

Impact of Data Centres on Power System

Sarasij Das

Associate Professor



Department of Electrical Engineering
Indian Institute of Science, Bangalore.

Indian Institute of Science

2

- Established 1909
- Research focused
- **Students:** 5000+
 - **Ph.D:** 60%
 - **Masters:** 30%
 - **Bachelors:** 10%
- **Professors:** 500+



Indian Institute of Science



At-a-Glance

BIO-DATA

- Associate Professor at IISc
- Ph.D (2014) University of Western Ontario, Canada
- Research Areas:
 - ✓ Electrical Equipment Protection
 - ✓ Grid Integration of Renewables
 - ✓ Cyber Security of Power Grids
- Industry Experience:
 - ✓ Schneider Electric India Pvt Ltd.
 - ✓ Power Research & Development Consultants Pvt Ltd.

PAST FUNDING PARTNERS



INTERNATIONAL COLLABORATORS

Europe

- Imperial College London
- TU Delft, Netherlands
- NTUA, Greece
- NTNU, Norway

USA

- Clemson University
- University of North Carolina
- Penn State University
- New Mexico State University

Canada

- University of Saskatchewan
- OntarioTech University

Motivation

The rapid buildout of generative-AI training and inference infrastructure is turning data centers into some of the largest and fastest-changing loads that power grids have ever had to serve.

Explosive demand growth

- Hyperscale AI campuses now request hundreds of megawatts to multiple gigawatts of interconnection capacity per site.

New dynamic behavior

- GPU-centric loads ramp, cluster, and interface with the grid in ways conventional industrial or commercial loads do not.

Planning & reliability risk

- Utilities and reliability coordinators need trustworthy simulation tools before these loads are widely interconnected.

Recent Investments



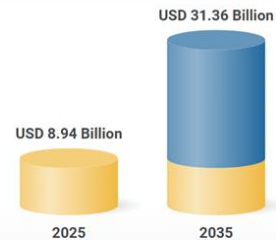
Adani to invest \$100 billion in AI-ready data centres by 2035

February 17, 2026



Indian Data Center Market

Market forecast to grow at a CAGR of 13.4%



<https://www.researchandmarkets.com/reports/6227141>

RESEARCH AND MARKETS
THE WORLD'S LARGEST MARKET RESEARCH STORE

Anant Raj to Develop Large-Scale Data Centre Infrastructure in Haryana with ₹20,000 Crore Investment

Anant Raj will invest ₹20,000 crore in Haryana data centres, boosting digital infrastructure, cloud services, AI growth, employment, and tech investments.

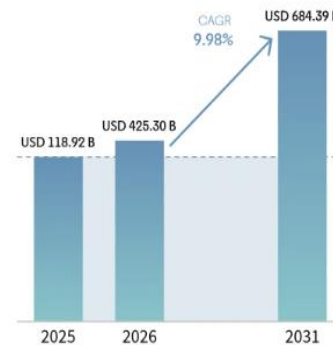
By [TRT Editorial](#) • [Save it](#)

Last updated: June 3, 2026

[Facebook](#)
Like

[Instagram](#)
Follow

Data Center Market Market Size in USD Billion



Source: Mordor Intelligence



AirTrunk to invest \$21 billion in India data centre

By Reuters

June 1, 2026 5:54 PM GMT+5:30 • Updated June 1, 2026



Reliance bets big on Andhra Pradesh with ₹1.08 lakh crore AI data centre push

By [Express Computer](#) — On May 20, 2026

NEWS | ARTIFICIAL INTELLIGENCE AI

Efficiency Metrics

Power Usage Effectiveness (PUE) :

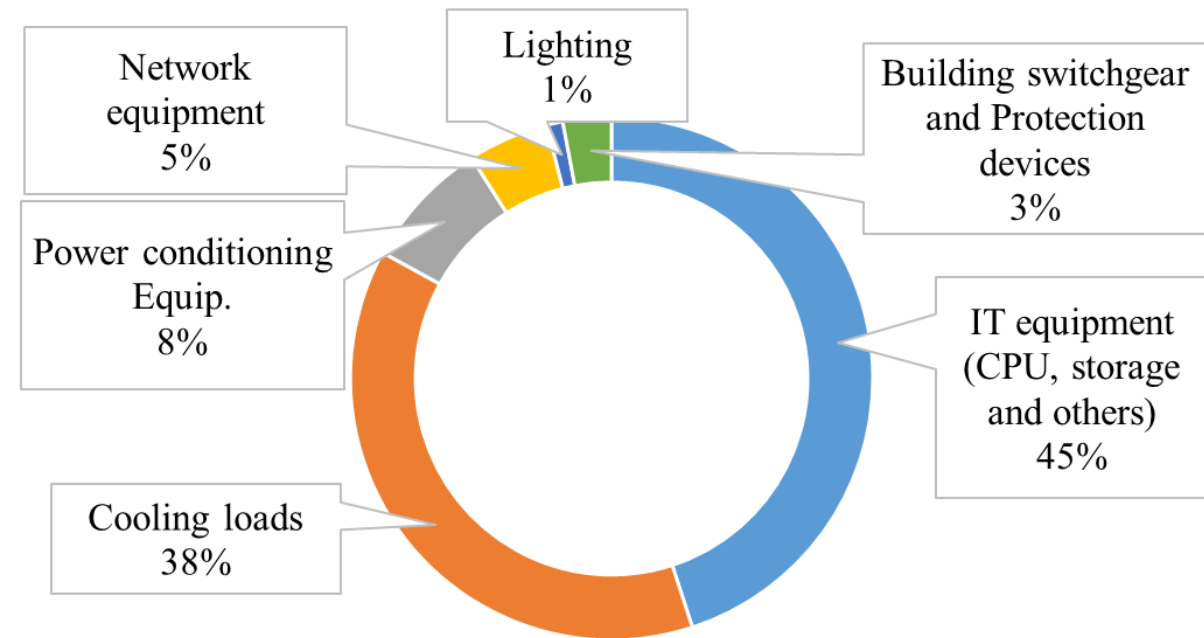
$$\text{PUE} = \text{Total Facility Power} \div \text{IT Equipment Power}$$

A PUE of 1.0 is the (unreachable) ideal

Every watt of facility power would go directly to computing — none to cooling, lighting, or distribution losses.

Modern purpose-built facilities: 1.2–1.4

Achieved with hot/cold aisle containment, free cooling, and efficient UPS/transformer selection.



Ahmed, Kazi Main Uddin, Math HJ Bollen, and Manuel Alvarez. "A review of data centers energy consumption and reliability modeling." IEEE access 9 (2021): 152536-152563.

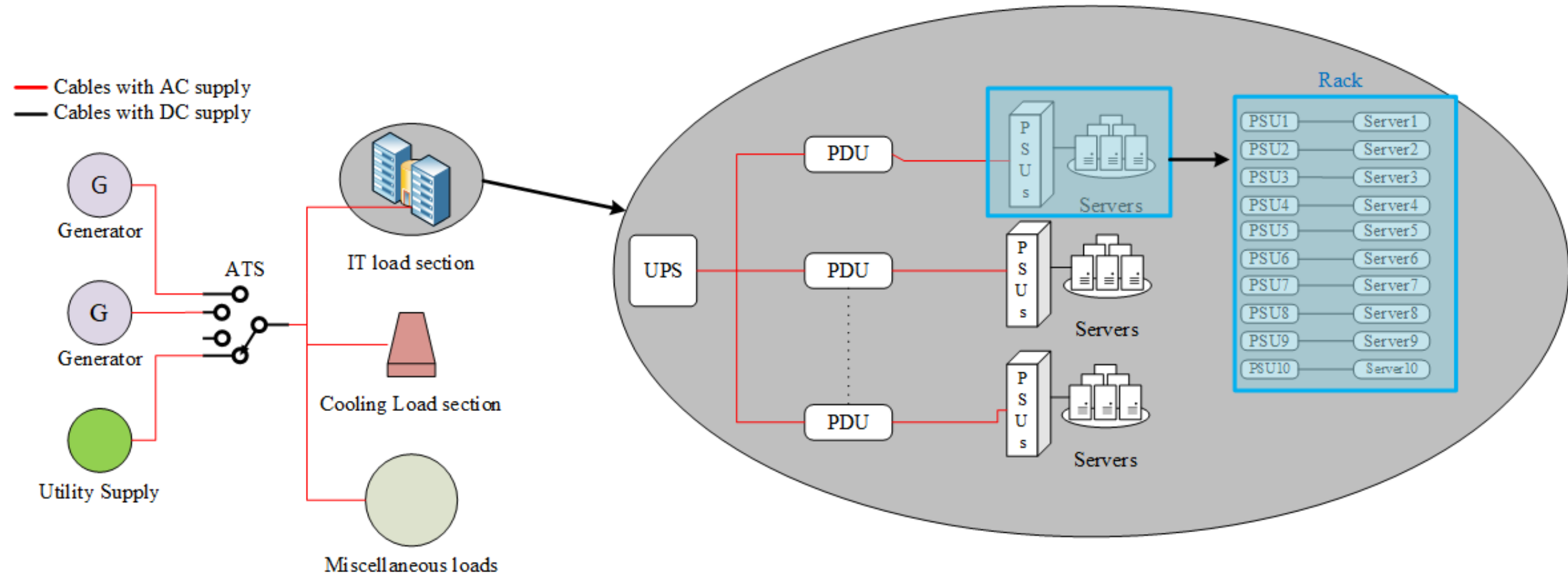
Data Center Types: Computations

Parameter	Conventional Data Center	AI Data Center
Primary workload	Enterprise IT, web hosting, storage, databases, cloud services	AI model training and large-scale inference
Compute hardware	CPU-based servers	GPU / accelerator clusters (GPUs, TPUs, ASICs)
Rack power density	~5-15 kW per rack	~30-150 kW per rack today; >200 kW planned
Facility size	~1-50 MW	~100 MW to 1 GW+ campuses
Load profile	Relatively steady; follows daily user demand	High utilization with fast, large power swings (synchronized training steps, checkpointing)
Cooling	Air cooling (CRAC/CRAH, hot/cold aisle)	Direct-to-chip liquid cooling or immersion; air alone insufficient
Typical PUE	~1.4-1.8	~1.1-1.3 (liquid-cooled, purpose-built)
Power distribution	208/415 V AC; lower current per rack	415/480 V AC busway; move toward higher-voltage DC (e.g., ± 400 V / 800 V DC)
Networking	Ethernet; mostly north-south traffic	InfiniBand / RoCE; heavy east-west GPU-to-GPU traffic, low latency
Redundancy approach	Tier III/IV for most critical services	Inference: high redundancy; training halls may accept lower redundancy (job checkpointing)
Energy storage role	UPS mainly for ride-through until generator start	UPS/BESS also used to smooth rapid load fluctuations
Grid-integration concerns	Moderate: capacity, power quality	Large load interconnection, voltage/frequency sensitivity, ride-through, sudden load loss, oscillations
Location drivers	Close to users / network hubs (latency)	Available grid capacity and low-cost power; latency less critical for training

