National Science Foundation Compute-Energy Nexus Workshop

Executive Summary

The National Science Foundation (NSF) sponsored the Compute-Energy Nexus Workshop, held in Chicago, December 8-9, 2025. This workshop was an invitation-based, two-day event to convene crucial stakeholders, articulate key challenges, and initiate an action plan. Approximately 75 invited attendees represented a wide range of stakeholders across industry, government, academia, and non-profit agencies. The wide range of participants added high value to the discussion. Many attendees, including those with deep domain expertise, were hearing for the first time about critical issues and practices in other domains.

National strategic imperatives require energy independence, artificial intelligence (AI) leadership, and superiority in advanced computing and data security. This report provides several examples, discussing critical issues going forward, and describing participant recommendations for next steps. The workshop's next steps include organization of a working group to prepare a national strategic roadmap for *integrated* solutions to the challenges. The report presents proposed immediate actions to address key issues.

Some key takeaways from the workshop are that growth of data centers, artificial intelligence (AI) workloads, and high-performance supercomputing continues to be constrained by:

- Energy and electricity grid infrastructure;
- Networking performance, latency, and power consumption;
- Computing hardware and related design issues;
- Software limitations;
- Thermal management: essentially 100% of energy delivered into a data center must be rejected as heat;
- Policy stakeholders and the combined aspects they must address in this context;
- Local community concerns – especially energy cost, property values, noise, water usage, and construction pollution, balanced against tax revenue.

Recommendations to NSF:

1. Create fully integrated teams, including industry, federal, and academic stakeholders, that will pursue combined energy and computing growth from a system engineering perspective.
2. Form roadmap development working groups that will look at tactical actions to address key issues.
3. Consider funding new interdisciplinary research on technologies that would enable data centers operating at 1 MW power per individual rack and 5000 MW overall. This goal will need new approaches ranging from power electronics research, dynamic grid integration, computer architecture chiplets, energy-heat aware data center hardware-software co-design, and energy-heat aware AI/enterprise applications design, implementation, and deployment.
4. Consider funding fundamental collaborative research between grid-side experts and compute-side experts to examine new operational models and time predictions intended to be much more responsive to the bursty power demands of computing. This will need to include research in power conversion, compute-utility interaction models, distributed AI models, thermal management techniques, and others.
5. Request a stronger community engagement from compute-energy-nexus stakeholders. Concerns about public distrust, weaknesses in community engagement, and the need for a community-linked version of the Open Compute Project (OCP), plus limited engagement with policy and government stakeholders, are substantial roadblocks. Often the main actors causing distrust (e.g. hyperscalers and big-tech companies) rely on third parties and are not involved in direct discussions.
6. Establish seminar series and regular workshops and conferences to bring together the diverse stakeholders in this arena.

More details and presentations are posted at
http://csl.illinois.edu/events/NSF-energy-compute-workshop-chicago

National Science Foundation Compute-Energy Nexus Workshop

Summary Report, September 2026

Investigators: R. Kumar, K. Nahrstedt, P. Krein, *University of Illinois Urbana-Champaign*

1. Introduction

On December 8-9, 2025, approximately 75 invited stakeholders met at the NSF-sponsored Compute-Energy-Nexus Workshop in Chicago to discuss a wide range of challenges at the intersection of power grids and data centers. The participants came from industry, federal agencies, and academia to discuss critical issues of energy and heat demands of ever-growing large-scale data centers, creating major concerns not only to the power grid and computing industry, but also to communities that host data centers.

This report outlines the goals of the workshop (Section 2), the workshop agenda (Section 3), provides a description of the participants (Section 4), summarizes challenges that have been discussed (Section 5), and provides summaries of findings (Section 6), conclusions, and next steps (Section 7). Appendix A includes summary discussions of each session, and Appendix B provides detail of the workshop agenda.

2. Goals

The organizers sought to invite a wide range of stakeholders across the power grid and computing industries, federal agencies, and academia, and engage them in emerging compute-energy nexus discussions of challenges and opportunities. The PIs aimed for workshop participants to articulate current and new challenges and opportunities and nucleate a team that will prepare a national strategic roadmap for integrated solutions. The roadmap will need to be feasible within the time frames of the computing industry but must realistically reflect energy industry implementation constraints. It must identify the most fruitful intersection and integration points for cohesion.

By the end of the workshop, volunteers were identified to serve as an active roadmap working group comprising cross-cutting stakeholder groups.

3. Workshop Agenda and Approach

The two-day workshop included keynotes, stakeholder sessions and break-out sessions. Sessions explored crucial research topics for the compute-energy nexus. The first day included two keynotes by Dr. Ruchir Puri from IBM and Prof. Prashant Shenoy, and four research themes:

- Stakeholders and team interaction,
- Dynamic grid integration of data centers and AI processes,
- Thermal management and thermal energy integration and recovery,
- Chip and system architecture with energy as a peer requirement.

The second day included three research themes and a discussion of the action plan.

- Power distribution within data centers,
- Datasets, modeling, and algorithms,
- Ultra-scale projects and resources,
- Summary and roadmap initiation discussion and action plan.

Each research theme session was followed by break-out sessions to collect feedback from participants. The detailed workshop agenda with titles of talks and speakers is specified in Appendix B.

4. Invitees

The wide range of stakeholders was a great benefit since many participants were hearing about critical issues and practices in other domains for the first time. Representatives from four groups were invited to the workshop.

- **Industry stakeholder representatives** included energy generators and utility players, data center HVAC vendors, data center implementers, data center operators, semiconductor companies leading AI hardware, enterprise software companies, and large enterprise IT developers and users.
- **Federal stakeholders** came from several agencies. Some had massive systems and deep expertise in computing and energy. Others came from domains where they are charged with defining and executing national strategic priorities related to computing.
- **Academic and industry researchers** focused on key research challenges across the compute-energy nexus. Researchers focused on grid integration of data centers and AI processes, power distribution within data centers, multiscale thermal management and heat recovery, chip and system architecture, and data center and energy-centric models and algorithms.
- **Leaders** from think tanks, regional policy experts, and other leading stakeholders with a broad viewpoint or policy expertise.

