed to operate at enormous scale



Cloud: a service delivery model that may run in enterprise, colocation or hyperscale facilities. Customers do not manage the underlying hardware.

AI data centers: facilities dedicated to supporting AI training workloads by both hyperscalers and organizations. These require much more concentrated power and cooling than other facility types.



The data center describes the environment. The workload describes what the IT inside it is doing.

Day 1 recap: Key points

Business model

affects who controls each risk and who may suffer loss.

Scale and interconnection

affect concentration and the reach of an outage.

Location

affects physical hazards, utility dependence and growth constraints.

AI

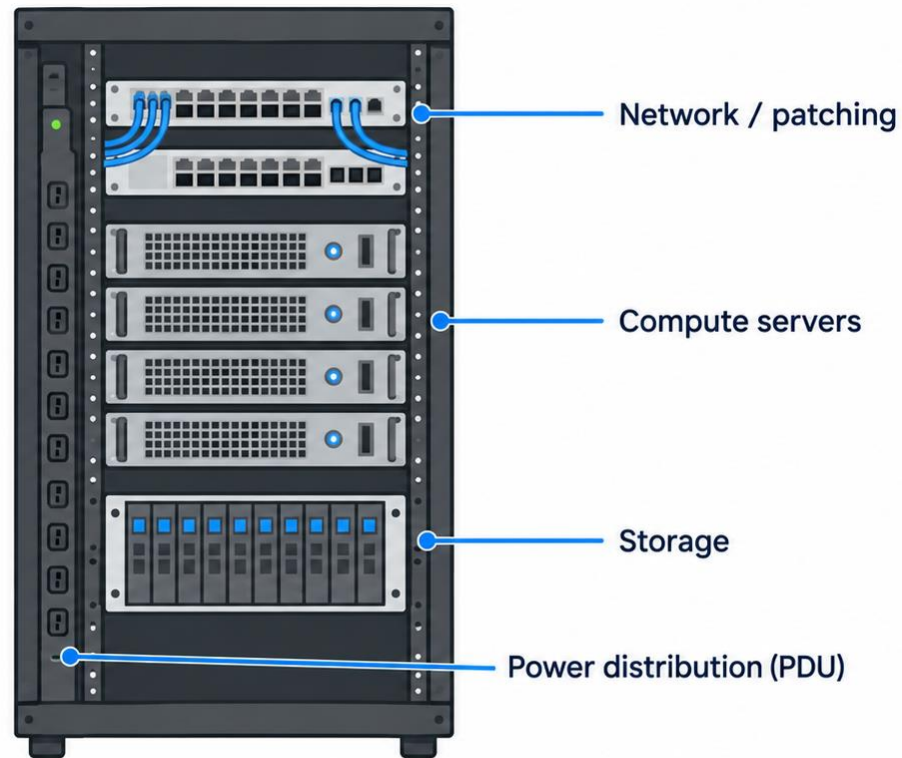
is increasing density, power demand and design pressure.

Data centers are service platforms where physical infrastructure, customer contracts, shared dependencies and operating responsibility combine to shape outage risk.

IT building blocks

- The inside of a server rack
- How a digital request moves through the stack
- Where IT failures can originate

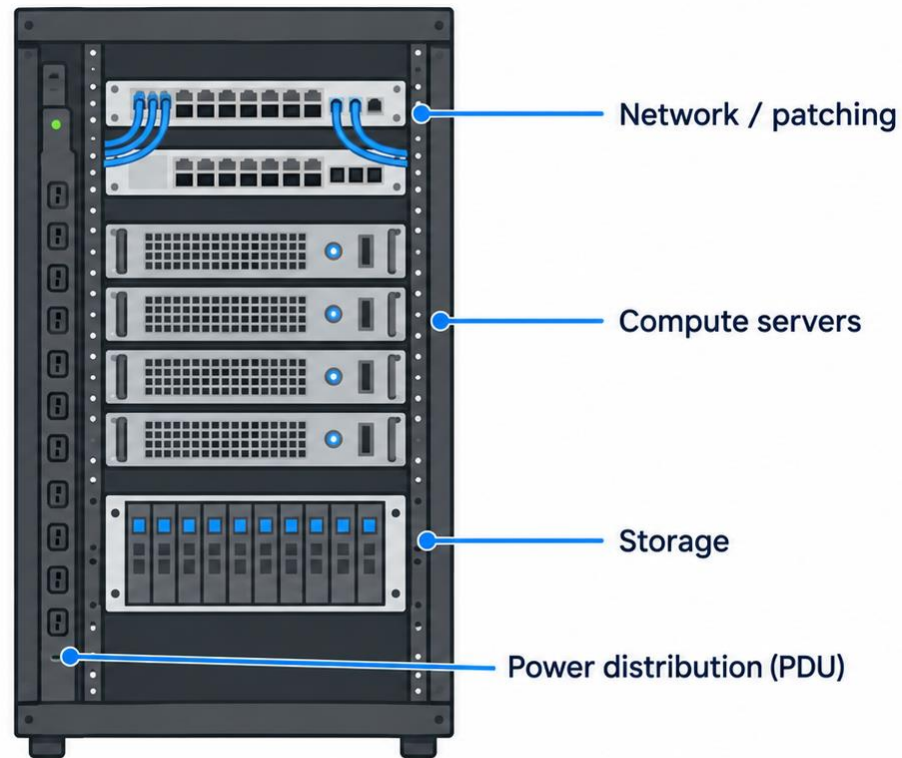
The inside of the server rack



Illustrative example only

- A server rack contains the IT equipment that performs the computing, storage and networking.
- Rack designs, layouts and power and cooling requirements vary widely.
- Rack density refers to how much power a single rack uses, determined by the processing power needs and # of servers
- A data center may contain hundreds or thousands of racks, depending on its size and purpose.