CRAH= Computer Room Air Handling

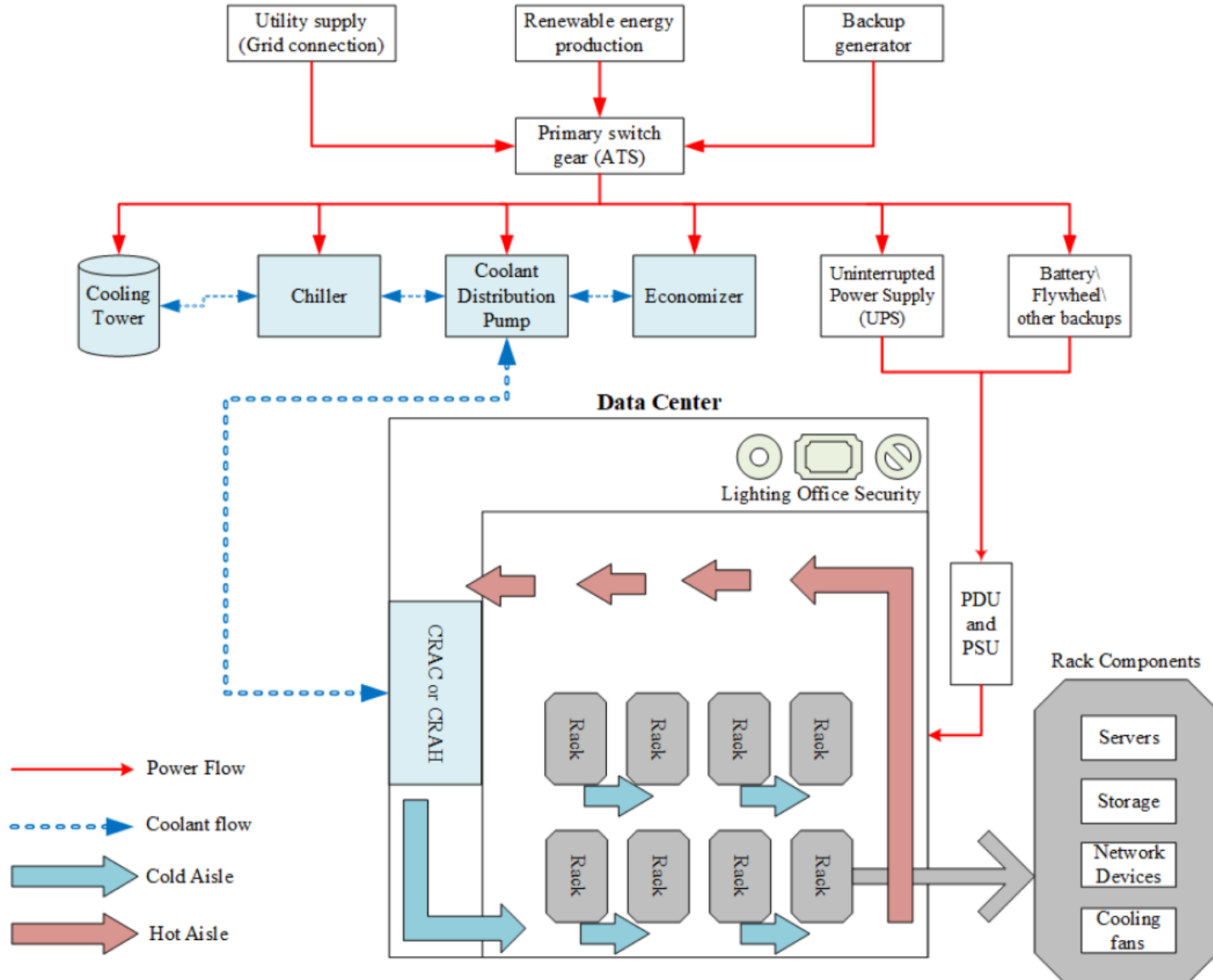
CRAC= Computer Room Air Conditioner

AI Data Center



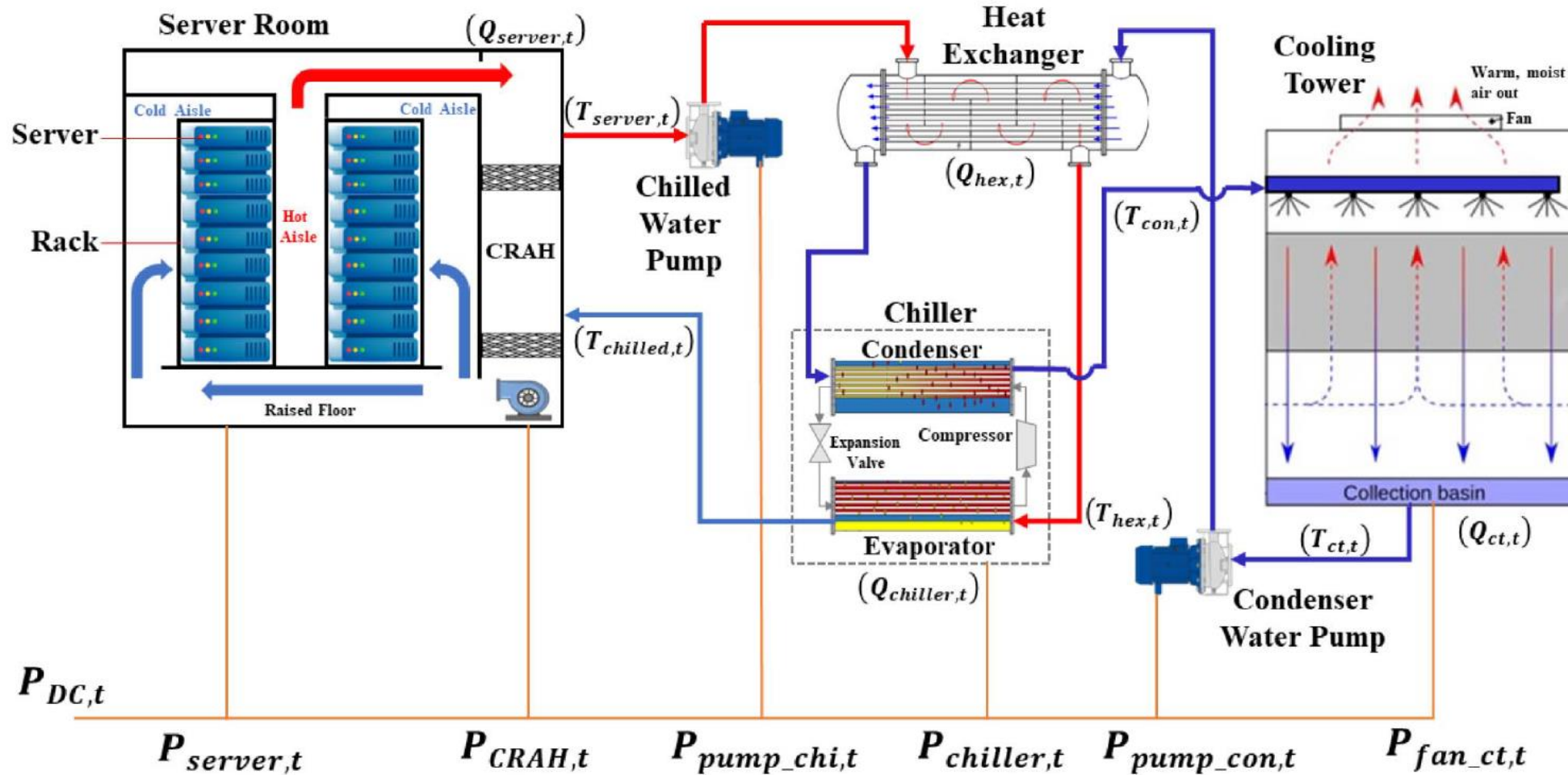
Ahmed, Kazi Main Uddin, Math HJ Bollen, and Manuel Alvarez. "A review of data centers energy consumption and reliability modeling." IEEE access 9 (2021): 152536-152563.

AI Data Center: Structure



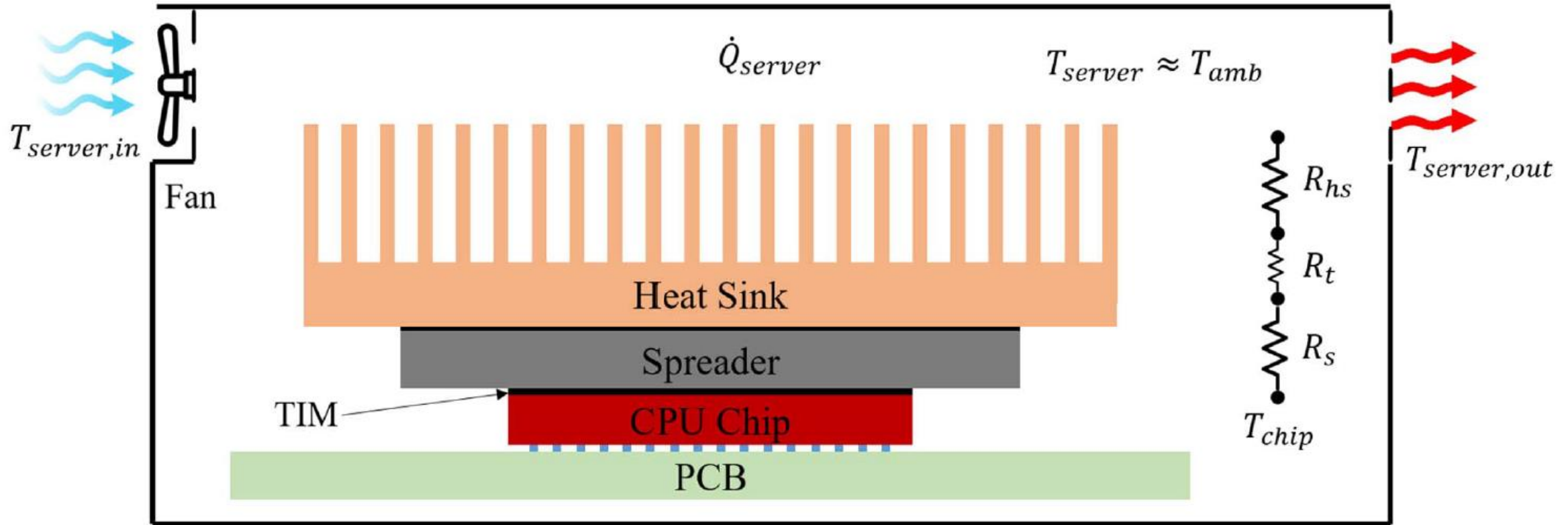
Ahmed, Kazi Main Uddin, Math HJ Bollen, and Manuel Alvarez. "A review of data centers energy consumption and reliability modeling." IEEE access 9 (2021): 152536-152563.

AI Data Center: Cooling System



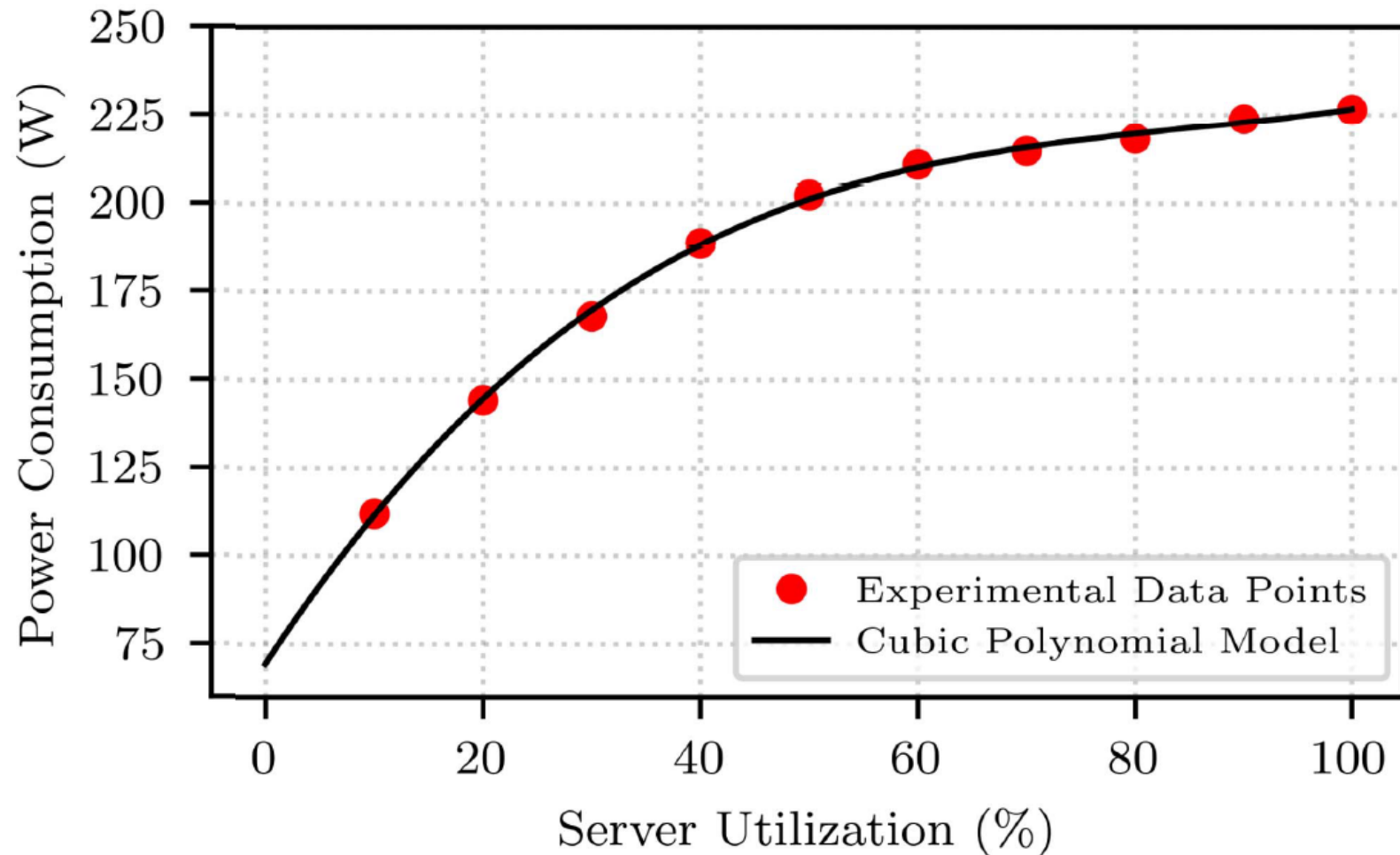
Gyang, Pam Paul, Pratyush Chakraborty, Lasantha Meegahapola, and Xinghuo Yu. "Dynamic modeling of a data center for power system stability studies." *IEEE Transactions on Power Systems* (2025).

AI Data Center: Server Thermal Management



Gyang, Pam Paul, Pratyush Chakraborty, Lasantha Meegahapola, and Xinghuo Yu. "Dynamic modeling of a data center for power system stability studies." *IEEE Transactions on Power Systems* (2025).

AI Data Center: Server Thermal Management



Gyang, Pam Paul, Pratyush Chakraborty, Lasantha Meegahapola, and Xinghuo Yu. "Dynamic modeling of a data center for power system stability studies." *IEEE Transactions on Power Systems* (2025).

