INNOVATION POLICY LAB | munk school OF GLOBAL AFFAIRS & PUBLIC POLICY

UNIVERSITY OF TORONTO

# REDEFINING ENERGY AND COMPUTE IN THE AI AGE

Leah Lawrence

INNOVATION POLICY LAB WORKING PAPER SERIES 2025-04

**Engineers for Climate - Getting new technologies to scale**

Engineers for Climate (EfC) aims to accelerate the adoption and diffusion of technologies that will decarbonize our data and energy systems at scale, bridging the gap between emerging technologies and real-world implementation. By fostering interdisciplinary collaboration and practical solutions, EfC seeks to redefine the relationship between AI infrastructure and energy systems, contributing to the achievement of our 2050 climate goals while addressing critical national security concerns.

This white paper is the summary of the first workshop of the EfC's Compute and Energy Working Group, held on October 29 and 30, 2024.

# Executive Summary

In our rapidly evolving digital landscape, data centres play a pivotal role, yet their burgeoning energy consumption presents a significant challenge as we seek sustainable solutions. This White Paper explores the intricate balance between advancing artificial intelligence (AI) and mitigating its environmental footprint. We delve into the impacts of AI on energy demand and the role of innovative technologies in transitioning to carbon-neutral digital infrastructures.

Key discussions, led by experts in the field, examine the limitations posed by our current technological parameters and propose strategic frameworks to reconcile the growing computational needs with the finite nature of our energy resources. Through workshops and expert panels, this document outlines potential solutions, such as improving energy efficiency, harnessing waste heat recovery, and moving computing operations to more efficient systems, including edge computing.

As the AI industry stands at a crossroads between vertical integration and an ecosystem-based approach, our analysis points to the importance of resource allocation, talent acquisition, and collaborative innovation in shaping a sustainable future. This White Paper serves as both a call to action and a guide for stakeholders looking to foster an energy-efficient and environmentally responsible AI ecosystem.

# 1. Introduction

In an era where data has become central to our economies and our lives, data centres and their demand for energy have been pushed into the spotlight. As artificial intelligence (AI) rapidly evolves from novelty to necessity, these technological powerhouses face an unprecedented challenge: an insatiable appetite for energy that threatens to outpace our ability to sustainably meet demand.[1]

This exponential growth raises critical questions about the future of computing and its environmental impact:

- How can we reconcile the seemingly limitless potential of AI with the very real limits of our energy resources?
- As we approach the theoretical boundaries of Moore's Law and Dennard scaling, what innovative solutions can bridge the gap between computational demands and energy efficiency?

[1] Bashir, Noman, Priya Donti, James Cuff, Sydney Sroka, Marija Ilic, Vivienne Sze, Christina Delimitrou, and Elsa Olivetti. 2024. "The Climate and Sustainability Implications of Generative AI." An MIT Exploration of Generative AI, March. https://doi.org/10.21428/e4baedd9.9070dfe7.

- Most importantly, what strategies can transition our digital infrastructure from energy-intensive to carbon-neutral?

## Workshop Overview and Objectives

A diverse group of compute and energy experts convened for a workshop to tackle this complex challenge at Harvard University in October 2024. Participants offered deep and complementary experience and knowledge in their relevant domains. Each participant was (and is) enthusiastic about tackling the problem of energy+data centres in new and more holistic ways.

This white paper is structured as follows:

- Section 2 examines the current energy landscape, including data centre energy demands, future consumption projections, and grid challenges.
- Section 3 discusses renewable energy in relation to data centres, covering corporate procurement strategies and variability issues.
- Section 4 explores the evolution of data centres, including trends in siting, behind-the-fence solutions, and cooling innovations.
- Section 5 delves into AI hardware advancements, focusing on processing units, memory, networking, and cooling systems.
- Section 6 analyzes the impact of AI workloads, their diversity, and implications for grid stability.
- Section 7 provides an overview of the AI stack and infrastructure, from cloud and chip layers to mobile applications.
- Section 8 highlights opportunities for AI in energy optimization, such as site selection and demand response integration.
- Section 9 outlines key challenges and considerations in the integration of AI and energy systems.
- Finally, Section 10 presents conclusions and future directions for addressing these complex challenges, emphasizing integrated planning and cross-disciplinary collaboration.

Note that while workshop participants recognize the importance of issues like cybersecurity and data privacy, these topics were beyond the scope of this workshop.

Appendix A includes a copy of the workshop agenda and list of participants.

# 2. The Energy Landscape

A decade ago in most jurisdictions, data centres seemed relatively small, even inconsequential in the context of the electric grid.[2] However, the shift from traditional on-premise computing to cloud computing, alongside the replacement of Central Processing Units (CPUs) with Graphics Processing Units (GPUs), has led to increased rack energy densities.[3] As a result, data centre energy requirements began to grow exponentially. The breakout year for Generative AI in 2023 and its subsequent adoption spike in 2024 compounded this trend,[4] [5] making power availability a primary concern.[6]

## Projections for Future Consumption

The Federal Energy Regulatory Commission (FERC) estimates that data centre power consumption increased by nearly 2,000 megawatts from 2023 to 2024 and is projected to reach 35,000 megawatts by 2030,[7] an increase of over 100 percent.[8]

The electricity generation capacity required to meet this demand will be significantly larger for several reasons. First, data centres operate 24/7, necessitating a constant power supply. Second, large technology companies are often committed to sourcing electricity for their data centres from renewable and/or clean energy sources. However, solar and wind energy don't provide a one-to-one match with demand because of their capacity factors, typically between 25 and 45 percent. This means that using solar or wind alone to fully power a data centre requires installing 2 to 4 times more capacity

---

[2] With the exception of leading data centre jurisdictions like Northern Virginia and Dallas, for example.

[3] Rack density, a key metric in data centre operations, refers to the amount of power consumed by equipment within a single server rack. It is typically measured in kilowatts per rack. Traditional non-AI workloads in modern data centres typically consume between 8 and 10 kilowatts per rack. However, AI-dedicated racks, which rely heavily on GPUs, are significantly more energy-intensive. Current AI workloads often require between 40 and 50 kilowatts per rack.

The energy demands for AI computing are rapidly escalating. Nvidia, a leading manufacturer of GPUs, has announced that its next generation of AI-focused processors will push rack energy densities to unprecedented levels. These new chips are expected to require up to 100 kilowatts per rack, with some projections suggesting densities could reach as high as 300 kilowatts per rack in the near future. This dramatic increase in power density presents significant challenges for data centre cooling, power distribution, and overall infrastructure design.

[4] Quantum Black AI by McKinsey, The state of AI in 2023: Generative AI's breakout year," August 1, 2023.

[5] Quantum Black AI by McKinsey, "The state of AI in early 2024: Gen AI adoption spikes and starts to generate value," May 30, 2024.

[6] Cushman & Wakefield Annual Data Centre Market Survey, 2024.