5. Challenges

Computing and energy have always been interwoven. Grid-scale electricity systems were among the earliest adopters of advanced computing. The earliest computers were intensive energy consumers, and requirements for power density increases and advanced cooling continue to expand. The two sectors differ in nearly every other respect.

During the workshop, speakers and participants outlined many challenges. Details of discussions are described in Appendix A. Here we highlight the most debated challenges:

A. Scale in Number of Racks and Data Centers: The scales of data centers are increasing. The scale increase happens at the rack level and at the overall data center campus level. This brings a major challenge for power delivery to racks and data centers. Many data center implementers envision near-term increases to 1 MW power delivery to individual racks, and data center/AI computation campuses are approaching 5000 MW. For perspective, a typical large power plant delivers 1000 to 2000 MW. Even the largest (nuclear) power plants in the US currently deliver less than 5000 MW.

B. Time Scale Differences: *Energy system projects* are massive in scale and years or decades in the making. Examples include building and upgrading power plants, routing and constructing large transmission lines, and reworking distribution systems. *Computing system projects* are driven in part by Moore's Law, which is synonymous with urgency, and which compresses time scales to months.

C. Rapid Growth of Dynamic AI Workloads in Data Centers: Rapid growth in computing systems for data centers and dynamic AI workflows magnify interaction issues and concerns.

D. Transparency and Data Interoperability: Software engineers are rarely asked to include energy optimization in their design processes *and may not have the granular energy measurements needed to support this.* Technical information about power ramp rates, ridethrough settings, and data for load models may be buried deep inside infrastructure specifications, hidden behind non-disclosure agreements, or stored in old archives that have not been digitized.

E. **Other Technical Differences**: Terminology differences, engineering considerations, planning, optimization, and reliability issues between power grid and computing infrastructures are distinct across the domains.
F. **Team Building**: The most fundamental challenges are to create fully integrated teams and pursue combined energy and computing growth from a systems engineering perspective. At the workshop, we discussed the need for teams and groups that will represent a comprehensive view of essential collective actions and build near-term tactical programs to address key issues and form a roadmap development working group.
G. **Flexibility, Renewable Resources, and Storage for AI Workloads**: AI computing has the potential for mass-scale flexible energy operation – although participants pointed to limitations on *flexibility*. Retiring coal plants offer interconnect points for distributed energy resources, new loads, and storage. Only an integrated effort can pursue the best strategies to integrate flexibility, renewable resources, storage technologies, and AI processes for cost-effective growth. We believe that the fundamental challenges are distinct from "get more energy, and quickly" or "build a different grid" perspectives.
H. **Funding Disparities**: Grid operators have been accused of underinvesting in equipment modernization and reliability improvement, but electrical distribution companies are tightly constrained by regulators and must justify any such investments in concrete ratepayer cost-benefit terms. Computing and artificial intelligence (AI) industries have experienced massive growth in capitalization and are likely to have access to orders of magnitude more funding than grid operators
I. **Regulatory and Societal Challenges**: The compute industry in general has far higher market capitalization than investor-owned utilities. Combined with tight regulatory regimes for utilities, the issues deeply constrain the ability of utilities to respond to rapid growth needs. For example, grid operators are concerned about adding additional supply with the risk that data center projects do not materialize – leaving them supporting new infrastructure without an end user. Many state regulators will not approve of large supply investments without customer surety. Some states (e.g., Ohio) have adopted requirements to prevent data center operators from dropping out late in the process. There is also the related issue that very large companies on the compute side are viewed as overwhelming potential small host communities, and communications are often sparse.

6. **Findings**

a. *"Bring Your Own Power" Concept*

One aspect that came out of the presentations and discussion is that the *"bring your own power"* concept is gaining traction. The basic idea behind this is that many data center companies plan to provision their own *local energy infrastructure* necessary to operate their facilities. A few hyperscalers are building out complete data campuses with energy resources right now. One example presented at the workshop included a 3000 MW natural gas turbine installation fully integrated into a data center campus footprint. In the context of participants, this highlights how lines are blurring across stakeholders. A few players are working to build cross-functional teams, primarily for data center design and construction.

An important related note is that *local power provisioning* potentially is governed by completely different regulatory regimes than public utilities, since customer and supplier distinctions may be lost. Power grid deregulation efforts from the 1980s and 1990s typically separated the electrical grid into generation companies, transmission companies, and distribution companies. Utilities in many ways are handicapped by this structure in competing with the "bring your own power" concept as they attempt to support large-scale new business.

b. *Disparities in Objectives, Constraints, Priorities, Terminology, and Mutual Understanding*

One of the most notable findings is that the various communities linked to these areas of expertise have disparate objectives, constraints, priorities, terminology, and mutual understanding. Examples ranged from the basis for long time horizons in the electricity grid to limited recognition of funding and investment sources. The disparities perhaps are reaching a critical level. Grid investment constraints from regulators discourage rapid change. When benefits of new equipment are too far into the future and seen to be speculative, regulators rarely approve such plans. At the other extreme, large-scale fund financial resources in the computing and AI industries cannot bypass regulatory constraints. An expedient such as "just purchase your grid company" does not avoid regulatory regimes and is too simplistic.

c. *Need for Detailed Operational Models of Computation for Grids*

Early in the discussion, a theme among the grid-side experts was the need for detailed operational models and predictable time behavior related to computing loads, analogous to conventional planning work for residential and industrial customers. Compute-side experts explained that the requested level of computational details is proprietary and often unavailable but offered assurances that full 100% equipment utilization is one objective. Multiple presenters showed Fig. 1, a sample test trace produced by Microsoft [1]. Computing experts agreed that this trace was representative. However, it is obviously not 100% utilization from an electricity grid perspective – and there was some acknowledgement from computing experts that 100% utilization at the grid connection is extremely unlikely given swings in power consumption in many computing workloads.