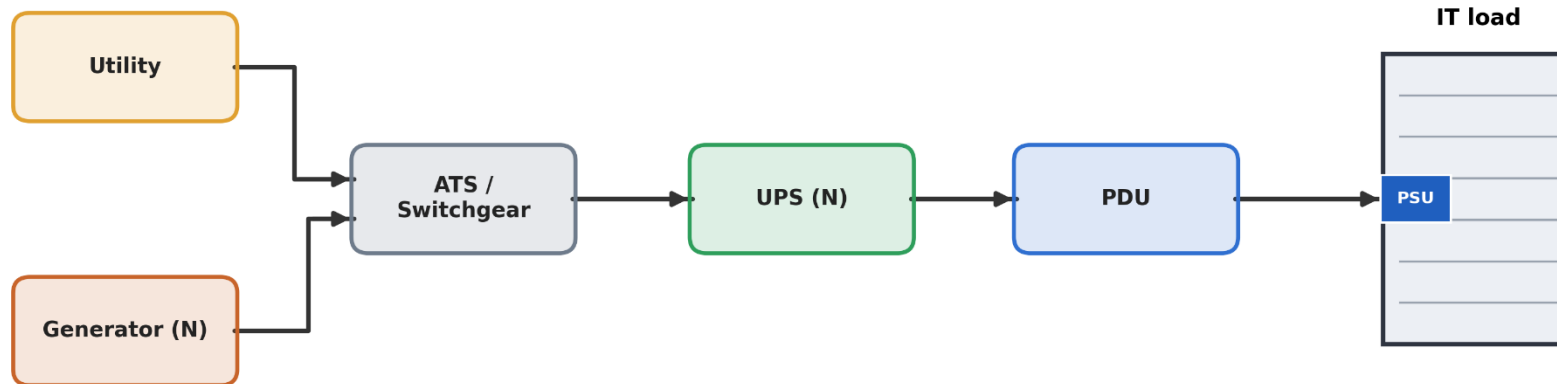
Data Center Types: Reliability

TIER I

Basic Capacity

Single path · no redundant components (N)

Tier I



*Every component is a single point of failure.
Any failure or planned maintenance → IT outage.*

Utility

Generator

Switchgear/ATS

UPS

PDU

--- dashed = redundant / standby element

Redundancy

N

Distribution paths

1

Concurrently maintainable

No

Fault tolerant

No

Typical availability

99.671% (~28.8 h/yr)

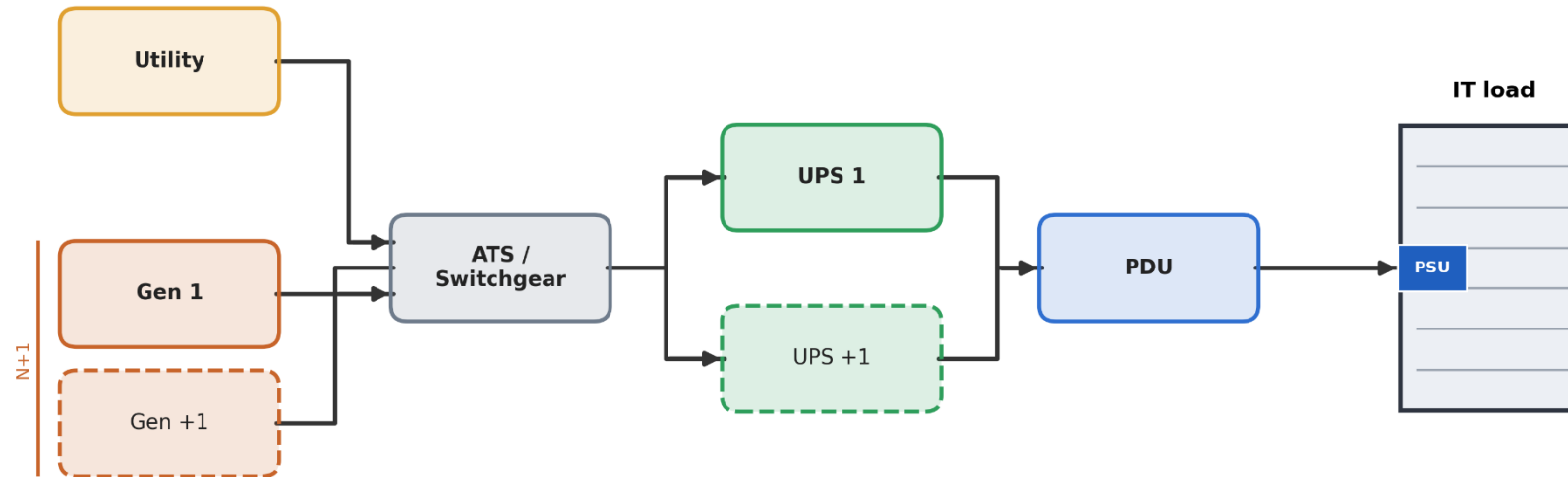
Data Center Types: Reliability

TIER II

Redundant Capacity Components

Single path · redundant components (N+1)

Tier II



*A single UPS or generator can fail or be serviced.
The distribution path (switchgear, PDU, cabling) is still a single point of failure.*

Utility

Generator

Switchgear/ATS

UPS

PDU

--- dashed = redundant / standby element

Redundancy

N+1

Distribution paths

1

Concurrently maintainable

No

Fault tolerant

No

Typical availability

99.741% (~22 h/yr)

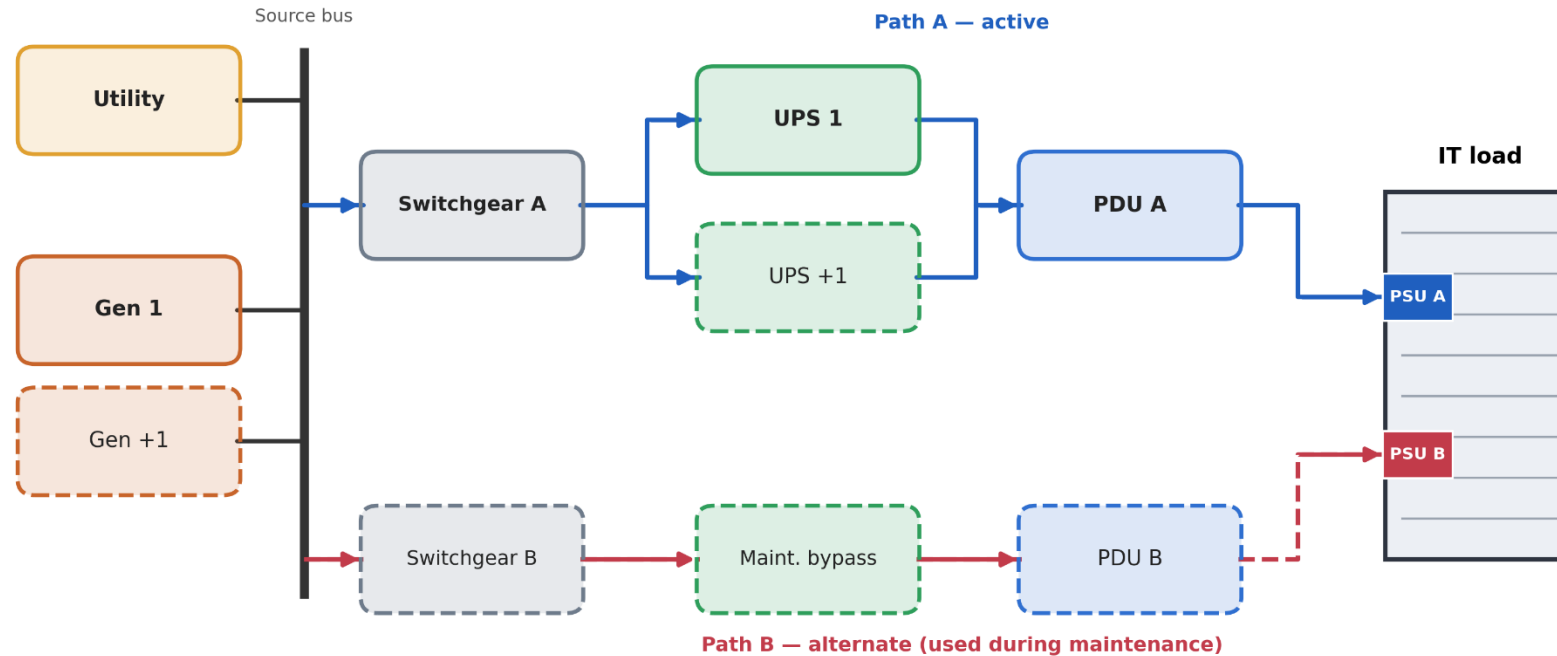
Data Center Types: Reliability

TIER III

Concurrently Maintainable

N+1 components · two distribution paths (one active, one alternate)

Tier III



*Any component can be removed for planned maintenance without IT shutdown.
An unplanned fault on the active path can still cause an outage.*

Utility

Generator

Switchgear/ATS

UPS

PDU

--- dashed = redundant / standby element

Redundancy

N+1

Distribution paths

2 (1 active)

Concurrently maintainable

Yes

Fault tolerant

No

Typical availability

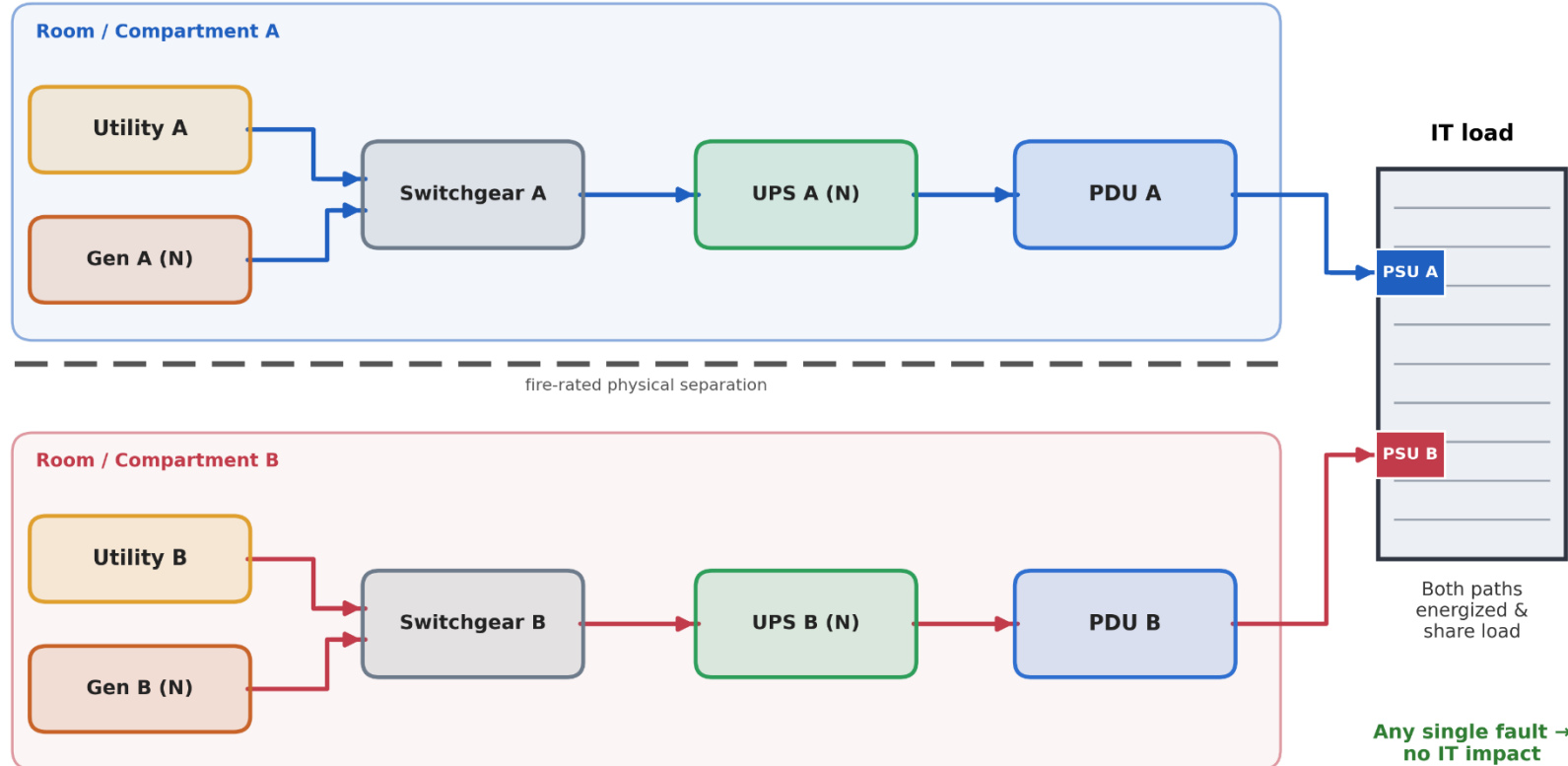
99.982% (~1.6 h/yr)

Data Center Types: Reliability

TIER IV

Fault Tolerant

2N shown (2(N+1) also common) · two simultaneously active, compartmentalized paths



Utility

Generator

Switchgear/ATS

UPS

PDU

--- dashed = redundant / standby element

Redundancy

2N / 2(N+1)

Distribution paths

2 (both active)

