Multiscale Grid Intelligence to Fight AI Data Centre Grid Defection:

Unlocking a Faster, Cheaper and Cleaner On-Grid AI Rollout

By Thomas Morstyn, Yihong Zhou, Ian Whitfield

Abstract

Rapid advances in AI computing capabilities have led to a race to build larger-scale data centres with escalating power demands. AI companies are now planning data centre sites with up to 5 GW capacity, and the International Energy Agency projects that data centre demand could grow to 1,700 TWh by 2035 (6% of current total global electricity demand). Power grid connections are becoming a major bottleneck, and it can now take more than 5 years to receive a grid connection for a new data centre in the US and Europe. The challenge of getting grid connections has driven a new trend towards grid defection, where data centres are being fully or partially supplied by local microgrids with dedicated generation assets. In some regions this is a voluntary choice, but in most cases, it is the only option available for building a new data centre. For most data centres, it is likely that a substantial grid connection would be beneficial from both an economic and sustainability perspective, due to how main power grids unlock generation economies of scale and enable lower cost renewable integration. However, inefficient AI data centre grid defection is being driven by a reliance on pre-existing technical approaches and institutional barriers to change. In this article, we propose "multiscale grid intelligence" as a new framework to fight AI data centre grid defection and support closer coordination between data centres and power grids. The proposed framework embeds power grid-focused analytics into how data centres are planned and operated, accounting for co-optimisation opportunities across spatial scales (from data centre rack-level distribution to national transmission) and time scales (from second-to-second frequency balancing to years-ahead network planning). Across these scales, we identify a range of opportunities for intelligent planning and operation which could offer significant value if implemented together. The article discusses the value multiscale grid intelligence can offer power system and AI stakeholders and proposes key areas for future research.

1. Introduction

Since the latest artificial intelligence (AI) boom was kicked off by the release of ChatGPT in November 2022, electric power grid demand forecasts and expansion plans have been thrown into disarray. The potentially transformative impact of AI across the economy has led to a race between AI companies to

build larger-scale data centres with escalating power demands. While data centres over 100 MW were previously considered “hyperscale”, AI companies are now planning sites with up to 5 GW capacity.

The rapid growth in AI data centre power demand comes at a critical time for power grid decarbonisation. Net-zero requires a large-scale build-out of low-carbon generation and transmission capacity to replace fossil fuels and enable the electrification of transport and heating. Renewable integration has been supported by the falling costs of solar and wind, but grid congestion, instability and balancing costs are growing challenges. Although sources of low-carbon firm flexibility are emerging, including battery storage and demand response, delivering security of supply with these technologies remains costly.

These factors have contributed to grid integration becoming a major bottleneck for AI data centres. Projects require complex grid connection studies to be carried out, and long delays are introduced if power infrastructure upgrades are required. A report by Ember found that the average grid connection time for data centres is 4 years in Europe, and 7 to 10 years in legacy hubs including the UK, Ireland and the Netherlands. Similar connection times are observed in US data centre hubs, with Bloomberg reporting data centre connection times of 4 to 7 years in Virginia.

The challenge of getting grid connections has driven a new trend towards grid defection, where data centres are being fully or partially supplied by local microgrids with dedicated generation assets. In some regions this is a voluntary choice, but in most cases, it is the only option available for building a new data centre. Building a data centre with a local microgrid can be faster than obtaining a grid connection and opens up sites where the grid is already heavily congested, as well as remote areas with limited grid infrastructure. Also, an advantage of data centre developers building their own power infrastructure is that it ensures financial risks are not shifted to transmission system operators (TSOs) and retail customers. The high costs and long lead times of grid expansions, coupled with the significant uncertainty around the evolution of AI technology, means that anticipatory grid buildout ahead of AI investments risks creating stranded assets.

However, there are reasons to be concerned by this trend towards data centre grid defection, both in terms of sustainability and economics. When built for speed, data centres tend to heavily rely on fossil fuel generation. A prominent example is xAI’s Colossus Data Centre, which was initially constructed in 112 days using 15 gas turbines to provide 150 MW of generation capacity at a site with an 8 MW grid connection. Major AI companies have had sharp emissions increases in recent years, including Google, which, despite being a pioneer in “carbon-aware computing”, reported a 51% rise in emissions between 2019 and 2024.

Data centre microgrids could be supplied by low carbon sources, for example solar/wind plus battery storage, biogas, geothermal or nuclear. However, apart from very remote sites, bypassing the grid is unlikely to be economically efficient due to the value the grid plays in unlocking economies of scale, connecting regions with supply/demand diversity and efficiently sharing reserve capacity. The grid has a particularly important role in renewable integration, since solar and wind benefit from being distributed across areas with high generation potential and being linked with energy storage and demand flexibility.

Even if built with fully renewable supplies, pushing data centres off-grid could still threaten the net-zero transition by creating underutilised infrastructure and clogged supply chains. At the same time, multi-year delays to data centre grid connections will prevent countries fully benefiting from the prosperity offered by the latest advances in AI and could lead to data centres migrating to places with low environmental standards. However, these are not the only two options. In this article, we argue that there is significant scope for more integrated and flexible approaches to data centre grid integration, which can unlock win-wins for TSOs and data centre developers.

This motivates our proposal, which is to fight AI data centre grid defection and support closer coordination between data centres and power grids, with a new framework we call "multiscale grid intelligence". The aim is to embed power grid-focused analytics into how data centres are planned and operated, accounting for co-optimisation opportunities across spatial scales (from data centre rack-level distribution to national transmission) and time scales (from second-to-second frequency balancing to years-ahead network planning). Across these scales, we identify a range of opportunities for intelligent planning and operation that remain fragmented and early-stage but could offer significant value if implemented together. We present the challenges posed by AI data centre grid defection for power grids, explain how the proposed framework for multiscale grid intelligence can help enable a faster, cheaper and lower carbon AI rollout, and discuss key areas for future research. Since our main concern is hyperscale AI data centres, we focus on high voltage transmission grid connections rather than smaller data centres in low or medium voltage distribution networks. However, much of the proposed multiscale grid intelligence framework would still be applicable.

2. The AI Power Grid Challenge

Data centres are used both for training AI models and running these models to complete inference jobs in response to user queries. AI training is time consuming and power intensive, with state-of-the-art models requiring 10,000+ graphics processing units (GPUs) running for several months. However, the overall energy demand of AI is dominated by inference, with AI companies now serving billions of queries per day. Despite the rapid growth so far, there is a high degree of uncertainty around future AI data centre demand. The International Energy Agency (IEA) estimated that global data centre demand could grow from 415 TWh in 2024 (1.5% of total global electricity demand) to be as high as 1,700 TWh by 2035 or remain as low as 700 TWh.

Key sources of uncertainty for AI energy demand growth are outlined in Figure 1. One major source of uncertainty is the trajectory of AI performance improvement. Given the current rate of improvement, Google DeepMind has forecast that artificial general intelligence (AGI) could be achieved as early as 2030, where AI models reach the competency of humans across most cognitive tasks. However, this is far from guaranteed. The current pace of AI data centre build-out is being driven by fierce competition between AI companies to be at the forefront of the technology, and AI investment is also being supported by governments based on its potentially vital future importance for national security, economic competitiveness and data sovereignty. These drivers could change quickly depending on how AI capabilities develop.