Indeed, grid experts pointed out that loads like these can excite sub-synchronous resonances, potentially causing physical failure of rotating machines [2], [3]. It was clear that few of the computing experts were aware of such issues, although the problem is pointed out in [1], and considerable discussion is included there. Choukse, et al. propose a combination of software/hardware activity smoothing and rack-based batteries to mitigate this behavior [1]. Power electronics experts in the audience pointed to task interleaving as an established strategy for smoothing pulsating power flows.

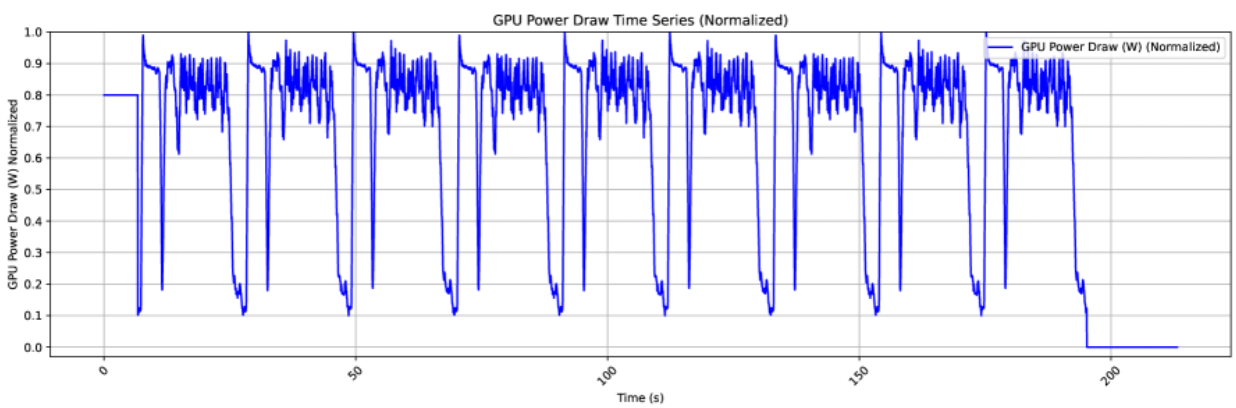

Fig. 1. Power readings from an at-scale training job on DGX-H100 racks. Reprinted from E. Choukse, et al., "Power Stabilization for AI Training Datacenters," Aug. 2025, arXiv:2508.14318v1. Licensed under CC BY-NC-SA 4.0.

d. *Global Competition for AI Development and Understanding of Future AI Loads*

From the computing perspective, AI development is perceived as a global competition and race. The need for fastest possible results is seen as a corporate and even national imperative. With investment levels reaching into trillions of dollars, revenue and worthwhile return on investment are essential. Any data center project does not generate return until it is up and operating. Together, these factors drive a relentless push for fast turnaround.

Current AI energy planning and expectations are dominated by training processes. It can be anticipated that future AI inference processes are likely to have much different energy behavior and perhaps lead to more distributed computation. Anticipated shifts from AI training to inference will alter energy consumption patterns, but AI systems run billions of agents, and expected 10x increases are likely to increase energy consumption despite any improvements in computing efficiency.

e. *Understanding of Electricity as an Instantaneous Exchange Resource*

From the grid perspective, electricity is not a commodity that can be shipped and stored, but instead an energy exchange medium that requires instantaneous matching. The grid must enforce moment-by-moment source and load balance. On the fastest time scales, any momentary mismatch links to the (limited) rotating kinetic energy of spinning generators. On slower time scales, dynamic adjustments of generators, combined with comprehensive statistical models of time-behavior of loads, allow operators to make the system work. Rapidly changing computing loads, unpredictable external sources, and other attributes of data centers are problematic in this context. The challenges are potentially untenable when loads scale to the size of large power plants.

f. *Understanding of Thermal Management Challenges*

From a thermal management perspective, the energy efficiency of computing systems is exactly zero. All energy supplied into a computing facility and process generates heat that must be managed and dissipated. The thermal engineering challenges of data centers range from the intense heat flux from individual devices (approaching 1 kW/cm^2) to the diffuse gigawatt-level heat generated across a complete facility.

g. *Understanding of Heterogeneity in Equipment, Energy Consumption, and Policies*

Computing, grid, and thermal management attributes of data centers are taken as self-evident by the respective communities but are not necessarily perceived as engineering requirements and constraints by the others. Participants added several elements along similar lines, such as design considerations across hardware and software, the fact that computing equipment is heterogeneous and difficult to model in terms of energy consumption, policy aspects that constrain grid development, community considerations, batteries and their costs, reliability expectations for energy supply, and others.

7. Implications of Challenges and Findings

The implications are rather extreme.

Implications of Different Perspectives: From the grid perspective, rapid load variations such as those in Fig. 1 must somehow be made up with a variable source, local battery storage, or combinations. Such behavior at scales of 1000 MW is far outside of prior experience or available technology. From the compute perspective, somehow resources need to be available and fully installed on time scales of months rather than years. From the thermal perspective, the heat dissipation from a 5000 MW data center campus is on the scale of that from a large nuclear plant, but typically the operating temperatures are intended to be quite low (on the order of 70°C), making heat removal or reuse especially challenging.

Implications of Funding Diversity: The huge cash flows and investments across the compute industry ultimately come from families and individual consumers – bills for streaming services and internet connections, margins added to products to cover advertising, electricity and water bills, and other cash streams. The discussion made it clear that the value of data centers is *not being conveyed effectively to these end-use customers*. Until quite recently (a year or so), it was difficult to raise investment funds for electricity grid hardware innovations. Today, some types of investments are being made almost negligently. There were many examples of large investments in early or even marginal technologies, and it is unclear whether the most promising innovations are getting support compared to the most dazzling. Since recent large-scale investment is driven by the computing industry, but reaches deeply into energy, thermal, and other contexts, the need for comprehensive systems considerations and system-targeted funding was made clear.

