

Addressing Constrained Grids with Onsite Generation

WHITE PAPER

by Rich Scroggins

Recent industry reports highlight the challenge that electric grids in the United States (US) are facing due to the rapid growth of data centers, accelerated by artificial intelligence (AI) workloads. Data centers are being developed at a pace that outstrips the utilities' ability to expand supporting infrastructure. One of the ways utilities are addressing this problem is to offer, or mandate, that operators agree to participate in grid support programs as a condition of their connection agreement. "Bring your own power" is becoming a common phrase to describe this. This paper reviews the latest research on US grid capacities and summarizes the challenges an engineer will face when designing onsite power generation systems to support a constrained utility.

SCOPE

Recent findings from an Electric Power Research Institute (EPRI) survey provided context for the scope of the problem. In the survey, 5 of 22 respondents indicate they have aggregated connection requests from data centers that exceed their current peak load (see Figure 1).

To be sure, not all these requests are going to materialize, and data centers will be years away from operating at full capacity, if they ever do. But,this does create uncertainty for utilities as they must consider the possibility of their capacity being oversubscribed.

In response to these emerging challenges, several studies have explored the potential of demand response programs, flexible connection agreements, and on-site generation strategies to help manage grid constraints. One recent study, Rethinking Load Growth from Duke University's Nicholas Institute, provides a detailed illustration of how small-scale curtailment could theoretically enable balancing authorities to accommodate significant new demand without major grid reinforcements. Figure 2 illustrates this concept using PJM during a representative winter week as an example.

The study models the integration of a constant additional load (shown in green in Figure 2) superimposed on the existing system load (blue). At certain times during this period, total system demand exceeds historical peak levels, marked by the red dashed line representing PJM's maximum recorded winter peak load. The portions where the added load surpasses this threshold (shaded in pink) illustrate the curtailed load necessary to manage peak demand during critical periods.

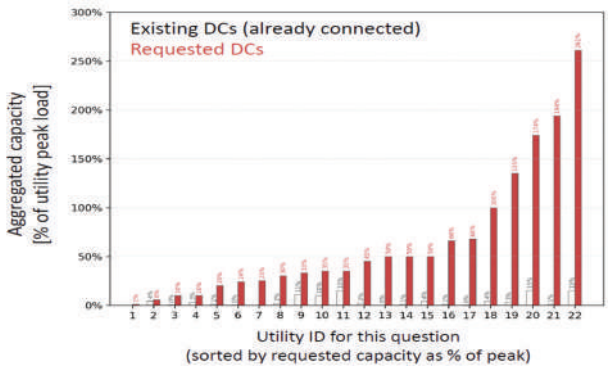


Figure 1. Data center utility connection request as a percentage of utility peak load. Source: EPRI – Utility Experiences and Trends Regarding Data Centers – 2024 Survey.

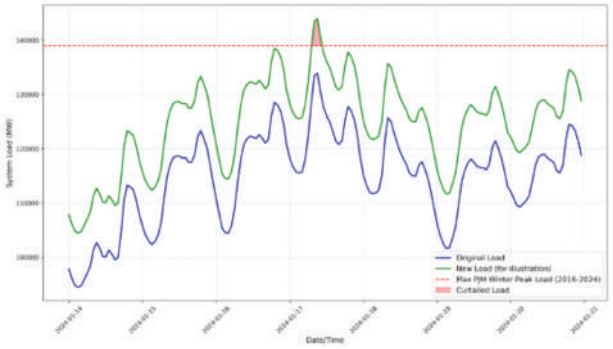


Figure 2. Illustrative Load Flexibility at PJM – From "Rethinking Load growth" by Nicholas Institute for Energy, Environment and Sustainability at Duke University.

Building on this concept, Figure 3 (corresponding to Figure 8 in the Duke study) quantifies the additional load – termed headroom – that could be supported by different balancing authorities if the load could be curtailed at a rate of 0.5%. Curtailment rate is defined as the amount of load in MWhrs that would need to be taken off the grid as a percentage of the total accumulated load over the year. Referring to Figure 2, this is the area shaded in pink divided by the total area under the green line.

