Recommendations on Powering Artificial Intelligence and Data Center Infrastructure

Presented to the Secretary of Energy on July 30, 2024

Data center power demands are growing rapidly. Connection requests for hyperscale facilities of 300-1000MW or larger with lead times of 1-3 years are stretching the capacity of local grids to deliver and supply power at that pace. A significant factor today and in the medium-term (2030+) is expanding power demand of AI applications. Advancements in both hardware and software have enabled development of large language models (LLMs) that now approach human capabilities on a wide range of valuable tasks. As these models have grown larger, so have concerns about sizeable future increases in the energy to deploy LLMs as AI tools become more deeply woven into society. With DOE's leadership role in energy efficiency, clean energy deployment, innovative grid technologies, and AI-related energy consumption and research, the department can play a central role in helping the nation meet these new, strategic energy needs.

The SEAB Working Group on Powering AI and Data Center Infrastructure has examined options for supporting these growing power demands reliably and affordably without harming existing customers and while limiting greenhouse gas emission impacts. The inquiry proceeded along three closely coordinated tracks:

1. Examination of energy efficiency and power dynamics in LLM training and inference.
2. Exploration of an operational flexibility framework to address current bottlenecks, based upon active collaboration between electricity companies and data center developers and operators.
3. Study of generation and storage technologies available today and in the future, examining approaches to more accurately project power needs, address supply chain constraints, and accelerate deployment at scale.

Methodology

The Working Group reviewed available information and reached out to a diverse set of stakeholders to solicit their views. These stakeholders included:

- Hyperscalers: Amazon, Google, Meta, Microsoft, OpenAI
- Data center developers/innovators: Blackstone/QTS Data Centers, Digital Realty, Verrus
- Technology providers: Fervo, General Electric, Hitachi, Intel, HPE, Long Duration Energy Storage Council, Nvidia
- Electricity companies: Associated Electric Cooperative, Constellation, Duke Energy, Evergy, NPPD, NextEra, PPL, Portland General, PSEG, Southern Company/Georgia Power, Vistra
- Independent system operators and regional transmission operators: CAISO, MISO, PJM, SPP
- Environmental NGOs: NRDC
- Researchers: Association for Computing Machinery, Brattle, Caltech, Carnegie Mellon, Department of Energy, EPRI, Johns Hopkins, IEEE, LBNL, MIT Lincoln Lab, NYU, UC-Santa Barbara, University of Chicago

Overarching Consideration

Before the detailed recommendations are made, it is worth noting some overarching considerations. The scale of the potential growth of both the electricity and the information technology sectors due to AI is extraordinary and represents the leading edge of projected electricity demand growth. This will invariably impact tribes and communities across our nation both as potential opportunities and challenges. It is important that the DOE and the electricity and information technology business sectors engage with local tribes and communities sufficiently early for planning and to address two critical issues: (a) to develop community benefits plans; (b) to streamline and mitigate risks for infrastructure development.

Highlighted Recommendations

Recommendations are provided for each track. We highlight six here, addressing immediate and longer-term challenges for each track:

Track 1 – Energy efficiency and power dynamics in LLM training and inference

- For immediate impact, the Secretary should direct relevant offices across DOE to explore opportunities for temporal and spatial flexibility in AI training and inference, and to demonstrate and publicize these capabilities in collaboration with the national labs and private partners.
- For immediate and longer-term impact, the Secretary should establish a data-center-scale AI testbed in DOE, which should be complementary to, but distinct from, the current set of high-performance computing facilities operated by DOE. This testbed can allow researchers from the national labs, academia, and industry to collaborate in development and assessment of algorithms for energy-efficient and/or energy-flexible AI training and inference, advancing the nation's AI capabilities and building on the success of comparable public-private efforts that have accelerated advances in high-performance computing.

Track 2 – Examine secure operational frameworks that allow data centers to optimize their energy consumption, contribute to grid peak load management, and provide other grid services.

- For immediate impact, the Secretary should convene energy utilities, data center developers and operators, and other key stakeholders to start active dialog on how to address current electricity supply bottlenecks, to advance understanding of real-time data sharing and protocols to govern data center operational flexibility (including both computational flexibility and backup power strategies), and to develop strategies for how to generate and deliver the power needed to sustain AI leadership into the future.
- For longer-term impact, the Secretary should work with other government agencies and the private sector to develop a standard taxonomy and framework for defining and orchestrating grid services for large energy users that is adaptable to local and regional circumstances and priorities.

