IMPERIAL

Department of Electrical and Electronic Engineering

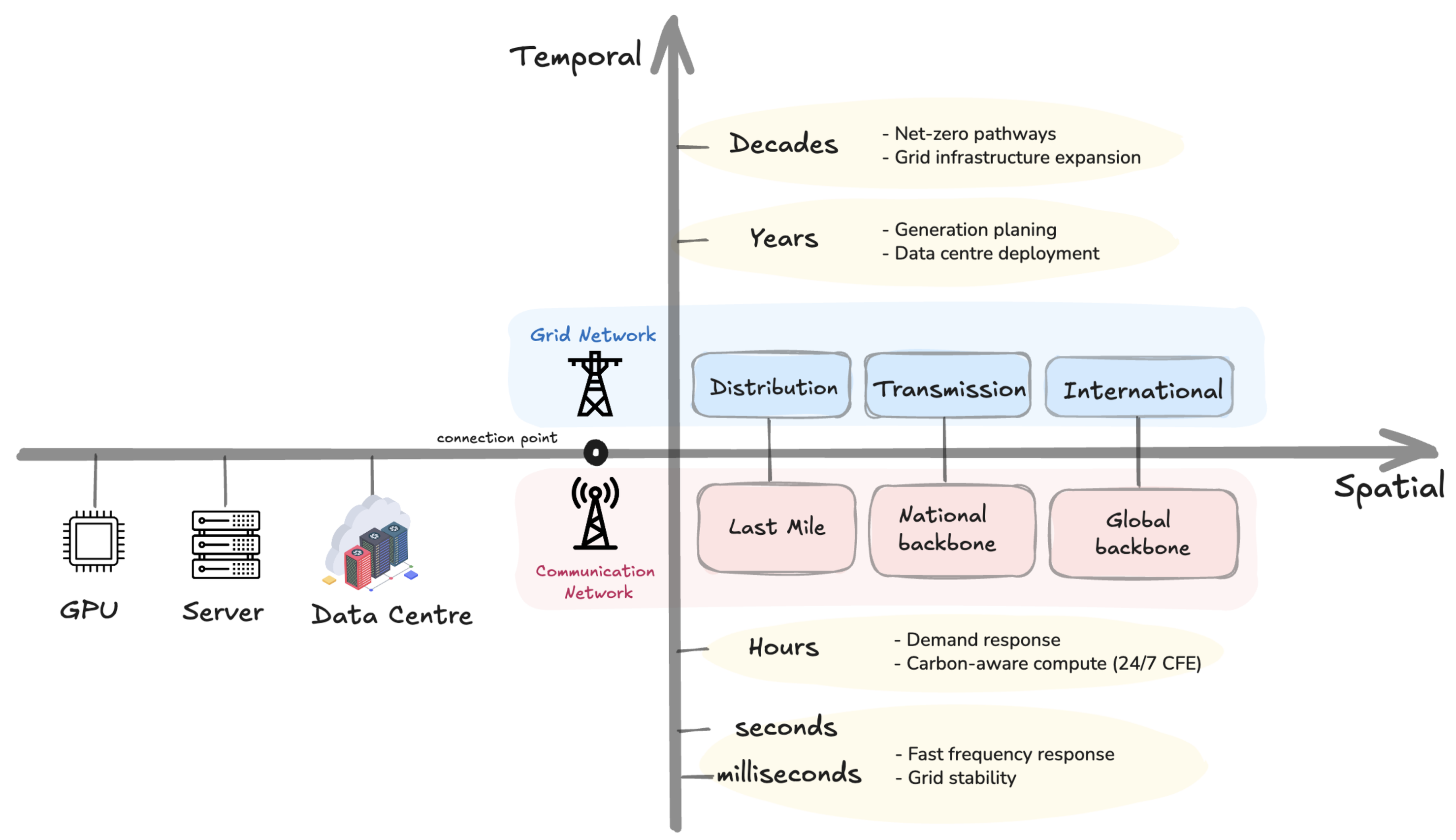
AI Data Centres in Power Grids: Multi–Time-Scale Pathways to Sustainable Integration

Challenges, Opportunities and Policy Implications

Ying Yu,
Dr Fei Teng, Prof Mark O'Malley,
Dr Yiren Zhao
Department of Electrical and Electronic Engineering, Imperial College London
Contact: Ying Yu, ying.yu24@imperial.ac.uk

Background and motivation

- The rapid expansion of artificial intelligence (AI) has driven exponential growth in data centres(DCs) and their associated energy consumption.
- Integrating AI data centres into the power grid spans multiple time scales, from millisecond-level control of servers to long-term planning of infrastructure and policy.
- Understanding these **multi-time-scale interactions** is essential for achieving sustainable and resilient integration of AI data centres into future energy systems.



Challenges

1 **Rising Power Intensity:** Increasing deployment of GPUs and TPUs leads to higher rack-level power density and overall energy consumption.

2 **AI Workload Dynamics:** AI workloads, particularly Large Language Model (LLM) training, can induce rapid power swings that may trigger forced or sub-synchronous oscillations (SSO), thereby stressing grid stability

3 **Evolving Grid Conditions:** Future power systems are becoming more decentralised and inverter-dominated, with reduced inertia and greater demand for reserve capacity.

4 **Regulatory and Interconnection Barriers:** Long interconnection queues and the absence of unified grid codes highlight the need for new standards, such as Fault-Ride-Through (FRT) requirements.

Main research objectives

To develop **multi-time-scale strategies** that transform DCs into grid-interactive, flexible, and reliable assets within future power systems.