Figure 3 shows that PJM (a grid operator who's territory includes the data center hub of Northern Virginia) could add 17.8 GW of new load if that load could be curtailed at a rate of 0.5%. The study indicates that based on existing load profiles a 0.5% curtailment could be achieved by taking some portion of the new load offline for 177 hours per year and for PJM the average duration of a curtailment event would be 3 hours.[1] Similar opportunities are identified across multiple grid operators, although potential headroom varies due to differences in system size, load profiles and existing grid flexibility.

This shows that utilities potentially have a problem with peak loads, not average loads. Utilities could add a substantial amount of new load if that load could be curtailed during times of peak demand, a small fraction of operating hours over the year. There are many variables at play, every utility will have different opportunities for curtailment enabled headroom. However, the study revealed opportunities for many utilities to add tens of gigawatts of load – if it could be curtailed for no more than a few hundred hours per year, for periods of 2-3 hours per event.

Utilities recognize that data centers have on site power available and are viewing them as potential assets to the grid in addressing capacity challenges. In some cases, we're seeing utilities mandate participation in grid support programs as a condition of their connection agreement or delaying interconnection unless their customers are willing to participate.

The level of curtailment required can reasonably be met with diesel or natural gas generator sets, as well as battery systems that are currently available. However, practical realization of such strategies depends not only on technical feasibility, but also on the operational, regulatory, and commercial willingness of large utility customers to participate.

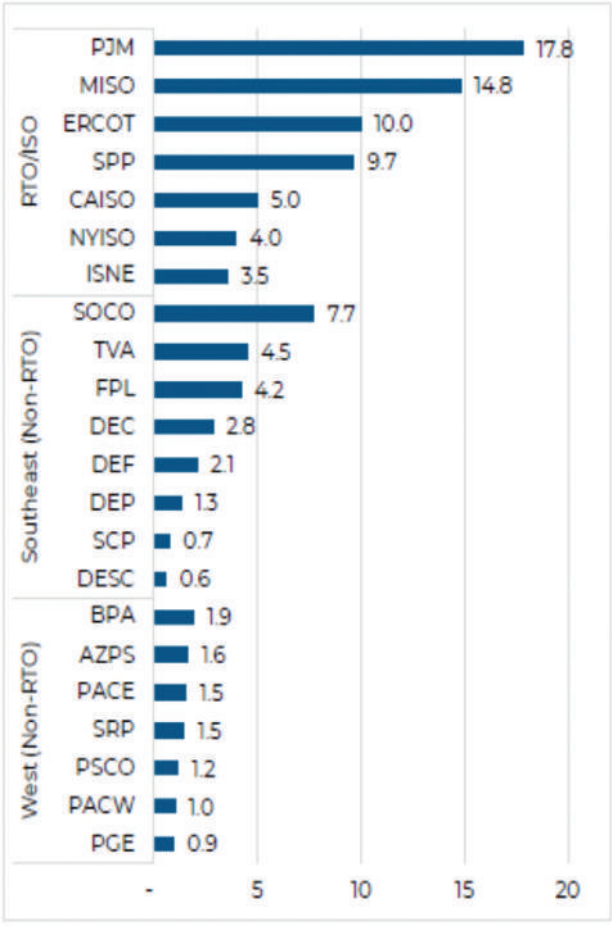


Figure 3. Curtailment Enabled Headroom.

[1] The Duke University study analysis is based on historical load profiles from 2016 through 2024 and does not account for transmission constraints or reserve margins, which could influence the actual headroom available on the grid

Onsite generation opportunities

DIESEL GENERATOR SETS

When diesel generator sets are used for load curtailment, the most challenging aspect is emissions regulations. In the US there are two broad areas for emissions regulations: EPA non-emergency operation and operating hours limitations based on Major Source thresholds.

NON-EMERGENCY OPERATION

In the US, the Environmental Protection Agency (EPA) defines emergency operation of a diesel generator set as running only during a utility outage. Any operation for load curtailment is considered non-emergency operation, which requires a Tier 4 certified engine. Tier 4 certification requires the manufacturer to certify the engine and any emissions controls equipment as a unit with the EPA. It is not sufficient to install a third-party emissions aftertreatment system, as is often done to reduce NOx emissions for emergency generator sets. Tier 4 certification also requires the generator set to shut down if the emissions performance is out of compliance. This makes the emissions aftertreatment system a single point of failure, which has been unacceptable to data centers. However, as participating in demand response programs becomes mandatory, there is a renewed interest in Tier 4 certified generator sets.