Implications on Electrification of Other Sectors: Electrification offers important benefits across many industries, including cement manufacturing, steel manufacturing, and other process-oriented industries. One consideration is whether data centers and the computing industry can impact this long-term evolution. It is not obvious that AI can support foundational changes in process industries, although load flexibility and high-performance computing are likely to help make inroads into industrial electrification. How can data centers and potential flexibility enhance the value of distributed and renewable resources?

8. Next Steps

Needs for stakeholder alignment, comprehensive co-design and operational coordination, broad engineering teams, and technical leadership are clearer following this workshop. Many of the attendees expressed willingness to be initiators of stakeholder alignment in their respective fields.

Building interfaces among technical and other stakeholder communities was seen as urgent. A critical issue is that grid operators, facility designers, cloud software engineers, power systems engineers, power electronics engineers, thermal systems engineers, computing hardware engineers, policy experts, and indeed all the stakeholders represented at the workshop speak different "languages," and must establish a common working base.

Identified short-term needs included:

- Fine-grained data collection and standards for it;
- Regulatory change to drive effective reporting about energy and water consumption;
- Developing accurate dynamic models for grid-interface data center operation,
- More comprehensive site planning with community engagement,
- Much more comprehensive collaborative teams. Often today the largest or loudest stakeholders dominate the setting of priorities. A broader decision base is needed that emphasizes the customer voice (including end-use customers), supplier voices, and community issues.

The group hopes to follow up by establishing one or more working bodies, seeking to develop a national roadmap that will help bring the stakeholders together. The group also hopes to establish a series of seminars and workshops to help better inform the communities involved.

References:

[1] E. Choukse, et al., "Power Stabilization for AI Training Datacenters," Aug. 2025, arXiv:2508.14318v1. Available: https://arxiv.org/abs/2508.14318.

[2] P. M. Anderson, B. L. Agrawal, J. E. Van Ness, "*Subsynchronous resonance in power systems.*" New York: Wiley-IEEE Press, 1990.

[3] "Reader's guide to subsynchronous resonance," *IEEE Trans. Power Systems*, vol. 7, no. 1, pp. 150-157, Feb. 1992, doi: 10.1109/59.141698.

Appendix A: Session Summaries

1. **Welcoming session.**

2. **Keynote Addresses.** The keynote talks addressed the future of business and society built around massive AI engines. The nature of AI workloads and efficiency considerations were presented. Three important developments emerged from the keynote talks and participant discussion:

   a. "Efficiency" requires a different perspective and understanding when incoming energy and power are constrained. The challenges need to be framed more in the sense of "What computational capability can be supported with a 1000 MW source? How can capability increase in an energy-constrained context?" This is consistent with a view based on the Jevons paradox: as individual computing device efficiency improves, there is increasing pressure to dedicate more resources to computing. *Efficiency improvements, for the time being, do not tend to reduce energy demands and may drive them up as systems get larger.*
   b. Current AI energy planning and expectations are dominated by *training processes*. It can be anticipated that future *inference processes* are likely to have much different energy behavior and perhaps lead to more distributed computation. The implications for hyperscale data center projects will not be clear for a few years, but agent-based AI usage is already accelerating dramatically.
   c. For energy consumption, the underlying principle in the electricity grid is that energy sources and loads must always match every moment: the energy balance process is continuous. Large generators cannot adjust instantly, but the basic operating requirement for moment-by-moment matching remains. Critical grid-related factors include power ramp rate limits, load variability limits, and procedures for operation through and recovery after faults and failures. *There are a few emerging load rescheduling and load throttling methods, but not at the dynamic rates associated with rapid shifts in computing processes.*

3. **Stakeholders and Team Interaction.** The session emphasized comprehensive local system design that includes a full team effort representing energy resources, storage, compute hardware, software, AI needs, networking and local communications, power distribution, power electronics, local power grid operation, cooling systems, reliability, and so on. The talks were chosen to touch on crosscuts from siloed topics, such as renewable energy, interconnection and energy congestion, network design and performance, data center dynamic energy usage, and water availability. Some of the highlights:

   a. For scale, it was pointed out that 100 MW of electrical base load must be linked to about 3 mi^2 of land (about 750 hectares) if it is to be covered with photovoltaics and batteries. Also, a 1000 MW load is typical of a large city, such as Philadelphia. Fault examples included full data center dropout in response to quick lightning disturbances, followed by full-on return. Many companies use onsite batteries or other "backup" power generation techniques like gas turbines and diesel generators. In the event of a power outage lasting more than a specific number of seconds, the companies will switch to these backup approaches to ensure continued operation. Abrupt switchovers cause large-scale swings in the power grid. *There is an urgent need to better define switchover and dropout behavior to reduce unanticipated large-scale grid disturbances.*
   b. There are potential game-theoretic approaches to data center interactions with the grid to *manage load flexibility, but markets and policies need to create incentives for flexibility implementation and control.*
   c. The anticipated spending in this context is expected to reach $3 trillion by 2028, of which roughly 10% is for electrical energy. Some of the newest AI engines have been using millions of GPU hours to train. Although there are multiple projects seeking to reduce energy needs per unit of

computation by 10x, the pace of growth exceeds the rate of improvement: Training requirements are growing about 10x every 18 months.

d. Anticipated shifts from AI training to inference will alter energy consumption patterns, but AI systems run billions of agents, and *expected 10x increases are likely to increase energy consumption despite any improvements in computing efficiency*.
e. Some concepts add intermediary resources with software control to (partially) decouple data center loads and the grid. The scale of storage is unclear but relatively high, although potentially modest compared to total data center capital and operating costs. In any case, *cooperation between computing and energy systems is essential, and flexibility is an important aspect*.