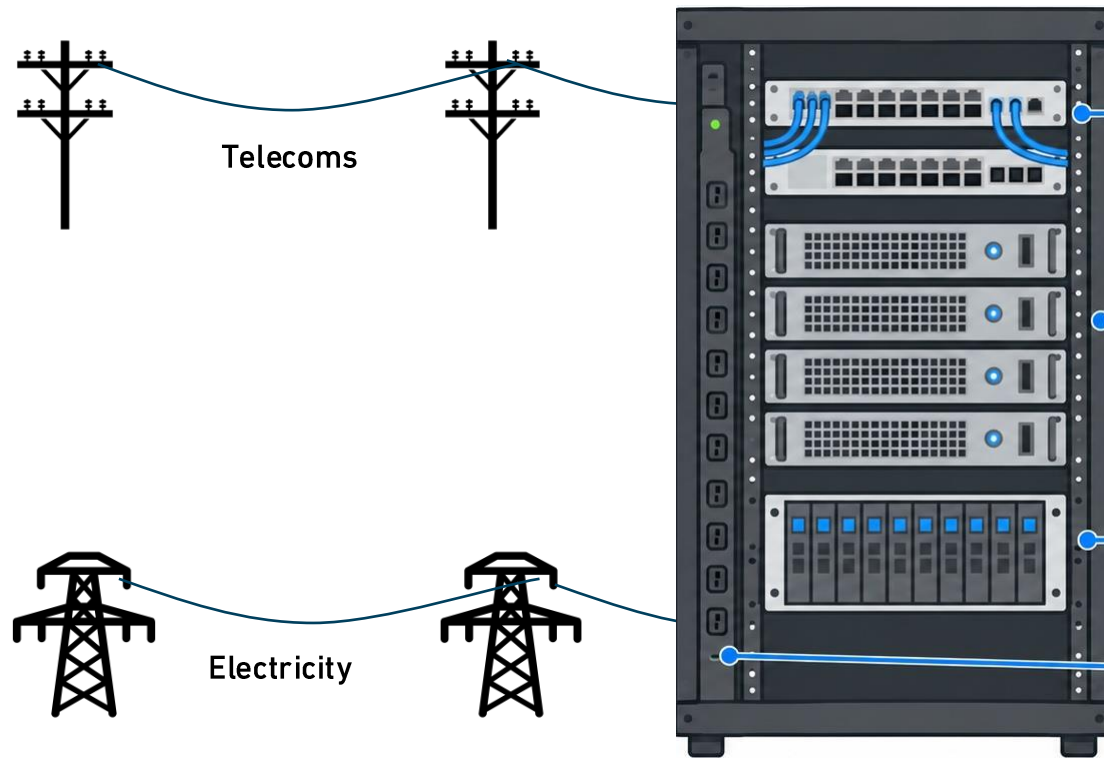
The inside of the server rack



Illustrative example only

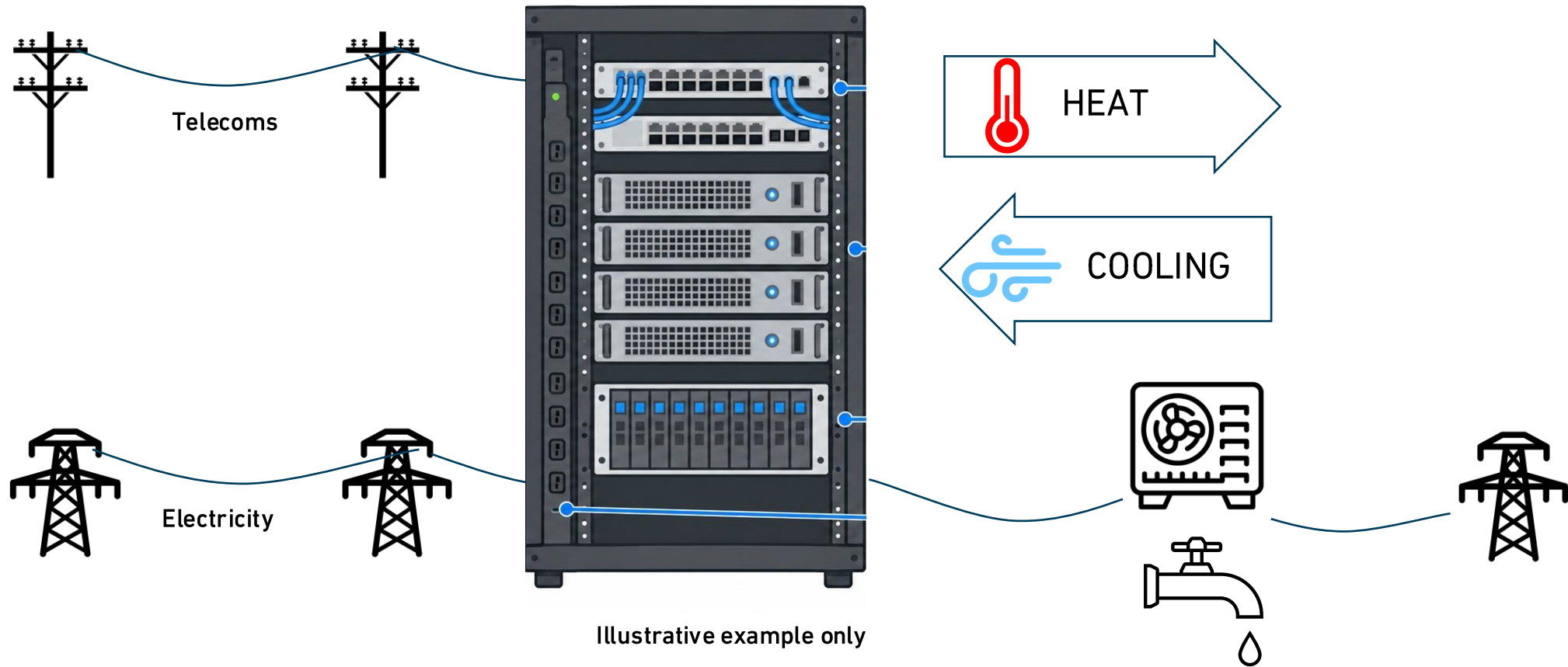
- Networking moves data between users, servers, storage and other sites. It includes switches, routers, fiber, network interfaces and external carrier connections.
- Servers run applications and process data. They contain processors, memory, power supplies, storage devices and network connections.
- Storage systems hold data that applications need. Storage may be local to servers, in shared arrays, or spread across many machines.