Track 3: Explore generation, storage and grid technologies to power data centers

- For immediate impact, all stakeholders emphasized the need for increased flexible, firm electricity supply to address current reliability concerns that are exacerbated by load growth. The Secretary should direct DOE and the national labs to conduct a rapid assessment of cost, performance, reliability, availability, and supply chain issues facing generation, storage, and grid technologies to support regional data center expansion. This technological and modeling assessment should include technology strategies (consistent with regulatory requirements) for addressing the economics and carbon footprint of new natural gas capacity additions, broader use of existing gas generation, delayed retirements of coal and nuclear, uprates of nuclear and hydroelectric facilities, as well as demand-side efficiency and flexibility improvements in data centers and other electric end-uses. Furthermore, the Secretary should direct DOE and the national labs to assess and deploy opportunities for reconductoring and integration of other grid enhancing technologies to increase their power carrying capacity of existing transmission rights-of-way.
- For longer-term impact, the Secretary should accelerate private investment in emerging technologies by supporting legislation that de-risks private investment in new technologies and by providing technical support to data center owners interested in making long-term financial commitments to next-of-a-kind technologies in nuclear, geothermal, long-duration energy storage, and CCS that are aligned with DOE liftoff reports.

The remainder of the report summarizes key findings from listening sessions and recommended actions for each track.

Track 1: Energy efficiency and power dynamics in large language model training and inference

Led by John Dabiri (California Institute of Technology), track 1 focused on opportunities to leverage improvements in how AI models are trained and queried ("training" and "inference", respectively) to mitigate stresses on the energy grid or even to provide new mechanisms of load-balancing.

Findings

1. Predictions of future energy demand are fraught with uncertainties due to: (i) lack of visibility into proprietary private sector planning for new model training; (ii) speculative and duplicative requests for new data center capacity from third party vendors that may ultimately go unfulfilled; and (iii) possible future breakthroughs in energy efficiency of training and inference that could reduce energy demand below current projections.

2. While many LLMs are trained at a single data center, some large models are now being trained across geographically distributed data centers. This regional distribution, while largely static, can alleviate spatially concentrated energy loads during AI training if planned appropriately. There is flexibility in siting of training centers because they are not purposed to serve large population centers like other cloud computing and AI inference tools are. Also, flexibility opportunities and reliability requirements at these AI training centers may differ from data centers supporting LLM inference or non-AI applications, with some arguing that their loads may be more like high-performance computing facilities.

3. LLM inference (i.e., creating responses to user requests) is amenable to real-time, geographic distribution of individual queries according to local grid load and renewables penetration, with limited negative impacts for user experience when response latency is not critical. Reliability is essential for these customer-facing functions.

4. Researchers in the private sector, academia, and government are actively exploring a diverse set of hardware and algorithmic improvements to further reduce AI energy consumption. Private sector investment far outweighs other funding and there is limited visibility into private sector progress. Public investment tends to be more forward-looking, aimed at developing the next generation of technology, and increasingly focused on public-private partnerships to accelerate progress.

5. Private industry is concerned about energy procurement timelines within the U.S. and is considering locating outside the U.S. if energy cannot be procured domestically. Siting of large AI training facilities can be more flexible than siting of data centers that need to be located near population centers, but their siting is somewhat constrained by national and regional laws governing data storage.

Recommendations

1. Gain better understanding of power needs through transparent energy use data and bottom-up scenario analysis. To address Finding 1, the Secretary should charge the Industrial Efficiency and Decarbonization Office (IEDO) to benchmark current data center energy use by center type and function. Recognizing the strategic importance of AI to the U.S., the Secretary should ask Congress to request routine collection of data that would allow quarterly tracking of trends in new data center commissioning and (to the extent possible) actual energy use for AI training and for AI inference, to refine models for more accurate projection of future AI energy needs and load shapes. DOE and the national labs should perform scenario analysis of data center power needs that address plausible

scenarios of computation demands, compute efficiency, and algorithmic efficiency, including possible feedbacks where efficiency gains help drive greater computational demands.