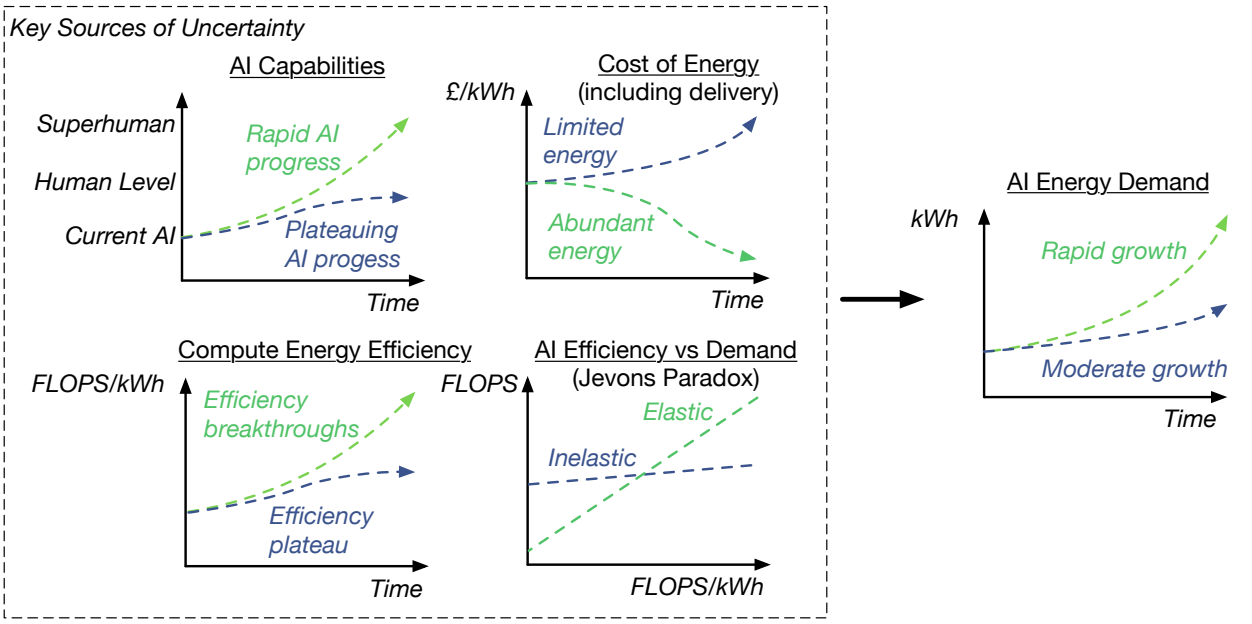

Figure 1: Overview of key sources of uncertainty for future AI energy demand growth. These include the rate AI capabilities improve, the full cost of energy including generation and transmission, the rate of compute energy efficiency improvement, and the extent to which energy efficiency improvement leads to greater AI usage (i.e. the strength of Jevons Paradox).

The power intensity of future AI technology is also highly uncertain. Analysis by Epoch AI on scaling suggests that power will be the key bottleneck up to 2030, rather than chip production, data scarcity or model training latency. However, AI algorithms and hardware are both seeing rapid innovation which could change this. On the algorithmic side, it has been shown that model distillation can significantly reduce power consumption, with only moderate performance penalties. However, chain-of-thought techniques, which boost performance by breaking down queries into intermediate reasoning steps, significantly increase energy use per query. Current assumptions could also be changed by hardware advances in areas including edge computing, optical computing, and quantum computing. Finally, Jevons Paradox–the phenomenon that energy efficiency improvements can counterintuitively lead to an overall increase in energy demand due to increased economic value–makes it difficult to predict the impact of AI energy efficiency improvements.

Compared with other sources of load growth, AI data centres place a particularly heavy burden on the grid due to their concentrated power demand and high utilisation rate. Because latency is a particular concern for inference-focused AI data centres, they are often sited near communication hubs where power grids are already congested. It has also been found that AI demand can have large transient variations linked to sudden multi-GPU training starts and stops, as well as swap-overs to backup power.

The high level of uncertainty, and potential for rapid AI data centre demand growth, are major challenges for power grid planning. Power grid upgrades must go through complex design, permitting and construction phases and key pieces of equipment have multi-year supply chains with limited spare capacity. According to the IEA, for the US and Europe, the average lead times and costs for transmission

cables and transformers have nearly doubled since 2021, and it takes between 5 and 15 years to build new transmission grid infrastructure.

Building AI data centres ahead of adequate transmission and generation investments could lead to increased grid congestion, higher balancing costs, and greater instability risk, as well as delays to the electrification of transport and heating. However, expanding the grid in anticipation of AI demand poses difficult questions for how investment costs are recovered, particularly if demand falls short of forecasts. In many liberalised electricity markets, costs associated with transmission upgrades, flexibility services, renewable support schemes and capacity remuneration mechanisms, are recovered through energy-based charges on customer bills, or through general taxation. This creates fairness issues and financial risks, particularly if forecast AI data centre demand fails to materialise. Box 1 provides a UK and Republic of Ireland-focused industry perspective on AI power grid challenges from Pure Data Centres Group.

Box 1: An Industry Perspective on AI Power Grid Challenges from Pure Data Centres Group

Data centres have turned pockets of the UK and Ireland into digital strongholds, but the power network that feeds them is creaking. The M4 corridor from London to South Wales now hosts Europe's premier data centre cluster because of its dark fibre backbone. Dublin's T50 loop has created a similar major data centre market in Ireland. This in turn encouraged investment from cloud computing providers to centre their infrastructure around those areas.

While the edges of these data zones will creep outwards to accommodate less latency dependent instances, major AI and cloud providers still prefer proximity to those cloud hubs as AI inference services emerge. This means that demand in some constrained areas is red-hot. Grid utilisation at peak periods can exceed 95% in these areas. In the UK, new data centre projects are being given grid connection target dates into the mid-2030s, while Ireland faces similar capacity availability pressures.

Power infrastructure built over 40 years ago, for a different era, is now struggling to cope with burgeoning digital enterprises and use cases covering every aspect of our lives. Unless we unlock new capacity, Britain and Ireland are at risk of hampering their digital economy and inadvertently encouraging hyperscale customers to look overseas.

To make things even more difficult, governments are racing to decarbonise power generation. By 2030, the UK is committed to 95% low-carbon electricity and Ireland to 80% renewable electricity. This is putting a strain on ageing transmission networks, creating congestion and curtailment that punishes both power generators and consumers. Britain already spends close to £1 billion a year paying wind farms to switch themselves off because the network cannot move energy from the windy north to the demand-hungry southeast at all times of generation. National Energy System Operator (NESO), the UK's TSO, has warned that, without radical reform, constraint costs could quadruple by the end of the decade. Ireland's electricity-related non-compliance fines are already estimated into the billions of Euros under EU rules. In addition, national energy strategies in both the UK and Ireland mandate that 20 to 30% of demand must be flexible by 2030, heightening the urgency for new data centre operating models.

3. Grid Defection

Data centres are designed to be highly reliable due to the value of the computing infrastructure they host, and the high cost of failures, data loss and downtime. The Uptime Institute's Tier Classification System provides a widely adopted standard for evaluating and comparing data centre reliability. Data centres used for high value computing applications, including AI inference and training, generally target a Tier 3 (99.982% uptime) or Tier 4 (99.995% uptime) reliability target.

These are higher than power grid reliability targets, which in the US and Europe vary between 99.90% and 99.97% (8 to 2.4 hours loss of load expected per year). To achieve these targets, data centres need to be able to continue autonomous islanded operation with local backup power supplies during grid faults. The Uptime Institute recommends that Tier 3 data centres be built with 72-hour backup power and that Tier 4 data centres be built with 96-hour backup and two fully independent power distribution paths.