Concurrently maintainable

Yes

Fault tolerant

Yes

Typical availability

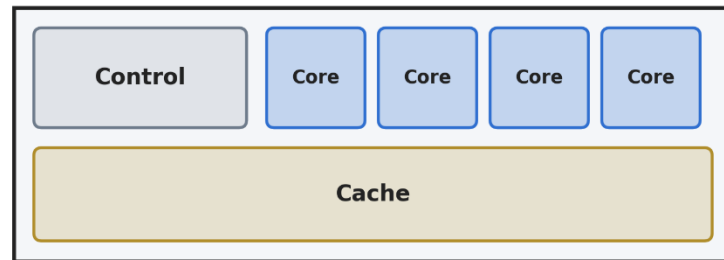
99.995% (~26 min/yr)

AI Computations

CPU-based vs GPU-based Computer Architecture

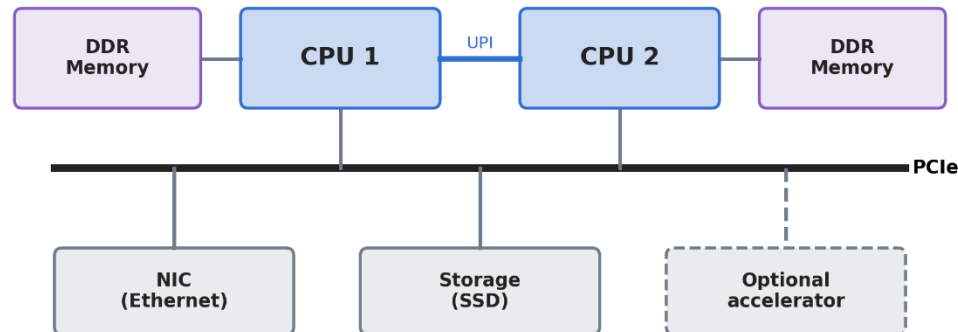
CPU-based Computer (general-purpose)

① Processor (chip level)



Few large, complex cores

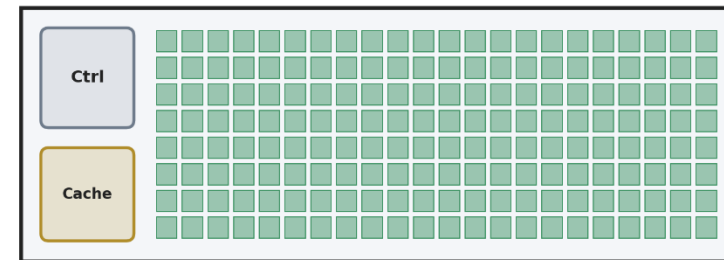
② Server (system level)



- Few powerful cores tuned for sequential, branch-heavy logic
- Large shared DDR memory; accelerators attach over PCIe
- Workloads: databases, web, virtualization, enterprise IT
- Typical server power: ~0.5-1 kW

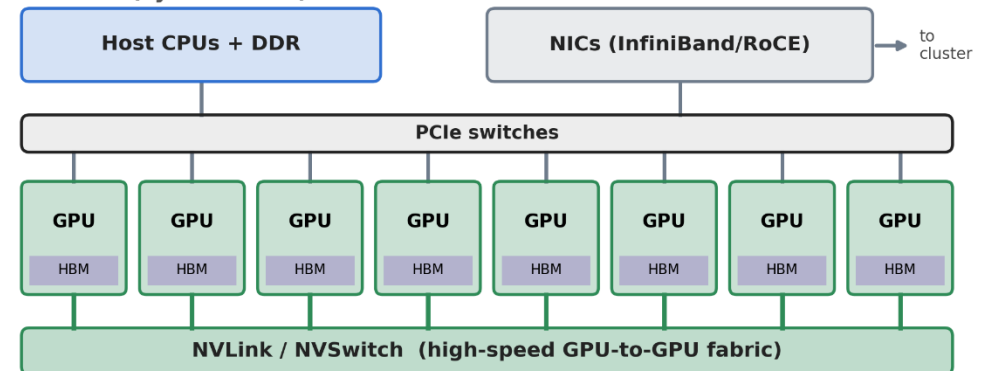
GPU-based Computer (AI accelerated)

① Processor (chip level)



Thousands of small, simple cores

② Server (system level)



- Massively parallel cores for matrix / tensor math
- High-bandwidth memory (HBM) packaged with each GPU
- NVLink inside the server; InfiniBand/RoCE across servers
- Workloads: AI training & inference; ~10 kW per 8-GPU server

AI Computations

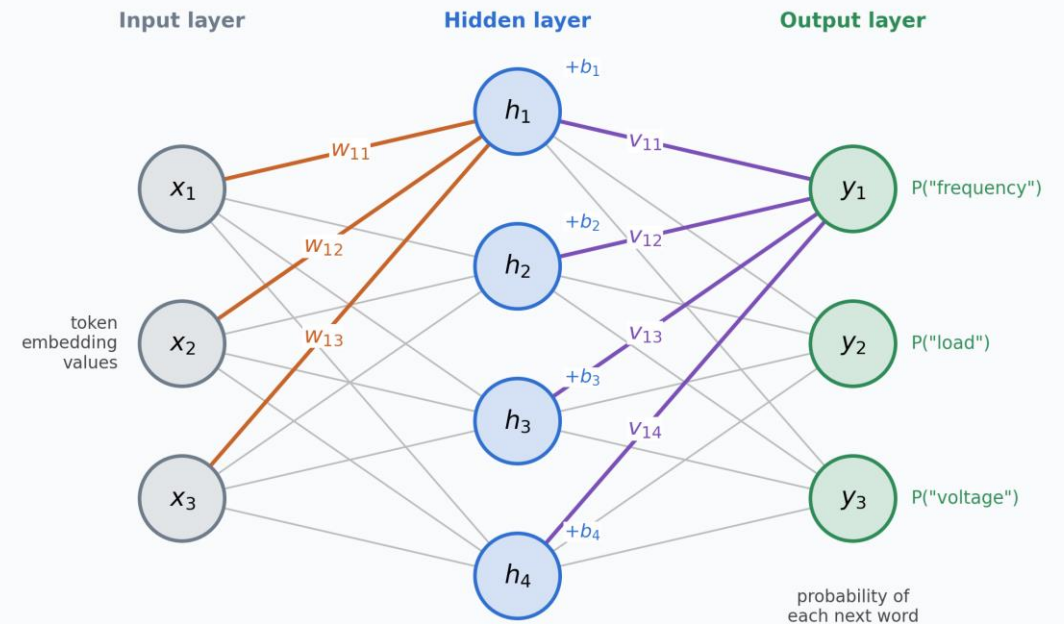
1. Training:

- Model trains on vast data set
- Power-intensive stage with high GPU utilization

2. Inference

- Application of trained models to new data
- Less-power intensive stage
- User driven
- Bursts of power consumptions

A small feed-forward network (building block of an LLM)



Each neuron = weighted sum + bias, then a non-linear activation f (e.g., ReLU/GELU):

$$h_1 = f(w_{11}x_1 + w_{12}x_2 + w_{13}x_3 + b_1)$$

$$\text{Matrix form: } \mathbf{h} = f(\mathbf{W}\mathbf{x} + \mathbf{b}), \quad \mathbf{y} = \text{softmax}(\mathbf{V}\mathbf{h} + \mathbf{c})$$

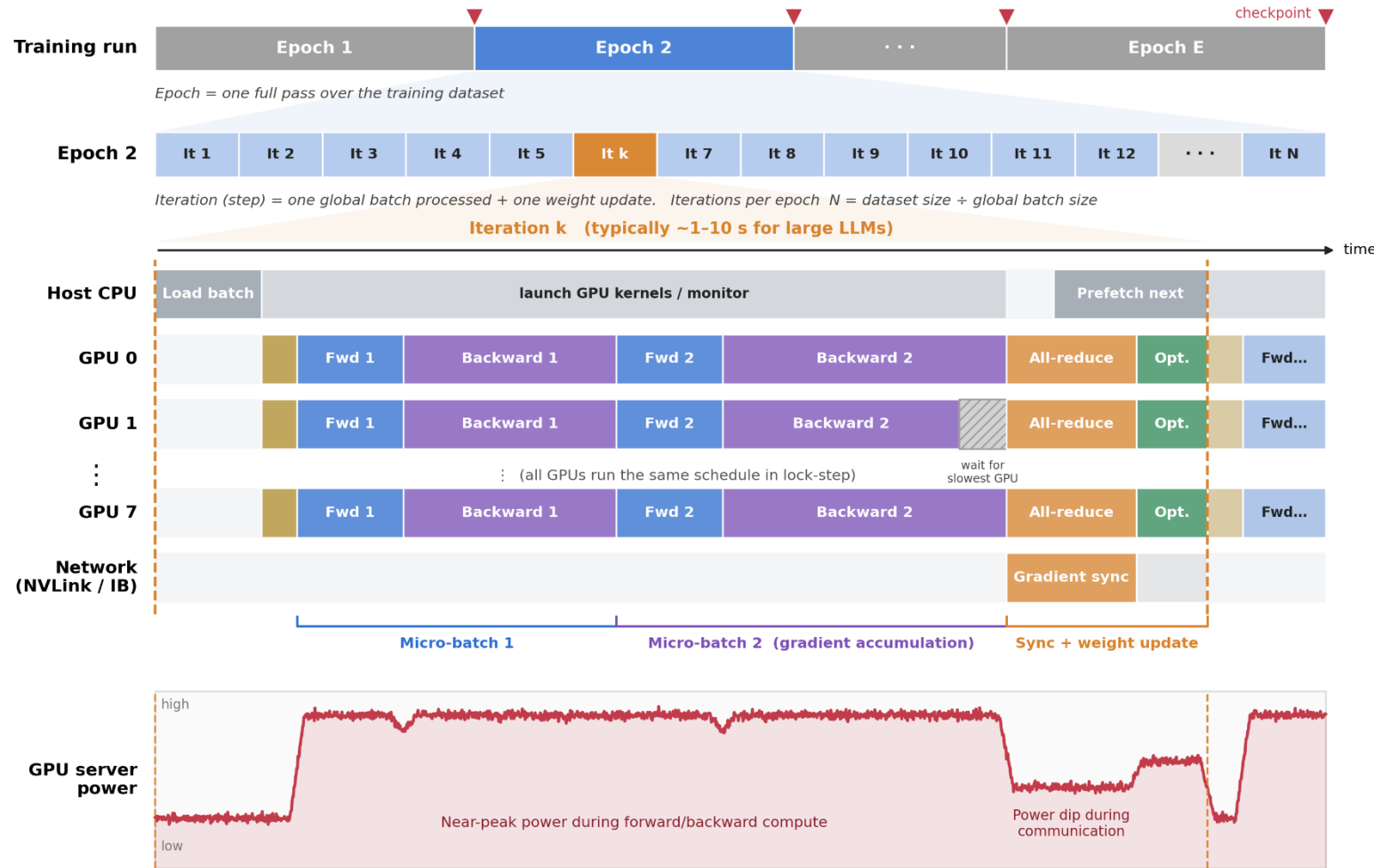
This network: $3 \times 4 + 4 + 4 \times 3 + 3 = 31$ variables

AI Computations

Rough LLM-scale numbers:

- Dataset: about 10 trillion tokens.
- Global batch: about 4 million tokens.
- That gives about 2.5 million iterations per epoch.
- At about 5 s per iteration, one epoch takes about 5 months on a single cluster. Real runs shorten this with more GPUs.

Timing of One Training Iteration of an AI/LLM Model on a GPU Server



Terminology

Epoch

One complete pass over the training dataset (LLM pre-training often uses ≈ 1 epoch over trillions of tokens)

Iteration / step

Process one global batch and update the model weights once

Batch / micro-batch

Samples (token sequences) per update; a micro-batch is the slice that fits in GPU memory

Forward pass

Compute predictions and loss layer by layer

Backward pass

Back-propagate to compute gradients ($\approx 2\times$ forward compute)

All-reduce

GPUs average gradients over NVLink / InfiniBand so all model copies stay identical

Optimizer step

Update weights using gradients (e.g., Adam)

H2D copy

Host-to-device transfer of the batch into GPU memory

Checkpoint

Periodic save of model state to storage for failure recovery

Batch size	Iterations per epoch
8,760 (all data at once)	1
876	10
24 (one day per batch)	365

This pattern repeats every iteration across thousands of GPUs $\rightarrow$ large, periodic MW-scale load swings seen by the grid

Global batch size = micro-batch size $\times$ accumulation steps $\times$ number of GPUs (data-parallel)

