Introduction

The United States is embarking on a multi-trillion dollar energy transition to replace existing fossil generation assets with utility scale wind, solar, and batteries to electrify our economy in a more sustainable manner. 36 states and the District of Columbia have established renewable portfolio standards (RPS) to measure performance against achieving clean energy production targets.[1] However, these standards were established in an era of relatively flat load growth, and since then there have been three landmark movements that have caused an increase in forecasted electrical demand, the likes we have not seen since the [industrial revolution / WWII / 1970?].

The first is in the face of increasing geopolitical turmoil, the United States has passed ambitious legislative acts (CHIPS and Science Act; the Inflation Reduction Act) to bring more manufacturing onshore to bolster the country's domestic supply chain. Since 2021, there have been approximately $465 billion worth of semiconductor, EV and battery factory projects announced, projected to consume significant amounts of power (estimated to be [ ]).[2]

The second is a continued and rapid increase in data center network buildout. A strong digital backbone has been the foundation for the United States to remain a global leader in innovation, supported by the construction of 17 GWs of data center capacity over the last [20] years. Building capabilities in artificial intelligence is the next extension of this innovation, but that requires more even compute intensity (and thus power). As a result, electricity consumption at US data centers alone is poised to triple from 2022 levels, to as much as 390 terawatt hours, or 40 GWs, by the end of the decade. This could equal up to 7.5% of the nation's projected electricity demand.[3]

The third is the electric vehicle transition for both residential and commercial vehicles. [(Put in estimated load addition for next 10 years)]

The core question now is not just whether or not we can decarbonize, but can we even power the confluence of these massive industrial projects at the scale envisioned? Private and public sector leaders are starting to publicly question the ability of existing grid infrastructure and current development practices to meet this demand, and the day-to-day execution on these large scale projects is falling downstream to state and local constituents to find solutions, most of whom are not fully equipped to do so. Forgoing or delaying these project development opportunities will have significant national security and economic development considerations.

There are several known options that stakeholders can consider in meeting these power demand requirements, each with its own set of challenges from a sustainability, affordability, and reliability perspective: (1) extending the retirement of thermal generation plants (e.g., coal), (2) building additional new natural gas plants, and (3) building large scale transmission to tie in new utility scale renewable generation to the grid.

However, a fourth option exists that is currently more seldomly evaluated but which can be an important solution to utilities. Data center growth is the largest driver of industrial demand growth, and historically has been built with under-utilized energy infrastructure (i.e., diesel generators) and drawing continuous power from the grid. With other energy infrastructure (e.g., battery energy storage) becoming increasingly cost-effective and reliable as a backup power source, and the ability to have flexibility in the dispatchability of certain AI workloads (e.g., training), data centers can provide large-scale flexibility to the grid that can help utilities manage their peak-demand challenges. This approach has several benefits to utilities, such as improving resource adequacy without having to build existing transmission capacity or generation, managing peak demand in a more sustainable and affordable manner, and supporting offtake for curtailed and/or additional renewable energy generation resources.

---

[1] EIA; Renewable Energy Explained
[2] Bloomberg; AI Needs So Much Power That Old Coal Plants Are Sticking Around
[3] Boston Consulting Group; The Impact of GenAI on Electricity

The purpose of this initial white paper is to raise awareness among the utility sector on the potential that a collaborative approach with data center developers and the flexibility of data centers can bring to its system planning. As we build the next several $100 billions of data center infrastructure domestically, it can and should be done through the lens of supporting energy infrastructure both locally and more broadly. Starting with what is possible today will enable discussion and planning at the project execution level.

This paper is supported by the [Coalition for Data Center Flexibility], representing a group of utility and digital infrastructure leaders, with the ambition to implement these practices at scale to help meet the electrical demand important for United States national security, innovation, and economic development in a way that also benefits utilities.

Status Quo Practices Have Overwhelmed the Grid

A data center is a specialized warehouse that hosts IT infrastructure (i.e., servers), providing primarily two services: cooling, in order to keep servers from overheating, and electricity, to provide continuous power to the servers. Generally, the IT hardware is either owned by an enterprise that has an abundance of compute requirements and the sophistication to manage its own infrastructure, such as Apple or Meta, or a cloud services provider that manages the IT infrastructure on behalf of other enterprises. These organizations, such as Amazon, Microsoft, Google, or Oracle, are often referred to as “hyperscalers.”

In an increasingly digital world, enterprises rely and value uptime in order to continuously run their operations, which is why data centers have been located and built the way they have historically. Locationally, facilities have been built in areas with strong grid stability and low natural disaster risk. Additionally due to the benefits of clustering data centers around each other from a technical perspective (e.g., latency and data gravity), a significant amount of capacity has aggregated around several core markets. From a build perspective, in the event there is a grid outage, data centers have been built with diesel generators and other mechanical equipment to run the data centers during grid outages and power the servers so that no loss of operations is experienced by the end user. However, since data centers are generally built in reliable locations, these diesel generators are almost never used.