4. **Breakout session.** Attendees broke into crosscutting teams, sought to identify stakeholder groups, outlined their roles and needs, and discussed the stakeholders and ways to establish much more interaction across domains.

   a. Who are the stakeholders? All groups pointed to an extended list. Beyond the most obvious (data center builders, operators, corporate users; grid operators in distribution, transmission, and generation companies, independent system operators; chip makers and AI component makers; power electronics conversion and integration engineers), the groups pointed to communities, policy makers and the governments linked to them, equity investors, data center and compute facilities across a wide size and power spectrum, and many others. A few attendees sought to remind everyone that individual end users are key stakeholders and the ultimate source of the money involved.
   b. "Data centers" range from small in-house facilities to massive GW-scale "plants," and users across this spectrum have contrasting concerns. It was argued that small and medium data center sizes make up the majority of computing infrastructure.
   c. The compute industry in general has far higher market capitalization than investor-owned utilities – even though utilities have massive installed capital in the grid. Combined with tight regulatory regimes for utilities, the issues deeply constrain the ability of utilities to respond to rapid growth needs. For example, grid operators are concerned about adding additional supply with the risk that data center projects do not materialize – leaving them supporting the infrastructure without an end user. Some data center developers can afford to initiate multiple projects, intending to finish only a fraction. Many state regulators will not approve of large supply investments without customer surety. Some states (e.g., Ohio) have adopted requirements to prevent data center operators from dropping out late in the process. There is also the related issue that very large companies on the compute side are viewed as overwhelming potential small host communities, and communications are often sparse.
   d. Issues linked to non-disclosure agreements (NDAs) and whether they have any real value in helping close gaps between large data centers and grid operators were mentioned.
   e. Although there are major challenges in workforce development, the compute industry perceives that the rapid pace of development is a mismatch to academia and government.
   f. Along all these lines, there is an urgent need for new frameworks that support rapid engagement and can be effective in linking wider stakeholders.

5. **Dynamic Grid Integration of Data Centers and AI Processes.** The presenters considered large digital loads both from AI infrastructure perspectives and from grid perspectives. Data center energy patterns are much different from existing aggregate loads. The "bring your own power" concept means that some data center operators are becoming, in effect, their own utility operators – which also reduces time to build because they can avoid connection to the power grid. AI processes potentially are more flexible than today's data center workloads, but economic incentives will be important. Generation technologies to reduce supply response time, optimized alongside energy storage, have large potential impact. Some of the highlights include:

a. *Renewable energy and workload migration* are central to future development. Since solar and wind energy vary over time and geography, it is plausible for workloads to migrate toward locations with abundant energy.
b. There is a significant need for *fine-grained telemetry and modeling* in this space, as it is challenging to optimize for operation when so much of the data are proprietary. Starting with DOE or NSF clusters, which may be more amenable to releasing detailed operating information, was suggested as one starting point.
c. *Backup power integration and control* have been problematic. An example was a data center load at the scale of 1.5 GW that switched over to local diesel backup following a lightning arrestor failure and later switched back to the grid at full power. The resulting large dynamic frequency (i.e. generator speed) and voltage fluctuations were almost enough to drive major grid failure. Data centers need to be able to ride through conventional grid fault protection activity. Any ability to *support reasonable down or up power ramp rates* could help. Some utilities are engaging actively to specify ramp rates, backup switchover behavior, and dynamics. A few are seeking to bring dynamic requirements into contracts.
d. The basic issues of *return on investment* (ROI) are highly siloed. ROI propositions are much different among grid operators, data center operators, companies building AI training centers, and so on. One example is trying to determine *who* invests and how to get ROI on energy storage installations, dynamic energy management, and renewable integration.
e. During discussion, one question raised was "*Given access to a spare 150 MW, could you use it to build a data center?*" The answer seemed to suggest that larger players in computing are beyond this scale, with 300 MW mentioned as a lower level. They implied there is little interest below that level.
f. The question of "*Who should pay for power variability*?" was answered in terms of the causing agent. However, costs for power variability are not fully quantified yet, although some participants have begun preliminary efforts on this.

Parallel Sessions, each with three 20-minute invited talks

6A. **Thermal Management and Thermal Energy Recovery**. Data centers convert essentially 100% of incoming energy to heat. This session discussed local and chip thermal management, including l*iquid, two phases, and immersion cooling*, *reuse of rejected heat*, and *thermal energy sto*rage. Some of the highlights:

a. Data centers generate large volumes of lower-grade heat, rarely more than 70°C. Liquid cooling can help concentrate this, and the level is plausible for district heating, certain carbon capture processes, and hothouse agriculture.
b. Water availability is a major concern. Some data center developers have gone to closed-loop liquid cooling systems that ultimately reject heat to the atmosphere to reduce pressure on water consumption.
c. Liquid cooling requires comprehensive management of coolant materials and quality. It is hard to make direct use of reclaimed or brackish water. Advanced cooling methods typically require high water quality.
d. Cooling system sizing and reliability are critical issues. As racks scale to 1 MW, liquid cooling is essential and must be extremely robust – and two-phase and immersion cooling are being investigated. At the mechanical densities required, there are extensive open research challenges.
e. A few data center integrators are using ground-source technology as effective thermal storage, but ground-source heat pumps have scale limitations. For example, Epic Systems has a facility in Wisconsin, likely the world's largest, involving more than 6,000 boreholes over about 5.5 km^2 and able to cool a 3.5 MW data center. It is hard to envision scaling up such a facility by a factor of 1000.