While the reliability tier dictates the backup supply requirements, there are also important design choices for the primary power supply (which provides power during normal conditions). The first decision is whether to build a grid connection, with the alternative being a fully autonomous microgrid with dedicated generation/storage capable of continuously supplying the data centre. If a grid connection is built, then the follow-up question is what the power rating of this connection should be, and whether it should be supplemented by dedicated local generation assets. These options are outlined in Figure 2, and Table 1 provides a high-level comparison between them.

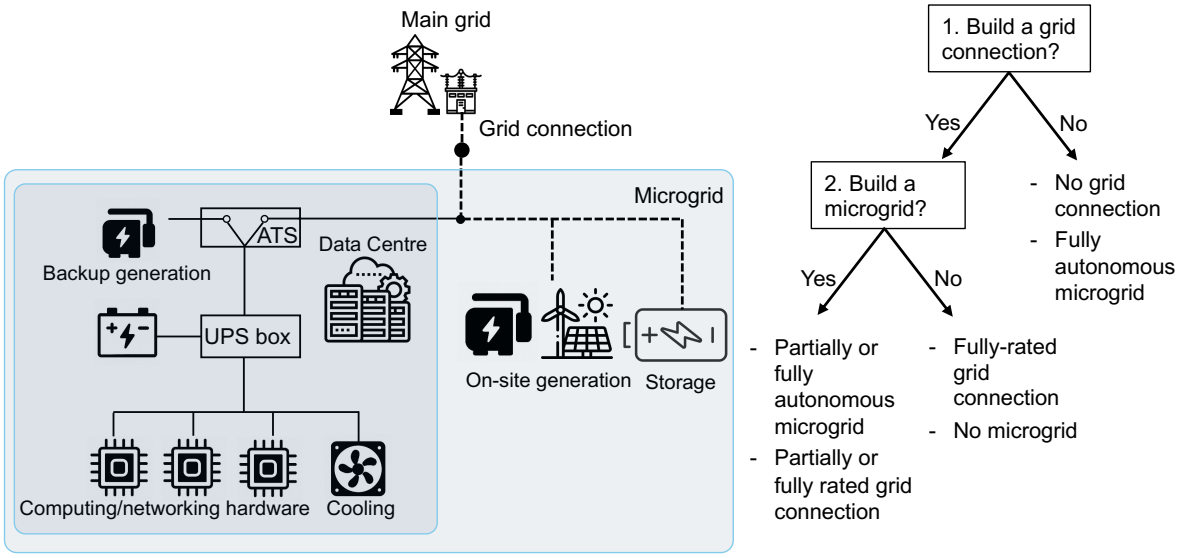

Figure 2: Data centre grid connection and microgrid design options.

Table 1: High-level comparison between data centre grid connection and microgrid options. The ratings for each feature are qualitative and relative. For each set of option, features are rated as either the same as a grid connection with no microgrid (=), higher (↑), much higher (↑↑), or possibly higher or lower depending on detailed design (~).

Grid Connection	Microgrid	Construction Cost	Resilience	Export Capacity	Connection Speed
Fully rated	Fully Autonomous	↑↑	↑↑	↑↑	=
	Partially Autonomous	↑	↑	↑	=
	None	=	=	= (zero)	=
Partially rated	Fully Autonomous	↑	↑	↑	↑
	Partially Autonomous	~	~	↑	↑
None	Fully Autonomous	↑	~	=	↑↑

The AI boom, coupled with the difficulties of obtaining grid connections, is driving a trend towards partially and fully off-grid data centres. Bloom Energy forecasts that by 2030, 27% of data centres will primarily rely on onsite generation, up from only 1% today.

To get up and running quickly, the xAI Colossus data centre initially operated almost fully autonomously with 150 MW of onsite gas generation supplementing an 8 MW grid connection. Subsequent upgrades have expanded the grid connection by 150 MW and expanded its computing hardware from 100,000 to 200,000 GPUs. OpenAI's Stargate data centre campus in Abilene Texas includes plans for a 360 MW onsite natural gas plant. Outside of New Braunfels in Texas, CloudBurst and Energy Transfer are planning a data centre site with a dedicated 1.2 GW natural gas plant. In Tolar Texas, Sailfish has plans for a 5 GW 2,600-acre data centre campus, where the first 600 MW will be supplied by onsite natural gas, which will later be supplemented by a grid connection, and onsite solar and nuclear generation. Box 2 describes a 68 MW off-grid data centre microgrid developed by Pure Data Centres Group in Dublin.

A grid connection already offers lower carbon emissions than dedicated gas generation in many European and US data centre hubs, and this gap will continue to grow if power grid decarbonisation programmes are sustained. Combined cycle gas turbines have emissions from 320 to 450 gCO_2/kWh (Timera estimates procurement lead times of 3 to 4 years), while gas reciprocating engines have emissions from 400 to 600 gCO_2/kWh (with lead times from 1 to 2 years). In 2024, the estimated average electricity carbon intensity was 19 gCO_2/kWh in France, 125 gCO_2/kWh in the UK, 256 gCO_2/kWh in Ireland, 321 gCO_2/kWh in Germany, and 370 gCO_2/kWh in the Netherlands. In the US, 2024 estimated electricity carbon intensity was 138 gCO_2/kWh in California, 332 gCO_2/kWh in Texas and 384 gCO_2/kWh nationally.

Although most off-grid data centres are being developed with dedicated gas generation, some projects are focusing on low-carbon sources. Soluna is developing a 770 MW portfolio of data centres collocated with wind and solar. Outside of Houston, ECL's TerraSite-TX1 project has plans for a 1 GW hydrogen powered data centre. Google has partnered with Kairos Power with plans to deploy 500 MW of advanced nuclear reactors for data centres by 2035, and Amazon is working with Energy Northwest towards the development and deployment of small modular reactors.

In theory, there are multiple advantages to building data centres with substantial main grid connections. Firstly, power grids can accommodate larger generation plants which benefit from economies of scale and allow renewables to be more flexibly sited based on generation potential. They can also help smooth out capacity requirements by connecting regions with more temporally diverse demand and supply, and by connecting a wider set of flexible resources. These factors mean that a well-designed power grid should be able to provide electricity more cheaply, cleanly and reliably than an autonomous microgrid. In

addition, even for data centre microgrids with the potential for full autonomy, a partially rated grid connection is likely to have significant value, both as an additional source of redundancy, and by enabling excess electricity to be exported. For full autonomy, a microgrid's generation capacity needs to cover the highest possible coincident demand from the data centre's computing, networking, and cooling systems, which will vary with workload patterns, computing job phases (e.g. data loading, training, validation) and weather conditions. The ability to export excess generation is particularly valuable for sites with wind and solar generation.

A high-power grid connection may be uneconomical for certain remote sites, but this does not apply to most AI data centres, particularly those used for inference, which are usually located near communications hubs to reduce latency. In many cases, inefficient AI data centre grid defection is being driven by a reliance on pre-existing technical approaches and institutional barriers to change, rather than power grid fundamentals. The 5+ years it can take to get a grid connection for a new data centre in the US and Europe is a non-starter for companies aiming to be at the forefront of AI technology. Also, grid connected sites are often subject to additional charges above their wholesale energy costs, related to the costs of network investments, flexibility services, capacity remuneration mechanisms and renewable support schemes. These charges are often allocated between sites based on simplified mechanisms (e.g. volume of energy imported or grid connection capacity), rather than being reflective of the actual cost of supplying the site given its demand profile, network location and level of flexibility.