The inside of the server rack

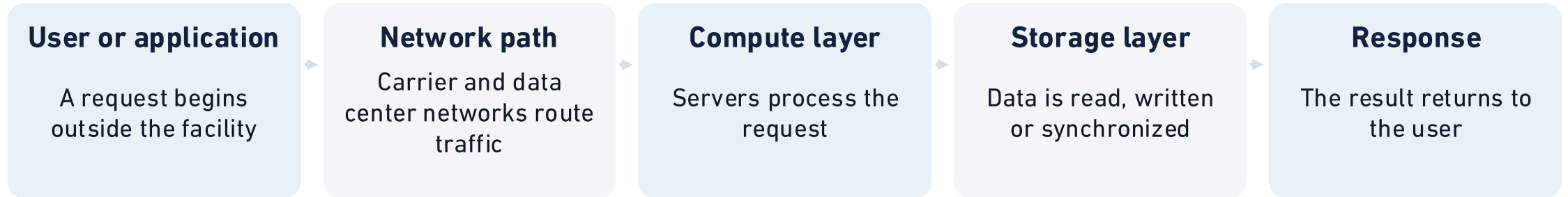


Illustrative example only

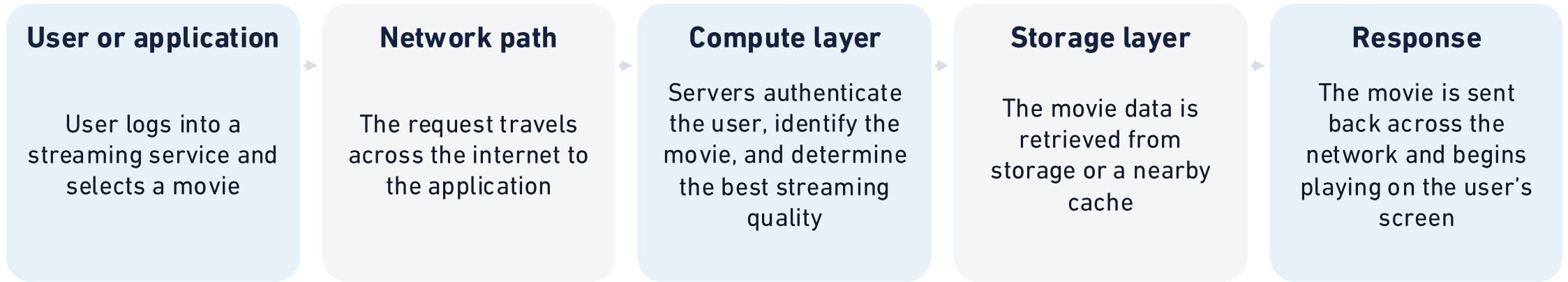
The inside of the server rack



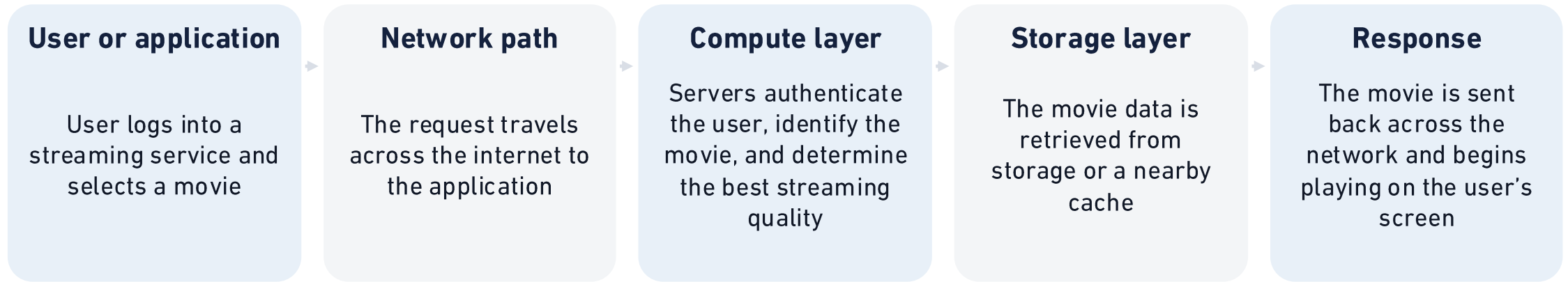
From user request to digital service



From user request to digital service: Example



From user request to digital service: Example



- Failure can occur at any point in the chain
- Some failures are technology or operations losses, not property losses.
- The customer usually experiences the failure as “the service is down.”

IT failures can be localized or systemic

Localized failure

A single server, disk, power supply or network link fails. Redundancy may absorb the problem, but maintenance and monitoring determine whether it remains contained.

Shared platform failure

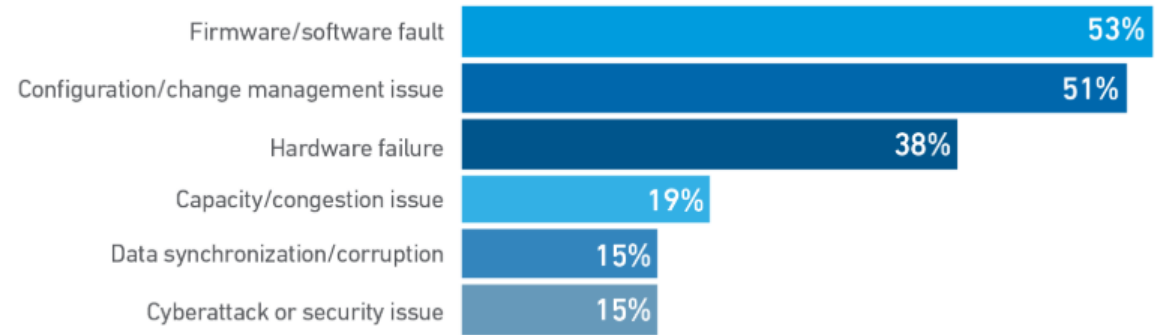
A common storage platform, network fabric, control plane or software system fails. The impact can extend across many customers or applications.

Recovery failure

The initial issue is resolved, but data restoration, failback, configuration errors or delayed parts extend the outage.

Takeaway: The severity of an IT incident is often driven less by the failed part itself and more by redundancy, controls, dependency mapping and recovery execution

Which are the most common root causes for the IT system/software-related outages at your organization's data centers over the past three years? Choose no more than three. (n=74)



(*Only respondents who indicated their organization experienced a significant, serious or severe outage caused by an IT system / software-related issue are included. Responses for "Other" and "Don't know" are not included.)

UPTIME INSTITUTE DATA CENTER RESILIENCY SURVEY 2026

uptime
INTELLIGENCE

uptime®
INTELLIGENCE

Why AI is different

- GPU-based compute and high-speed interconnects
- Higher density and larger power requirements
- Advanced cooling and specialized supply chains

Why AI workloads are different

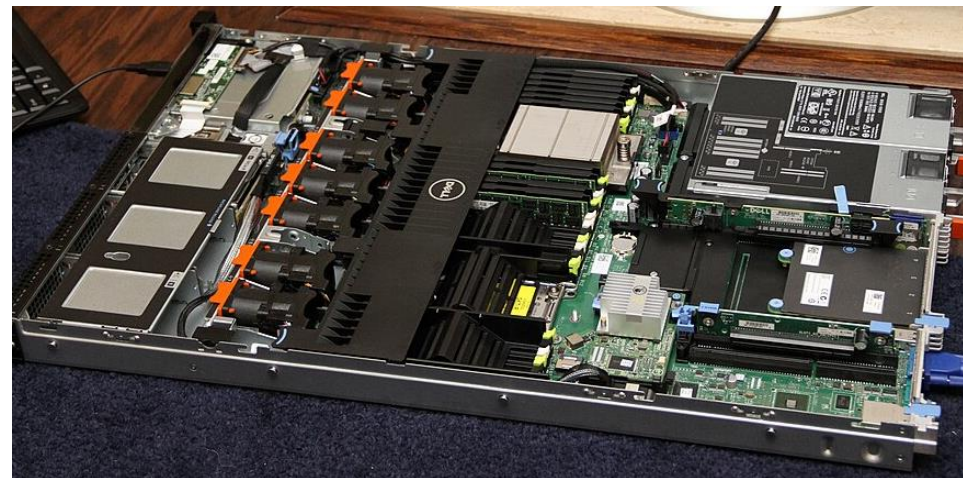
Artificial intelligence uses computer models to identify patterns in data and perform tasks such as generating text, recognizing images, making predictions or answering questions.