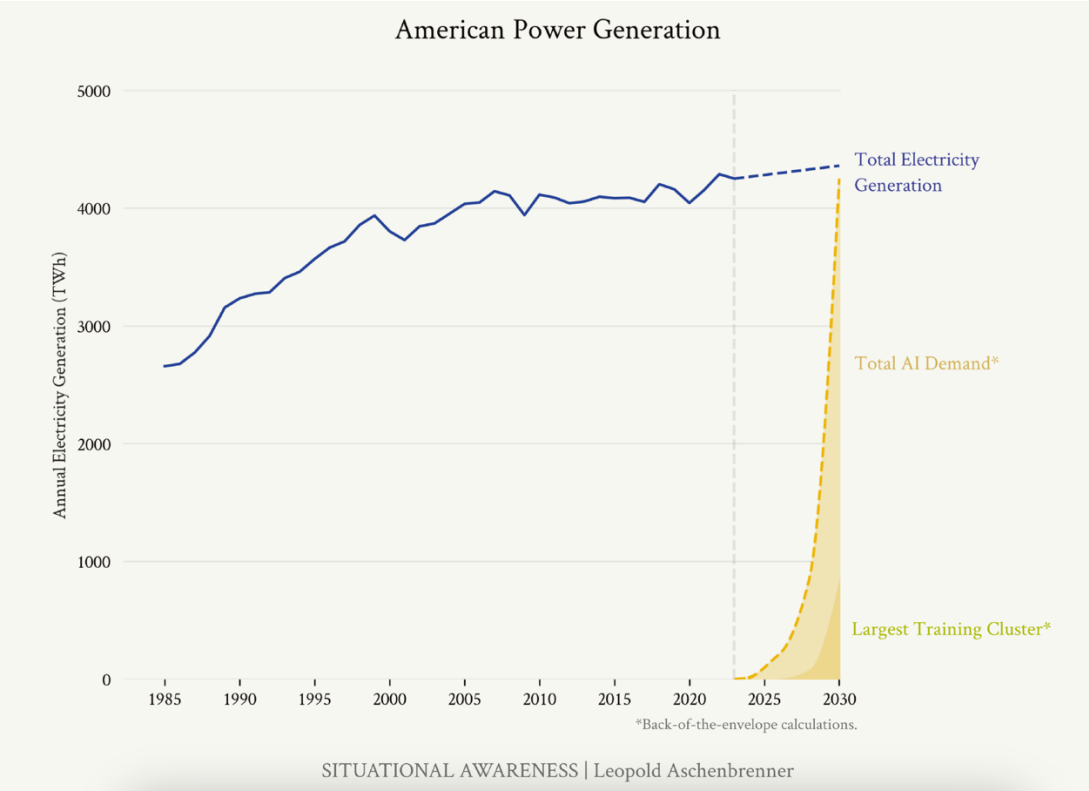
[7] Federal Energy Regulatory Commission, 2024 Summer Energy Market and Electric Reliability Assessment, May 23, 2024.

[8] Data centre consumption in 2022 was 17,000 megawatts.

than the actual consumption. For this reason, renewable energy is often “backed-up” by utilities and grid operators by natural gas power generation (see Section 3).

Notably, AI-driven consumption of electricity could be much higher, even approaching levels on par with all annual electricity generation in the U.S. in a very short time as envisioned in Leopold Aschenbrenner's "Situational Awareness" (see Chart 1). [9] Practically speaking, however, such timelines are likely to be constrained by the availability and procurement of equipment for both generation and transmission. In addition, the permitting process for new generation facilities and for electric and natural gas transmission infrastructure is typically lengthy, and construction schedules could further complicate deployment.

**Chart 1 Aschenbrenner’s “Back of the Envelope” Estimates of AI Electricity Demands**




Source: Aschenbrenner, Leopold, Situational Awareness: The Decade Ahead, June 2024

[9] Aschenbrenner, Leopold, *Situational Awareness: The Decade Ahead,* June 2024.

## Grid Challenges and Interconnection Issues

And so it is that utility companies and grid operators have been caught in the rush to meet requests for connectivity, outstripping available capacity for both electric transmission and natural gas supply and transmission to meet the demand for new natural gas generators. Applications to connect new energy generation facilities and data centre loads have spiked.[10] [11] In Alberta, Canada, for example the Alberta Electric System Operator (AESO) has over 6,000 megawatt of data centre applications.[12] To put this into context, the average total load demand in Alberta is approximately 10,000 megawatts, so if all the requests in the AESO queue would go ahead, the generation required to meet demand would increase overall average grid flows by 50 percent.

The backlog of projects in the U.S. continues to grow - nearly 2,600,000 megawatts of generation and storage capacity was actively seeking grid interconnection at the end of 2023.[13] The current backlog is more than twice the total installed capacity of the existing U.S. power plant fleet.

A big challenge for utilities and grid operators is that, if history is any guide, the vast majority of these proposed projects will never be built. Research undertaken by Berkeley Lab showed that only 19 percent of the projects (and just 14 percent of capacity) that were submitted from 2000 to 2018 had been built and were operating commercially by the end of 2023.[14] This leaves utilities and grid operators with the uncertainty of which projects will proceed and when, with significant risks to the system if the assumptions are incorrect. The result is delays in the decision-making process until more certainty is achieved.

In the big picture, one might argue that data centre electricity demand isn't that significant, especially after discounting projects that won't materialize. This view may seem reasonable at the macro level, but it's crucial to understand the rarity of a 1,000-megawatt load. Until recently, such requests for stable baseload demand came only once a year or every few years across an entire country. Now, as the data shows, they are appearing almost weekly and with the added complication of the significant load variability that accompanies AI workloads (see Section 6).

---

[10] Blum, Sam, “Warnings about an AI Buble are Growing. When Could it Burst?” Inc., July 10, 2024.
[11] Chow, Andrew R. and Billy Perrogo, “The AI Arms Race is Changing Everything”, *Time*, February 17, 2023.
[12] For more detailed information, see the AESP project list, updated monthly. See https://aeso-portal.powerappsportals.com/connection-project-dashboard.
[13] Berkeley Lab, “Grid connection backlog grows by 30% in 2023, dominated by requests for solar, wind, and energy storage”, April 10, 2024, https://emp.lbl.gov/news/grid-connection-backlog-grows-30-2023-dominated-requests-solar-wind-and-energy-storage.
[14] Ibid.

One of the major challenges is the insufficient development of new generation capacity. Many jurisdictions are struggling with supply adequacy concerns, where even a single large data centre can significantly disrupt the supply-demand balance. This disruption can increase both the risk to consumer reliability and the cost of energy, as prices tend to rise in the short term until new generation capacity becomes available. From a public policy standpoint, this poses affordability issues, as higher energy prices can lead to increased local inflation. If such high prices persist over the medium term, they may drive businesses away, ultimately affecting local employment. To address these challenges, some governments have implemented "bring your own generation" policies. These policies require data centre developers to meet their electricity demands through power purchase agreements for new generation or by developing their own energy sources.