The large-scale integration of renewable generation means that system costs are no longer driven by energy use, but instead by the peak generation and transmission capacity required during periods with low renewable supply. It is commonly assumed that AI data centres have limited potential for demand flexibility. However, this does not account for the high value of grid flexibility during periods of extreme stress and the potential for data centre flexibility to enable faster lower-cost grid connections. The case study in Box 3 shows how flexibility could particularly accelerate data centre grid connections in Great Britain.

Recent research has demonstrated that flexible AI computing could allow data centres to competitively offer grid flexibility services, and the scope for flexibility is further increased by also considering the cooling system and local generation/storage assets. Key unresolved questions include: how to unlock this flexibility at scale? how to incentivise data centre operators to deliver it? and, how to effectively integrate it into power system planning and operation?

Box 2: Dublin Off-Grid Data Centre Case Study from Pure Data Centres Group

Pure Data Centres Group acquired land in the greater Dublin metro area in 2017, zoned to support data centre development. Initial discussions with the electricity network operator suggested grid capacity would be constrained until the mid-2020's.

At this time, natural gas was widely considered part of the transition to a renewable energy future in Ireland and indeed, the large majority of power generation in Ireland was reliant on gas. Pure Data Centres Group decided to create a dedicated onsite microgrid to generate its own power. Using gas fuelled generators removed transmission losses and allowed the development of a data centre campus whilst a grid connection remained unavailable. A high-pressure gas connection

was secured from Gas Networks Ireland and planning secured for a three-building data centre campus. In late 2025, Pure Data Centres Group will be trialling biomethane as a primary energy source replacement for natural gas. Looking to neutralise the carbon footprint of natural gas, Pure Data Centres Group has engaged Irish and European based biomethane producers, acting as a catalyst for the formation of a new domestic Irish supply chain.

The site's single line diagram is shown in Figure 3. The development comprises three separate data centre buildings, each designed for high availability, collectively delivering a computing capacity of 54 MW. Continuous power supply is delivered through three generator buildings forming a 68 MW energy centre. The energy centre has been designed to operate in line with Ireland's national climate and sustainability objectives, demonstrating Pure Data Centres Group's commitment to supporting these goals.

The energy centre uses Wärtsilä dual fuel continuously rated 20-cylinder reciprocating engines operating in an N+2 arrangement. These machines are connected into a 20 kV concurrently maintainable ring arrangement which in turn provide 2N utility supplies to each data centre building.

While the gas transmission grid in Ireland has never failed, like any data centre with an electrical grid connection, service level agreements require that backup power is seamlessly maintained through liquid fuel stored on site. As part of its commitment to sustainability, Pure Data Centres Group selected hydrotreated vegetable oil (HVO) in lieu of diesel due to its significantly lower carbon footprint and lower airborne emissions.

The energy centre is controlled via dual redundant control systems and a redundant comms network ensuring all aspects of the microgrid can be maintained throughout its life whilst continuously delivering high quality power to the data centres. 20 MWh of battery energy storage with grid forming capabilities is also provided to maximise system availability, support stable power characteristics and enable efficient use of the engine load steps.

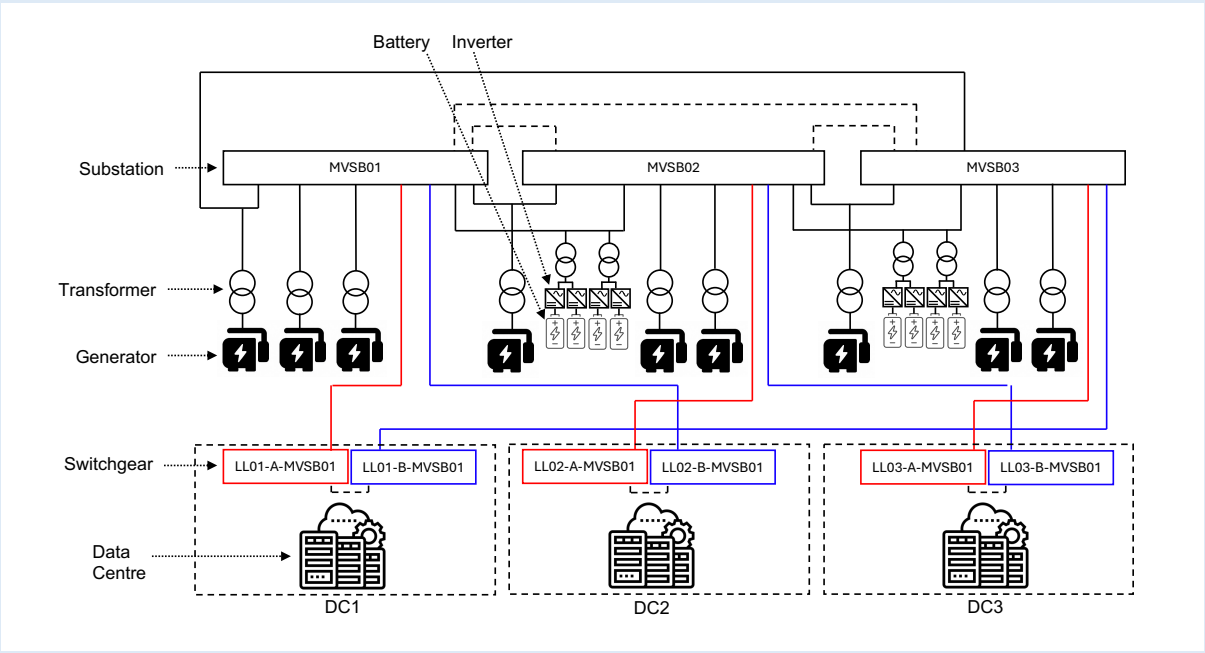

Figure 3: Single line diagram for Pure Data Centre's Dublin 68 MW data centre microgrid.

Box 3: Great Britain Case Study on Power Grid Connection Times and Flexibility

Figure 4 shows how flexibility could help accelerate data centre grid connections in Great Britain (GB). The GB onshore transmission network is managed by three network operators: National Grid Electricity Transmission (NGET), Scottish Hydro Electric Transmission Ltd. (SHE), and SP Energy Networks (SPEN). NESO's Electricity Ten Year Statement gives projections for future transmission network capacity in 2025, 2027, 2029, 2032, and 2035. As shown in Figure 4(a), NGET operates in the region with the highest data centre concentration. To provide an order-of-magnitude estimate, we use cumulative new transmission-level transformer capacity as a proxy for the demand that could be added to the transmission network. However, it should be noted that the hosting capacity for new demand will also depend on linked investments being made in transmission line reinforcements and new power plants, and it will also be impacted by localised grid power flow constraints and stability-related constraints. Figure 4(b) and Figure 4(c) show that for the whole of GB, and NGET's network, flexibility measures that reduce data centre peak demand by 30% allow the new data centre demand that could otherwise be hosted in 2035 to be brought forward to 2029, a time reduction of 60% (six years).