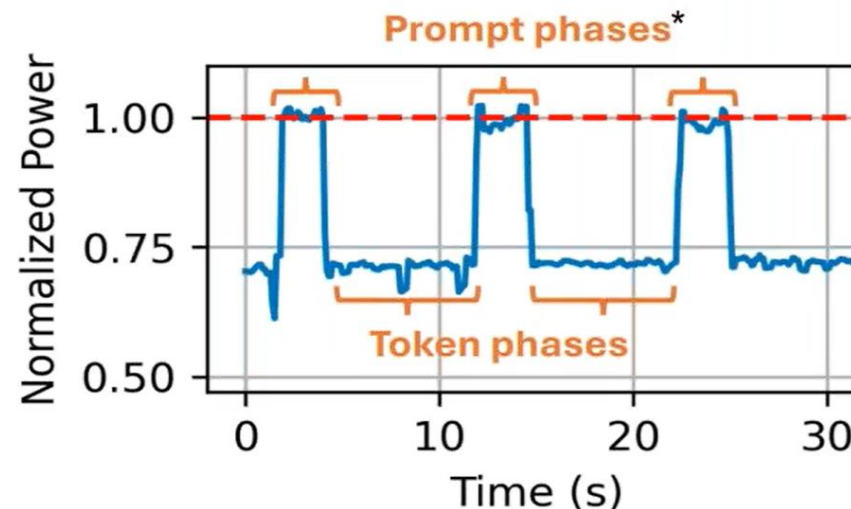
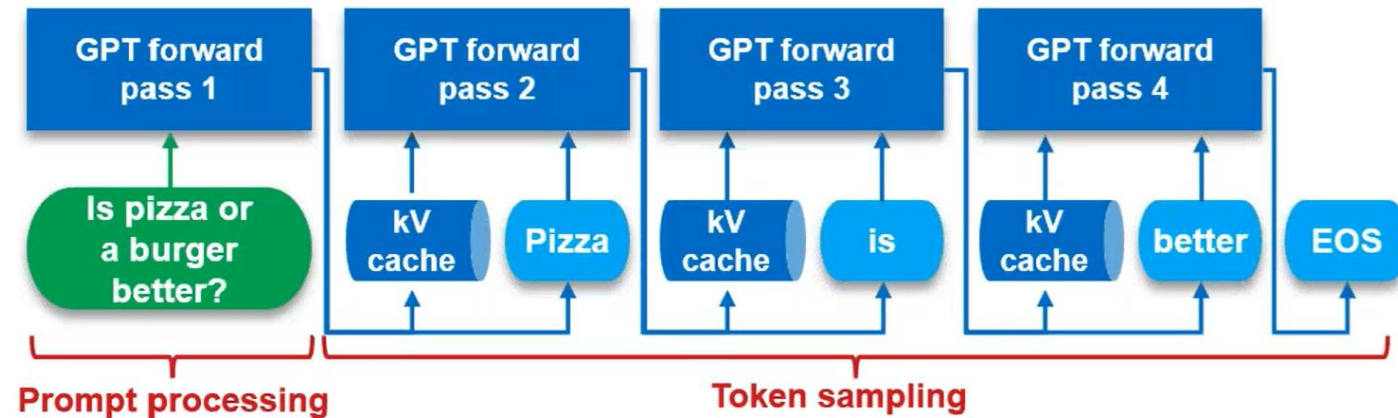
AI Computations

Inference Stage

In general, whenever you send a request to an LLM (ChatGPT, Copilot, etc.), the model will go through two phases:

- 1. Prompt phase:** At the start of each inference request, the system enters a compute-intensive phase where the prompt is processed using parallel calculation.
- 2. Token phase:** Once the prompt phase is processed, the system shifts into a more stable, lower-power state where it performs sequential calculations.

In data centers, inference requests are stochastic over a short time intervals. This means it is unlikely that multiple GPU racks serving inference will start or stop heavy compute at the same time.

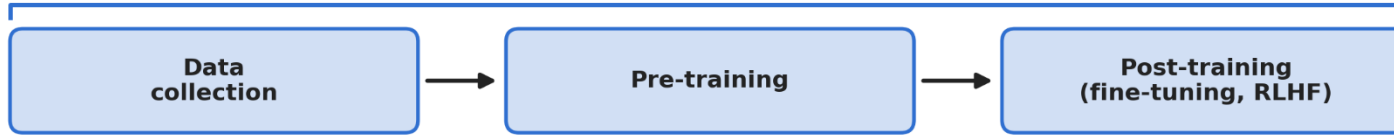


*Image taken from P. Patel et al. "Characterizing Power Management Opportunities for LLMs in the Cloud", Association for Computing Machinery, April 2024.

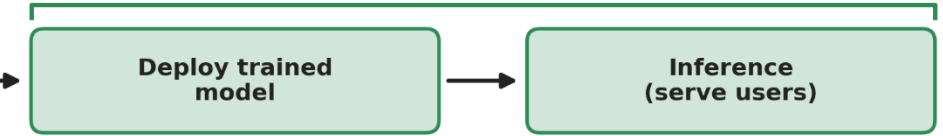
AI Computations

Training vs Inference: Two Stages of AI/LLM Workloads

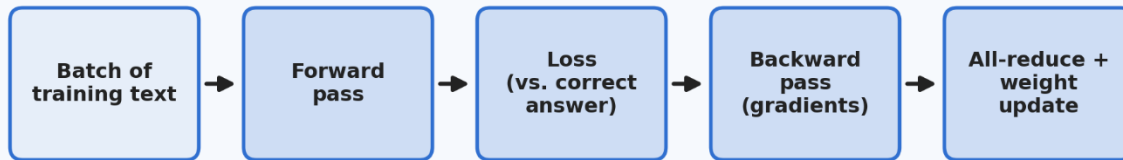
TRAINING — build the model (done once / periodically)



INFERENCE — use the model (continuous)

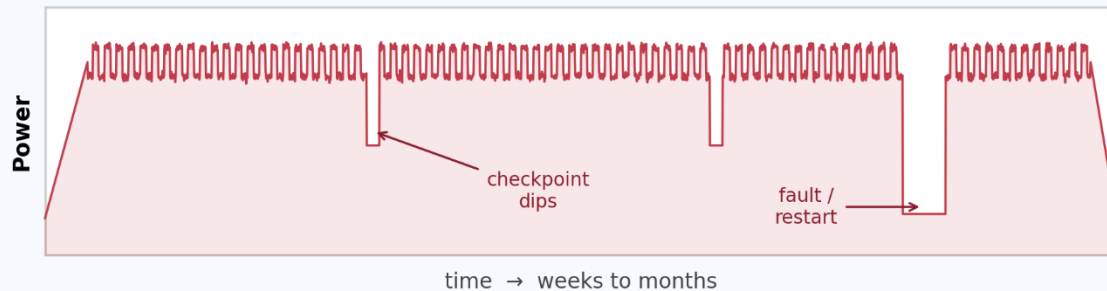


① TRAINING — model learns from data

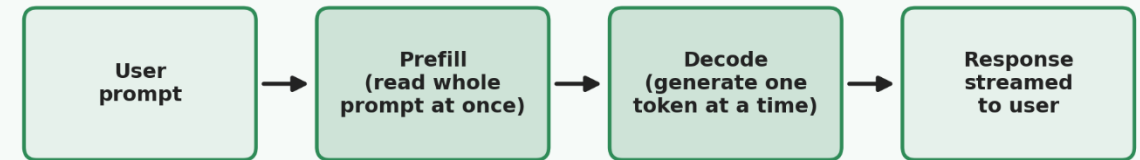


next batch → repeat for millions of iterations (weights change every step)

Sustained near-peak load with fast, synchronized swings every iteration

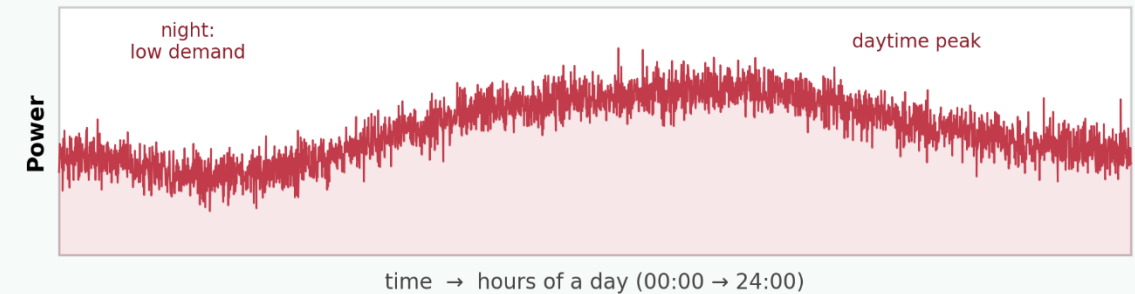


② INFERENCE — model answers requests

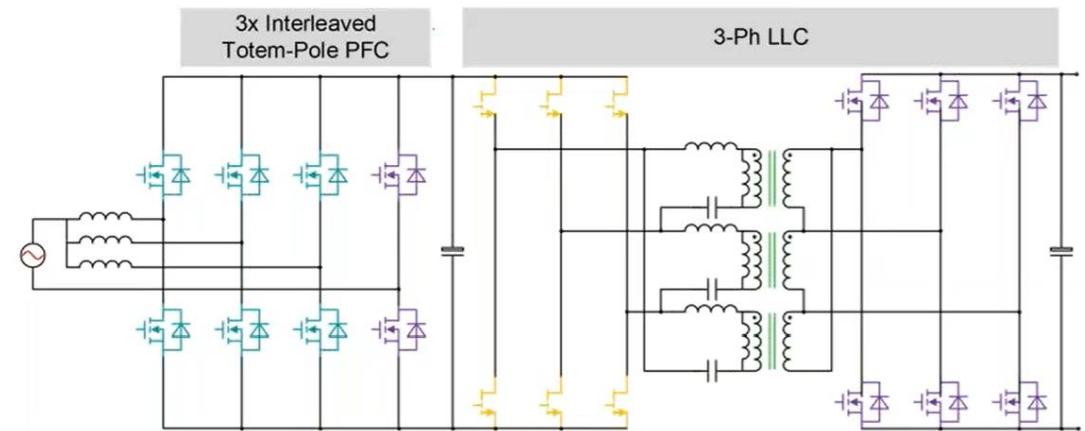
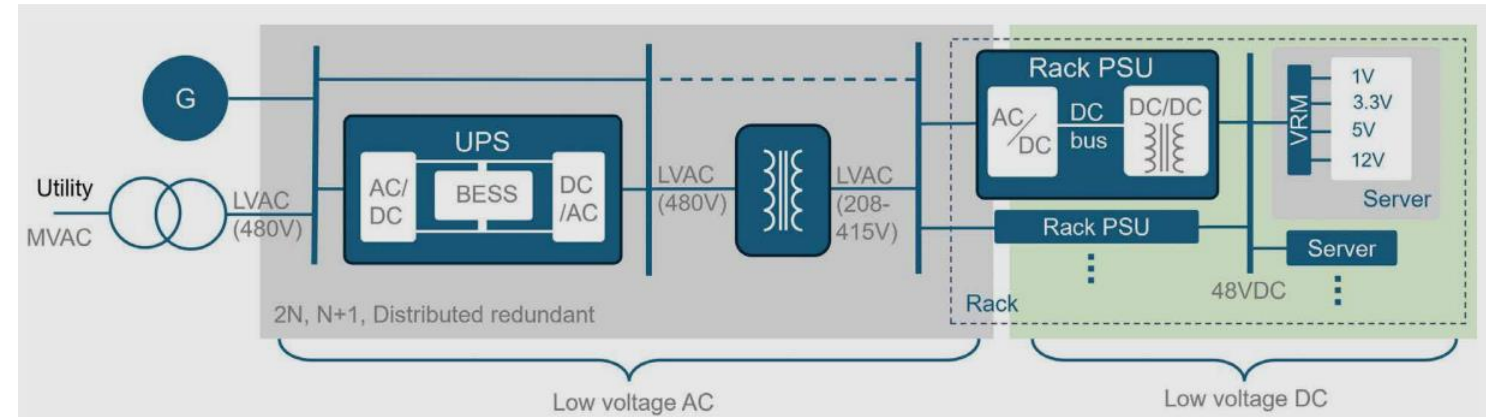
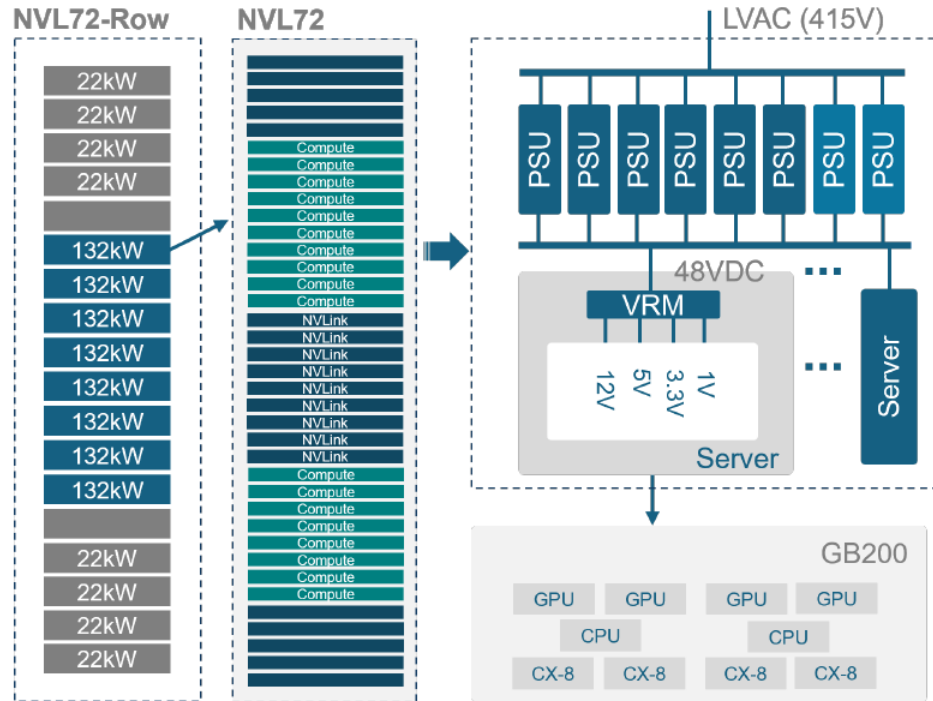