Historically, it has been both preferred and acceptable to build data centers in geographic clusters given the scale of facilities were in the tens of MWs each, and relative to the local utility’s generation and transmission system, relatively small. Now, forecasted data center projects are growing increasingly large in size, some as large as hundreds of or even more than a thousand MWs for a planned development, representing the same amount of power to energize hundreds of thousands of homes in a footprint as small as hundreds of acres. The result of this explosive demand growth is significant power constraints in core markets, often coupled with an abundance of underutilized diesel generator energy infrastructure. Areas such as Northern Virginia and Phoenix have had to put temporary pauses in new data center construction in order to more proactively plan and manage go-forward energy demand, with gigawatts of backup generators sitting idle.

Continuing to build with status quo practices will exacerbate the problem and lead to billions of dollars of stranded capital expenditures in diesel generators that are seldomly used.

New Technology Enables a Different Approach to Data Centers

It is highly unlikely that the demand for several $100 billions of new data center infrastructure will decrease, and as a result there must be a more proactive approach to understanding how the next decade can support the corresponding energy infrastructure on the backdrop of the energy transition. Fortunately, advances in technology

can enable a more bi-directional interactivity with the project development and the utility that can allow a data centers to be an asset to the grid. Broadly, this falls into two categories.

First is the actual data center facility infrastructure itself. Simplistically, instead of building diesel generators for backup, other technologies can provide backup to the IT infrastructure and run in the event of grid outages. Today, 4-6 hour lithium battery energy storage can provide sufficient backup to the IT infrastructure in most locations (given historical outage data) and can also provide grid services such as demand response in peak demand periods. The cost declines of this technology have improved by [ ] over the last several years, and are comparable on a $/W basis to diesel generators. Utility scale batteries have proven effective in several markets in alleviating peak demand constraints, such as CAISO and PJM[4], and will continue to play an important role as many grid mixes shift towards more intermittent generation sources. Furthermore, new technologies, such as long-duration energy storage and nuclear micro reactors, are on the cusp of commercialization and may ultimately be able to serve longer-duration backup and grid support use cases.

Second is the underlying IT infrastructure. Artificial intelligence is broken up broadly into two types. The first is the “training” phase. In this, a model is given a vast amount of information and patterns to learn. During training, the model then analyzes this data and adjusts its internal structure to make accurate predictions. Once fully trained, the model enters the "inference" phase, where it uses its learned knowledge to analyze new, unseen data and make predictions based on what it has learned. A user in this instance would be able to prompt the AI and receive a response during this inference phase.

Training these models can take vast amounts of power. [GPT-3 took up [ ]power]. However at the same time, this can be done in a much more flexible and time insensitive manner. In this instance, the owner of the model could theoretically schedule the provisioning of power to run the training model in accordance with certain agreed upon times from the utility in order to have it to be treated as a large scale, flexible asset to the grid. In the short-term, this data center load flexibility and battery storage can be directly dispatched by grid operators.

This Approach Can Benefit the Utility

Both the flexibility of the energy and of the IT infrastructure of a data center development are assets that can be considered as part of a utility’s Integrated Resource Plan (“IRP”). An IRP will take a longer-dated, 10+ year view on forecasted demand growth and generation mix, utilizing a bottoms-up build of potential in-development assets to develop a model to meet economic, reliability, and sustainability metrics. Shorter-duration 4-6 hour lithium ion batteries are valuable assets today that are often included in these IRPs, but in some markets face hurdles in bringing to operations given a myriad of development and economic considerations. Further, as renewable penetration increases and the effectiveness of short duration BESS declines, other technologies that are not yet proven at scale are required in order to meet sustainability and resource requirements.

A phased data center development can be a pathway to incorporating these assets at scale with synergies in the development, by virtue of having an existing demand source (the data center) versus having to implement new market designs in order to achieve development of these assets on a standalone basis. As shown in Exhibit [ ], a well-structured development plan can incorporate a myriad of assets “behind the meter” that can meet data center reliability requirements as well as provide value through a variety of grid services. Given the scale of these projects, the benefit can be substantial.

Exhibit [ ] walks through an example in more detail.

---

[4] WSJ; Giant Batteries Helped the U.S. Power Grid Eke Through Summer

This Requires a New Approach to Interconnection

Historically, the data center interconnection process has been a “one-way” process whereby developers request access to power and the utility works on power studies and then comes back with a feasible power solution. This was acceptable in an era with abundant power access and when data center capacity was not at the scale it is today. However, in the face of the current supply-demand imbalance, it is no longer feasible.

Moving forward, a more collaborative approach can serve as a “win-win” for both data center developments and the utilities, by enabling access to power for the project in exchange for that project providing valuable grid services. Said differently, not all megawatts are created equal, and providing the data center with electricity during periods of abundance and having the data center provide power during critical periods that could support the utility system planning would be valuable to both parties.

[Show ven diagram]

What’s Next

[TBD - let’s see what feedback we get on the consortium broadly]

Glossary

- 3PDC
- BESS
- Behind the meter
- Data gravity
- Dispatchability
- DR
- Hyperscaler
- Interconnection [Queue, Process]
- IOU
- IRP
- ISO/RTO
- Latency
- PUE
- RPS
- SLA