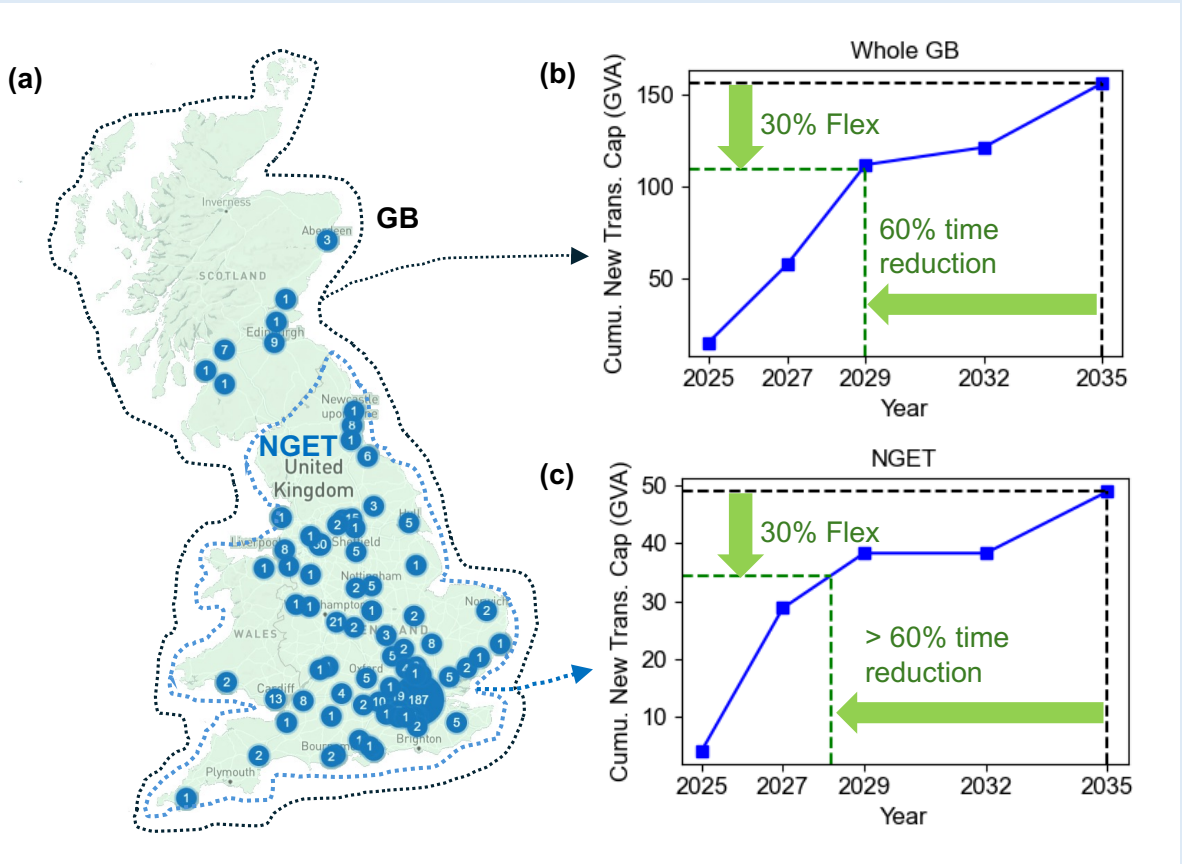

Figure 4: (a) GB's data centre density map (www.datacentermap.com) overlayed by the boundary of the GB transmission grid managed by NGET (the other parts of the GB network are managed by SHE and SPEN). The blue dots show the number of nearby data centres. (b) Cumulative new transmission-

level transformer power capacity for the whole of GB to 2035. (c) Cumulative new transmission-level transformer power capacity for NGET's network to 2035.

4. Multiscale Grid Intelligence

Data centres create close coupling between information and communications technology (ICT) and power infrastructure. Managing these infrastructures involves complex coordination across a wide range of spatial scales and time scales.

The ICT infrastructure of a hyperscale data centre is spatially divided between thousands of racks, each hosting 30 to 80 servers with computer processing units (CPUs), memory, storage and AI accelerators (e.g. GPUs, tensor processing units (TPUs)). For computing, servers are grouped into physical or virtual nodes which undertake jobs. To deliver AI services to billions of users, networks of data centres are operated together on a regional and/or global scale.

Data centres need extensive power distribution infrastructure to supply servers and cooling systems. A centralised uninterruptible power supply (UPS), or distributed rack-level/server-level UPSs, provide power quality filtering and continuous operation during faults with seamless swap over to backup generation. A data centre may operate within a larger microgrid, with primary generation and/or storage assets, and would normally be connected to the main transmission grid through a high voltage substation.

Data centre operations require the coordination of ICT infrastructure across a wide range of time scales. For computing jobs where latency is critical, such as AI inference, distribution to computing nodes is done on the timescale of milliseconds. Dynamic autoscaling of resources between computing applications is done every few seconds. For longer computing jobs requiring many resources, such as AI training, schedulers check resource availability and assign queued jobs every few minutes.

Power management also involves a wide range of timescales. Temperature needs to be monitored second-by-second to guard against overheating/cooling in the case of equipment failures. Cooling demand also varies due to diurnal and seasonal temperature changes and weather patterns. When importing power from the grid, power flow limits and voltage limits need to be managed at the transmission substation and throughout the data centre's internal network. Small short-term violations may be acceptable, but larger and more sustained violations may degrade equipment and cause catastrophic failures. While disconnected from the main grid, the data centre's power demand needs to be balanced second-to-second to maintain stable frequency and voltage. UPS batteries take over supply milliseconds after a grid fault, giving backup generators time to ramp up (10 to 30 seconds for diesel generators).

Planning occurs over longer timescales, since a new data centre takes months or years for site selection and acquisition, design, permitting, equipment procurement, construction and commissioning. The scale of AI data centres makes them a significant load from a power grid planning perspective. Major transmission and generation projects can take 10+ years, but modular technologies can be deployed much

more quickly, such as battery storage, solar generation, and Flexible AC Transmission Systems (FACTS) (power electronics-based grid control devices).

We propose multiscale grid intelligence as a new framework for AI data centre planning and operation to address the challenges driving inefficient grid defection. For operation, our idea extends previous work on data centre demand response and carbon-aware computing, where data centre power use is reduced during periods of grid stress or adjusted to match renewable generation. Multiscale grid intelligence goes beyond this by integrating a wider awareness of power grid operation and constraints across spatial scales and time scales into data centre coordination. Multiscale grid intelligence also recognises the importance of integrated planning between data centres and power grids. Effective planning needs to consider how design decisions flow-on to impact system operation. Box 4 presents an industry perspective on progress towards data centre grid intelligence from Pure Data Centres Group.

Across scales, we identify a range of opportunities for grid-intelligent operation and planning. Important examples of these, and the value streams they can create, are listed in Table 2. Figure 5 illustrates how these opportunities and value streams map to spatial scales, and the subsequent sections provide more detailed discussions.

Table 2: Opportunities for multiscale grid-intelligent AI data centre operation and planning, and potential value streams that can be created. These are broadly mapped to the (DC) data centre, (MG) microgrid or (TG) transmission grid spatial scales.

