Life Is On | Schneider Electric

Retrofitting Existing Power Systems for AI Clusters

White Paper 126
Version 1

Energy Management Research Center

by Stuart Sheehan
Allegia (Gia) Wiryawan

Executive summary

The IT industry is developing large generative AI models at a fast pace. These models can require megawatts of power at hundreds of kW/rack in data centers, known as AI factories. This paper explains the unique power requirements of these workloads and the challenges data center operators have in supporting them. Guidance is provided on how to provision *existing* data centers to support these loads.

RATE THIS PAPER ★★★★★

Key takeaways

Key takeaway 1

AI compute clusters rewrite the density rulebook: they will overwhelm legacy power paths unless retrofits account for the core challenges facing operators.

Key takeaway 2

Start with a load study to understand your true power capacity and then design head room and live monitoring into your systems.

Key takeaway 3

Then address the series of challenges data center operators must overcome to provision their existing data center for AI: overload, block size, arc flash, voltage, power distribution unit (PDU) limits, and variable frequency drive (VFD) harmonics.

Key takeaway 4

Your solution playbook must include: Increasing distribution block size, limiting fault current, bringing in liquid cooling, addressing harmonics, and adding continuous monitoring. Use validated reference designs that incorporate these and other design practices.

Introduction

AI workloads influence data center power systems differently than traditional IT workloads. Most data center electrical infrastructure was engineered for lower rack density, some by a factor of 10x or more compared to the 100+kW racks of today. Modernizing to meet evolving requirements – in a living operating data center - is a real obstacle to leveraging the benefits of AI factories.

We are still in the early stages of understanding how different AI workloads impact data center power systems. Most assume that power system challenges are limited to the pretraining and post-training (e.g., fine-tuning) of large language models (LLMs). However, within these categories, a range of variables can either increase or decrease the strain on power systems and we currently lack detailed power profiles[1] to quantify their effects.[2] The rapid evolution of AI research[3] makes supporting these workloads a moving target. For instance, newer, compute-intensive *inference* tasks-sometimes called "long thinking,"[4] may also present significant power challenges, but specific power profiles for these workloads are not yet available.

Power profiles are essential for predicting how a data center's power system will respond to specific AI workloads. While we may not have comprehensive profiles for every workload, we have identified **five key attributes and trends** that help us estimate the demands of a worst-case scenario. By designing data center power systems to accommodate these worst-case profiles, we can better verify readiness for future generations of AI workloads.

[1] The sum of all power drawn by a data center's IT loads charted over time.

[2] Examples of variables include precision, data batch size, model compression techniques, accelerator (type, generation, & cooling method), and workload orchestration.

[3] I.e., in-memory computation, compute-efficient algorithmic operations, algorithm evolution, and others.

[4] This is also known as test-time scaling, one of three scaling laws, but others may follow.

Every organization will need to address power system design challenges based on their unique AI workloads[5] and data center requirements. In this paper, we outline the **key attributes and trends** shaping AI power demands, define a representative worst-case power profile, and map the resulting challenges to data center power systems. We also present retrofit strategies to help existing sites meet these new standards.

Let's talk about five key attributes and trends:

1. **Accelerator network communication** – Accelerators like GPUs can process data and generate tokens at a much faster rate than the communication speed between them. This makes accelerator network communication latency a bottleneck that determines how fast you can complete a given workload. A cost-effective and power-efficient approach to decrease this latency within the rack, is to use copper network cables given the short distances. However, if we used copper to connect the racks (*inter-rack* communication), the longer distances would introduce untenable latency. This means we need faster, more expensive, and energy-intensive, *inter-rack* communication (e.g. fiber). Hence the incentive to maximize the number of accelerators in a rack. This is what leads to higher rack densities in AI clusters. See "Accelerator network communication" section of White Paper 110 for more information on latency.

2. **Thermal design power (TDP) of accelerators** – thermal power consumption, measured in watts, is commonly specified with TDP. TDP, and the associated compute performance, trends upward with new generations of accelerators. This means you can train models and infer (both of which generate tokens) in less time and with lower cost. This trend adds to increasing rack densities.