Requests for new load interconnections are not due to data centres alone. Onshoring and the expansion of manufacturing, as well as transportation electrification, are also playing a role. For example, the TSMC semiconductor facility in Arizona aspires to be 1,000 megawatt-sized, with a first phase of 200 megawatts.[15] This dramatic shift from occasional to frequent 1,000 megawatt-scale demands signals a major change in our energy landscape.

Finally, and as illustrated by the TSMC example, there is the question of phasing. One thousand megawatt-sized applications are unlikely to come on in one shot. The Power Purchase Agreement (PPA) between Talen Energy and Amazon in Pennsylvania (later rejected by Federal Energy Regulatory Commission) envisioned a data centre requiring 960-megawatts of capacity, however, the first phase is only for 120 megawatts. There also was an opt-out provision after 480 megawatts.[16]

As interconnection timelines extend, data centre developers are exploring options "behind-the-fence," such as collocating with nuclear power facilities. This strategic relocation raises questions about whether these shifts are temporary or indicative of long-term siting decisions.

[15] Jensen, Audrey, “Utility company makes progress on infrastructure for Taiwan Semiconductor project in north Phoenix,” Phoenix Business Journal, April 5, 2022.

[16] U.S. Energy Information Administration, “Data centre owners turn to nuclear as a potential electricity source”, Today in Energy, in-brief analysis, October 1, 2024.

# 3. Renewable Energy and Data Centres

Large technology companies have adopted energy procurement strategies that commit to ensuring their energy usage is backed by renewable energy, primarily wind and solar. According to S&P Global research, Amazon, Google, Meta, Microsoft, and Apple are responsible for 60 percent of all corporate-backed renewable energy globally, with the vast majority of this installed capacity built and operating in the United States (40,000 of 45,000 megawatts).[17]

## Variability and Grid Stability Challenges

As technology companies (and other companies) have looked to cover the energy usage of data centres with PPAs for solar energy, renewable generation has grown significantly year-over-year. The scale of solar and wind facilities has also increased, such that it is not unusual for utility-scale installations to have a capacity between 400 and 1,000 megawatts. The integration of renewable generation - primarily wind and solar energy - is inherently variable, however, with a fluctuation in energy of around 30 percent of generation or 120 to 300 megawatts for larger, more recent projects.[18]

To accommodate variability in renewable energy and semi-dispatchable generation, grid operators employ two main strategies: ancillary services and dispatchable backup power. Ancillary services are rapid-response mechanisms that maintain grid stability in real-time, including automatic generation control,[19] fast-ramping resource response, and inverter-based voltage regulation.[20] These services operate on timescales of seconds to minutes, and in traditional grids were supported by inertia from large synchronous generators. [21]

[17] Wilson, Adam, “Datacentre companies continue renewable buying spree, surpassing 40 GW in US”, S&P Global, March 28, 2023.
[18] Guang Chao Wang et al., “Maximum Expected Ramp Rates Using Cloud Speed Measurements,” Journal of Renewable Sustainable Energy 12, 056302 (2020).
[19] Automatic generation control is a real-time control system that automatically adjusts the power output of multiple generators at different power plants, in response to changes in the system frequency.
[20] Inverter controls refer to the electronic systems that manage how renewable and storage resources (such as solar photovoltaics and batteries) convert direct current (DC) into alternating current (AC) for the grid. Modern “advanced inverters” can also provide grid support functions such as voltage regulation, frequency response, and reactive power compensation—capabilities traditionally supplied by large synchronous generators.
[21] Grid inertia refers to the ability of a power grid to maintain stability and resist changes in frequency. It's derived from the rotating mass of large synchronous generators, such as those in coal, nuclear, and hydroelectric power plants. These generators naturally resist changes in their rotational speed due to their mass, thereby helping to stabilize the grid's frequency when there's a sudden imbalance between electricity supply and demand.

By contrast, dispatchable backup power refers to generation sources that can be turned on or off relatively quickly to meet longer-term changes in electricity demand, typically operating for hours or days. The same resources often supply ancillary services when scheduled to do so.

Traditionally, fossil fuel plants, particularly natural gas, have provided both ancillary services and dispatchable backup power. These plants offer fast-ramping capabilities and contribute inertia, which helps maintain grid frequency. However, relying on fossil fuels to support renewable energy generation creates a tension between the need for reliable power and the desire to minimize carbon emissions.

Clean, semi-dispatchable resources are increasingly being used to provide both ancillary services and backup power. Large-scale hydroelectric facilities, where available, can offer fast-ramping response, though their capacity depends on reservoir levels and can vary seasonally. Nuclear power, while typically used for baseload generation, can in some cases be designed to operate flexibly, as seen in Ontario, Canada, providing a clean alternative for grid support. [22]

AI-focused data centres are anticipated to drive an increased demand for ancillary services, raising the unresolved question of who will bear the costs for these additional services in many jurisdictions. Accommodating the heightened loads from data centres, alongside a growing reliance on renewable energy, will require significant investment in electric grid infrastructure. A key policy debate revolves around how these costs should be allocated to ensure that tariffs for data centres remain just and reasonable, while also safeguarding the interests of existing customers.

As the grid evolves, new technologies such as battery storage, demand response, and resources with advanced inverter controls are emerging as alternatives for providing ancillary services and even some forms of dispatchable power, helping to balance the variability of renewable sources while reducing reliance on fossil fuels. Because of their unique characteristics, data centres are also increasingly being utilized as providers of ancillary services. For example, a data centre in Calgary, Alberta has been participating in a local Demand Response Ancillary Service program since 2014. During dispatch events, which occur only a handful of days each year, the data centre can disconnect from the grid and run on its on-site backup generators.[23]

---

[22] International Atomic Energy Agency, “Non-baseload Operation in Nuclear Power Plants: Load Following and Frequency Control Modes of Flexible Operation,” IEA Nuclear Energy Series, No. NP-T-2.23 (2018).
[23] Enel X, “How Data Centres Support the Grid with Ancillary Services?”, https://www.enelx.com/tw/en/resources/how-data-centres-support-grids.

## Grid Supply Overbuild Effect

The integration of variable renewable energy and semi-dispatchable resources into the grid has led to a "grid supply overbuild effect." This effect refers to the need for excess generation capacity to be built beyond peak demand requirements to accommodate for transmission line losses, maintenance, variability of generation and/or load and to ensure grid stability. While the pre-renewable era required a 20 to 25 percent overbuild, current levels typically range from 25 to 50 percent varying by jurisdiction (see Table 1). This increased overbuild addresses the variability of renewable sources, ensures peak demand coverage, accounts for transmission losses, and maintains grid stability.