Weights frozen · forward pass only · many users batched together

Load follows user demand: daily pattern plus random request bursts



Case Study

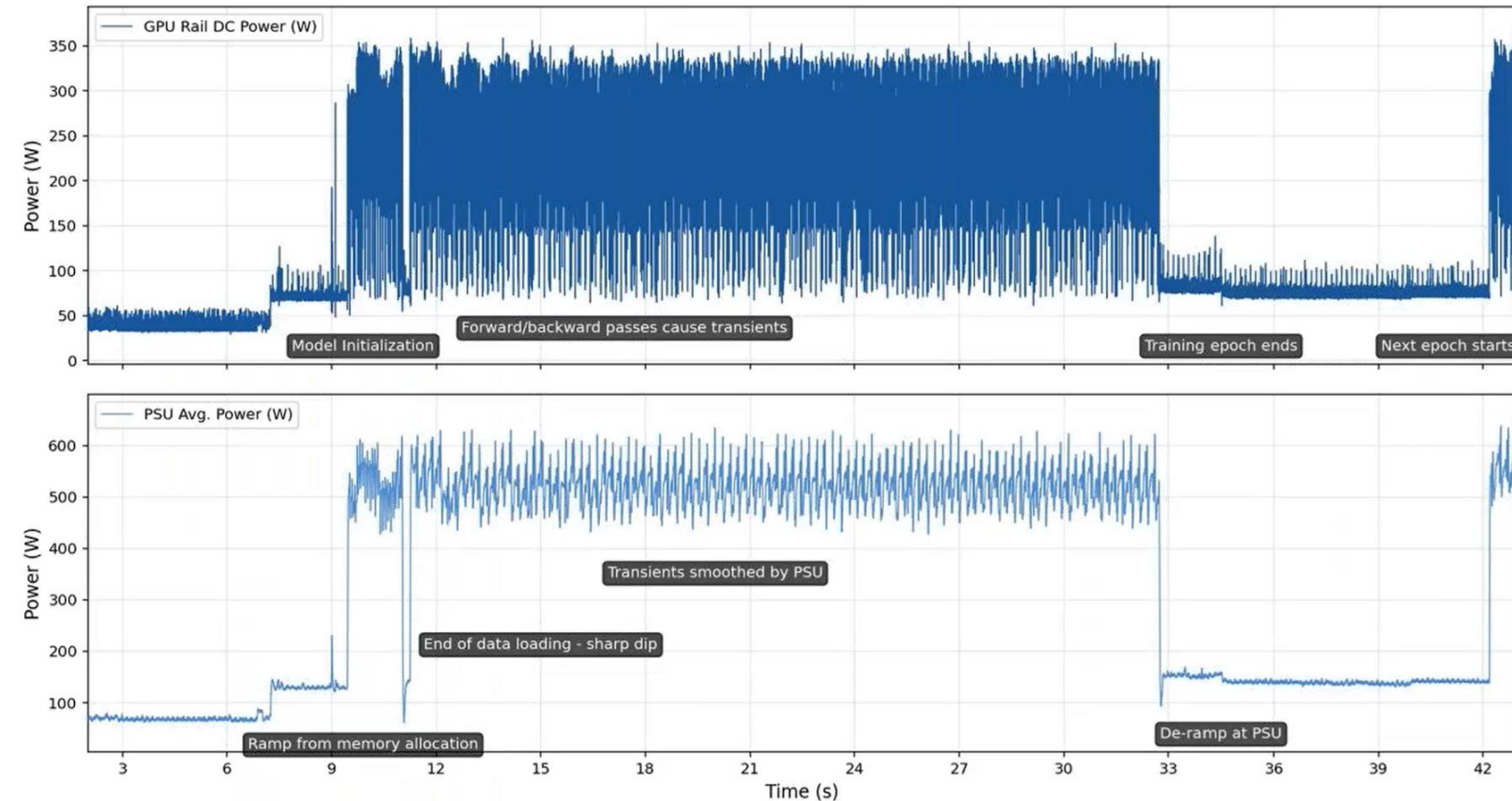


8.5 kW PSU architecture from Navitas

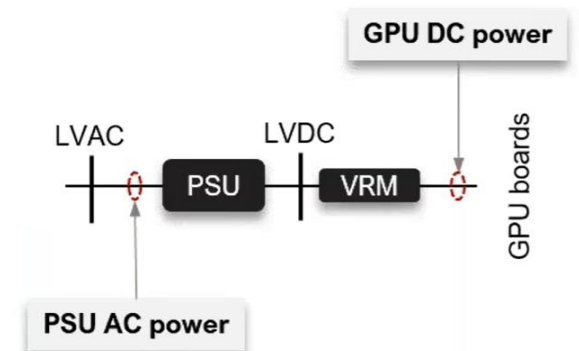
Case Study

LLM Load Examples – Training Epoch

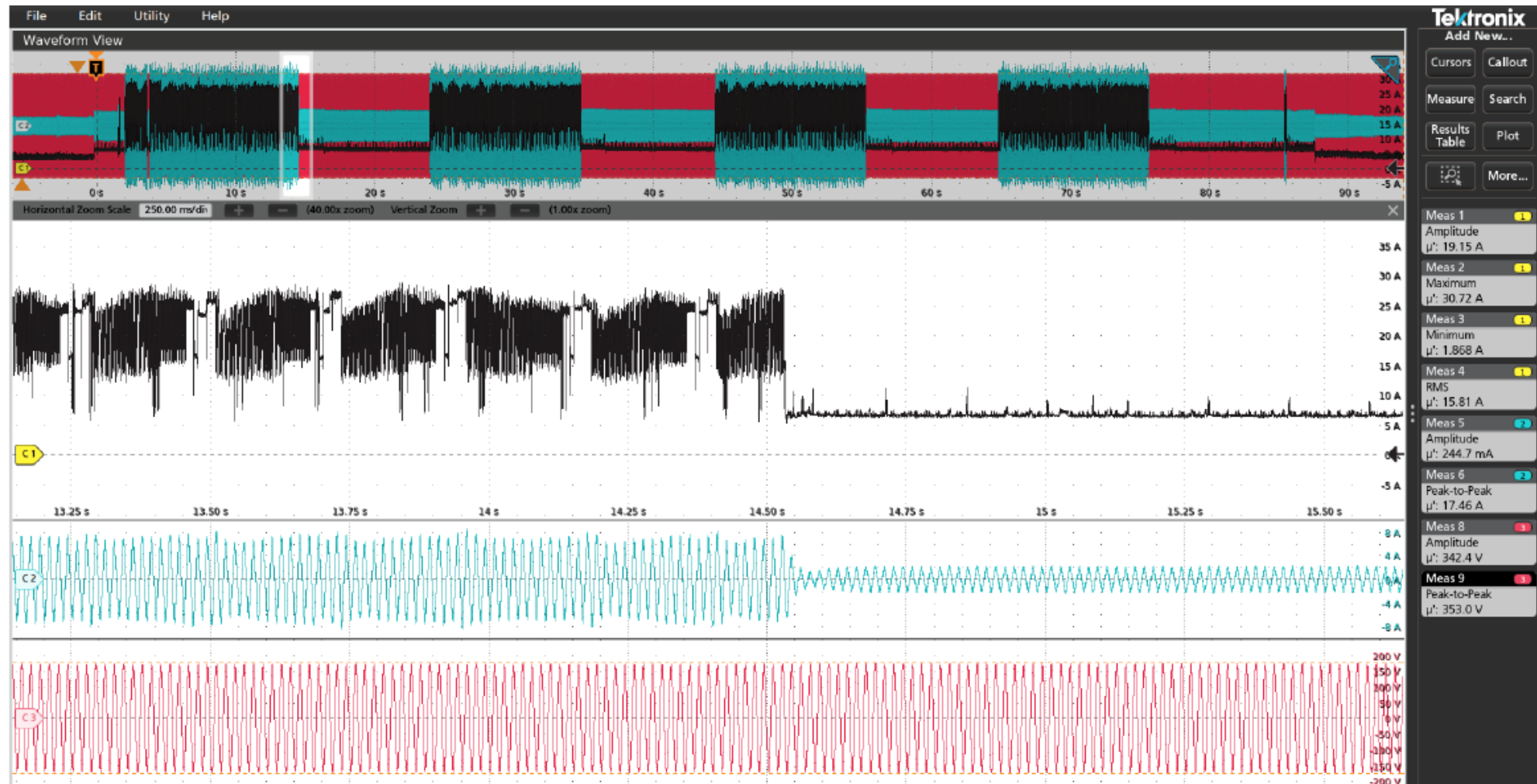
Power Draw During an LLM Training Epoch



- Data is collected from an **NVIDIA RTX 4090 GPU, 1000 W MPG PCIe 5.0 Gold (80+ Gold)** power supply unit¹.
- Measurements captured with an oscilloscope at an 80 μs^2 sampling interval.
- Training performed on LLaMa-3.1 (8B) local Meta model.



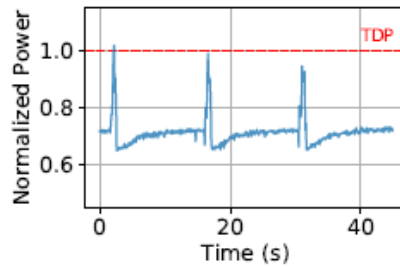
Case Study



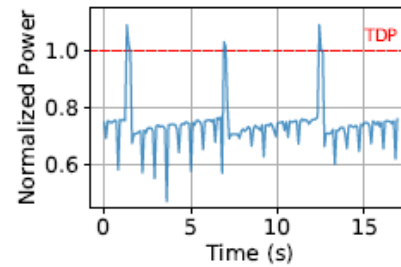
Li, Yuzhuo, and Yunwei Ryan Li. "AI load dynamics—a power electronics perspective." *IEEE Power Electronics Magazine* 13, no. 3 (2026): 94-104.

Case Study

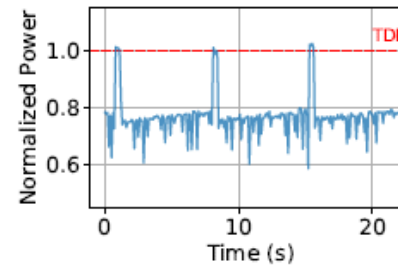
LLM Load Examples: Inference



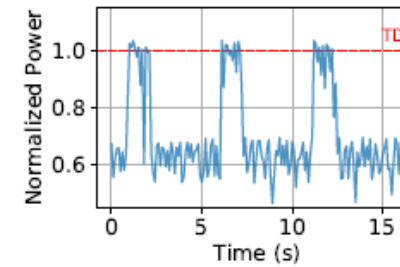
(a) Flan-T5.



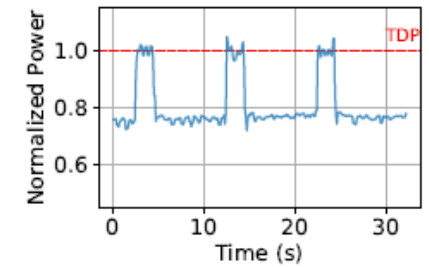
(b) GPT-NeoX-20B.



(c) OPT-30B.



(d) Llama2-70B.

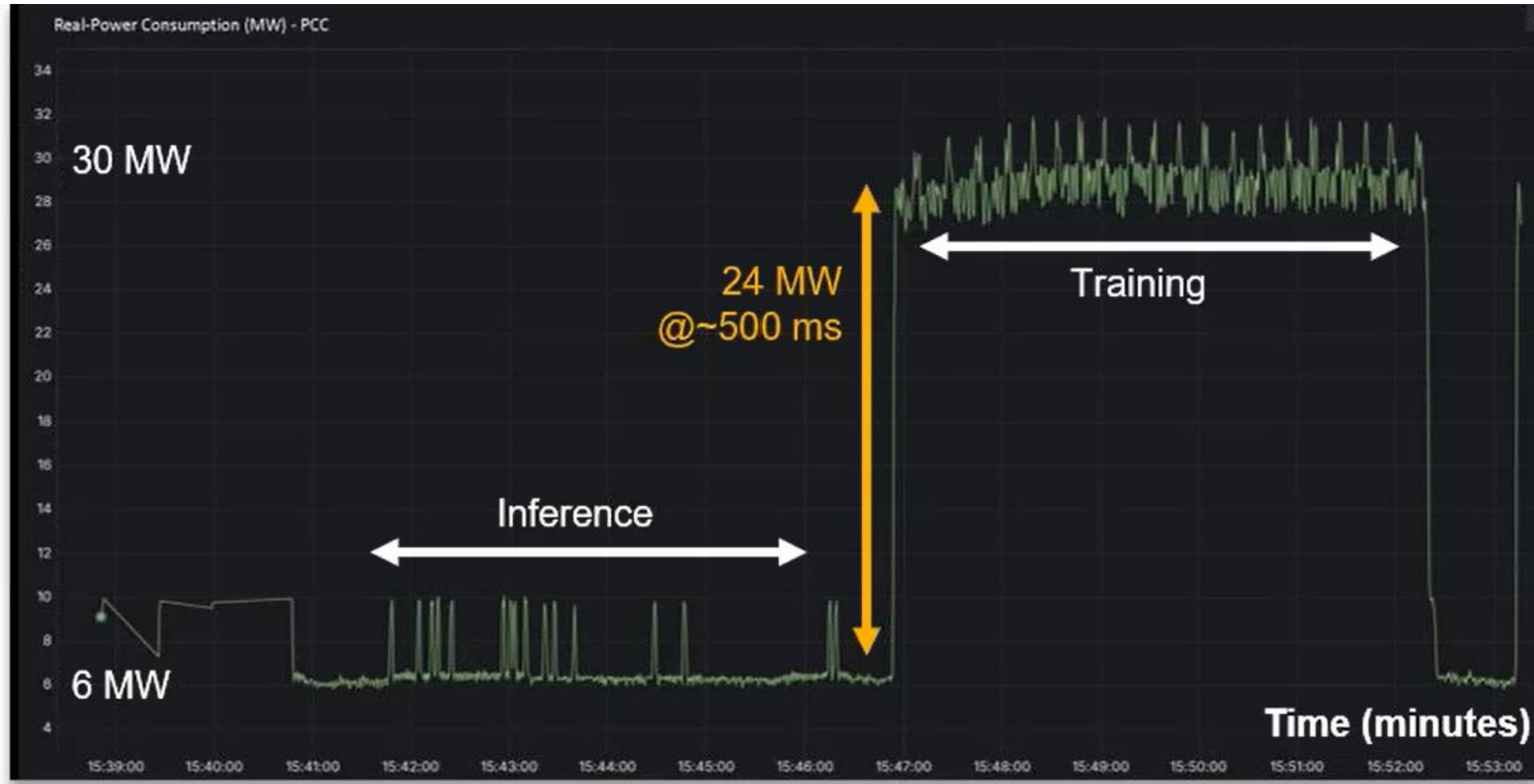


(e) BLOOM-176B.