2. Characterize training and inference flexibility. To address Findings 2 and 3, the Secretary should expand the Frontiers in Artificial Intelligence for Science, Security, and Technology (FASST) initiative to direct relevant offices across DOE to explore opportunities for temporal and spatial flexibility in AI training and inference, and to demonstrate and publicize these capabilities in collaboration with the national labs and private partners.
3. Develop public-private efforts to advance computational technologies. The Secretary should request additional funding from Congress to support DOE labs, Office of Science, and National Nuclear Security Administration efforts to partner with industry to advance compute efficiency and algorithmic efficiency for AI, advancing the nation's AI capabilities and building on the success of comparable public-private efforts that have accelerated advances in high-performance computing.
4. Establish AI testbed in DOE. The Secretary should establish a data-center-scale AI testbed in DOE, which should be complementary to, but distinct from, the current set of high-performance computing facilities operated by DOE. This testbed can allow researchers from the national labs, academia, and industry to collaborate in development and assessment of algorithms for energy-efficient and/or energy-flexible AI training and inference.
5. Improve training and inference methodologies. To activate private sector and academic researchers in the context of Finding 4, the Secretary should task DOE with developing a benchmark LLM and creating a funded prize challenge for open-source, energy-efficient training and inference of LLMs and other large AI models.

Track 2: Examine secure operational frameworks that allow data centers to optimize their energy consumption and contribute to grid peak load and critical stress management.

Led by Maria Pope, CEO of Portland General Electric, track 2 focused on the cross-industry collaboration needed to transition data centers from passive purchasers of power to active participants in the grid. Specifically, it explored approaches to operationalize the flexibility opportunities identified in Track 1 and models for grid utilization of data center backup power.

Findings

1. Hyperscalers and technology providers state that temporal and spatial computational flexibility is possible if they are given appropriate signals. Despite this perception of technical capability, we identified no examples of grid-aware flexible operation at data centers today other than the carbon-minimizing geographic optimization that Google has employed for several years, recent efforts to respond to energy shortages in the European Union resulting from the Russian-Ukraine war, and flexibility requirements in Ireland. This lack of flexible operations in the U.S. may result from the fact that electricity providers only recently started having to say no to data center interconnection requests.
2. There is no standard terminology in the U.S. for flexible operation of any type of assets, including data centers, which is a significant impediment to rapid the scale-up of flexibility programs even when multiple parties want to cooperate.
3. In transmission-constrained locations, electricity providers often can accommodate the energy and capacity requests of a data center for (say) 350 days but need to find a win-win solution for the remaining 15 days. Data center backup generation and storage could provide a solution, but with typical permits only allowing emergency operation of diesel backup generation, this would likely require use of alternative fuels or installation of advanced backup technologies. Data centers are experimenting with or considering natural gas, renewable natural gas, batteries, clean hydrogen, and other technologies to address this challenge.

Recommendations

1. Convene key stakeholders to accelerate data center interconnection. The Secretary should convene energy utilities, data center developers and operators, and other key stakeholders to start active dialog on how to get through the current energy supply bottlenecks, to advance understanding of real-time data sharing and protocols to govern flexibility, and to develop strategies for how to provide the clean power needed to sustain AI leadership into the future.
2. Develop a flexibility taxonomy and framework that explores the financial incentives and policy changes needed to drive flexible operation. To address Finding 2, the Secretary should work with other government agencies and the private sector to develop a standard taxonomy and framework for defining and orchestrating flexibility services that is adaptable to regional circumstances and priorities. The working group developed a starting point for a taxonomy (Appendix A). Standard interconnection requirements could be a practical way to operationalize this idea. Building support for the idea requires: 1) analysis to demonstrate the benefits and costs of flexibility, 2) the flexibility taxonomy, 3) policy and contractual advances to enable and support flexibility, and 4) model tariffs for data centers and other large loads that incentivize both efficiency and flexibility/demand response capabilities. Due to accelerating investments in data centers, the sooner the better for standard requirements.
3. Provide technical, business model and permitting support for novel backup power strategies. To address Finding 3, the Secretary should provide data center support via DOE staff and the national labs to help the centers, utilities, and state-and-local governments to explore and demonstrate innovative, flexibility solutions. Drop-in fuels provide a large opportunity for existing data centers

and additional options are possible for new construction. DOE's work on virtual power plants and microgrids could provide an important starting point.