**Table 1 Grid Supply Overbuild Effect in Four Example Jurisdictions**

| State / Province | Installed Capacity (GW) | Peak Demand (GW) | Renewable Resource (GW) | Renewable / Clean Resources Installed (%) | Overbuild (% of Installed) |
|---|---|---|---|---|---|
| Alberta* | 21 | 12.5 | 7 | 33% | 40% |
| Ontario | 40 | 30 | 36** | 90% | 25% |
| Texas | 155 | 78 | 69.5 | 45% | 50% |
| California | 83 | 46 | 33 | 40% | 45% |

*Note that each of these jurisdictions have very different market structures. For example, in Alberta, the market is fully deregulated and independent power producers make generation planning decisions based upon market conditions and electricity market frameworks. There is no centralized planning, and the grid operator does not mandate an installed capacity or portfolio generation.

**This number includes significant large-scale hydro and nuclear energy resources. As noted above, these semi-dispatchable resources have a different, non-variable generation profile than other renewable and clean resources.

As renewable generation increases and grid management becomes more complex,[24] grid overbuild is expected to continue, supported by natural gas power generation. Alternative solutions such as direct air capture of carbon dioxide and long-duration energy storage are emerging. However, these technologies still require significant

[24] MISO's Renewable Integration Impact Assessment (RIIA), Summary Report - February 2021.

commercialization efforts to bring costs down to levels comparable with natural gas power generation coupled with carbon capture and storage.[25]

# 4. Data Centre Evolution

While traditional data centre hubs like Northern Virginia, Dallas, Chicago, Phoenix, and Northern California remain popular, developers are experimenting with new siting strategies, including co-location with industrial facilities and cooler-climate locations. These new strategies include:

- **Co-location with industrial facilities**: Data centres are being built adjacent to nuclear plants[26] and natural gas power generation facilities. [27]
- **Cool climate locations**: Developers are choosing sites in cooler regions to leverage natural cooling benefits.[28]

This shift in location strategy aims to ensure reliable power supply and reduce energy costs. However, it comes with a trade-off in terms of increased latency.[29] The long-term viability of this trend remains uncertain.

Ninety percent of electricity used by a data centre is for cooling, server operations, and network equipment (see Table 2). As data centre developers look further afield for access to power, they are also looking to increase energy efficiency and reduce cooling requirements.

[25] Brick, Jamie et al., "The role of natural gas in the move to cleaner, more reliable power", McKinsey & Company, September 1, 2023, https://www.mckinsey.com/industries/electric-power-and-natural-gas/our-insights/the-role-of-natural-gas-in-the-move-to-cleaner-more-reliable-power.

[26] Garver, Rob, "Tech firms increasingly look to nuclear power for data centre," Voice of America News, October 15, 2024.

[27] Sandy Segrist. Behind the Hype: The ' Jaw-dropping' Expectations for AI, Natural Gas. *Hart Energy,* September 5, 2024. https://www.hartenergy.com/exclusives/behind-hype-jaw-dropping-expectations-ai-natural-gas-210333.

[28] Municipal District of Greenview, "Media Release: World's Largest AI Data Centre Industrial Park 'Wonder Valley' coming to the Greenview Industrial Gateway", https://mdgreenview.ab.ca/media-release-ai-data-centre-gig.

[29] Latency is the delay between sending and receiving information over a network

**Table 2 Energy Use Comparison of Traditional vs. Hyperscale Data Centres***

| Category | Traditional Data Centre | Hyperscale Data Centre |
|---|---|---|
| Cooling | 40% | 25% |
| Server Operations | 30% | 50% |
| Networking Equipment | 20% | 15% |
| Storage Drives | 5% | 5% |
| Other | 5% | 5% |

*Note that these numbers are approximations and depend on the method used to embed Network Interface Controllers (NICs). The ambiguity arises from the fact that NICs can be viewed as both part of the server hardware and as networking equipment.

## Industrial Co-location Benefits

Locating data centres near large industrial sites offers several advantages:

- Fiber connectivity
- Existing industrial permits, including water permits
- Access to pre-existing grid interconnections or bypassing the electric grid entirely
- Access to natural gas
- Potential integration of waste heat from industrial facilities for data centre cooling needs

## Cooler Climate Location Benefits

Siting data centres in cooler regions provides significant advantages:

- Natural heat dissipation, reducing energy consumption for cooling
- Alignment with free cooling principles, minimizing chiller usage
- Potential for waste heat recovery systems (using absorption chillers), offering up to 10% additional energy use reduction[30]

In a case study comparing siting a data centre in Alberta versus Texas, analysis indicated that this benefit of siting at a cool-temperature location could be as much as

[30] Estimate by a workshop participant.

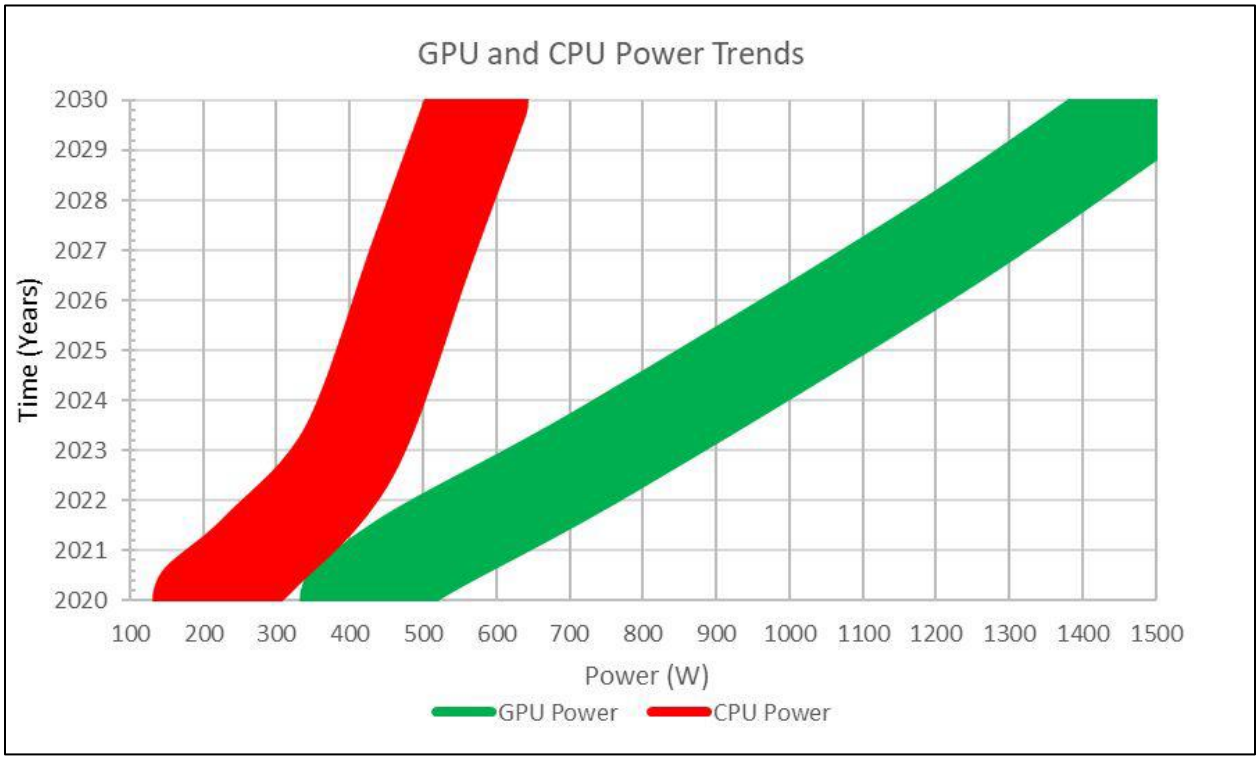
20 percent. Energy savings are assumed to be transferred from cooling to computing tasks, thus also theoretically reducing a data centre user's cost of compute.