GPU power usage timeseries for multiple inference models. It shows distinct power usage patterns in prompt (spiky) vs. token phase (longer, more stable, and lower). The phases in each model take different amount of times.

Case Study

Bulding-level Aggregated Data Center Load:



Impact of Grid Disturbances on Data Center

Typical UPS Sequence of Operations

When there is a significant power quality issue, the following happens:

UPS disconnects from the utility and supplies IT loads from a small battery (~5 minutes).

After ~1 minute, **data center will switch over to generators.**

Data center operators will stay on the generators until the power quality issue is over. But . . .

Data centers are very cautious to return to utility; they may run on generators for several minutes to hours.

1

2

3

4

5

If the **power quality issue last longer than ~2–3 minutes**, the data center will automatically startup the generators.

UPS will come off the battery and run on the generators.

The utility provider may see a large load loss, which can cause grid stability issues.

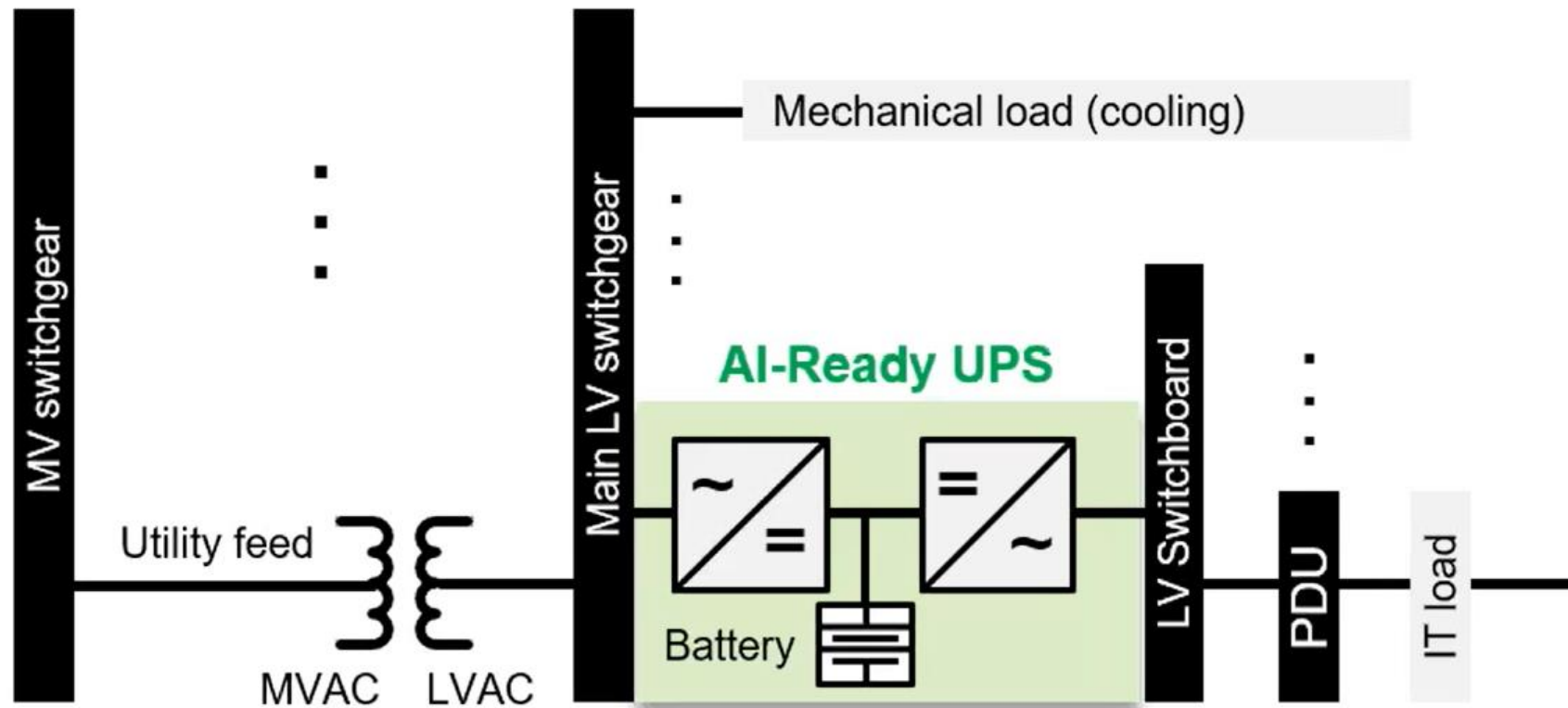
Impact of Grid Disturbances on Data Center

Northern Virginia data center disconnect causes 3 GW power loss on PJM grid

Published on 24 July 2026 at 8:16 am • Written by [Chris Martin](#) • Reading duration : 3 minutes • read 682 times



Impact of Grid Disturbances on Data Center



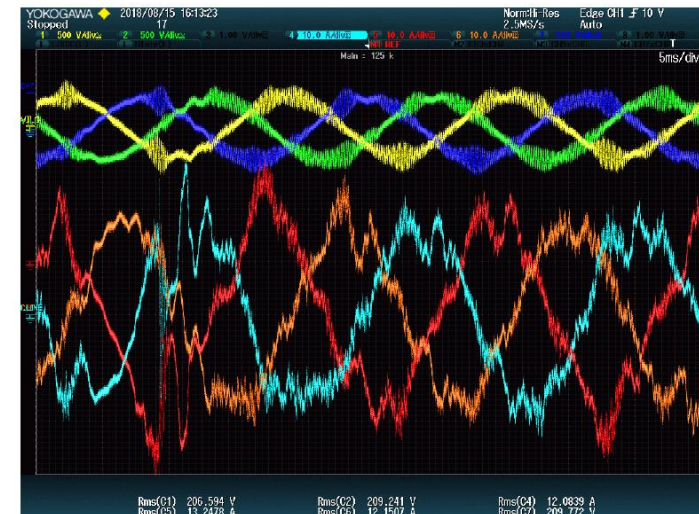
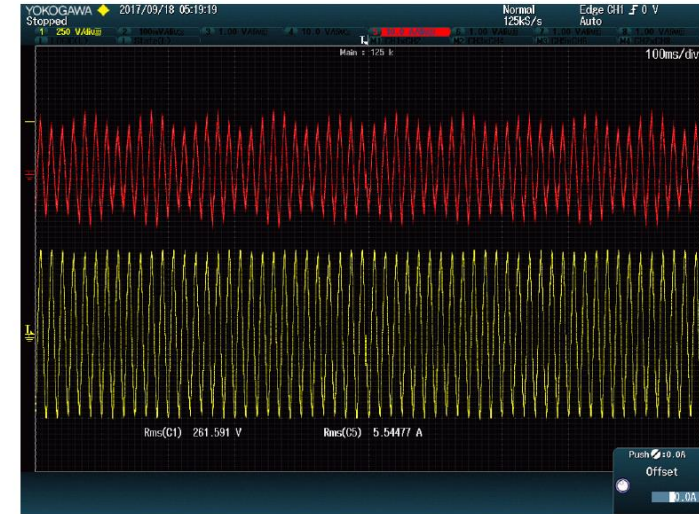
Impact of Data Center on Grid

- Harmonics
- Voltage flicker and temporary over voltages
- Frequency variations
- Sub-synchronous Oscillations

Impact of Data Center on Grid

- An oscillation at 11 Hz is seen in the amplitude of both the current and voltage measured at the tap box to a server rack. Fourier analyses indicated strong harmonics at 49 Hz and 71 Hz in all three phase currents and voltages.
- In some other data centers, high frequency resonances in the 5–10 kHz range were also encountered.

Sun, J., Xu, M., Cespedes, M., & Kauffman, M. (2022). Data center power system stability—Part I: Power supply impedance modeling. *CSEE Journal of Power and Energy Systems*, 8(2), 403-419.



Minimizing Impact of Data Center on Grid

Two solution categories:

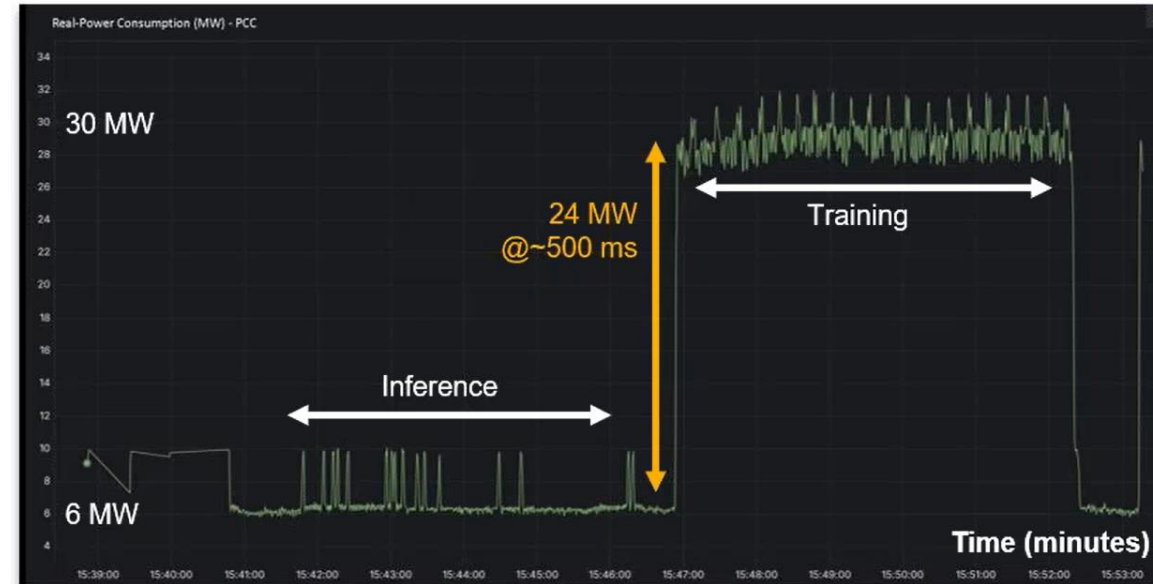
- 1. Software-defined solutions:** Manage power and energy usage across tiers (e.g., software-driven power capping, frequency locking, idle power management, power oversubscription).
- 2. Grid-compatible solution:** Apply grid-interactive control strategies to mitigate operational challenges from LLM workloads.
 - On the grid side (front-of-the-meter)
 - On the load side (behind-the-meter).

Supercapacitors (e.g., E-STATCOM)

Enhanced UPS/PSU

Grid-forming BESS

Hybrid Solutions



*MW load at the building level in an AI data center**

Minimizing Impact of Data Center on Grid

NVIDIA GB300 NVL72: A Hybrid Rack-Level Power Smoothing Solution

- **Combines** software-based power management and rack-level energy storage.
- **Smooths** GPU power ramps during startup, steady operation, and shutdown.
 - Power controller caps GPU ramp-up at workload start.
 - Power-burn mode mitigates ramp-downs at the workload end or during checkpoints.
- **PSU capacitors (65 J) buffer energy:** charge at low GPU utilization, discharge during high utilization.