Workload	Definition	Infrastructure
Conventional compute	Runs business applications, databases and general digital services.	Often more mixed and distributed across standard server environments.
AI training	Builds or refines models using large datasets and many GPUs working together.	Requires dense clusters, high-speed networking and large synchronized power/cooling capacity.
AI inference	Uses a trained model to respond to user or application requests.	Can be more distributed, latency-sensitive and continuously scaled with demand.

CPU server

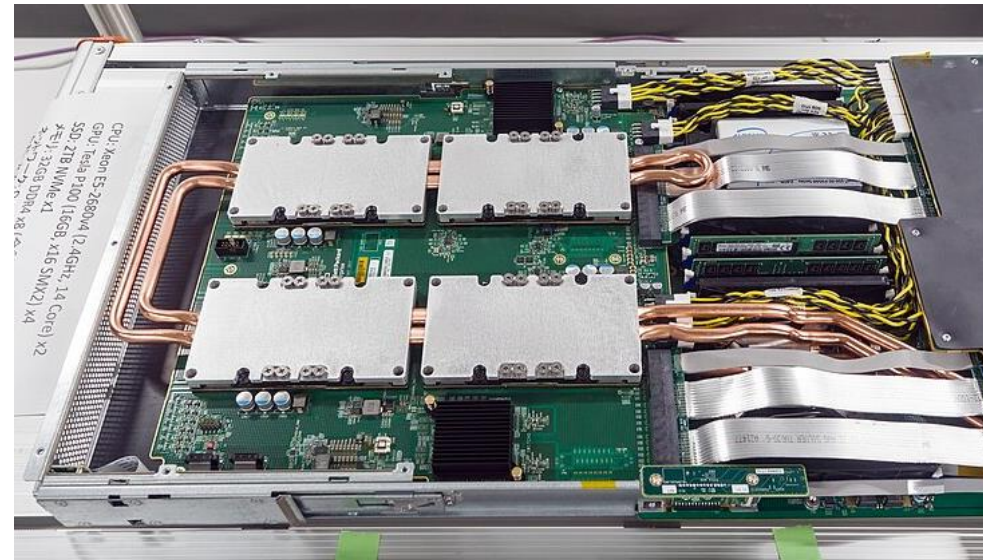
- General-purpose processing
- Relatively compact hardware
- Often lower power density
- Air cooling commonly sufficient

Inside of a 2013 web-server (Dell)

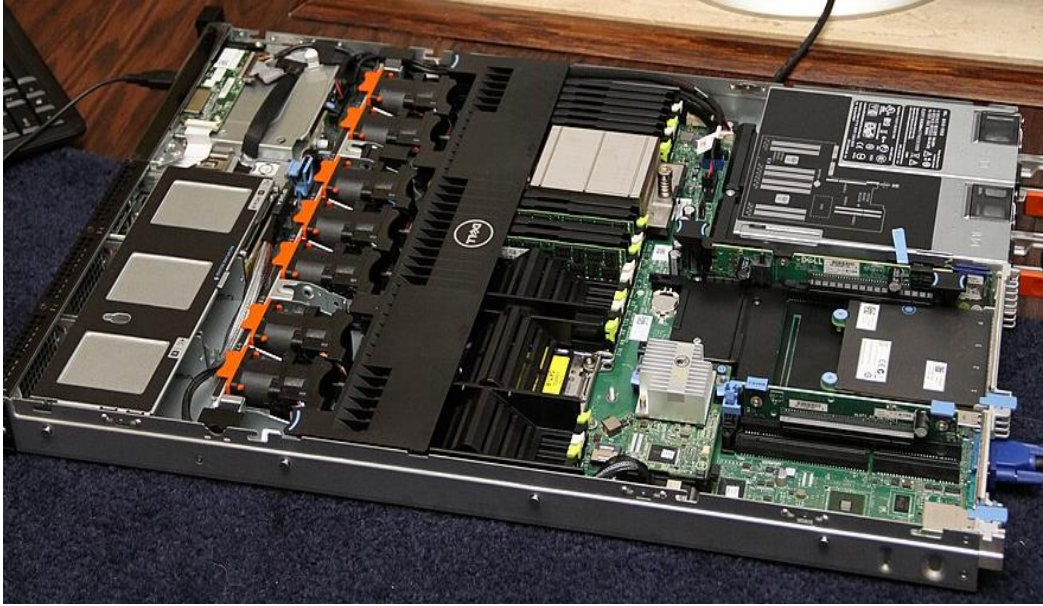


GPU server

- CPU plus multiple GPUs
- More accelerators and high-speed connections
- Much greater power and heat concentration
- Increasing need for advanced or liquid cooling



Inside of a GPU server of a TSUBAME 3.0 supercomputer (2017)



Dell web server (2013)

- ~\$7,000 fully configured (per server)
- ~\$400,000 for the entire rack
- ~0.25 kW power draw (per server)
- ~10 kW power draw (per rack)

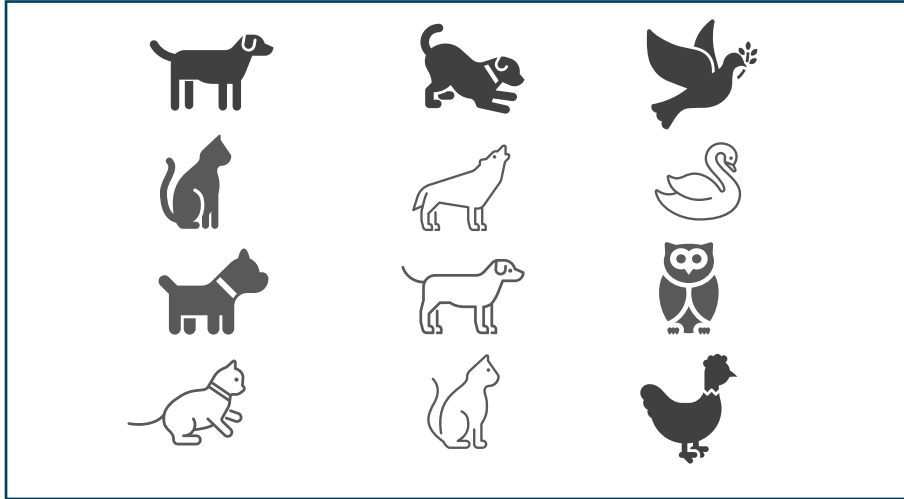


TSUBAME 3.0 Supercomputer GPU server (2017)

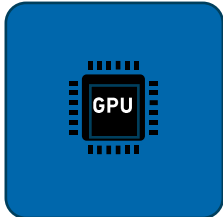
- ~\$41,000 for GPUs/CPUs alone
- ~\$80,000 fully configured (per server)
- ~\$2.5M for the entire rack
- ~1.5 kW power draw (per server)
- ~55 kW power draw (per rack)