3. **Peak power** – While nearly all IT workloads exhibit peak power consumption, chips used in AI workloads may exceed their TDP multiple times per second and later idle. This higher threshold is referred to as the electrical design point (EDP). These transients may exceed the steady state TDP (e.g., 50%) for tens of milliseconds while not violating the average long-term TDP thermal limits. **Figure 1** illustrates an example of power peaks that exceed TDP and later fall to an idle power state. The profile, duration, and frequency of these peaks and lows will vary depending on some key variables. These include IT hardware (i.e., GPUs, power supplies, storage, and network), AI workload, and software limits imposed on loads (i.e., power limits). The magnitude of these peaks and valleys will be lower when measured at the data center level (due to other IT hardware used in an AI cluster such as network switches and storage).

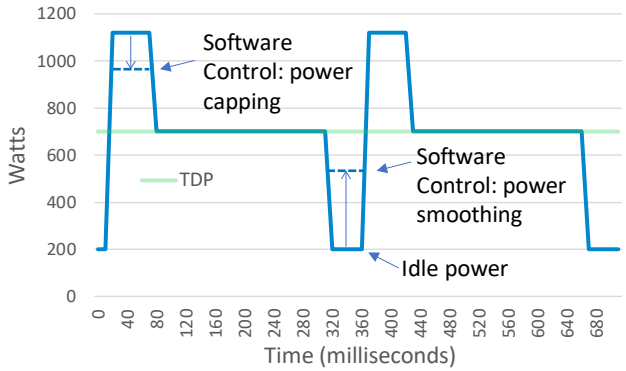

Figure 1

Example of accelerator peaks and idle states shown in millisecond timescale (green line represents TDP)

[5] AI workloads can include inference and training. Training encompasses a spectrum of techniques including pre-training (training from scratch), fine-tuning, prompt tuning, etc. The AI clusters required for these workloads may range from a data center with 100's of megawatts of racks to several racks.

4. **Synchronous computation** – The power consumption pattern for virtually all IT equipment and workloads resembles a series of peaks and valleys over time. This pattern for traditional IT workloads is asynchronous, meaning that the power consumption peaks occur at different times and don't coincide (i.e., loads are diversified). For example, while individual servers may have a power variance of 60% between idle and full load, the aggregate load from all servers (as seen by a UPS) will have a lower variance. The probability that all these peaks occur at the same time is very low. This asynchronous pattern is what allows data center designers to "oversubscribe" power and cooling systems like UPS and chillers.

   In contrast, certain AI workloads may result in a synchronized power consumption pattern for all AI servers. This means that peak power draw occurs at the same time multiple times per second, acting like quick step loads. **Figure 2** provides a hypothetical illustration of how the *sum* (blue line) of all the workloads varies only slightly for asynchronous workloads but for synchronous workloads, the sum is highly variable.

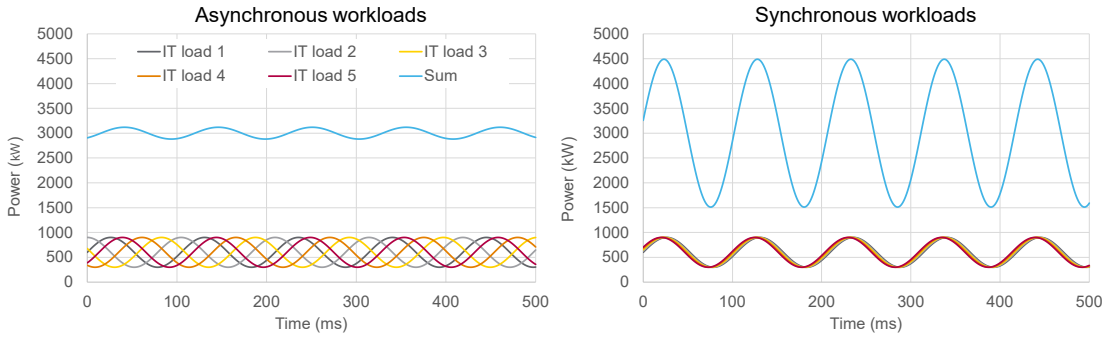

Figure 2

Hypothetical comparison between asynchronous and synchronous workloads

5. **AI cluster size** – certain AI workloads can be large, parallel processes that extend beyond single servers, potentially utilizing thousands of accelerators. For example, pre-training LLMs can require a dedicated data center loaded to nearly 100%. Depending on the data center size, even a modest AI training cluster represents a significant percentage of data center load.