6B. **Chip and System Architecture.** This session presented computer and system architectures, methods, software, and energy integration to reduce energy per unit of computation by orders of magnitude. It considered *adaptable heterogeneous systems and secure support for legacy hardware reuse*. Some of the highlights:

a. *Large language models (LLMs) impose high costs and latency*. Tradeoffs can be substantial, for instance adding a bit of latency and slowing response by about 10% can provide a 5x reduction in compute energy per token.
b. *Pathways to 10x and 20x improvements* are plausible but require *new algorithms and more flexible hardware*. There is an inverse relationship, however, between token throughput of a device and token interaction with a user.
c. Some LLMs are now managing more than a trillion parameters, but the fast memory requirements scale with this demand and are now also above 1 TB. The need for high-speed memory is creating shortages and price increases in DRAM and related markets.
d. Architectures now integrate FPGA and processors directly with high-bandwidth memory and SRAM. Chiplet configurations support this. More specialized heterogeneous chiplet implementations imply compilers that break code into modules suitable for each accelerator or network management device. The pathways lead to near-memory and in-memory compute architectures.
e. In some cloud computing contexts, hardware is reported to be greatly underused (in contrast to earlier claims about 100% utilization). This implies opportunities for dynamic load control and implies a need for *comprehensive energy state management* so that *idle hardware draws no power*.
f. *Demand response concepts* are in use in some existing locations to alter energy use over time frames of minutes to hours. The grid signals, pricing models, or emergency situations are quantified in some jurisdictions. The EPRI Flex DC (flexible data center) program is an emerging example of industry attempts to implement flexibility in time and geography.
g. *Sustainability*, including embodied carbon content, operational energy consumption, water, and other factors tend to be highlighted at the intersections of design, operation, grid awareness, and suitable models for hardware components, generation mix, and other factors.

7. **Breakout session.** Attendees broke into crosscutting teams, assigned at random and different from the first grouping, with a charge to identify critical roadblocks.

a. *Concerns about public distrust, weaknesses in community engagement*, and the need for a community-linked version of the Open Compute Project (OCP), plus limited engagement with policy and government stakeholders are substantial roadblocks. Often the main actors causing distrust (e.g. hyperscalers and big-tech companies) rely on third parties and are not involved in direct discussions.
b. Major roadblocks include *materials limitations and supply chain bottlenecks*, such as rare-earth materials and copper, magnetic steels and power transformers, manufacturing limitations for generation and electrical equipment, and advanced electrical construction and installation expertise.
c. *Problems of transparency and data related to interoperability* are major bottlenecks. For example, software engineers are rarely asked to include energy optimization in their design processes and *may not have the granular energy measurements needed to support this.* Technical information about power ramp rates, ridethrough settings, and data for load models may be buried deep inside infrastructure specifications, hidden behind non-disclosure agreements, or even in old archives that have not been digitized.
d. The *necessary workforce* is far beyond computer scientists, and includes electrical engineers and power experts, mechanical engineers and thermal experts, civil engineers and construction experts, and the technicians and skilled workers to install and operate major projects. Experts in direct current (dc) power systems are rare, for example. However, support for workforce development and active mutual engagement have been weak.

e. Are there concepts such as one-minute ridethrough storage, closed-loop cooling with net zero water use, dynamic energy pricing structures, or investment further down supply chains, that could alleviate some of the roadblocks?

This marked the conclusion of the first day of the event. Subsequent sessions took place on the second day.

8. **Power Distribution within a Data Center.** This session addressed dc distribution and emerging concepts within data centers and among data center clusters. Implications for computing architecture and hardware changes to enhance energy supply matching were described. Some of the highlights are:
   a. Many data center designers are exploring a transition to internal dc distribution. Dc distribution can eliminate a few conversion stages and provide better integration with energy storage and renewable resources. Emerging solid-state transformers (SST) technology can help support this transition even though the technology is not fully mature
   b. Enterprises seeking to "win" the AI race are turning to direct on-site power, with emphasis on natural gas. This has added to supply chain pressure – gas turbine order books are said to be full for at least the next four years. However, rapid growth, fast load swings, and rising power densities require combined technology solutions that involve turbines, SSTs, batteries, microgrids, and data/AI campus designs to levels of 5 GW and up.
   c. There are already hyperscale campuses operating above 1 GW of IT load, and the expected increase to 1 MW per rack is being provisioned. Such racks are heavy, and single-story slab-on-grade construction is typical today. Microgrid designs are in use for data campuses. Dc distribution at 400 V and at ±400 V (and equivalent 800 V) are near deployment, although there are not enough qualified manufacturers and vendors above 400 V dc. There was limited visibility, and some misconceptions, about the distinctions between unipolar and bipolar dc distribution (e.g. 800 V dc compared to ±400 V).
   d. Speakers argued that futureproofing requires flexible, upgradable designs. New data campuses are being designed with options for voltage upgrades, provisions for onsite generation, modular cooling, and other flexible attributes.
   e. Internal power levels are on a path to 1 MW per rack. Presenters pointed out that this implies hundreds of kilograms of copper per rack unless there are fundamental changes in operating voltages and methods.
   f. In a massive data center, even components with million-hour mean time between failures (MTBF) will experience a few daily failures based on huge numbers of devices. Failure identification, management, isolation, and repair must be planned into designs.

9. **Datasets, Modeling, and Algorithms.** The session discussed benchmarking load requirements. Quantifiable results related to data centers and operation. Energy data and performance, energy site integration, models and algorithms for deployment across geographical sites. Some of the highlights include:
   a. Data centers are not the only new load coming online. Drivers include data centers, crypto currency, industrial electrification, and transportation electrification. Many of these are large discrete loads, and a challenge is that these do not benefit from statistical distribution over time and space. Smaller, more distributed loads with gradual growth (some of this is experienced with transportation) can be calibrated into conventional forecasting models.
   b. At GW scales, transmission capability comes into play, but transmission networks are a complex policy environment. Conventional grid planning methodologies are inadequate for many emerging types of grid-enhancing technologies.
   c. Relevant datasets include technology operation information, granular weather information, load dynamic operation, and functional implications of load interconnection. Note that detailed data about computing centers as loads is not available. The audience discussed how data availability

might be encouraged or even required. *Databases about load-side dynamic power needs and operation represent an important opportunity for future work.*

d. AI system operation can be characterized by flexibility, energy efficiency, and performance. In general, pick any two, and the third must follow (e.g. if the operator prioritizes flexibility and energy efficiency, performance will be constrained, and so on). However, flexibility is especially difficult, since a compute process is likely to have an inflexible completion time.