## Cooling Innovations

Meanwhile, the power required by CPUs and GPUs to handle AI workloads continues to grow (see Chart 2). The more power used; the more heat generated. This necessitates the use of innovative liquid and immersion cooling technologies, as traditional direct air cooling will no longer suffice at the IT component level.[31] Technologies under development include:

- **Advanced Liquid Cooling:** This technology offers superior heat dissipation compared to traditional air cooling, reducing energy consumption associated with thermal management.
- **Immersion Cooling:** This approach submerges servers and IT equipment in a non-conductive dielectric liquid to efficiently dissipate heat, reducing energy consumption and increasing computing density.

**Chart 2 GPU and CPU Power Trends**




Source: Open Compute Project

[31] Chen, Cheng et al., OCP OAI System Liquid Cooling Guidelines, Open Compute Project, October 1, 2024.

# 5. AI Hardware Advancements

In the race for AI supremacy, the need for speed has become paramount, driving an urgent demand for hardware innovation. As AI systems grow increasingly complex and data-intensive, traditional computing architectures face significant challenges. These include limitations in processing capacity, bottlenecks caused by copper interconnects, and thermal constraints that force systems to operate below their full potential. The laws of physics themselves present barriers to advancement, pushing researchers and engineers to explore novel solutions. These challenges have become the primary catalysts for innovation in AI hardware.

To overcome these obstacles, the industry is witnessing a shift towards specialized AI hardware. Advanced GPUs, Tensor Processing Units (TPUs), and custom-designed chips are being developed to handle the unique computational demands of AI workloads. Moreover, innovations in hardware design are not limited to processing units alone. Advancements in interconnect technologies, cooling systems, and energy-efficient architectures are equally crucial in pushing the boundaries of AI capabilities.

## Processing Units

AI servers increasingly integrate advanced processing units to handle complex AI workload demands. These include:

- CPUs and GPUs: Traditional CPUs are often paired with GPUs, which are optimized for parallel processing tasks common in AI applications.
- AI Accelerators: Specialized hardware such as TPUs and Field-Programmable Gate Arrays (FPGAs) are being deployed to enhance computational efficiency and performance per watt.

These advancements aim to address the significant energy demands of AI workloads by improving processing efficiency and reducing power consumption.

## Memory and Storage

The memory and storage components of AI servers are critical for managing large datasets and ensuring rapid data access:

- High Bandwidth Memory (HBMs): HBMs are like stacked memory chips that can transfer data super-fast, making computers more powerful and energy-efficient than those using regular memory. Integrated inside GPUs, this is the fastest memory with lowest level of latency in existence.
- High-Capacity Random Access Memory (RAM): Essential for holding models and datasets in memory, enabling faster processing.

- Fast Storage Solutions: Solid State Drives (SSDs) and Non-Volatile Memory Express (NVMe) drives provide the speed necessary for efficient data retrieval and storage operations.

Innovations in this area focus on reducing latency and improving throughput, contributing to overall energy efficiency.

## Networking Components

AI workloads require robust networking solutions to facilitate rapid data transfer:

- GPU Networking using NvLink: This technology is crucial for scaling compute power beyond a single chip, with copper connections supporting up to 72 GPUs and optical interconnects extending this capability to thousands of GPUs, essentially allowing them to act as a single, massively powerful chip.
- High-Speed Interfaces: Technologies like InfiniBand and high-bandwidth Ethernet support the fast data transfer rates needed for AI applications.
- Switches: High-performance network switches are crucial for maintaining efficient communication within server clusters.

These components help minimize bottlenecks in data flow, thereby optimizing energy use across the network infrastructure.

## Silicon Photonics

Silicon photonics is an emerging technology that leverages light for data transmission within and between chips, offering several advantages over traditional electronic interconnects, such as copper:

- Higher Bandwidth: Silicon photonics can potentially achieve data transmission rates of terabits per second, significantly enhancing the speed of data transfer in AI systems.
- Lower Latency: By using optical signals instead of electrical ones, silicon photonics reduces signal delay, which is crucial for high-performance AI applications.
- Reduced Power Consumption: Optical interconnects consume less power than electrical counterparts, especially over longer distances, making them a promising solution for improving energy efficiency in data centres.

Copper interconnects, widely used for linking components like GPUs, face limitations in reach and performance as data capacity increases. At high capacities, copper requires "retimers" or electronic repeaters to maintain signal integrity, which introduces significant power consumption and latency due to signal regeneration. Silicon photonics

overcomes these limitations as a miniaturized optical technology built in the same material as semiconductor integrated circuits which allows it to be co-packaged with other IT components. This technology offers an extended reach of hundreds of meters, enabling larger GPU domains compared to copper interconnects, while also significantly reducing power consumption and cost over those distances.

The integration of silicon photonics into AI hardware could lead to substantial improvements in energy efficiency and performance, particularly in environments where data movement is a significant energy consumer.

# 6. AI Workload Diversity and Its Impact

AI workloads can be broadly categorized into two main types: training and inference. Both types consume significant amounts of energy, but their patterns of consumption differ. Training involves creating the model's knowledge base and capabilities, while inference is the application of a trained model's knowledge to specific tasks.

Recent studies reveal substantial heterogeneity in resource demands and performance profiles across various AI applications. *The resource profile of AI workloads varies not only between training and inference, but also across specific applications.* This diversity extends across multiple dimensions, including:

- Compute Intensity
- Memory Capacity
- Memory Bandwidth
- Network Latency Sensitivity
- Network Bandwidth

These diverse computational and networking requirements ultimately translate into distinct patterns of electricity use. To understand the full system impact, it is necessary to examine how AI workloads drive fluctuations in power demand at the grid level.

## Power Demand Fluctuations

As Large Language Models (LLMs) and other AI workloads become more prevalent within data centres, new patterns of electricity consumption and the potential for significant transient behavior are starting to emerge. Recent engineering studies have highlighted the unique challenges posed by these workloads, particularly in terms of their rapid power demand fluctuations and heterogeneous resource requirements.

A study[32] on the impact of AI workloads on electric grid stability focused on the rate of change of power consumption in large-scale GPU clusters. Example calculations for a data centre with 10,000 GPUs and 50,000 GPUs, assuming a worst-case scenario where GPUs transition from idle (10% thermal design power) to full load (100% thermal design power) within a one-second timeframe, showed that the calculated rates of change, particularly for the 50,000 GPU scenario, exceed typical minute-level ramping capabilities of distribution grid designs (see Appendix B).