Workflow

Open-access trace

Power trace of own experiments

Power Pattern Analysis
- Profile and predict AI-driven power consumption in DCs

scale-up & scale-out to hyperscale DCs

DC Load Modelling
- Develop multi-time-scale models of DC

System Integration
- Evaluate grid impact and design appropriate integration strategies

Category	Component	Flexibility Source	Response Time
Compute	CPU/GPU Dynamic Voltage Frequency Scaling (DVFS)	Fine-grained power modulation through dynamic voltage and frequency scaling; enablesfast frequency response (FFR)	μs – ms
	GPU Power Capping	Enforce an upper bound on GPU board power for fast, bounded load reduction (using clock control)	ms – s
	Workload Scheduling/ Throttling (Focus of the case study)	Defer, throttle, or migrate workloads within Service Level Agreement(SLA) constraints to manage power	s – h
	Geo-distributed Load Shifting	Reallocate workloads across sites for energy, price, or carbon optimisation	min – h
Power	Uninterruptible Power Supply (UPS)	Rapid charge/discharge for grid-support services; enables FFR	ms – s
	Backup Generator/ Transformer	On-site generation or standby asset to modulate load or provide backup at longer time scales	few minutes
Cooling	Chillers/ Compressor Control	Modulate compressor frequency or speed to adjust cooling load	30 s – 1 min
	Air Fans Control	Adjust fan speed or static pressure for secondary cooling flexibility	2–10 min

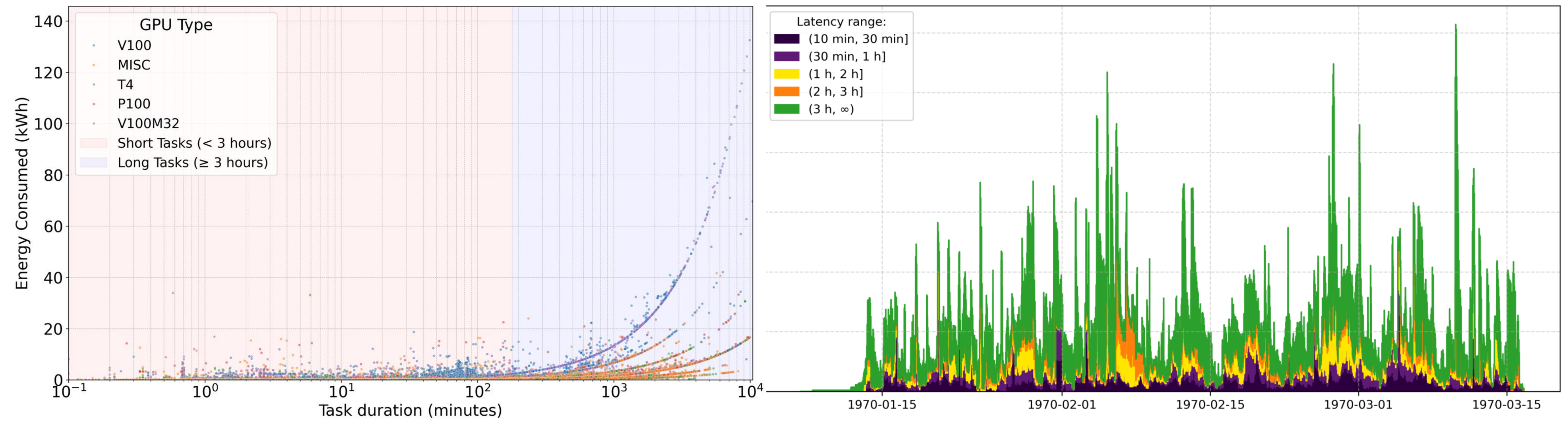
Summary of Flexibility Sources in DCs and Their Response Times (summarised from [1])

Case Study

- Analyse the two-month real-world workload trace from Alibaba: AI workloads submitted to **6742 GPUs** of the Platform for AI production cluster
- Non-time-sensitive tasks with queuing latency (the delay between task submission and execution) can be flexibly scheduled and coordinated through an aggregator to provide grid-support services without violating DCs' Quality of Service.
- DC power consumption model (Simplified)

$$P_t^{DC} = \sum_{i=1}^{N} G_i \cdot P_i^{GPU,nom} \cdot \bar{\chi}_{it}$$

- G_i: number of GPUs of **type** i deployed.
- $P_i^{\mathrm{GPU,nom}}$: **nominal power** of a **single GPU** of type i.
- $\bar{\chi}_{it} \in [0,1]$: **average utilization** of all GPUs of type i at time t.



Overview of task duration and energy consumption in the trace (figure left).
Stacked time series of total power consumption categorised by latency ranges (figure right, note: all original date and time information was anonymised and reset, even though the data was originally recorded between 2019 and 2020) [2]

Results

- High-latency tasks (≥10 min) account for only 5.5% of tasks but over 20% of total power use.
- The [3 h, ∞) category alone contributes 10.1% of power while representing just 0.7% of the workload
- Rescheduling these long-running jobs can significantly reduce peak power demand.

References:
[1] S. Zhang, R. Addanki, T. Shankar, et al., "SustainDC: Benchmarking for Sustainable Data Center Control," Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS), Vancouver, Canada, Dec. 2024.
[2] A. Caprara, Y. Yu, F. Teng, A. Junyent-Ferre, E. Bullich-Massagué, and M. Aragues-Penabna, "Flexibility potential assessment of Data Centers using non-critical IT workloads: a quantitative case study based on Alibaba Cluster Data," manuscript submitted to International Journal of Electrical Power & Energy Systems, Aug. 2025.

LEVERHULME TRUST