e. Metrics to reduce carbon footprint are feasible and being tested. An example was given of a metric that can drive a data system to 33% carbon footprint reduction with about 5% increase in time to perform computing tasks.

f. At present, about 20% of data center power needs are linked to AI processes, and this is growing quickly. A major problem is that energy use and environmental impact are hidden "black box" issues to end users. One presenter suggested that 90% of AI power consumption goes to inference, although many others pointed to multi-month training sequences as the dominant energy use.

g. Metrics such as *power usage effectiveness (PUE)* (total facility energy as a ratio to IT equipment energy), water usage effectiveness, and carbon intensity factor have been defined at the data center level but do not necessarily help end users understand the impact of their queries. Analysis suggests that the average energy consumption of one popular LLM is on the order of *11 MWh per household per year*. This is about *the same as average annual household electricity consumption!* PUE is staying relatively flat (and is expected to stay flat in the near future) even as more data centers and capacity are being added, worsening the existing problem.

h. The *Jevons paradox* is kicking in: since energy supply becomes a limiting factor, better compute efficiency leads to more hardware installed in a data center and more software work, i.e. improving energy use by 10x leads to at least 10x increase in the amount of active computing performed. The total energy consumption does not fall and might even grow to keep up with accelerating demand, especially as AI agents become widespread.

i. There was recognition that *cooperation between data centers and the power grid is necessary* – the mindset of *"consume as much power as desired, whenever it is desired"* is no longer an acceptable operating regime given the constraints.

10. **Breakout Session.** The final breakout session emphasized critical short-term and long-term actions that need to be instigated and executed. Attendees broke into crosscutting teams with a charge to consider actions and action plans. A key theme was the need to create a broad facilitation group.

a. Building interfaces among technical and other stakeholder communities is urgent. A critical issue is that grid operators, facility designers, cloud software engineers, power systems engineers, power electronics engineers, thermal systems engineers, computing hardware engineers, policy experts, and indeed all the stakeholders represented at this workshop speak different "languages," and must establish a common working base. Co-design is vital.

b. *Broad short-term needs* identified included fine-grained data collection and standards for it, regulatory change that is driving effective reporting about energy and water consumption, developing accurate dynamic models for grid-interface data center operation, more comprehensive site planning with community engagement, and much more comprehensive collaborative teams.

c. Hardware developers, software developers, and end users have limited energy use data and little feedback or incentives to enhance energy performance. *Can there be "green certification" methods and metrics?*

d. Often today the largest or loudest stakeholders dominate the setting of priorities. A broader decision base is needed that emphasizes the customer voice, supplier voices, and community issues.

e. There is an urgent need for data center design and operation to make the best use of an available power source, scaling down to a few megawatts rather than massive GW campuses. The near-term economies of scale for energy growth favor locations with modest excess energy, although perhaps the scales for the computing industry favor huge installations.

f. Regulatory regimes and investor expectations are barriers to funding for power grid enhancements, and ratepayers end up in the middle. More direct ways to support comprehensive grid improvements that do not burden ratepayers are needed urgently.
g. The computing industry needs to consider *workforce issues* and long-term challenges, since these require thoughtful longer-scale investments. The issues include supplier and support technologies such as dc protection and circuit breaker hardware, medium voltage components at scale, and suppliers of magnetic steels. Some participants voiced the opinion that "the market will drive this," but there is a chicken-egg situation in many such supplier markets with limited willingness and ability to invest billions of dollars in development or production capacity based on uncertain and potentially short-term future product orders.
h. High-fidelity models for data center and grid interaction are lacking.
i. What do re-use, re-work, recycle and circular economics issues look like for large data centers? The operating life might be dictated more by AI functionality than by hardware operation.
j. Can this group establish *regular conference and interactive communication channels*? How can we develop *open-source engagement*?

11. **Ultra-Scale Projects and Resources**. The session discussed large-scale data center and AI projects from a federal perspective and in the context of national strategic priorities. It addressed industry perspectives on energy efficiency in ultra-scale systems. There was discussion of policy and strategy for AI infrastructure growth. Some of the session highlights include:
a. Compute cluster costs are growing quickly in scale from billions of dollars to trillions of dollars. The rate of change has created bottlenecks in power transformers, switchgear, distribution hardware, and other power grid equipment. Power contracts and economic approaches are also a challenge. However, from the perspective of companies using data centers, these costs are not viewed as the biggest concern. At present, they are willing to pay significant "startup" costs if it helps them "win" the AI race. Minimizing the time from "design to power on" of a data center is viewed by these companies as the biggest challenge.
b. Power needed to train LLMs is likely to increase to about 100 GW by 2030, at which point it is more than 20% of U.S. electricity consumption. For reference, present national capacity is about 1280 GW.
c. Wafer production (almost all from Taiwan) will need to increase by 4x to meet demand.
d. Natural gas supplies about 45% of electrical energy in the U.S. at present. Although there is limited spare capacity, rapid growth relies on additional natural gas production.
e. ARPA-E has been studying how to accelerate deployment of nuclear power and geothermal power. Small modular nuclear reactors (SMRs) in particular are seen as a potential near-term option with significant potential, but there is no consensus on whether SMRs are realistic in time frames faster than ten years.
f. Chiplet, systolic arrays, custom inter-chip interconnections, and many other hardware innovations are being explored and developed to reduce energy requirements.
g. Software developments to address energy range from integer numerical algorithms to dynamic workload profiling and include efforts to optimize at the compiler level. A relevant consideration is that AI training lends itself to lower computing precision compared to detailed engineering simulation tools.
h. Hardware-software co-design is an important area of emphasis. Metrics that include the "occupied time" (fraction of time when a chip is performing a job) and "on-duty time" (the fraction of time when a chip is performing useful operations) are important to drive improvements. However, fully coordinated engineering teams across hardware, software, and energy are rare even within individual companies, let alone at a coordinated national level.
i. Load flexibility is potential extremely beneficial from a grid perspective but is a difficult challenge on the hardware/software side, especially when thousands of chips are working together. Compute-side impacts of flexibility on reliability are unknown.