- **Comparable 2026 version:** ~\$200,000/server
~14.3kw/server, 57 kW/rack
- **Leading edge 2026 version:** ~\$500,000/server
14.5 kW/server, ~100 kW/rack

Training from scratch



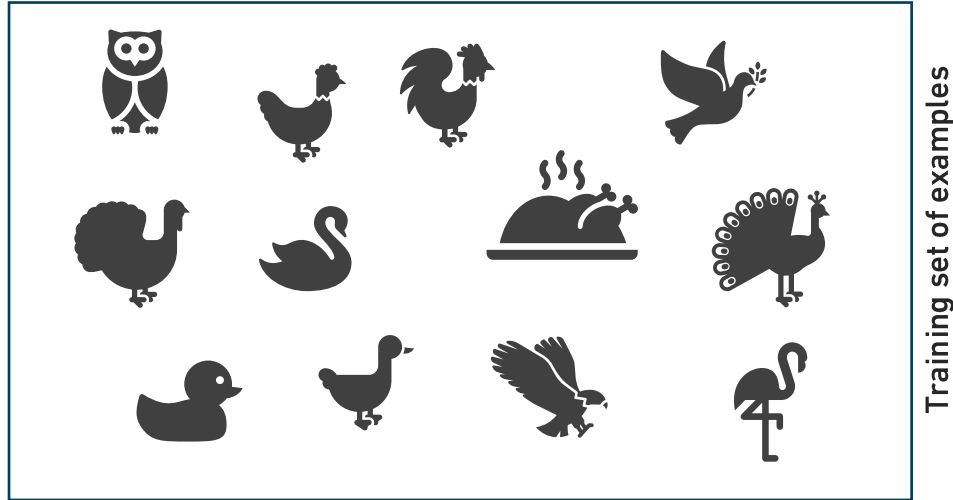
Training an animal image classifier



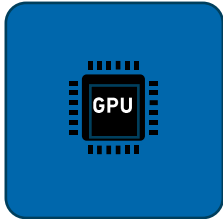
≠ CAT

- Infrequent batch process
- No real-time interactions with end-users
- Does not need to be located near to end-users
- Performed in advance of value generation
- Usually demands specialist chips to perform training in useful time frame

Fine-tuning



Training a bird image classifier

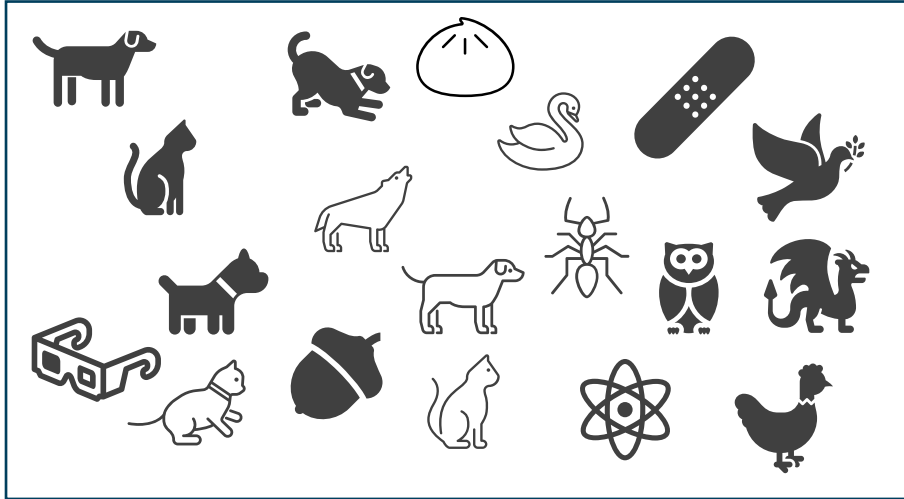


Found-
ation
Model

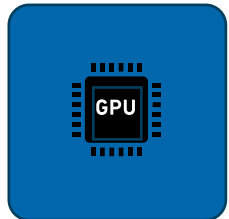
- A foundation model is a general-purpose model pre-trained by a third-party
- The model is trained once, and used by many as a baseline for further fine-tuning

Inference

User requests



Using the model for predicting and classification



Inference System



Trained model

- Interactive process
- Output is embedded in an application
- Must be located close to application
- The response and availability of the application affects the user experience
- Less resource intensive than training, suitable for CPUs/ASICs

Training a language model

Training set of examples

What's my current balance?	Show my last five transactions
Why was I charged a fee?	Freeze my card, I've lost it
How much did I spend on groceries?	Send £50 to Sam
Cancel my gym direct debit	Am I on track to pay my rent?
What's my mortgage rate?	Is this payment a scam?

The sentence fragments in this training set are tokens

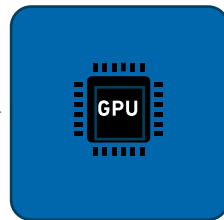
Generally, more tokens = a wider education

The variables “inside the model” that are updated during training are called parameters

Generally, more parameters = greater knowledge

Training a language model

Every token...



...updates every parameter

$$\int (a + b + c + d + \dots)$$

New AI technologies stretch and stress operators



Density & Cooling

- **10–20×** higher Density per rack
- **Liquid cooling** replaces or supplements air
- **Cooling failures** can cause immediate shutdown



Physical & Structural

- **Size of facility** increases dramatically
- **New rack designs**, data center layouts
- **Floor load** must support heavy liquid-cooled racks
- **Mechanical electrical** space expands dramatically



Power storage and distribution

- **On site (behind the meter) generation and storage** adds to investment, skills and risks.
- **Erratic power** profiles adds complexity, risks
- **800V DC** requires technical change, safety risks