## Implications for Grid Stability

The rapid fluctuations in power demand from AI workloads pose significant challenges for grid stability. Traditional grid designs and technical interconnection requirements are typically limited to ramping capabilities of 10 percent per minute. The potential for sudden, large-scale changes in power consumption from AI-focused data centres could strain existing grid infrastructure and require new approaches to power management and distribution.

Grid operators are responsible for ensuring the reliability, stability, and resilience of the electrical system. If interconnections have the potential to negatively impact this imperative, grid operators require the interconnection to self-manage to meet technical requirements. Self-management could include managing variability with battery storage or off-grid self-generation to minimize impact on grid stability. However, for a 1,000-megawatt data centre, suitable battery storage options do not currently exist, forcing data centres to adhere to technical grid requirements.

Data centres currently have the capability to manage their demand and data use, including the ability to support peak load shaving (reducing load during peak periods). However, the current offering of ancillary services and grid interconnection requirements does not provide incentives for data centres to support grid operators in fulfilling their mandate.

These challenges underscore the need for more sophisticated grid modeling, predictive infrastructure planning, and potentially new strategies for demand response and load balancing to accommodate the unique characteristics of AI workloads.

[32] Li, Yuzhuo et al., “The Unseen AI Disruptions for Power Grids: LLM-Induced Transients”, September 9, 2024, https://arxiv.org/html/2409.11416v1. Presented to IEEE Subcommittee on Big Data and Analytics for Power Systems.

# 7. The AI Stack and Infrastructure

The AI-stack, which includes the cloud and chip layer, database layer, LLMs, middleware layer, and mobile and web applications, provides a holistic view of energy consumption across the entire AI infrastructure. Each layer contributes differently to the overall energy usage.

## Cloud and Chip Layer

At the foundation of the AI stack, cloud infrastructure and specialized AI chips form the computational backbone. These components, including GPUs, TPUs, and other AI-specific hardware, provide the massive parallel processing power essential for AI tasks. The energy demands here are substantial, highlighting a key area for efficiency improvements. Performance per watt has become a critical metric in this layer, driving the development of more energy-efficient AI chips.

## Database and Storage Layer

Built atop the hardware, this layer manages the vast datasets crucial for AI. Efficient data storage and retrieval systems are vital for optimizing AI workloads, with energy considerations extending to data centre cooling and power management for large-scale storage.

Energy-efficient data management techniques are gaining traction. For example, data compression can reduce storage requirements and energy consumption by up to 80 percent, while intelligent caching mechanisms can minimize data movement, further reducing energy use.

## Large Language Models

LLMs represent the core AI engines, requiring immense computational resources for training and inference. The energy intensity at this layer, particularly during extended training periods, underscores the need for more efficient training algorithms and hardware utilization.

## Middleware Layer

Acting as a critical bridge, the middleware layer manages Application Programming Interfaces (APIs), load balancing, and data preprocessing. While less computationally intensive, its role in orchestrating data flow and resource utilization is crucial for overall system efficiency.

Efficient resource allocation through advanced scheduling algorithms can lead to energy savings of up to 30 percent in large-scale AI deployments. Additionally, workload distribution optimization can reduce idle time and improve overall energy efficiency.

## Mobile and Web Applications

The top layer comprises user-facing applications that leverage the underlying AI infrastructure. Though less energy-intensive, optimizing data transmission and local processing on devices remains important for system-wide efficiency.

The balance between on-device processing and cloud offloading is crucial for energy efficiency. Edge AI, which processes data closer to the source, can reduce energy consumption by up to 50% compared to cloud-only solutions for certain applications.

## Interdependencies

The layers across the AI stack are interdependent and, as AI workloads grow, each layer faces increased demands, highlighting the need for scalable solutions across and between components. For this reason, there is an increasing emphasis (or reemphasis) on hardware-software codesign which can lead to energy efficiency improvements.

# 8. Opportunities for AI in Energy Optimization

## Optimized Site Selection

AI algorithms can be used to analyze complex datasets to identify ideal data centre locations, considering factors such as:

- Power availability and grid stability
- Proximity to renewable energy sources and other critical infrastructure
- Climate conditions for efficient cooling
- Availability of water for cooling
- Network connectivity requirements
- Proximity to users, latency
- Local regulations and incentives

This data-driven approach can lead to more energy-efficient and sustainable data centre placements.

## Demand Response Integration

Data centres have the potential to play a role in grid demand response by intelligently managing their energy use. This can involve:

- Shifting non-urgent computing tasks to off-peak hours
- Adjusting workloads based on real-time grid conditions
- Participating in automated demand response programs

It is worth noting that while such solutions have been available for several years, there has been limited implementation of these approaches to date as there is a lack of incentives for ancillary services.

## Integrated Grid Modeling

AI can facilitate the development of sophisticated digital twins for regional and national power grids, enabling:

- Real-time monitoring and optimization of power distribution
- Improved load balancing and demand response strategies
- Enhanced integration of renewable energy sources
- Predictive maintenance of grid infrastructure
- Rationalization of actual versus designed or planned system models

These models can lead to more efficient and resilient power systems.

## Predictive Infrastructure Planning

Leveraging AI's predictive capabilities, data centre operators can:

- Forecast future power demands based on AI workload trends
- Optimize power distribution for specific AI tasks
- Model renewable energy availability and required grid responses
- Reduce grid overbuild by accurately predicting energy needs

AI-driven simulations can also help identify and mitigate potential negative impacts of data centre operations on energy and water resources at local, regional, and national levels.

## Cooling System Optimization

AI can optimize data centre cooling systems by:

- Predicting heat generation based on workload patterns
- Adjusting cooling in real-time to match actual needs
- Identifying opportunities for waste heat recovery and reuse

These optimizations can significantly reduce energy consumption associated with cooling, which accounts for a substantial portion of data centre energy use and cost.

## Workload Scheduling and Resource Allocation

AI can improve the efficiency of data centre operations through:

- Intelligent workload scheduling to maximize energy efficiency
- Dynamic resource allocation based on real-time energy availability
- Optimization of server utilization to reduce idle energy consumption

# 9. Challenges and Considerations

## Skills Gap in Holistic Integration

The integration of AI in data centres and electric grids faces a significant skills gap, particularly in merging expertise from power engineering and computing to develop innovative solutions. This gap is evident in the need for professionals who can navigate both domains effectively, fostering interdisciplinary collaboration and systems thinking to address complex challenges. Bridging this skills gap is crucial for creating holistic approaches that enhance energy efficiency and sustainability in both sectors.