4. Expand technical support of state energy planning departments to include analysis of infrastructure investments for data centers and other large, strategic loads, including assessments of their potential impacts on people and communities. The Secretary should direct DOE and the national labs to develop new planning tools for projecting future infrastructure needs to support data centers and other large loads (recognizing both potential load growth and flexibility) and to expand technical support to state energy planning departments. Key to this effort is an evaluation of the potential impacts of this rapid buildout of infrastructure on people and communities. In addition, the Secretary should convene the Power Marketing Administrations to develop strategies, where appropriate, to support these large strategic loads.
5. Better utilize the existing grid. For regions limited primarily by transmission (rather than generation), grid-enhancing technologies (GETs) may provide a bridge to the future. DOE's efforts to advance GETs provide a starting point, but studies examining the specific needs of data centers would be valuable. It is recommended that DOE explore model tariffs that have data centers share or pay in full for required grid upgrades, to reduce the cost impact on other ratepayers.
6. Revisit Federal Power Act 202(c) authorities to leverage data center backup capacity. As part of an on-going reassessment of FPA 202(c) authorities, the Secretary should consider the potential strategies of allowing data center and other large sources of backup generation to provide grid services to meet public needs in emergency situations and provide technical support to guide state's efforts to enable broader use of backup generation in emergency situations.
7. Promote facility level solutions. To address Finding 2, DOE should increase awareness and encourage adoption of leading-edge building efficiency technologies and energy management best practices by increasing private sector participation in on-going programs such as DOE's Better Buildings Initiative and the Center for Expertise for Energy Efficiency at LBNL. As an example, installation of advanced cooling technologies demonstrated by these programs could reduce load significantly to help the system ride through heat waves. To encourage experimentation, DOE should create a prize to explore innovative technologies and methods for reduction of power, water, utilization of waste heat, and facility level electricity supply. In addition, the Secretary should support legislation to develop a specific demonstration program for innovative flexible solutions for technologies and methods deployed at the data center facility level to maximize data center flexibilities, avoid infrastructure requirements to serve peak demand, and ease grid integration.

Track 3: Explore generation, storage, and transmission technologies to power data centers

Led by Shirley Ann Jackson, former President of RPI and former chair of the Nuclear Regulatory Commission, track 3 focused on the generation and storage technologies needed to power data center growth while maintaining and enhancing U.S. leadership in AI. Specifically, it explored approaches to operationalize the flexibility opportunities identified in Track 1 and models for grid utilization of data center backup power.

Findings

1. With data centers largely relying on diesel generation for backup, there is little experience with the cost and performance of cleaner backup technologies that provide similar site reliability.
2. A broad concern about resource adequacy and reliability of today's grid was expressed by electricity providers, data center customers, and other large customers that were interviewed. This concern existed prior to consideration of increases in data center power demand. Almost uniformly, they recommended accelerating generation and storage additions, delaying retirements, making additional investments in existing resources (e.g., uprating and relicensing of existing nuclear and hydroelectric facilities), and demonstrating new clean, firm, affordable, dispatchable technologies as soon as possible. In most cases, they see new natural gas capacity additions – in addition to solar, wind, and batteries -- as the primary option available today to maintain reliability.
3. There is limited knowledge about the cost and performance of emerging technologies -- such as batteries, renewable natural gas, long-duration energy storage, small modular reactors, enhanced geothermal – nor reliable estimates of how long it might take for them to deploy at scale. These technologies will be critical to electricity companies to maintain reliability as they shutter coal plants and to data centers for on- or near-site power with minimal transmission build.
4. There is uncertainty regarding the broader grid cost, adequacy and reliability impacts of supplying large data centers demands. We need to better understand the costs and benefits of both behind-the-meter and grid-supplied alternatives that are emerging. While most operating data centers are grid connected, lengthy lead times to construct new high voltage transmission lines has increased interest in co-location for larger data centers seeking connection today.
5. Growing data center load exacerbates on-going supply chain concerns for electrical equipment (e.g., transformers, switching equipment, generation equipment, advanced transmission technologies) both in the near- and longer-term.
6. The power needs of future data centers are unclear both in terms of magnitude and temporal shape. For a large, flat load, characteristic of many data centers today, technologies such as nuclear or gas with CCS may be preferred. If data center computational activities increasingly have flexible or fluctuating requirements, other generation and storage technologies may be preferred.
7. The historical time that it takes to move from technology demonstration to widespread deployment of new generation technologies is measured in decades. Approaches are needed to de-risk new technologies more broadly to encourage increased testing and, if successful, adoption.