The type of workloads we refer to in this paper rely on scale-out computing (a large number of machines running in parallel). The AI servers are assembled into an array of racks known as an AI cluster which essentially operate as a single computer. Each compute rack in a cluster could be over 100 kW with direct-to-chip liquid-cooled servers. Traditional power distribution architectures are unable to support these densities without changes to the existing electrical system.

While White Paper 110, *The AI Disruption: Challenges and Guidance for Data Center Design*, provides high-level recommendations to address these challenges, this paper provides more detailed guidance for those operators that are ready to implement in *existing* data centers. **Table 1** provides the challenges and their mapping to four power subsystems.[6] **Figure 3** illustrates the electrical flow of the subsystems (from input power to critical rack distribution) and calls out the challenge(s) associated with each subsystem. **For convenience, each challenge in the table is hyperlinked, so you can easily navigate to that section.** Also, every page has a "home" symbol on the upper right which returns you to **Table 1**.

[6] See White Paper 61, *Electrical Distribution Equipment in Data Center Environments*

Table 1

Impact of challenges on power subsystems

	Challenge	Onsite power sources	Main SG	UPS systems	Critical dist.
1	Lack of load diversity and peaky loads increase risk of overload	⚠	⚠	⚠	
2	Small power distribution block sizes are impractical				⚠
3	Increased risk of arc flash hazard complicates work practices				⚠
4	120/208 V distribution is impractical to deploy				⚠
5	Standard 60/63A rack power distribution units (PDU) impractical to deploy				⚠
6	CDU pump VFD harmonics risk causing equipment malfunctions			⚠	⚠
7	* High rack temperatures increase risk of failures & hazards				⚠

** Not discussed in this paper as it is thoroughly covered in White Paper 110.*

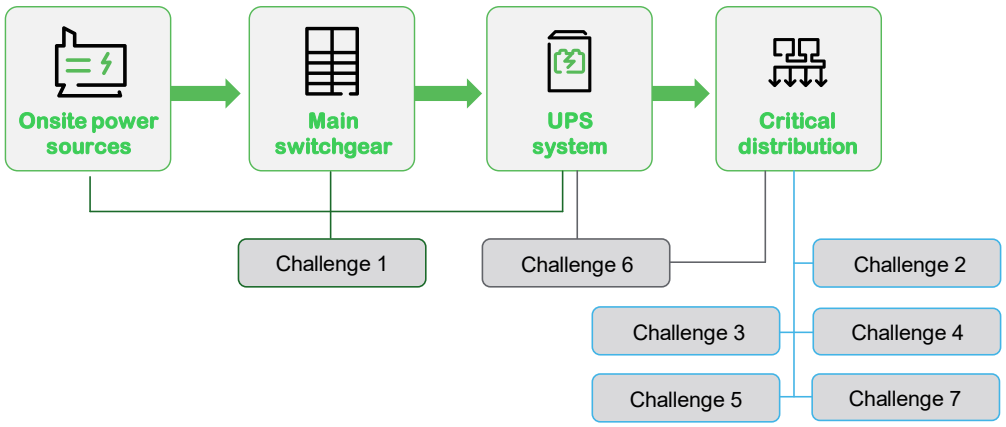

Figure 3

A data center's four power subsystems

These four data center subsystems are described below:

- Onsite power sources – Includes power sources like generators and the associated switchgear.
- Main switchgear – Includes the main breakers that feed subsystems such as cooling, UPS, and load banks.
- UPS system – Includes, not only the UPS, but also the batteries and the UPS output switchgear, which includes the maintenance bypass.
- Critical distribution – Includes step-down transformer(s)[7], distribution circuit breakers, branch breakers, and rack power distribution units (rPDU).

[7] International Electrotechnical Commission (IEC) countries don't typically use step-down transformers in data center critical distribution.

1. Lack of load diversity and peaky loads increases risk of overload

Onsite power sources	Main SG	UPS systems	Critical dist.
⚠	⚠	⚠	

Average power

The average accelerator power consumption we refer to is over a period of a few seconds, not over days or weeks. It is over this worst-case timespan that we determine how much of the data center capacity is utilized. If we were to average over a day or two, the cluster's idle periods would lower the average, leading you to believe that the data center utilization was lower.