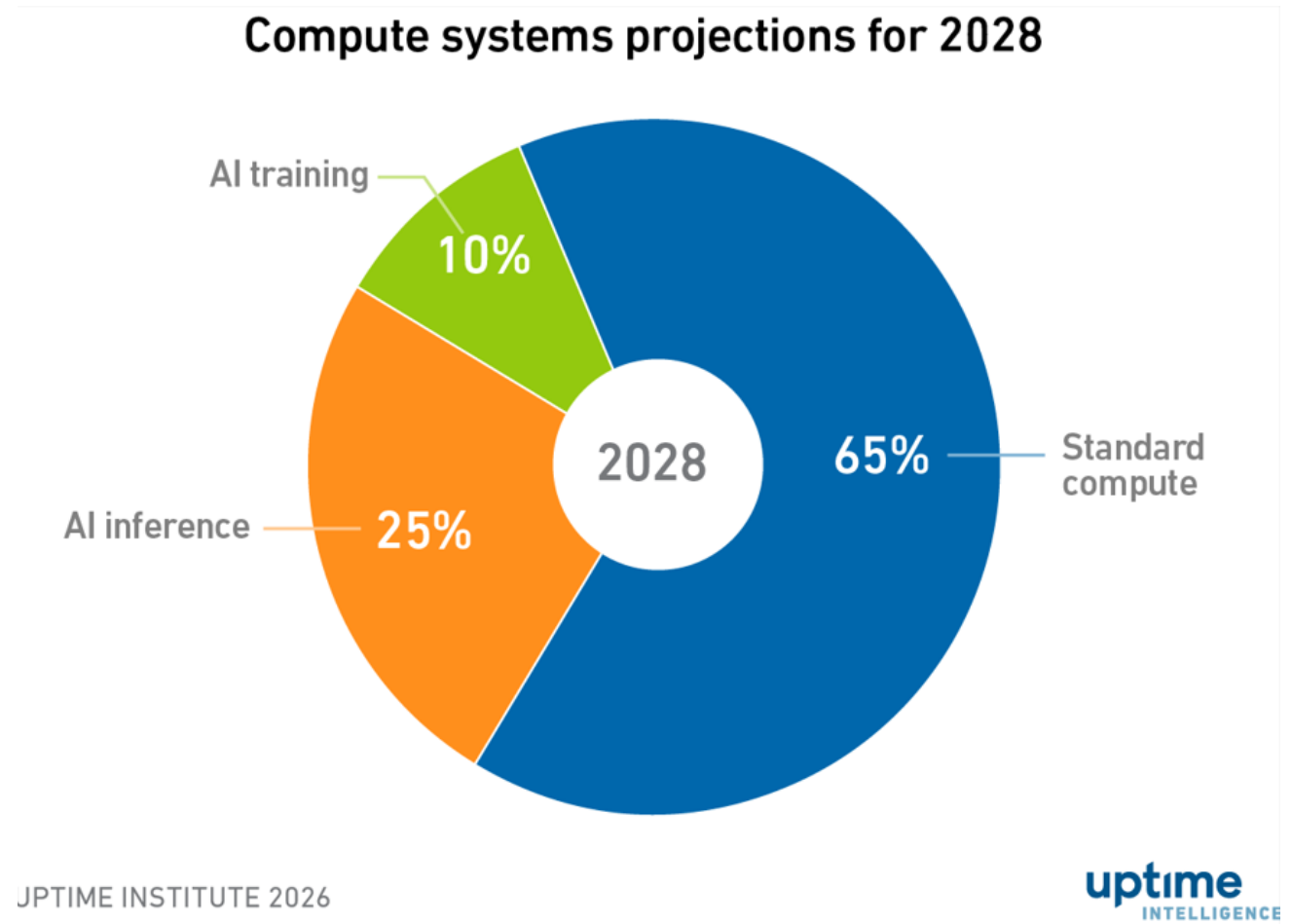


Large loads and the grid

- **Behind the Meter power** requires major investment, skills, adds complexity, risk.
- **Large loads** must interact with the grid
- **Environmental concerns** mean scrutiny, compliance

Rapidly changing, incompatible and unpredictable workloads

- **Training** – High performance, GPU based, High density, liquid cooled. Dominated by AI and hyperscale giants.
- **Inference** – Uncertain venues, uncertain power needs, high availability, low latency.
- **Traditional (CPU) compute:** Low density, all air cooled, high available, mostly low latency.

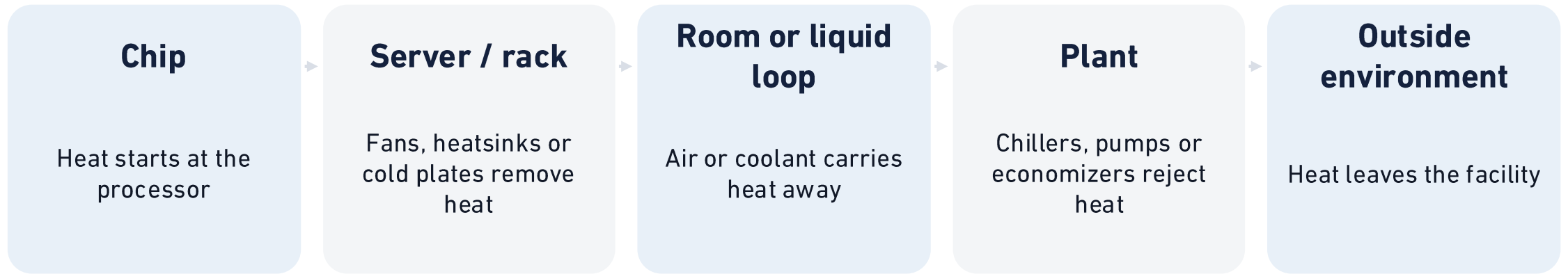


** By power demand, based on various global estimates, and Uptime analysis*

Power and cooling keep IT running

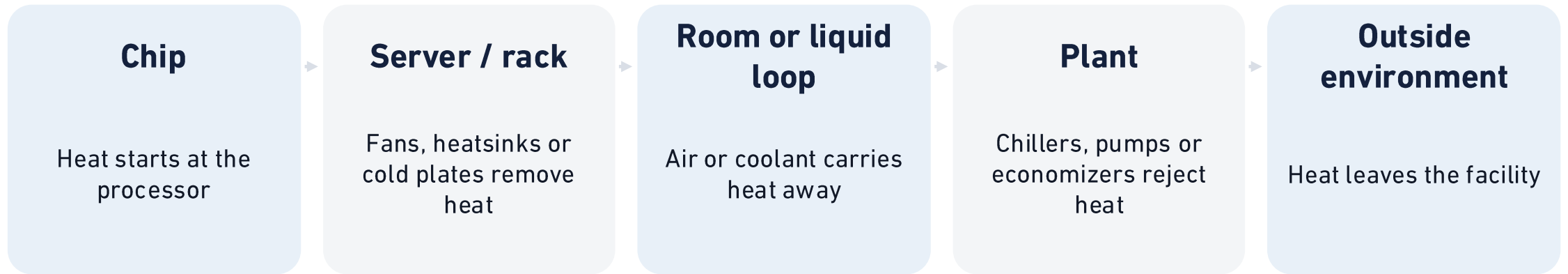
- Cooling removes heat before equipment throttles or fails
- Power systems condition and distribute electricity to racks
- Grid interaction and backup power shape outage resilience

Cooling removes heat from chips, racks and rooms



- Nearly all power consumed by IT equipment becomes heat that must be removed continuously.