Recently, the EPA issued an interpretation of their regulation governing onsite generation (RICE NESHAP) allowing emergency generators to be used for up to 50 hours per year if they are dispatched by a local authority to "avert or reduce the risk of local power supply interruptions...". This will simplify use of diesel generator sets for load curtailment programs, limited to 50 hours per year.[2]

MAJOR SOURCE THRESHOLDS

Under Title V of the Clean Air Act the EPA has defined annual, cumulative thresholds for various emissions constituents above which an installation would be considered a major source of emissions, subject to fees and monitoring requirements. The Major Source Threshold applies to cumulative emissions at site level, not per individual generator set.

The EPA has also identified counties, known as non-attainment areas, in which concentrations of regulated pollutants exceed levels set by the National Ambient Air Quality Standards (NAAQS). Thresholds vary depending on the level of non-attainment. The default major source threshold for NOx is 100 tons per year. For counties designated to be in Serious non-attainment, the threshold is 50 tons per year. As a point of reference, most counties surrounding the Dallas and Los Angeles markets are considered serious non-attainment areas. For severe non-attainment, the threshold is 25 tons per year.

Local air quality boards issue permits that limit the number of operating hours for generator sets to keep the cumulative NOx emissions below the major source threshold. For large sites, particularly those located in non-attainment areas, permitted operating hours may be limited to a level that is too low to participate in a demand response program.

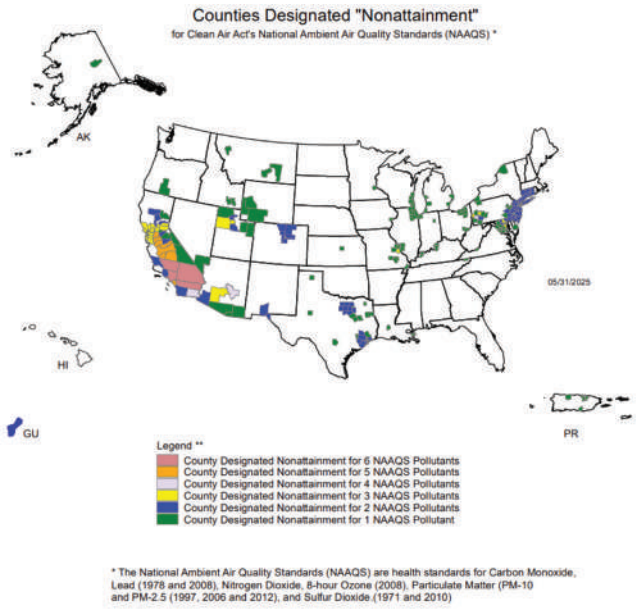


Figure 4. EPA non-Attainment Areas

[2] EPA's rule interpretation can be found here: Recent EPA 50 hour demand response interpretation.

Natural gas generator sets

Natural gas generator sets offer advantages over diesel generators for demand response operation. In addition to lower NOx emissions, they also have the advantage of lower fuel costs, so at higher operating hours, the Total Cost of Ownership is reduced. However, there are some challenges associated with natural gas.

PIPELINE AVAILABILITY

Natural gas pipeline capacity is a potential challenge for data centers wishing to use natural gas. Pipelines of sufficient capacity to supply data centers in the hundreds of MW to GW range aren't available everywhere, particularly in rural areas where large data centers are being built. Costs of extending gas pipelines can exceed $2M/mile. There is also a challenge in seismic areas where shutoff valves are required.

In situations where natural gas is inaccessible, but the project's duration justifies its use, a virtual pipeline system can be deployed. This system uses a modular approach with Compressed Natural Gas (CNG) or Liquefied Natural Gas (LNG), transported via highway, railway, and waterways. Virtual pipelines bridge the gap in areas lacking direct natural gas infrastructure, enabling efficient fuel delivery to remote sites well before a conventional pipeline is built, addressing immediate energy needs while awaiting permanent solutions.