## Data Quality and Availability

Both data centres and electric grids increasingly rely on vast amounts of data for efficient operation. Data centres need high-quality datasets for AI training and inference, while electric grids require accurate, real-time data for load balancing and demand response. Ensuring data quality, managing large volumes of information, and extracting meaningful insights are shared challenges.[33]

## Switching Costs

Both data centres and the electric grid face significant switching costs when adopting new technologies. For data centres, transitioning to AI-optimized infrastructure involves substantial financial investments in new hardware and software, as well as potential disruptions during the implementation phase. Similarly, electric utilities face high costs when integrating renewable energy sources or upgrading grid infrastructure, which can include both financial expenditures and operational challenges.

## Balancing Efficiency and Resilience

Data centres strive to maximize computational efficiency while maintaining high availability. Similarly, electric grids aim to optimize energy distribution while ensuring grid stability. Both sectors must balance the drive for efficiency with the need for robust, reliable systems that can handle peak demands and unexpected disruptions.[34]

---

[33] Rueda, A.R., Henri van Soest and Hye Min Park, “The Promise and Peril of AI in the Power Grid”, *The National Interest*, January 25, 2024.

[34] Accomondo, J, et al., “The Intersection of Energy and Artificial Intelligence: Key issues and future challenges”, Morgan Lewis, August 12, 2024.

## Infrastructure Integration

Both data centres and the electric grid face challenges in integrating new technologies with legacy systems. For data centres, this involves incorporating AI-specific hardware like GPUs and novel cooling systems into existing IT infrastructure. Similarly, electric grids struggle to integrate renewable energy sources, smart meters, and advanced control systems with older grid components.

## Incentive Structures

The incentive structures in both technology companies and the energy sector often do not align with a holistic approach to energy efficiency. In technology companies, incentives may prioritize performance improvements and rapid deployment over long-term energy efficiency gains. Similarly, in the energy sector, existing regulatory frameworks and market dynamics may not adequately incentivize the adoption of innovative AI technologies or sustainable practices. Adjusting these incentive structures could encourage more sustainable practices and investments in both sectors.

## Institutional Barriers

Both data centres and the electric grid face significant institutional barriers that can hinder the adoption of new technologies. These barriers include:

- Regulatory Hurdles: Existing regulations may not be well-suited to accommodate innovative technologies, creating delays and additional compliance costs for both sectors.
- Organizational Resistance: There can be resistance to change within organizations, where established practices and legacy systems are deeply entrenched. This resistance can slow the adoption of new, more efficient technologies.
- Lack of Cross-Sector Collaboration: Effective integration of AI and energy technologies often requires collaboration across different sectors and industries. However, siloed operations and communication gaps can impede this necessary cooperation.

# 10. Conclusions and Future Directions

The convergence of artificial intelligence, data centres, and energy systems presents both unprecedented challenges and opportunities for innovation. As AI continues to evolve and expand, its energy demands are reshaping the landscape of computing and power generation. This white paper has explored the multifaceted issues surrounding this transformation and highlighted several key areas for future development.

The exponential growth in data centre energy consumption, driven by the rapid adoption of AI technologies, necessitates a paradigm shift in how we approach energy and compute. Traditional models of grid management and data centre operations are being pushed to their limits, requiring novel solutions that span multiple disciplines and industries.

Several critical themes have emerged from this analysis:

1. **Multi-Path Innovation and Integration**: The future of energy efficient AI computing will likely unfold along two parallel, but tension filled tracks. One emphasizes coordinated integration of energy systems and data centres through site selection, demand response, and infrastructure planning to achieve systemic efficiency. The other relies on distributed, often disruptive innovation driven by individual actors across the value chain. These tracks are not automatically complementary; they may compete for resources, attention, and regulatory support. As Clayton Christensen's work[35] reminds us, constraints in one domain often catalyze breakthroughs in another. For example, permitting delays for renewable energy may spur faster adoption of more efficient AI hardware. Policymakers and investors should therefore adopt a portfolio mindset, encouraging both systemic coordination and decentralized innovation as parallel, mutually reinforcing strategies.
2. **Technological Advancements**: Continued innovation in AI hardware, cooling systems, and renewable energy technologies will be central to reducing the environmental impact of AI workloads. Emerging technologies such as silicon photonics and advanced liquid cooling show particular promise in improving energy efficiency at scale.
3. **Grid Stability and Flexibility**: The unique power consumption patterns of AI workloads pose significant challenges to grid stability. Addressing these issues will require sophisticated modeling, predictive infrastructure planning, and new approaches to load balancing that can accommodate sharp and irregular demand spikes.
4. **Cross-Disciplinary Collaboration**: Bridging the skills gap between power engineering and computing is essential for developing comprehensive solutions. Interdisciplinary collaboration and systems thinking should be fostered across

[35] Clayton M. Christensen, *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail* (Boston: Harvard Business School Press, 1997).

academia, industry, and government to align technical, regulatory, and operational innovations.

5. **Adaptive Infrastructure**: Future data centres and energy systems must be designed with flexibility and scalability in mind, capable of adapting to rapidly changing AI workloads and evolving energy availability. Infrastructure that can expand or contract responsively will be better positioned to meet both current demands and long-term sustainability goals.

While these channels represent promising directions, they differ substantially in feasibility, impact, and time horizon. Table 3 summarizes the main approaches discussed in this paper, highlighting their relative feasibility, potential impact, and time horizon.

**Table 3 Summary of Approaches to Reduce Energy Use and Emissions**

| Channel | Feasibility (near-term) | Potential Impact | Time Horizon | Notes / Examples |
|---|---|---|---|---|
| **AI hardware innovation** | High | High | Near–mid | Specialized accelerators, HBM, silicon photonics; strong commercial momentum. |
| **Cooling innovation** | High | Medium | Near | Liquid and immersion cooling; proven, reduces thermal overhead, enables density. |
| **Siting & integration** | Medium–High | Medium–High | Near–mid | Co-location with generation, cooler climates, waste-heat use; depends on permitting and power access. |
| **Data centres as grid services** | Medium | Medium | Near | Demand response, ancillary services (e.g., Calgary example); policy incentives needed. |
| **Battery storage & advanced inverters** | Medium | Medium | Near–mid | Mature technology: good for fast-response services but not scaled to full-load backup. |
| **Flexible clean backup** | Medium (situational) | High | Mid–long | Hydro and flexible nuclear (Ontario model); geography and regulation critical. |
| **Renewable procurement/expansion** | Low–Medium | High | Mid–long | PPAs, wind/solar buildout; bottlenecked by interconnection and permitting. |

| | | | | |
|---|---|---|---|---|
| **Integrated grid modeling & planning** | Medium–High | Medium | Near–mid | Digital twins, predictive siting; improves efficiency, reduces overbuild. |
| **AI software & workload optimization** | High | Low–Medium | Near | Scheduling, compression, orchestration; incremental gains, complements hardware/cooling. |
| **Long-duration storage & DAC** | Low (today) | Potentially High | Long | Cost and commercialization hurdles; needed for deep decarbonization. |

As we move forward, addressing these challenges will require a concerted effort from industry leaders, policymakers, and researchers. Systemic coordination can accelerate sustainable AI infrastructure, while distributed innovation ensures resilience and diversification of approaches.