Cooling removes heat from chips, racks and rooms



- Cooling failure can cause hot spots, server throttling, emergency shutdowns or equipment damage.
- Cooling resilience depends on airflow, controls, redundancy, maintenance access and operating discipline

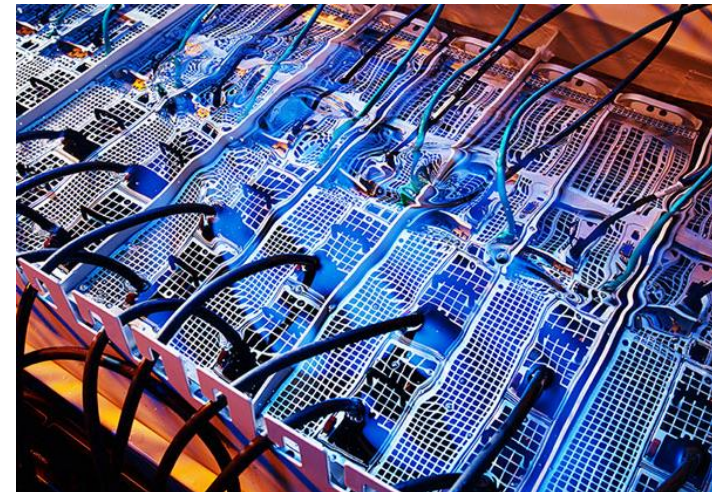
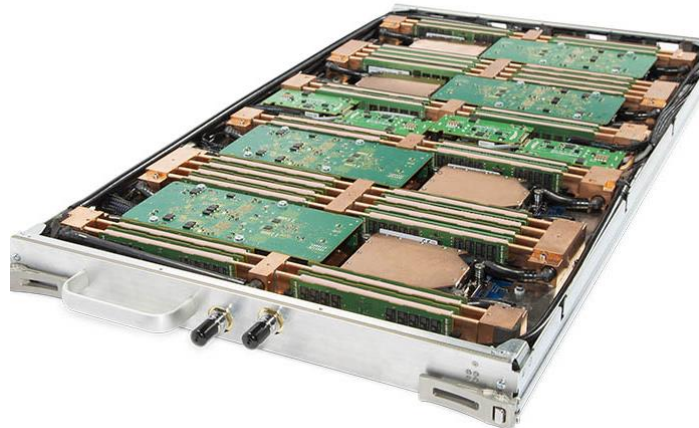
Cooling technologies range from air to liquid

Cooling method	Description	Notes:
Perimeter air cooling	CRAHs, CRACs or fan walls cool the room; airflow carries heat from racks.	Common and well understood but can struggle as rack power rises.
Close-coupled air cooling	Cooling units are placed closer to racks or rows.	Useful for higher-density zones; can complicate maintenance in IT areas.
Direct-to-chip liquid cooling	Coolant is delivered to cold plates attached to hot components.	Improves heat removal but introduces piping, CDUs and new service boundaries.
Immersion cooling	IT equipment is submerged in dielectric fluid.	Specialized design; operational practices and hardware compatibility matter.

- Higher rack densities require cooling closer to the heat source

Liquid cooling

Liquid cooling refers to a set of technologies that circulate a liquid coolant inside the server chassis to remove heat. Two major types:



Cooling type:

Design principle:

Form factor:

Cold plate ("direct to chip")

Closed coolant loop cools components that produce the most heat

Conventional server chassis + racks

Immersion

Coolant bath (non-conductive) cools all components

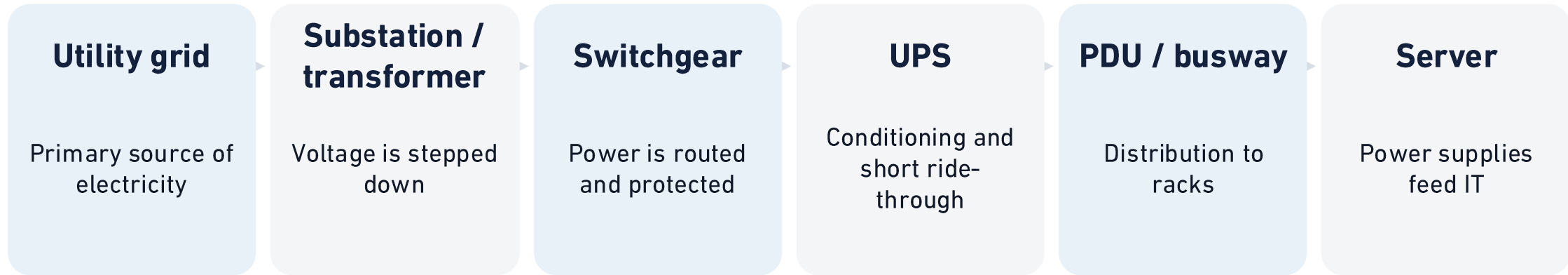
Vertical servers in horizontal vat

Cooling plants seek to utilize free cooling

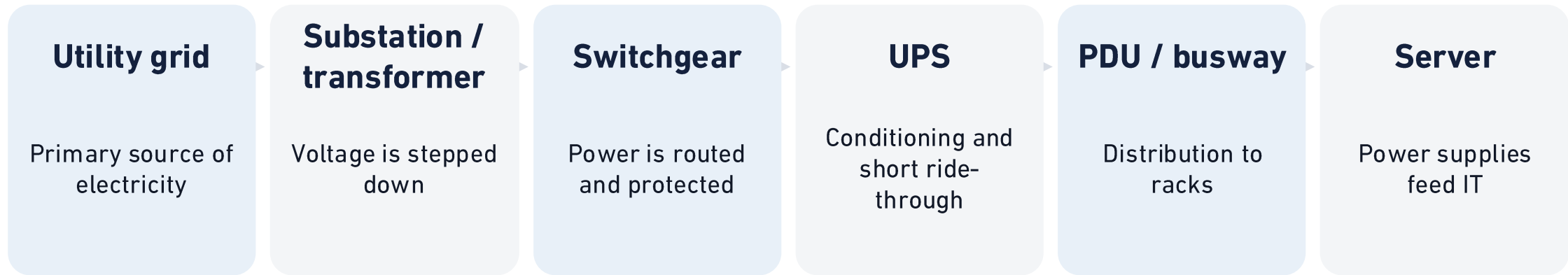
Cooling method	Description	Notes:
Mechanical cooling	Typically uses electrical compressors to remove heat (like air conditioning)	Universally applicable but consumes significant amounts of electricity.
Free-air cooling	Leverages cold outside air to improve cooling efficiency	Popular choice (multiple variants) depending on climate.
Adiabatic / Evaporative cooling	Uses water mist to improve cooling efficiency (not applicable in wet climates)	Efficient but can consume significant amounts of water.
Once-through water cooling	Improves cooling efficiency using sea, river, or lake water	Only possible at certain sites

- Water availability, climate and environmental constraints influence cooling design choices.

The data center power distribution chain



The data center power distribution chain



- Electrical failures can originate in utility supply, switchgear, UPS systems, PDUs, cabling or server power supplies.
- Protection systems are intended to isolate faults, but miscoordination can trip more equipment than intended.
- Most impactful data center outages involve power issues, maintenance activity or transfer failures rather than a simple lack of utility supply.