	Opportunities for Grid-Intelligent Operation	**Scale**
O1	CPU/GPU dynamic power capping	DC
O2	Job scheduling	DC
O3	UPS battery control	DC
O4	Inverter reactive power control	DC
O5	Backup generator control	DC
O6	Cooling system control	DC
O7	Microgrid generation, storage and flexible demand control	MG
O8	Connected waste heat network control	MG
O9	Inter-data centre job routing	TG
O10	Data centre aware grid dispatch	TG
	Opportunities for Grid-Intelligent Planning	**Scale**
P1	ICT hardware selection and sizing	DC
P2	Backup generation selection and sizing	DC
P3	Backup generation export path	DC
P4	UPS selection and sizing	DC
P5	Cooling system design	DC
P6	Thermal storage	DC
P7	Data centre monitoring and control systems	DC
P8	Data centre AC vs. DC distribution	DC
P9	Primary generation and storage selection and sizing	MG
P10	Microgrid loads	MG
P11	Grid connection sizing	MG
P12	Microgrid architecture design	MG

P13	Waste heat-based heat network	MG
P14	Microgrid monitoring and control systems	MG
P15	Grid-aware data centre siting	TG
P16	Strategic data centre aware power grid planning	TG
P17	Data centre focused electricity market design	TG
	Value Streams from Multiscale Grid Intelligence	**Scale**
V1	Water use reduction	DC
V2	Energy efficiency improvement	DC
V3	Energy cost reduction	DC
V4	Quality of service improvement	DC
V5	Carbon emissions reduction	DC
V6	Reliability improvement	DC
V7	Anchor load for renewables and electrification	MG
V8	Waste heat provision	MG
V9	Grid connection speed and cost improvement	MG
V10	Land use reduction	MG
V11	Grid flexibility service provision	TG
V12	Grid upgrade deferral	TG
V13	AI hosting capacity increase	TG

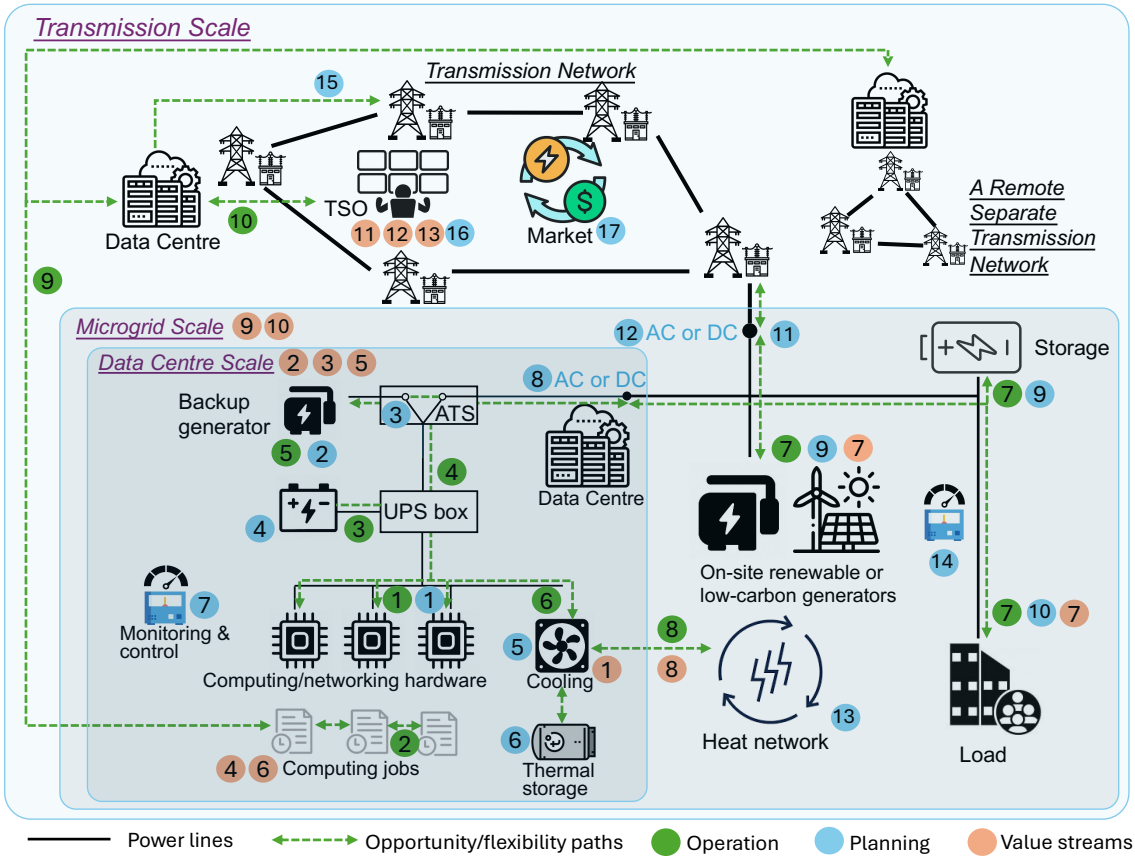

Figure 5: Mapping of the opportunities for grid-intelligent operation and planning and potential value streams from Table 2 across spatial scales.

Box 4: Industry Perspective on Progress Towards Data Centre Grid Intelligence from Pure Data Centres Group

In the UK, Labour's recent Modern Industrial Strategy described two key measures to address AI data centre power challenges:

1. The Connections Accelerator Service to fast-track grid access for "strategic demand" like AI campuses.
2. The British Industrial Competitiveness Scheme to cut standing network charges by up to 90 per cent for energy intensive sectors, with a consultation underway on extending the relief to AI data centres and advanced manufacturing.

Those are welcome initiatives, but data centres are still being treated mainly as consumers, not contributors. Data centres can play a much bigger role as part of the energy solution. By coinvesting in network reinforcements, participating in demand flexibility services and deploying on-site battery energy storage systems, data centres can offer genuine grid support. At Pure Data Centres Group we are introducing battery systems which could dramatically reduce evening peak draw. The same batteries can absorb surplus renewables during periods of curtailment and then discharge to steady frequency and voltage, delivering the fast-acting response services TSOs now prize. We are also implementing dynamic load scheduling and virtual-utility controls that switch between grid, storage and low-emission generation in milliseconds.

Beyond this, alternative commercial models, such as time-of-use tariffs, long term green Power Purchase Agreements (PPAs) paired with on-site storage, and non-firm connection agreements, can give operators financial incentives to offer flexibility and avoid bypassing network charges.

Analysis from the National Infrastructure Commission shows that maximising demand-side flexibility could trim whole system costs by around 15%, sparing consumers billions in future network charges. Energy UK's Powering the Cloud report adds that smarter siting and faster connections for data centres could unlock £44 billion in extra Gross Value Added (GVA) between 2025 and 2035 and keep the UK's digital infrastructure growing instead of migrating to lower cost grids abroad.

The 68 MW data centre microgrid described in the Box 2 case study is being used to test new grid intelligent solutions. To support a future grid connection application, the project has been designed to meet the dispatchable generation criteria specified by Ireland's CRU2021-124 direction. The data centre buildings and energy centre also have the capability of connecting to a district heating scheme on the perimeter of the site. Pure Data Centres Group has engaged with a heat offtake partner targeting 60 GWh heating demand in the area with potential to achieve a reduction of 12,000 tCO_2 if all heating needs are met by the district system.

4.1. Data Centre Scale Grid Intelligence

At the data centre scale, grid intelligence focuses on unlocking and coordinating demand flexibility. This can enable the provision of grid flexibility services, reduce peak demand allowing for less expensive grid connection, increase renewable utilisation and improve reliability by helping with supply–demand balancing during islanded operation.

One way to unlock demand flexibility is computing job scheduling, where jobs are delayed or completed with serial rather than parallel resources. Job scheduling software is already a core part of data centre operations to optimise the use of ICT infrastructure, and there is now a growing literature on the use of job scheduling for grid flexibility provision. For AI inference, where short jobs are continually arriving and need to be processed with minimal delays across thousands of computing nodes, the computational complexity of optimal scheduling is a significant challenge. For larger AI training jobs, overheads are the more pressing challenge. When pausing training, checkpoints need to be saved to avoid losing progress, and then data must be reloaded to restart training. This can take more than a minute for large language models (LLMs). These issues make it difficult for job scheduling to deliver fast flexibility services which require bursts of power at short notice, such as primary frequency response.