Recommendations

1. <u>Perform critical evaluation of generation and storage technologies commercially available for on-site backup power today</u>. To address Findings 1 and 5, the Secretary should direct DOE and the national labs to conduct a rapid assessment of cost, performance, reliability, availability, space requirements, emissions, and supply chain issues for current technologies, including renewable diesel, natural gas, renewable natural gas, fuel cells, battery storage, enhanced geothermal, long-duration energy storage, and other potentially viable technologies available to support regional data center expansion. This assessment should explore operational modes in which backup power provides grid services.

2. Evaluate how best to increase flexible, firm generation on the grid while minimizing emissions. To address Finding 2, the Secretary should direct DOE and the national labs to conduct a rapid assessment of cost, performance, reliability, availability, and supply chain issues facing current and emerging generation and storage technologies that could address data center demand in the near term. This assessment should include technology strategies and an assessment of potential for expanding use of existing, low-carbon generation, (e.g., through operational improvements, uprates and relicensing of nuclear and hydroelectric facilities), assessments of storage technologies, evaluations of demand-side efficiency potential, and evaluation of flexibility improvements in data centers and other electric end-uses. The strategy can also consider approaches for addressing the carbon footprint of new natural gas capacity additions, broader use of existing gas generation and delayed retirements of coal consistent with existing and proposed regulatory approaches. Furthermore, the Secretary should direct DOE and the national labs to assess and deploy opportunities for reconductoring and integration of other grid enhancing technologies to increase their power carrying capacity of existing transmission rights-of-way.
3. Develop technology strategies to limit the carbon footprint of new gas generation. The Secretary should commission an assessment of technological approaches to limit the lifetime carbon emissions of natural gas capacity additions, e.g., through use of renewable natural gas, hydrogen-ready, CCS-ready, and other means. The Secretary should develop a strategy for DOE to provide assistance to state regulators, state energy offices, and grid operators in assessing and planning for the role of new natural gas facilities, to help ensure that new gas capacity can transition to lower utilization over time, shift to low-carbon fuels, or install control technologies consistent with regulations. Managed well, new gas additions are consistent with a least-cost, net-zero emissions future. Most published net-zero scenarios conducted with models that consider both energy and capacity project new natural gas additions on the least-cost path to net-zero emissions by 2050. The new gas additions provide both energy and capacity value in the near-term and primarily capacity value in the 2040s when they are seldom used if their emissions are uncontrolled.
4. Assess lead times for emerging technologies to reach scale. To address Finding 3, the Secretary should direct DOE to perform realistic assessments of technology timelines, focusing both on general grid applications and examining any issues specific to onsite application at data centers. This assessment should build upon the excellent "Liftoff" living documents that DOE has developed and include supply chain development necessary to support new energy systems (e.g., clean hydrogen production, transport, storage, and use, or needed infrastructure for CCS or advanced nuclear at scale).
5. Improve regional and national projections of load and load flexibility from data centers. Wise investment hinges on robust plans that anticipate where and when power demand will occur, but also recognizes the potential impacts of efficiency gains. Recent DOE assessments for national transmission planning and EIA assessments of load growth provide a starting point but need to be updated to reflect current realities and future likelihoods and the opportunities for both temporal and spatial flexibility in compute needs, including recognition of the market and regulatory drivers for where these facilities may be located. The Secretary should ask Congress to provide new authority for DOE or EIA to collect and maintain a confidential database of prospective large electric demand requests to improve efficient power system planning and address speculative and possible double counting. This would help the Department provide more public information about where load is likely to emerge, when it will emerge, and how flexible it will be. With rapid change in the electric industry, regional plans that integrate generation, transmission, distribution, and load flexibility are increasingly valuable.
6. De-risk investment for first movers by advocating approaches to deal with non-firm pricing for emerging technologies (e.g., through innovation hubs). The Secretary should support the development of strategies, and if needed, legislation that de-risks private investment in new technologies. Efforts to