The backup power sequence

Utility event

Grid supply is interrupted or disturbed

UPS ride-through

Batteries bridge the initial gap

Generator start

Engines start and stabilize

Load transfer

Critical load moves to backup source

Sustained operation

Fuel, cooling and maintenance continue



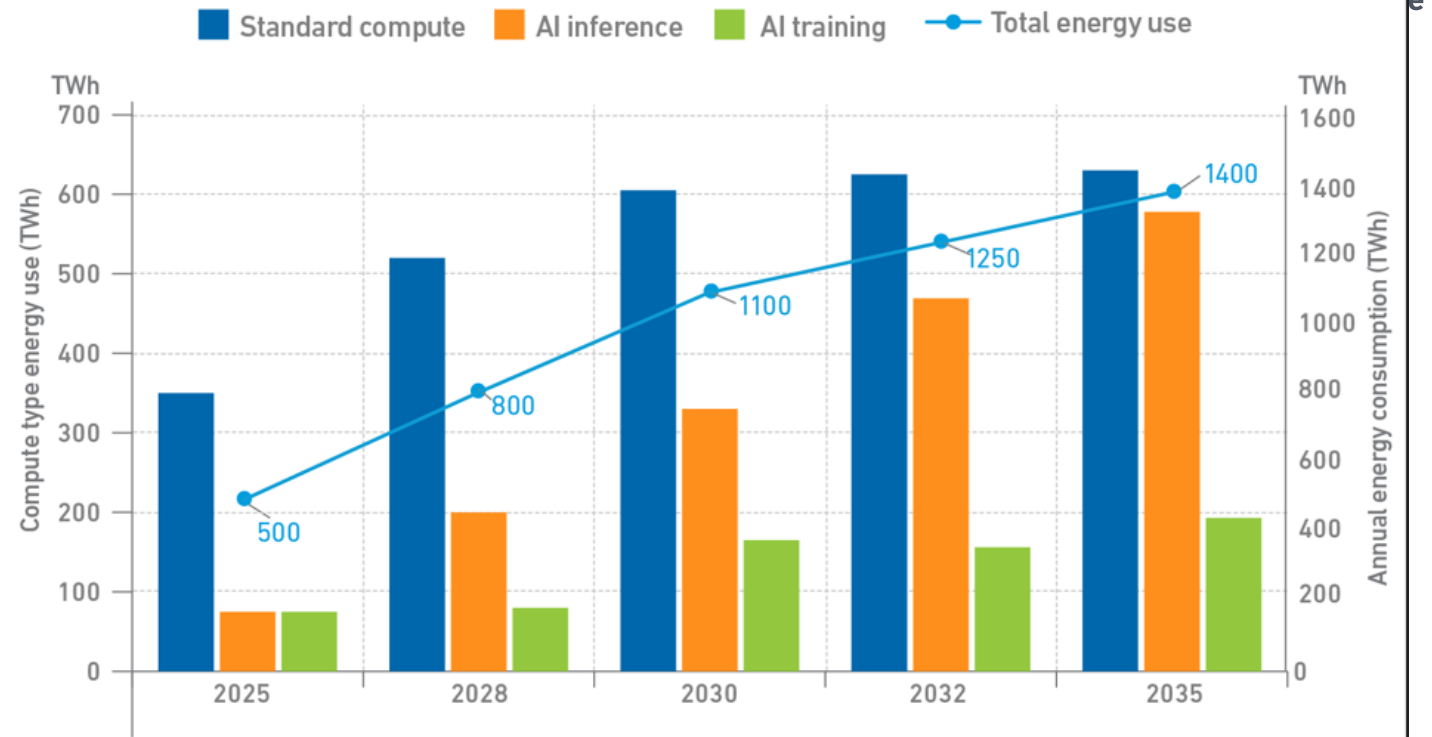
Costs, forecasts and future pressure

- Energy demand is expected to rise sharply
- New technologies may improve capacity but add operational complexity
- Build and operating costs depend on both IT and facility systems

AI growth increases pressure on power and design

- AI creates new business and infrastructure challenges.
- Standard compute is expected to remain large, while AI training and inference add new demand.
- Utility timelines may not match data center development timelines.
- Power pressure can influence site selection, design choices and operating headroom.

Projected data center global energy consumption 2025-2035



Uptime Intelligence estimate drawing on data from seven major studies and market deployment knowledge.
Estimates include on-site generation and grid power.

UPTIME INSTITUTE 2026

uptime
INTELLIGENCE

Takeaway: Power constraints link the AI trend directly to location, growth and resiliency discussions.

uptime®
INTELLIGENCE

Emerging technologies are responses to constraint

Technology	Driver	Risk
Liquid cooling	Supports higher rack densities and AI hardware.	Who maintains the liquid loop, and how are leaks or component failures isolated?
Onsite generation and storage	Can address power availability, cost or resilience goals.	How does it interact with grid supply, fuel logistics and emissions obligations?
AI-assisted operations	Can improve monitoring, anomaly detection and staffing efficiency.	What decisions are automated, and what remains under human control?

Takeaway: New technology can reduce one constraint while introducing new dependencies, maintenance needs or failure modes.

Build and operating costs split across IT and facilities

IT costs

Servers, GPUs, storage, network switches, optics, software and refresh cycles. These costs are highly sensitive to workload type and technology generation.

Facility / M&E costs

Electrical systems, mechanical systems, cooling plant, UPS, generators, switchgear, fire protection, building shell and commissioning.

Operating costs

Energy, maintenance, staffing, service contracts, spare parts, water, taxes, leases, insurance and lifecycle replacement.

Takeaway: A facility can be well built but economically stressed if utilization, energy costs, equipment refresh or maintenance requirements are challenging

What worries data center management:

High costs, unpredictable demand, the power crunch

Very concerned Somewhat concerned Only slightly concerned Not at all concerned

Cost issues



Forecasting future data center capacity requirements



Power availability



Supply chain disruptions



A lack of qualified staff



Meeting AI infrastructure needs



← Uncertainty of workloads is a commonly repeated concern of operators in all geographies

(Only the top six response categories are shown. All figures rounded)

Day 2 Key points

IT availability depends on servers, storage, networks and software — not only on the physical facility.

AI infrastructure increases density, power demand, cooling complexity and reliance on specialized components

Power and cooling are linked systems; failures often occur during transitions, maintenance or control interactions

Grid constraints, equipment supply and cost pressure are becoming central to both development risk and operational resilience.

Visit www.uptimeinstitute.com for more information.



Uptime Annual Survey
Executive Summary



Uptime Intelligence
Free Trial Request

uptime[®]
INTELLIGENCE

Uptime Institute is the Global Digital Infrastructure Authority.
Uptime Institute is headquartered in New York, NY, with offices
in Seattle, London, Sao Paulo, Dubai, Singapore, and Taipei.

©2026 Uptime Institute, LLC. All Rights Reserved.

Uptime Institute
405 Lexington Avenue
New York, NY 10174