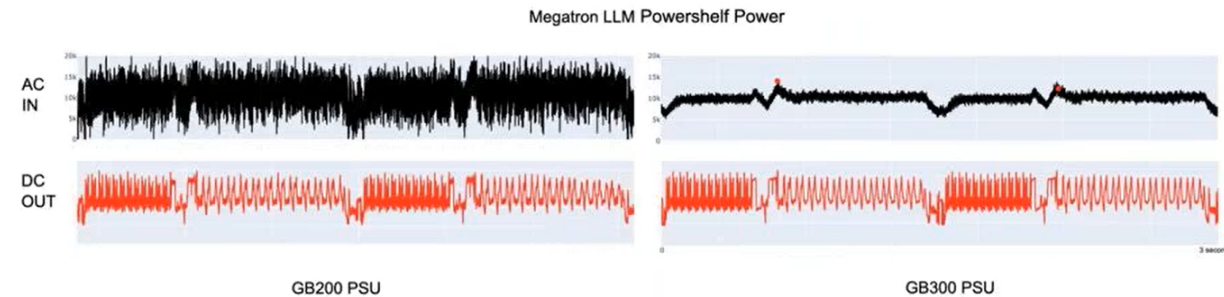
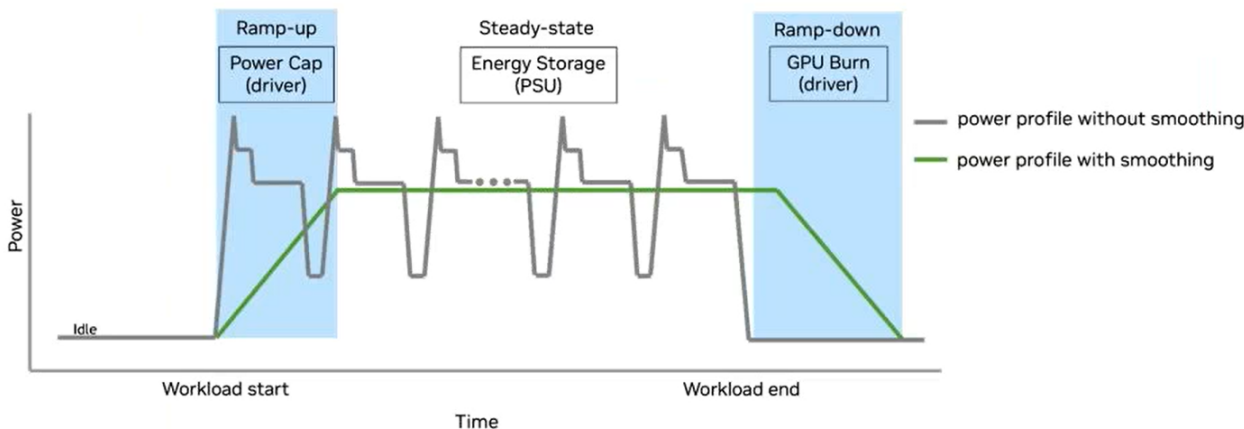


Image from <https://developer.nvidia.com/blog/how-new-gb300-nvl72-features-provide-steady-power-for-ai/>

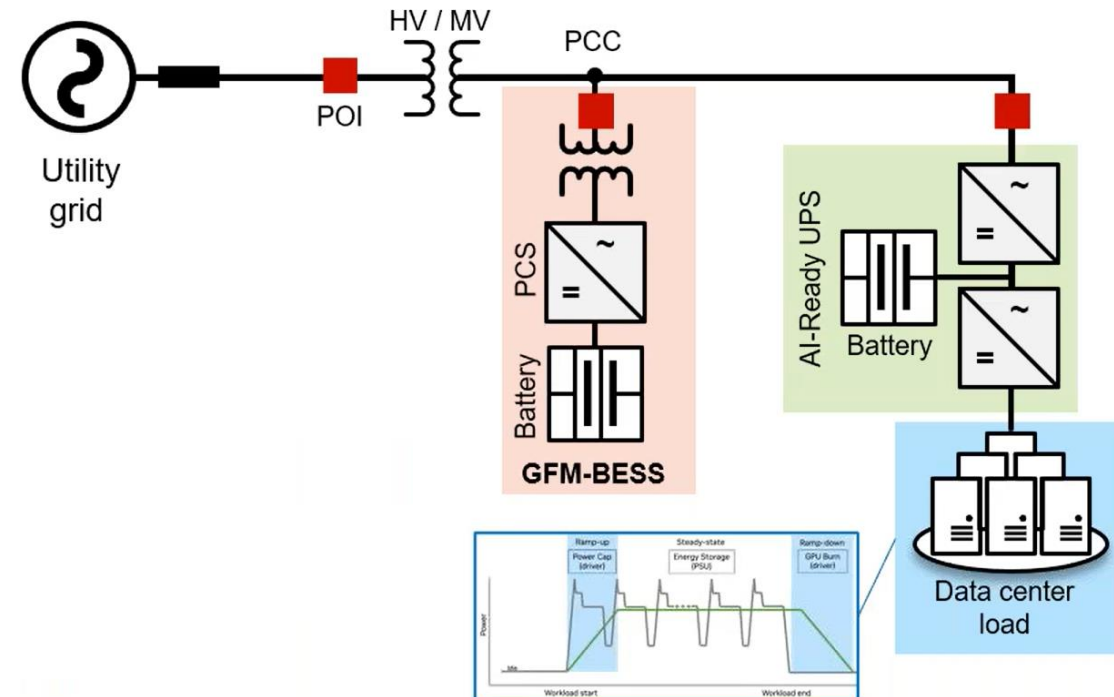
“Understanding AI Load Profiles and Impacts on Power Systems”, by Quanta Technology

Minimizing Impact of Data Center on Grid

Layered “Rack-Level” + “Grid-Level” Solutions

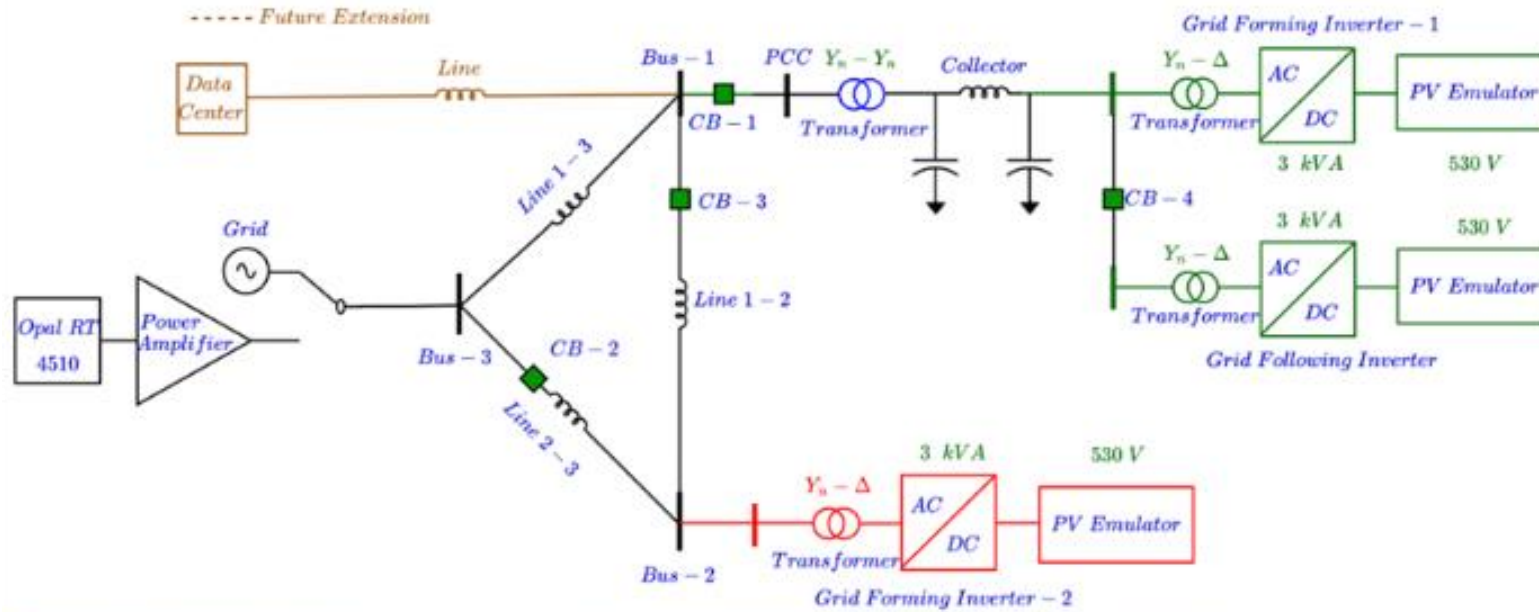
- **Complementary operation:** Combines software enhancement, AI-ready UPS (fast rack-level response), and BESS (grid-level smoothing and energy capacity).
- **Improved reliability:** Provides load shaping, demand flexibility, and extended ride-through for AI workloads.

Benefits: +	Challenges: -
Load smoothing/shaping	Higher cost
Demand flexibility	Increased footprint
Grid services	



“Understanding AI Load Profiles and Impacts on Power Systems”, by Quanta Technology

Facility at Indian Institute of Science



Air-cooled Data Center being Developed



Server 1

Server 2

- **Resiliente GPU Server**

- 1 x AMD EPYC 9124 (16C, 32T, 3.0 GHz, 64MB, 200 W)
- 1 x 4U- 2GPU capable server with 2x 1600 W power supply
- 4 x 32GB DDR5 RDIMMS @ 4800 MTs
- 1 x 1TB NVMe M.2 Enterprise SSD
- 1 x Nvidia RTX 6000 Ada 48GB (GPU)

COURSE FEE

Students: INR 10000+GST
Post-Docs: INR 12000+GST
Faculty: INR 14000+GST
Industry: INR 16000+GST

WHO CAN BENEFIT

Engineering students, post-docs, academic faculty, and industry professionals in electrical power, data centers, and AI infrastructure seeking expertise in grid integration, power delivery, cooling systems, stability, and simulations.

CONTACT US

Centre for Continuing Education (CCE)
Indian Institute of Science (IISc) Bengaluru
Phone: 080-2293-2055/2491/2247
www.cce.iisc.ac.in



OBJECTIVES

This workshop explores the grid integration of AI data centers, covering power delivery, cooling systems, stability challenges, cybersecurity, market tariffs, and practical simulation sessions to prepare participants for future demands.

ELIGIBILITY

B.Tech/B.E

CONTENTS

Refer to page 2



WORKSHOP ON GRID INTEGRATION OF AI DATA CENTERS: CURRENT STATUS, AND PATH AHEAD

Date : 17 to 19 December, 2026
Time : 9:00 AM to 06:00 PM

Course Coordinator:
Prof. Sarasij Das, Department
of Electrical Engineering, IISc



SPEAKERS

- Prof Vladimir Terzija, Newcastle University, UK
- Dr. Deep Deka, MIT, USA
- Prof Saurabh Amin, MIT, USA
- Prof Sukumar Kamalasadan, University of North Carolina at Charlotte, USA
- Prof Sairaj Dhople, University of Minnesota, USA
- Prof Nilanjan Ray Chaudhuri, Penn State University, USA
- Prof Parikshit Pareek, IIT Roorkee
- Prof Pramod Kumar, Indian Institute of Science
- Prof Parimal Parag, Indian Institute of Science
- Prof Praveen Ramamurthy, Indian Institute of Science
- Prof Kaushik Basu, Indian Institute of Science
- Prof Vinod John, Indian Institute of Science
- Prof Sarasij Das, Indian Institute of Science

SESSIONS

- Data Centres in the Modern Grid: AI Load Growth, Demand Trends and Grid Challenges
- Energy Aware AI Algorithm Scheduling and Running in Data Centres
- Power Delivery Architecture and Electrical Load Characteristics of AI Data Centre

- Cooling Systems and Cooling Loads in Data Centres
- Energy Water Connection in Data Centre and Impact on Environment
- Data Centre Planning, Siting and Flexibility
- Medium-Voltage Converters and Solid-State Transformers for Data Centres
- Power Electronics for AI Data Centre Servers
- Power System Dynamics and Stability Issues with Data Centre Loads
- Grid Codes and Modelling of Data Centres for Grid Connection Studies
- Limiting the Impact of AI Data Centres on Fatigue Life of Thermal Turbine Generators in the Grid
- Cybersecurity of Data Centre Infrastructure
- AI/ML Models for Data Centre Control
- Tariffs, and Electricity Markets for Data Centre Loads
- Building a Co-Simulation Suite for AI Data Centre-Power Grid Interactions: Design and Early Results
- Hands-on Session on MW-scale AI Data Centre Simulations

HANDS-ON SESSIONS

Hands-on Session on MW-scale AI Data Centre Simulations will be based on MATLAB Simulink

ACCOMODATION

Accommodation is limited and is available on first come first serve payment basis.

First 14 participants will be accommodated at JVH guest house (IISC Campus)

Double occupancy rooms are there at the guest house. The tariff is 1000+ GST per person per day.

Next 20 participants will be accommodated at Hotel Shri Durga International, bus stand, 35/1, Gayathri Temple Rd, near Yeshawanthpura, Dr.Ambedkar Nagar, Yeshawanthapura, Bengaluru, The tariff is 1200 + GST

IN-PERSON WORKSHOP

Scan QR to know more



[Registration link](#)

**Registration
Deadline: 15
November 2026**

VENUE

Department of Electrical Engineering, Javanica Marg, Indian Institute of Science (IISc), Bengaluru, Karnataka 560012

Conclusions

- Data Centre load has unique characteristics
- More study required to understand evolving server technologies
- Data Centre design should consider impacts on the grid
- Grid Codes need to specify requirements for Data Centres