NATURAL GAS PRICING

Natural gas is less expensive than diesel, however there are some pricing risk factors that need to be considered when assessing the long-term financial performance of a project. While natural gas prices have remained relatively stable in recent years, historically there has been some volatility in natural gas prices which should be considered. Also, utilities typically have performance requirements for customers participating in grid support programs, driving customers to enter firm natural gas contracts - often coming at a premium over standard contracts.



TRANSIENT PERFORMANCE

Natural gas generators have historically been used primarily for base load operation, rather than standby. Because of this, manufacturers have made design decisions to optimize efficiency and fuel consumption rather than start up time and load acceptance time. In recent years there has been more interest in natural gas in standby applications, for data centers, and in commercial and industrial markets. To serve those markets, manufacturers have developed fast start natural gas generator sets. These generators are capable of meeting the NFPA 110 ready to load requirement of being up to rated frequency and voltage within 10 seconds of starting. They are also able to accept full load in a single step, as opposed to a conventional lean burn natural gas generator set that takes a minute or more to accept load.

Even fast start gas generator sets that can meet the 10 second start time will still take longer than diesel generator sets to accept a block load and recover to rated speed and voltage within an acceptable tolerance. This is where it becomes important to understand what loads can tolerate in terms of how long they can be without power.

DATA CENTER LOADS

At a high level you can divide data center loads into IT loads and mechanical loads, which can include cooling and general house loads. In some cases, there are separate generator sets for different loads, while in others, gensets will serve both IT and mechanical loads.

IT LOADS

IT loads are supported by batteries, either as part of a conventional UPS or mounted in the server rack. UPS batteries typically have minutes of autonomy, so it is feasible for UPS to support IT loads for a longer period if it takes a natural gas generator set longer to accept load than a diesel genset. This can be done by either extending or delaying the load ramp. There are practical limitations to this however, as drawing more energy from batteries could shorten battery life. Also, colocation providers may have stipulations in their Service Level Agreements (SLA) which state how quickly generator sets must accept load and return voltage and frequency within tolerance, as well as how long it is acceptable to support the load with batteries.

MECHANICAL LOADS

Mechanical loads are typically not supported by UPS so it is important to know how long these loads can be without power before there are consequences to the operation. With air cooled server racks the question is how long can the servers operate without cooling air before temperatures begin to reach their thermal limits?

The other concern regarding cooling loads is that they have controls that will go through a software reset if they are without power for long periods of time. They may have supercaps that can maintain power for about 30 seconds but if that time is exceeded it will be several more seconds before power to the cooling equipment is restored and thermal runaway becomes a risk.

FOOTPRINT CHALLENGES

The other technical challenge is the power density of gas generator sets, in terms of kW/sq ft, is less than that of diesel. As footprint is always a challenge, designers will have to think creatively about how to deploy natural gas generator sets if they want to maintain a common site layout as they have for their sites where they are using diesel backup. Paralleling smaller generator sets allows a data center to install the same kW of capacity of natural gas generation in a similar footprint as they would use for diesel. Smaller generator sets have some logistical advantages over larger generator sets , as they are easier to transport and may have shorter lead times. Figure 5 highlights an example of a 4 MW module consisting of two paralleled 2 MW generator sets stepped up to medium voltage, packaged in a single enclosure of a similar width to a 3.5 MW diesel enclosure. The packaged generator set enclosure is longer than a typical 3.5 MW diesel enclosure, but the widths are very similar allowing the gas generator sets to be deployed without disrupting the data center's electrical yard design.