An alternate method for obtaining computing flexibility is CPU and GPU dynamic power capping, which is a standard feature of modern devices. Voltages and clock frequencies are controlled to maintain power consumption below an adjustable power limit at the cost of slower computation. Power caps can generally be adjusted in less than a second, enabling fast timescale flexibility to be obtained. Device manufacturers usually specify a minimum power cap level around 50% for both CPUs and GPUs, above which stable operation and accurate computation will be maintained. A complexity is that the impact of a given power cap on job completion time will depend on the specific job and the hardware being used. For example, research on the EcoFreq software tool for carbon intensity-based CPU and GPU power scaling found that, for various high-performance workloads, capping power at 70% of the nominal rating was associated with 3% to 15% slower computation for an NVidia A40 GPU and 6% to 20% slower computation for an Intel CascadeLake CPU.

This motivates the need for granular coordination strategies which can dynamically adjust CPU/GPU power caps for individual servers to deliver data centre level demand flexibility with minimal overall quality of service loss.

Data centre job scheduling and CPU/GPU dynamic power capping both result in slower computation. The value of the flexibility obtained needs to be high given the cost of AI computing hardware and the value placed on computing speed. This means computing flexibility is most likely to be viable for high value grid flexibility services, which includes fast services, like primary frequency regulation, and demand reduction during periods of extreme grid stress. Our previous work assessing usage patterns for AI data centres and the value of computing speed, inferred from cloud computing platform pricing, found that AI data centres could provide flexibility at around 50% lower cost than traditional CPU-heavy high performance computing data centres, and that their flexibility could be cost competitive in the German, Australian and UK electricity markets.

Grid flexibility can also be provided by UPS battery storage. UPS batteries need to keep sufficient energy ready for unexpected disconnections, but there is an opportunity to oversize UPS batteries and make use of the inverter power capacity (which is needed regardless) to enable grid import/export and flexibility service provision. Even without making use of UPS batteries, a UPS inverter's reactive power capability can be used to provide voltage regulation and power factor correction.

In standard data centre network architectures, an automatic transfer switch (ATS) is used to swap over to backup generators when a grid fault is detected. However, if built with a suitable export path, backup generators can also be used to help power the data centre and to deliver grid flexibility services during normal operation. This increases backup generator utilisation and therefore requires careful reliability assessment. However, a certain amount of regulator backup generator use can help detect idling problems, such as fuel degradation and moisture buildup.

Traditionally, lead-acid battery UPSs and diesel backup generators have been preferred due to the low cost of lead-acid batteries, the relative ease of storing diesel fuel and the fast startup times of diesel generators. However, given the rapidly falling costs of technologies including lithium-ion batteries, redox flow batteries, and photovoltaic generation, there is an opportunity to fundamentally re-think the design of data centre backup systems.

A related design option is to have DC rather than AC power distribution within the data centre. DC distribution involves more complex protection systems and less established supply chains. However, it has the potential for lower cost, higher efficiency and higher reliability. This is because DC distribution has lower losses than AC for the same cabling, and battery storage and ICT equipment can be connected to the DC bus without the need for DC/AC power conversion.

Another potential area for operational flexibility is the data centre cooling system. With data centres becoming more power dense, there is a trend away from air cooling towards chip-level liquid cooling and dielectric fluid immersion. Using predictive modelling of the cooling system thermal mass, upcoming workloads and ambient temperature, flexibility can be obtained without violating equipment temperature limits through a combination of preemptive and delayed cooling. Flexibility can be significantly enhanced with the addition of thermal storage, which can use various technologies including underground heat-transfer fluid pipes, pebble beds, or phase change materials. Both intra-day and inter-seasonal heat storage is relevant since data centre cooling requirements will be higher during the day than at night and higher in summer than in winter.

Data centre cooling system design and operation have focused on Power Usage Effectiveness (PUE), which is an efficiency metric defined as the ratio of total energy consumption to ICT energy consumption. However, the growth of zero marginal cost and zero marginal emissions renewable generation means that a less efficient system, but with greater flexibility, may be more beneficial from both an economic and environmental perspective. Another important consideration for AI data centre cooling systems is water use. Evaporative cooling can reduce power demand by up to 90%, but uses large amounts of water, which can be a problem for environmental impact and social acceptability. Closed-loop cooling systems provide a near zero-water use alternative, but at the cost of higher power consumption. The indirect water use of generation assets also needs to be considered.

4.2. Microgrid Scale Grid Intelligence

Grid intelligence at the microgrid scale focuses on how grid awareness can enhance the design and operation of a partially or fully autonomous microgrid built around a data centre. Technology selection and sizing of primary generation and storage assets is a key consideration. Primary generation and storage

can get the data centre running at partial load while waiting for a grid connection and can supplement a lower cost partially rated grid connection. Microgrid generation and storage assets can provide grid flexibility services and can enhance data centre reliability if configured to support islanded operation during grid faults. The technology mix and sizing of local generation and storage assets can be chosen to optimise trade-offs between energy cost, reliability, environmental impact and land use, while accounting for computing workloads, grid energy prices, emission intensity and upstream transmission constraints.

Building a data centre microgrid also creates the opportunity to connect other local loads, such as residential communities, electric vehicle charging stations or industrial sites. The data centre can serve as an anchor load supporting the business case for a high reliability microgrid, which can then support complementary business cases around renewable generation, energy storage, and electrification. Local loads are beneficial when they provide temporal diversity and flexibility, enabling greater utilisation of local generation and transmission infrastructure and assisting with supply-demand balancing.

A data centre's cooling system will generate significant amounts of waste heat. There is the potential for this to become a source for a heat network distributing hot water or steam. Combined heat and power generation units are another potential source of waste heat. Data centre waste heat can also be used to generate local electricity through a Rankine cycle system.

Another set of design decisions relate to the microgrid architecture, which includes topology selection and a choice between AC and DC distribution. A radial topology will be lower cost, but a ring or meshed topology will provide higher reliability. The choice between AC and DC distribution at the microgrid scale has similar trade-offs to the choice between AC and DC data centre distribution, namely that AC has more established supply chains and simpler protection systems, while DC distribution can offer greater efficiency, higher reliability, and fewer power conversion stages, particularly for battery storage, ICT loads and renewable sources.

4.3. Transmission Scale Grid Intelligence

At the transmission scale, grid intelligence aims to create co-optimisation opportunities for data centres and power grids across operation and planning. Large AI companies may operate multiple data centres spread across a transmission grid and may have other data centres in further geographic regions with separate transmission grids. This creates an opportunity to route computing jobs between data centres based on local grid conditions, as long as quality of service requirements are not violated. For heavily constrained transmission grids, moving computing jobs, and the associated power demand, between sites will lead to different generators being dispatched, affecting energy cost, emissions intensity and grid reliability. Jobs can also be transferred between data centres in regions with separate transmission grids, which will have different generation mixes, demand patterns (particularly if in different time-zones), reliability levels, and congestion patterns.

In countries with liberalised electricity markets, TSOs are responsible for procuring flexibility services and dispatching power grid assets to balance supply and demand at lowest cost, while ensuring network constraints are not violated. Traditionally, TSO dispatch has focused on generation plants, but the transition to renewable generation has made demand-side flexibility increasingly important. The scale of

AI data centres, and the flexibility that can be unlocked from their ICT, cooling, UPS storage, backup generation and microgrid assets, motivates their integration into TSO dispatch systems. This is aligned with broader trends towards TSO dispatch automation and greater use of demand side flexibility. Granular monitoring, communications and control infrastructure at the data centre and microgrid scales are important for enabling this.