The path ahead is not a choice between integration and disruption, but a recognition that both will unfold in parallel. Together, coordinated planning and disruptive experimentation can shape a more sustainable AI-energy ecosystem.

The road ahead is complex, but it also presents an opportunity to redefine our relationship with technology and energy. By prioritizing efficiency, fostering innovation across multiple channels, and promoting collaboration while embracing diversity of approaches, we have the opportunity to build a foundation for sustainable AI growth, one that secures environmental benefits while meeting society's expanding computational needs.

# Appendix A - Workshop Agenda and Participants

**Future of Energy + Data Centres**
**October 30, 8:30 a.m. to 2 p.m.**
**5th Floor, 14 Story Street**

**Purpose and Approach:**
As we think about the energy and climate transition, there is a need to convene cross-sector experts with deep expertise to talk about how we get from here to there. The focus of this first conversation is co-creating the next generation of data centres.

This invitation-only event brings together compute and energy experts (operators, entrepreneurs and researchers). Each brings deep and complementary experience and knowledge. Each is enthusiastic about the challenge and is thinking about it in different ways (i.e. hardware, software, energy, etc.)

The discussion is Chatham House Rule.

**Agenda:**

| | | |
|---|---|---|
| 08:00 - 08:30 | Gather | Light breakfast will be served |
| 08:30 - 08:45 | Welcome and Context Setting<br><br>Discussion Leader:<br>Leah Lawrence | Key question:<br>What are you curious about learning more about today? |

| | | |
|---|---|---|
| 08:45 - 09:45 | Weird World of Energy Generation for Data Centres<br><br>Discussion Leader:<br>Terri Steeves<br>Ian MacGregor | Key questions:<br>Is all compute brown compute (aka underpinned by some amount of fossil fuels)? How do we change it to green compute?<br><br>What are the impacts of large load additions like data centres to the electric systems, in particular for reliability?<br><br>What might be achieved through waste recovery or different approaches to cooling? |
| 09:45 - 10:15 | Break | |
| 10:15 - 11:15 | Computing Energy Efficiency and What Happens When it is Moved to the Edge<br><br>Discussion Leader:<br>Jim Waldo<br>Jack O'Brien | Key questions:<br>Where is computing energy efficiency today and where is it going? |
| 11:15 - 12:15 | Nuts and Bolts of Data Centre Operations and their Hardware<br><br>Discussion leaders:<br>Harold Moss<br>Hamid Arabzadeh | Key questions:<br>How do data centre operators think about energy?<br><br>Where is the "hardware" of compute going from an energy demand standpoint? |
| 12:15 – 13:00 | Lunch | |

| | | |
|---|---|---|
| 13:00 - 14:00 | AI, Innovation and the Grid<br><br>Discussion leaders:<br>Le Xie<br>Rob Davidson | Key questions:<br>What are the key concerns from the perspective of the electric gird?<br><br>How might new technologies and innovation address these concerns and enable the energy transition? |
| 14:00 - 14:30 | Discussion and Debrief for Next Steps | |

**Participants**

| | |
|---|---|
| Chris Biegler | https://www.linkedin.com/in/chris-biegler-816767a2/ |
| Claudia Lopez Hernández | https://www.advancedleadership.harvard.edu/2024-fellows-and-partners/claudia-lpez-hernndez |
| Dino Shiatis | https://www.advancedleadership.harvard.edu/2024-fellows-and-partners/constantinos-shiatis |
| Doug Sutcliffe | https://www.linkedin.com/in/dougsutcliffe/ |
| Glenn Dixon | https://www.advancedleadership.harvard.edu/2024-fellows-and-partners/glenn-dixon |
| Hamid Arabzadeh | https://www.linkedin.com/in/hamidarabzadeh/?originalSubdomain=ca |
| Harold Moss | https://www.linkedin.com/in/hmoss/ |
| Ian MacGregor | https://www.linkedin.com/in/ian-macgregor-354533129/?originalSubdomain=ca |
| Jack O'Brien | https://www.linkedin.com/in/thejackobrien/ |
| Jim Waldo | https://www.hks.harvard.edu/faculty/jim-waldo |

| | |
|---|---|
| Klara Jelinkova<br>(dinner only) | https://evp.harvard.edu/people/klara-jelinkova |
| Le Xie | https://xiele00.github.io/ |
| Leah Lawrence | https://www.advancedleadership.harvard.edu/2024-fellows-and-partners/leah-lawrence |
| Mitsuki Suda | https://www.linkedin.com/in/mitsuki-suda-40054b10a/ |
| Noman Bashir | https://impactclimate.mit.edu/people/noman-bashir/ |
| Rob Davidson | https://www.linkedin.com/in/rob-davidson-p-eng-6857127/?originalSubdomain=ca |
| Terri Steeves | https://www.linkedin.com/in/terri-steeves-994a9b2a9/?originalSubdomain=ca |
| Treyden Chiaravalloti | https://www.linkedin.com/in/treyden/ |

# Appendix B - Understanding Potential Transient Impacts of AI Workloads on the Grid

Using the methodology presented in the study Li, Yuzhuo et al., “The Unseen AI Disruptions for Power Grids: LLM-Induced Transients”, September 9, 2024, example calculations were made to understand the aggregate impact as seen by the grid for a data centre with 10,000 GPUs and 50,000 GPUs.

**Assumptions**

1. GPU Model: We'll assume the use of high-end GPUs similar to NVIDIA A100 or H100, with a Thermal Design Power (TDP) of 550W per GPU1.
2. Utilization Range: We'll consider a scenario where GPUs transition from idle (10% TDP) to full load (100% TDP).
3. Transition Time: Based on the paper's mention of rapid power changes, we'll assume a transition time of 1 second for the entire cluster.

**Calculations**

Case 1: 10,000 GPUs

- Total GPUs: 10,000
- Power per GPU: 550W
- Idle Power: 10% of 550W = 55W per GPU
- Full Load Power: 100% of 550W = 550W per GPU

Aggregate Power Swing:
(550W - 55W) * 10,000 = 4,950,000W = 4.95 MW
Rate of Change:
4.95 MW / 1 second = 4.95 MW/s

Case 2: 50,000 GPUs

- Total GPUs: 50,000
- Power per GPU: 550W
- Idle Power: 10% of 550W = 55W per GPU
- Full Load Power: 100% of 550W = 550W per GPU

Aggregate Power Swing:
(550W - 55W) * 50,000 = 24,750,000W = 24.75 MW
Rate of Change:
24.75 MW / 1 second = 24.75 MW/s

**Summary Table**

| Scenario | Number of GPUs | Aggregate Power Swing | Rate of Change |
|---|---|---|---|
| Case 1 | 10,000 | 4.95 MW | 4.95 MW/s |
| Case 2 | 50,000 | 24.75 MW | 24.75 MW/s |