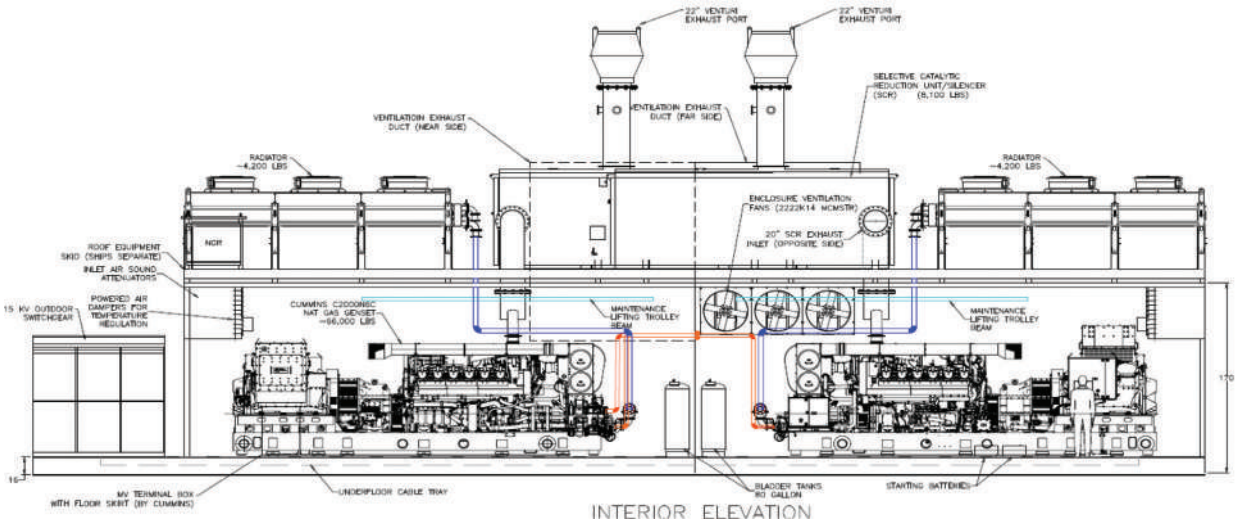


Figure 5. Two 2MW Natural Gas generator sets package in an 80' X 14' enclosure. A typical 3.5 MW diesel enclosure will be on the order of 54 ft by 15 ft.

PARALLELING ARCHITECTURES

As onsite generators become adopted for flexibility rather than standby operation, there are advantages to using a paralleling architecture. Figure 6 shows two high-level data center power system designs. On the left is a low voltage, modular direct coupled design. In this design, each generator set is dedicated to backing up a specific server room. The gensets don't parallel. This is the most common and preferred design in data centers as it is simple to both commission and maintain. The scenario discussed in the previous section where two smaller gas generator sets were paralleled to replace a single larger generator set could still use this basic architecture. Each generator set shown in this drawing could actually represent a pair of paralleled generator sets, so this overall architecture is maintained.

The right side of Figure 6 shows a paralleled genset design. Although this design is less common in data centers, it does offer some advantages over the direct coupled design. The advantages come from the fact that the sources and loads are aggregated rather than discrete. For one thing, it allows for flexibility in the size of the generator sets. In the modular design, the generator sets need to be sized for the loads in each specific cell or server room.

Aggregating sources and loads enables a reduction in stranded capacity as system capacity can be sized closer to the average load. Consider a system where most of the modules never go above 50% capacity, but one or two of them are at 90% capacity. In a modular system, most of the generator sets would be underutilized while one or two of them would be pushed to their maximum. In a paralleling design, all of the generator sets could be loaded at the same level. By distributing the peaks across the entire bus, the total capacity could be reduced where allowed by the SLA.

Aggregating loads can also smooth out a volatile load profile. Consider a facility serving both AI and conventional cloud loads. AI load profiles are very volatile and will be challenging for a generator set to follow the load while keeping frequency and voltage within tolerances. Aggregating the loads will smooth out the load profile any generator set sees, making the system more stable.

A paralleling system enables fuel savings in demand response operation, as generator sets that are not needed to carry the load can be shut down. Fuel costs can be significant in grid support operation, where the generator sets may be running 500 hours per year, rather than 50 hours per year that you might have in a standby operation.

Also, for peak shaving applications, where the generator sets take only part of the load off the grid, operation is simpler with a paralleling system as a single set point can be set for the entire system, rather than individual set points for each generator set. Therefore, while modular, direct coupled power systems are generally preferred in data centers for their simplicity, paralleled systems offer some advantages in providing grid support functions.