Beyond incorporating data centres into existing electricity market mechanisms, there are also opportunities for further value to be unlocked through electricity market reforms and new power grid planning processes which directly address the needs and capabilities of AI data centres. Power grid planning and operation are closely tied to government policy objectives due to the importance of affordable and reliable electricity for human health and prosperity. Decarbonisation targets have led to policy objectives related to transitioning to renewable generation and simultaneously expanding power grid capacity to electrify heating and transportation. Further reforms are now motivated with AI emerging as an additional area of policy priority due to its potentially critical role for future national security and industrial competitiveness.

Electricity market reforms could include new grid flexibility services co-designed by TSOs and data centre operators to have delivery requirements (e.g. ramp-rates, durations, non-delivery penalties) that AI data centres can meet without excessive quality of service loss, offering the potential for greater flexibility provision. Another opportunity is to leverage the islanding capabilities of data centres provided by their UPSs and backup generation. TSOs could design capacity remuneration mechanisms for data centres that allow themselves to be curtailed or disconnected during periods of severe grid stress.

The escalation in AI data centre power requirements makes it increasingly impractical to conduct siting and sizing separately from power grid transmission and generation planning. A key challenge is the conflicting timelines of major transmission grid upgrades (5 to 15 years) versus data centre development (6 to 36 months). However, modular technologies, such as FACTS devices and battery storage, can be deployed on a similar timescale to data centres to increase grid capacity. In addition, if data centres can demonstrate reliable flexibility, there may be the potential to bring them online within the network capacity headroom provided by grid upgrades that are already in the pipeline.

In regions where grid connections face long delays and uncertain approvals, a growing problem is “ghost projects”, where data centre developers submit speculative grid connection requests, many of which will not actually go forward. This makes it difficult for TSOs to accurately forecast data centre demand and prioritise transmission grid upgrades. To address this, new mechanisms are needed for data sharing and collaborative planning between data centre developers and TSOs. Of particular value would be standardised tools for predictive modelling which TSOs can trust when assessing grid impacts and upgrade requirements, thereby streamlining planning studies and permitting. The goal of fast-tracking AI data centre grid connections also motivates new organisational approaches and business models for power grid planning. With suitable regulations and governance arrangements in place, grid studies and network upgrade projects could be accelerated through co-funding and closer collaboration between AI data centre developers and TSOs.

5. Conclusions and Future Research

Multiscale grid intelligence aims to provide a holistic framework for coordination between AI data centres and the power grid, creating opportunities to accelerate AI deployment, reduce energy costs and increase sustainability. Recent advances in AI capabilities have motivated large investments by industry, and governments are focused on AI as a potentially critical technology for future economic competitiveness and national security. However, without a strategic approach to grid integration, there is a major risk that AI data centres will either not get built or will be built with costly and emissions intensive autonomous microgrids.

Multiscale grid intelligence addresses this challenge by providing a structured approach for linking the planning and operation of ICT and power infrastructure across scales, from individual data centre servers, up to national transmission grids and networks of data centres delivering AI applications. We have presented a range of examples where multiscale grid intelligence can offer value, but these are illustrative rather than exhaustive, and new opportunities will emerge given how quickly AI and power system technology is advancing.

Future research is needed in three key areas to realise the full potential of multiscale grid intelligence for AI data centres:

1. Data and Modelling: There has been a global effort to make power grid data openly available to support research and innovation, including by the IEEE Power & Energy Society. Datasets related to AI data centre computing workloads are starting to become available, but there remains a lack of open datasets with linked power consumption data. Data centre power demand and flexibility modelling are also challenging, due to the complex interactions between ICT equipment, computing workloads, weather patterns and cooling systems. New modelling approaches are needed that can balance model fidelity against computational complexity based on the specific spatial scales and time scales relevant for specific control and planning tasks.

2. Scalable and Robust Data Centre Computing Flexibility: Research on AI data centre workflows and computing patterns has shown that computing flexibility could be valuable, particularly for fast grid flexibility services. However, accessing this flexibility at scale is challenging due to the complexity of AI computing hardware and workflows. Also, the high value of computing means that data centre operators have been cautious about approaches that could reduce quality of service and reliability. Previous work on data centre computing flexibility has focused on job scheduling, but GPU power capping offers a promising alternative for AI data centres. This approach is suited to the delivery of fast grid flexibility services since power caps can be adjusted in less than a second without incurring the overheads associated with pausing and restarting jobs. Controlling thousands of computing nodes so that in aggregate they deliver a specific grid flexibility service is analogous to the operation of virtual power plants which aggregate grid-edge devices (e.g. electric vehicles, heat pumps), where there is already a substantial

literature on distributed optimisation for scalability, chance constrained optimisation for robustness and reinforcement learning for adaptivity.

3. Socioeconomics and Social License: Public acceptance and approval are of critical importance for the development of AI data centres. Without social license to operate, developers may face community opposition, delays in planning approval, and heightened regulatory scrutiny. Building public trust and support is complicated by the potential for AI data centres to increase electricity bills, carbon emissions, and water use, as well as broader uncertainty around how AI will affect employment and the economy. Multiscale grid intelligence could offer solutions, for example by repositioning data centres as low-cost suppliers of waste heat for district heating systems or anchor loads for local renewable generation projects. There is a need for interdisciplinary research spanning computer science, power engineering, social science and economics to understand the actual and perceived socioeconomic impacts of AI data centres, and how these can be shaped collaboratively by communities, developers, TSOs and policymakers.

For Further Reading

- A. Radovanović, R. Koningstein, I. Schneider, B. Chen, A. Duarte, B. Roy, D. Xiao, M. Haridasan, P. Hung, N. Care, S. Talukdar, E. Mullen, K. Smith, M. Cottman, W. Cirne, “Carbon-Aware Computing for Datacenters”, *IEEE Transactions on Power Systems*, vol. 38, no. 2, 2023.
- A. Tsiligkaridis, P. Andrianesis, A. K. Coskun, M. C. Caramanis, I. C. Paschalidis, “Distributed Economic Dispatch in Power Networks Incorporating Data Center Flexibility,” *IEEE Transactions on Sustainable Computing*, vo. 10, no. 4, 2025.
- C. Crozier, M. Liska, “The Potential of Data Center Energy Demand to Provide Grid Flexibility” *Current Sustainable/Renewable Energy Reports*, 2025.
- O. M. Kozlov, A. Stamatakis. "EcoFreq: Compute with Cheaper, Cleaner Energy via Carbon-Aware Power Scaling”, *ISC High Performance*, 2024.
- W. E. Gnibga, A Blavette, A. Orgerie, “Renewable Energy in Data Centers: The Dilemma of Electrical Grid Dependency and Autonomy Costs” *IEEE Transactions on Sustainable Computing,* vol. 9, no. 3, 2024.
- Y. Zhou, Á. Paredes, C. Essayeh, T. Morstyn, “AI-focused HPC Data Centers Can Provide More Power Grid Flexibility and at Lower Cost”, *arXiv*, 2025.
- Y. Zhou, Á. Paredes, C. Essayeh, T. Morstyn, “Evaluating and Comparing the Potentials in Primary Response for GPU and CPU Data Centers”, *IEEE Power & Energy Society General Meeting*, 2024.

Biographies

Thomas Morstyn and ***Yihong Zhou*** are with the Department of Engineering Science, University of Oxford, Oxford, OX1 3PJ, United Kingdom.

Ian Whitfield is with Pure Data Centres Group, 5 Fleet Place, London, EC4M 7RD, United Kingdom.