Note that when we talk about the total data center power, the total is composed of more than the totality of accelerators. The other server components, data storage, and networking, all combine to dull the impact of the transients.

How this challenge impacts the power system

Three of the key AI attributes and trends discussed in the introduction (*peak power*, *synchronous computation,* & *AI cluster size*) can be combined to form a worst-case power profile. This profile increases the chance of overload and potentially reduces the life of upstream infrastructure. Note that the word "average power" related to this challenge, implies a specific time scale. (See sidebar)

When you combine *accelerator power peaks* with *synchronous computation*, the resulting peak load on the physical infrastructure is proportional to the accelerator quantity. Fortunately, the collection of other IT components inside and outside the server dampen the magnitude of the peaks. This is where the *AI cluster size* attribute comes in. The power profile of the AI cluster, relative to the data center's IT design capacity (kW), determines the impact on the data center's physical infrastructure. Assuming a worst-case power profile, if a 1N data center's only load is an AI cluster, and its *average* load approaches the data center's IT capacity, the data center will likely experience consistent overloads. The greater the proportion of traditional IT workloads, the less likely the data center will experience this challenge. This is because the asynchronous load will dampen the effect of the synchronous EDP transients. **Since we don't yet have power profiles for different AI workloads, we will assume the worst-case profile for this challenge, which does not include traditional loads.**

Even though the accelerator peaks are of short duration (tens of ms), they may impact upstream power infrastructure. If the magnitude and frequency of the peak exceeds the specifications (rated current, overload specifications, and rate of change) of upstream infrastructure such as UPS and generators, their performance can be compromised. **Note that the peaks may lead to two distinct and independent challenges, each with their own risk: overload and step load.** It's possible that equipment is sized such that it isn't overloaded yet is unable to support the step load. Potential impacts of exceeding design limits are:

Onsite power sources[8]

- Degrading generator power quality (frequency & voltage) below acceptable limits for loads and UPS. Frequency and voltage regulation is especially vulnerable to accelerator transient step loads (e.g., peak to idle states), especially when heavily loaded. If these step load thresholds are breached (magnitude & rate of change), the generator may automatically shut down to safeguard itself and connected loads. This could happen even when the peaks do not overload the generator.
- Step changes could create oscillations or resonances with site power infrastructure such as distribution equipment.

Electrical switchgear equipment

- Increased risk of tripping circuit breakers. Circuit breakers with electronic trip units often have a thermal memory function to prevent conductors from overheating during cyclic loading. Continuous peaks may lead a breaker to open.
- Thermal stress and aging of distribution components: insulation, wire terminations, and fuses.
- Nuisance tripping of protection devices.

[8] The Llama 3 Herd of Models, July 23, 2024, p. 14, "During training, tens of thousands of GPUs may increase or decrease power consumption at the same time, for example, due to all GPUs waiting for checkpointing or collective communications to finish, or the startup or shutdown of the entire training job. When this happens, it can result in instant fluctuations of power consumption across the data center on the order of tens of megawatts, stretching the limits of the power grid. This is an ongoing challenge for us as we scale training for future, even larger Llama models."

Uninterrupted power supply (UPS) system

- Switches to bypass if inverter limits (instantaneous & thermal) are exceeded.
- Draws energy from battery if UPS AC input current limits are exceeded.
- Degrades battery state of charge (SOC) if overloads are significant enough to draw from battery and frequent enough to not allow full battery recharge between peaks.
- Shortens life of batteries (i.e., reduced state of health) through cycling and thermal wear (i.e., increased temperatures).
- Decreases life of semi-conductors and fuses if design limits are exceeded.

How to address this challenge in existing data centers

To address this challenge, we assume that your data center has enough spare capacity to support the *average* AI cluster load.[9] The optimal solution to this challenge depends on several factors related to the AI workload power profile and the electrical distribution topology. Prior to adding the AI cluster, the first step is to perform a detailed load study for the upstream infrastructure. This should include:

1. Anticipated AI load – Estimate the expected frequency and magnitude of any power peaks. While not foolproof, the nameplate power rating (watts) of IT power supplies can be more indicative of the AI load's AC power peaks than TDP. This can be a useful conservative estimate if the server manufacturer data does not specify peak values. Note that power supply capacitance helps "absorb" power peaks; more capacitance will better limit power peaks. This capacitance varies among different IT power supplies. Other loads like storage and networking will dilute the magnitude of a peak (lower peak to average). Remember to account for the power required to cool the AI cluster.
2. Available electrical capacity – Use power quality meters[10] to monitor the power draw on the main breaker over the course of a week to determine the data center's average and peak load. Document the nameplate capacity and available capacity of components that will supply the AI cluster including transformers, circuit breakers, generators, and UPSs. Verify that the power infrastructure's redundancy meets the original design specifications (e.g., standby generators, UPS, PDUs, etc.).