advance CCS, LDES, and advanced nuclear over the past few decades have had limited success. Rather than cost-sharing a few demo projects, DOE could develop strategies to address the challenge that investors cannot get a firm cost contract for key, emerging technologies. One shared-risk approach that builds upon the strong response to the hydrogen hub concept would be to develop comparable hubs for data center flexibility, small modular reactors, long-duration energy storages, and other key technologies. In addition, the Secretary can direct DOE to provide technical support to data center owners interested in making long-term financial commitments to next-of-a-kind technologies in nuclear, enhanced geothermal, fuel cells, long-duration energy storage, and CCS that are aligned with DOE liftoff reports. Additionally, the Secretary can direct the national labs to verify and communicate the cost and performance of emerging technologies, such as enhanced geothermal, to help support investment decisions by other entities.

Appendix A – Flexibility Taxonomy

Taxonomy for Data Center Power Flexibility

The electric grid is undergoing great change with renewable generation sources replacing traditional dispatchable carbon emitting resources and now, new large loads associated with data centers. Grid flexibility is more important than ever to balance the needs of the system. Grid flexibility refers to the ability of the grid to respond dynamically to variability in electricity supply and demand. The flexibility needs will vary regionally based on generation resource mix, storage, and active load profiles. The needs generally increase with rising penetration of variable renewables and differ significantly for solar- and wind-dominant regions. However, the magnitude of new large loads across the nation far exceeds the typical load growths of the past 2-3 decades. Understanding the characteristics of the load combined with the regional dynamics will determine methods available to orchestrate the flexibility of these new large loads.

Data centers can potentially provide energy flexibility either through varying the processing of their transactions based upon timing and/or location. The following is proposed as a common taxonomy to characterize flexibility needs of the grid and to explore opportunities for data centers to participate.

General Data Center Power Supply Characteristics

- **Location on grid:** point at which data center is connected (e.g., generation bus or physical location).
- **Demand magnitude/shape:** MW maximum demand and load shape characteristics.
- **On-site supply:** local generation (type, capacity, fuel constraints) and storage assets (type, capacity, duration).
- **Multiple locations and the data center ecosystem:** location of data centers to provide operational flexibility for large learning models. Different designs can provide different usage patterns. So the ability to move and time transactions provides flexibility.

Factors that Influence When Data Center Flexibility Is Dispatched

- **Leadtime:** driven by data center characteristics and by grid operation.
  - **Response speed:** how quickly the data center can respond to a signal.
  - **Notification time:** time lapse from notification of need to delivery of services, e.g., hour-ahead, day-ahead, or other to correspond with times steps for grid management.
- **System level price:** flexibility triggered by the price to which the load is exposed.
- **Contractual:** flexibility triggered by a non-price measure such as reserve margin, excess generation, or other dispatch construct.
- **Data Center ownership & operations:** many data centers are owned and operated by 3rd parties with service level agreements with companies processing AI. Management of the data processing is through the AI processing company and not the owner operator of the facility.

Flexibility Characteristics – Both for On-site and Geographic Load Management

- **Available flex capacity:** available MW considering load shape and other potential dependencies (e.g., if response depends on price, contract, time of day, etc.).
- **Response frequency:** maximum number of calls or frequency of calls for flexibility.
- **Flex response duration:** maximum consecutive minutes or hours flexibility can be provided.
- **Ramp rate up-and-down:** maximum rate at which response can be delivered.
- **Additional on-site generation permitting limits:** e.g., limits on hours/seasons of operation.
- **Ability to provide ancillary services**: e.g., frequency regulation.
- **Ability to shift load**: ability to adjust time of load vs. curtailing demand completely.
- **Performance under extremes**: response limits due to extreme weather or data processing load.
- **Uncertainty in response magnitude/speed:** any conditional factors limiting response.

These characteristics have to be known by the energy provider and operational coordination is critical between the energy provider and the AI processing companies. Similar to other industrial demand response programs of the past, such as the aluminum industry, these operating parameters must be understood at the plant and by the grid operators.