j. A fundamental challenge is to consider critical underlying data center and AI questions: "How fast?", "For what?", "How much?" in full context. Industry wants speed, ambition, and scale. Local communities are unconvinced about ambitions. Federal and state authorities are concerned about speed and scale. Alignment is needed across many stakeholders.
k. Policy stakeholders with distinct perspectives include the federal government, state governors, state regulators, and local communities. Federal programs at present seem to promote rapid, large-scale AI infrastructure deployment. Governors seek to balance economic development goals and legislator concerns, and for example might veto legislation curbing data center development. State regulators must balance established utility planning frameworks against rapid growth of large, concentrated loads, and at the same time protect grid operation and stability. Ratepayer protection is important. Examples include a new rate class for data center customers in Virginia and long-term payment guarantees in Ohio. Local communities worry about electricity cost, property values, water usage, and construction pollution, balanced against tax revenue.
l. How much computing is enough? What is AI really for? The answers are unknown in general, and public perception of AI is limited but often negative.
m. A framework for an integrated national AI strategy is key, and needs to be built on alignment across policy makers. Clarity and priorities for societal benefits are essential. The DOE Genesis Mission is a recent example of an attempt to address the issues in part.

12. **Summary and Roadmap Initiation Discussion.** Following the sessions and discussion, a summary was presented. A few additional points were highlighted:
    a. All stakeholders need to keep in mind that the huge cash flows and investments across the industry ultimately come from families and individual consumers. The value of data centers is not being conveyed effectively to these end-use customers.
    b. Until very recently (a year or so), it was difficult to raise investment funds for grid hardware innovations. Today, investments are being made almost negligently, and it is unclear whether the most promising innovations are getting support compared to the most dazzling.
    c. Electrification offers important benefits across many industries, including cement manufacturing, steel manufacturing, and other process industries. How can data centers and the computing industry play into this long-term evolution? How can data centers enhance the value of distributed and renewable resources?
    d. Needs for stakeholder alignment, comprehensive co-design and operational coordination, broad engineering teams, and technical leadership are clearer following this workshop. Can our attendees help to be initiators of stakeholder alignment in their respective fields?

    Attendees were encouraged to participate in preparing the workshop report and in future road mapping efforts.

13. **Concluding Remarks.** Wrap-up by organizers and sponsors.

Appendix B: Schedule, Speakers and Talks:

Date, time	Activity summary
Dec. 8	
8:00	Breakfast and welcome
9:00	Keynotes: Ruchir Puri (IBM) "Dawn of a new Era: Reimaging Business & Society with Intelligent Machines" Prashant Shenoy (UMass) "Data Centers, AI Workloads, and Efficiency: A Systems Perspective"
10:00	Break
10:15	Stakeholder session: Howard Gugel (NERC) "Reliability Impacts of Inverter Based Technology" Andrew Chien (UChicago) "Facing Computing's Challenges at the End of Technology Scaling – the Power/Energy-Compute Nexus" Alaa Youssef (IBM) "Where are we heading with Generative AI?" Panel Discussion
11:30	Breakout Session 1: Identification of Stakeholders and Needs Report back
12:15	Working lunch
1:15	Grid integration session: Jian Huang (UIUC) "Towards Sustainable AI Infrastructure: Chips, Servers, and Racks" Nathaniel Rice (Dominion) "Utility Perspective of Large Digital Loads" Brian Zahnstecher (Power Rox) "Keeping Both the Grid and Load Happy in Multi-MW Data Center Implementations" Panel discussion
2:30	Break
2:45	Parallel sessions
	Session A: Thermal and water management Kashif Nawaz (ORNL) "Turning Heat into Opportunity: Next-Generation Waste Heat Recovery for Data Centers and Computational Ecosystems" Davide Ziviani (Purdue) "Impact of Degradation Phenomena on Reliability and Energy Efficiency of Data Centers" Praveen Kumar (Illinois) "Beyond the Compute–Energy Nexus: Integrating Regional Water Availability into the Infrastructure Equation"
	Session B: Chip and system architecture Karu Sankaralingam (NVIDIA) "Serving Intelligence: The Unseen Limits of LLM Inference and the Hardware Race to Overcome Them" Çan Hankendi (Boston University) "Enabling Sustainable and Grid-Aware AI Data Centers" Josep Torrellas (Illinois) "Addressing Energy Inefficiencies in Datacenter Infrastructure"
3:45	Breakout Session 2: Identification of Critical Roadblocks Report back
4:30	Day 1 closing remarks
5:00	Social hour

Dec 9	
8:00	Breakfast
8:30	Power distribution and management session: Srdjan Lukic (NCSU) "Advancing power distribution in data centers through solid-state transformer (SST) technology" Rajesh Gopinath (Oracle) "Gigawatts at the Speed of AI: Power System Realities of the Next-Generation Data Center Campus" William Schaumann (Burns & McDonnell) "Evolution of Datacenter Designs" Panel Discussion
9:45	Break
10:00	Datasets. modeling, and algorithms session: Bruce Tsuchida (Brattle) "Loads of Loads" Mohammad Hajiesmaili (UMass) "Algorithmic Foundations of Grid-Integrated AI Infrastructure" Abdeltawab Hendawi (URI) "How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference" Panel Discussion
11:15	Breakout Session 3: Critical Short- and Long-Term Actions to be Taken Report back
12:00	Working lunch
1:00	Ultra-scale systems session: Isik Kizilyalli (Avalanche Thinking) "Data Centers and Artificial (General) Intelligence Energy Consumption in the Next Decade" Sushma Prasad (Google) "Towards Ultra-Scale Energy Efficiency" Tanya Das (Formerly Bipartisan Policy Center) "Aligning America's AI Infrastructure Strategy" Panel Discussion
2:15	Break
2:30	Summary and road mapping next steps
3:00	Concluding remarks
3:30	Adjourn