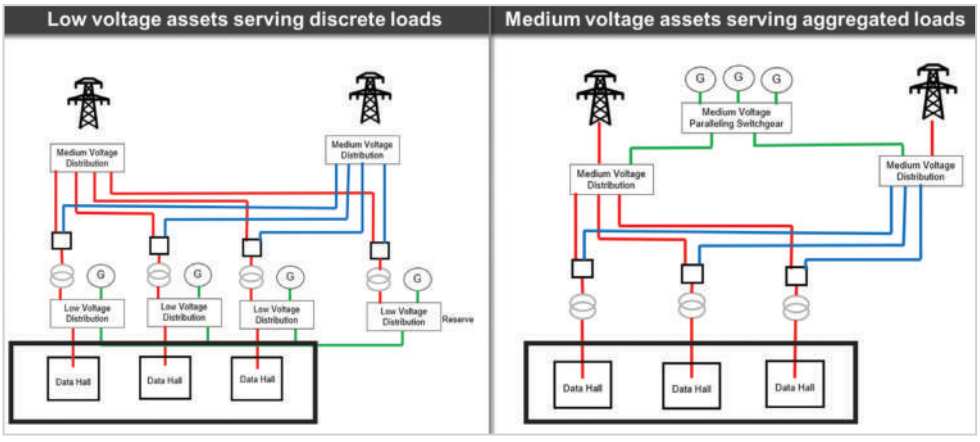


Figure 6. Data Center Power System Architecture

BATTERIES

Lithium Ion batteries are generating a lot of interest in onsite generation in many different applications as technological improvements are increasing energy density and life expectancy. Battery Energy Storage Systems (BESS) are being deployed both as a means to firm power from renewable sources and to provide grid support functions. Batteries have an advantage in this area over engine driven generators in that they have a greater capacity to stabilize grid voltage by producing and absorbing VARs.

The 2-4 hour expected duration of demand response events fits well within the capability of the available technology. Batteries can support these events without the emissions concerns of a diesel generator set and they don't rely on gas pipeline infrastructure making batteries a promising solution.

The challenge of course is that batteries by themselves can't provide the 24-48 hours of resiliency commonly required by data centers. Hybrid designs are being deployed with batteries and generator sets, often all paralleled on a common high voltage bus similar to the architecture shown in Figure 6. Batteries may be used for curtailment events and initially during outages with generators only coming online for outages that last more than a few hours. This will require a sequence of operations that coordinates functions between the system level control, dispatching assets and setting their operation mode, and the asset level (BESS) control which executes the functions of the particular asset. This type of distributed control architecture, where BESS level functions such as synchronizing and control of charge and discharge rates are executed at the asset level and supervisory functions are executed at the system level, minimizes single points of failure and enables a resilient power system.

In these scenarios, data center operators will have an opportunity to decide whether all workloads are critical enough to require backup power for outages lasting more than a few hours. To be sure, many data centers will continue to require 24-48 hours of resilience (and it may be required in their service level agreements with their customers). However, where data centers can shut down or migrate some of their workloads during a rare extended outage, there will be an opportunity to reduce the number of generator sets, which will reduce capital expense and simplify emissions permitting.

Conclusion

The rapid increase in power requirements for data centers will create challenges for utilities as aggregated connection requests exceed current maximum loads. Analysis shows that flexibility of loads, the ability to curtail loads during peak demand, will allow data centers to add substantial load without immediately expanding capacity. Because of this, utilities are in some cases mandating data center operators to participate in grid augmentation programs as a condition of their connection agreement. On site diesel and natural gas generator sets and battery energy storage systems can be used to support these programs, but each asset presents challenges. Diesel generator sets create emissions permitting challenges. Natural gas generator sets present challenges around pipeline capacity, performance and footprint. Battery energy storage systems solve emissions challenges but offer limited autonomy. Addressing constrained grids will require creativity and consideration around power system architectures and operating sequences.

About the author

Rich Scroggins
Applications Engineering Technical Advisor



Rich Scroggins is a Technical Advisor in the Application Engineering group at Cummins. Rich has been with Cummins for 18 years in a variety of engineering and product management roles. Rich has led product development and application work with transfer switches, switchgear controls and networking and remote monitoring products and has developed and conducted seminars and sales and service training internationally on several products. Rich received his bachelors degree in electrical engineering from the University of Minnesota and an MBA from the University of St. Thomas.

Cummins Inc.
Box 3005
Columbus, IN 47202-3005
U.S.A.

cummins.com

Bulletin 6644586 Produced in U.S.A. 8/25
©2025 Cummins Inc.