With the details from the load study, you can take a more informed design approach. We recommend the following solutions listed in order of most feasible for a production data center.

- Size the AI cluster to match the spare data center capacity – This solution addresses *overloads*. The least disruptive solution for a production data center is to decrease the number of proposed AI servers and supporting IT gear such that the resulting peak power is equal to or less than the spare capacity. This solution represents a conservative approach and should be evaluated in light of business goals. This solution does not require changes to your current facilities operation (e.g., maintenance schedules, generator testing, etc.).

- Accelerator software control – Using software to cap accelerator power consumption may allow your proposed AI cluster to operate within your data center's spare capacity (addresses *overloads*). This necessarily means a tradeoff between performance and energy consumption dependent on the extent of

[9] The spare capacity of the main switchgear and UPS is greater than or equal to the anticipated AI cluster's *average* load (including cooling load).

[10] Unlike other meters, power quality meters are able to capture sub-cycle wave forms.

the capping.[11]

Accelerator power smoothing software addresses *step loads* by injecting extra load, thereby increasing the idle power. Raising the idle power decreases the step load seen by the electrical infrastructure. This reduces the negative impact of rapid step loads such as with generators.

- Install a specialized energy storage system – This solution addresses both *overloads* and *step loads*. Examples of energy storage include high-cycle batteries and super capacitors. For *overloads*, the stored energy discharges during the peaks, acting like a peak shaver such that the upstream systems do not see the peaks. For *step loads*, energy storage charges during the idle points to decrease the step load, acting like a shock absorber such that the upstream systems see a smoother load. In both cases, the controls determine when to charge and discharge thereby providing power smoothing.
- Use existing UPS for power smoothing – This solution addresses *step loads* only. If pressed to address step loads, your existing UPS may be capable of performing a power smoothing function but requires investment. Existing batteries must be replaced with high-cycle batteries or super caps. The UPS controls would also need to be altered to perform this function.
- Increase capacity of switchgear, conductors, and UPS – This solution addresses *overloads*. The most disruptive action is to increase the capacity of the power system in a production data center. In essence, this means rightsizing only the power system to accommodate the workload's peaks. Assuming the power system is sized to the average AI cluster load, increasing the power system capacity by 6% to 15% represents about 1% to 2.5% of the total cost of the IT kit.[12] It's a small price to pay for peace of mind. Note that in this example, your data center is still rated for its original capacity because the cooling system capacity didn't change. Only the power system capacity increases.

In all of these solutions, we recommend, at a minimum, power quality metering at the main breaker and the sub-feed breakers feeding the cluster. Also, across all designs, we recommend electrical power monitoring software (EPMS). EPMS polls the meters for the data. Data center operators can then monitor the power using the EPMS user interface or the EPMS can share the data with other software such as data center infrastructure management (DCIM) and building management system (BMS) software. Among other things, monitoring software allows you to set threshold alerts when critical levels are approached. This will allow IT admins time to assess which loads to temporarily throttle to avoid exceeding infrastructure limits.

The risks of allowing infrastructure to "absorb" the peaks (overloading)

You may be familiar with the idea of maximizing the power system's capacity such that the EDP peaks overload system components like UPS. The typical rationale for this approach is that you can maximize the AI workload with every available watt. On the surface this sounds like a great idea, if we ignore the stress placed on the power system. Consider the perspective of a data center operator who invests millions of dollars in an AI cluster to accelerate workloads (e.g., parallelization, high-speed interconnects, mixed precision, etc.). It becomes illogical, then, to jeopardize the very goals operators seek to achieve by knowingly overloading the underlying physical infrastructure.

Physical infrastructure, such as circuit breakers and UPS systems, is designed and tested to withstand intermittent power transients above its rated capacity. However,

[11] Zhao, *et al.*, *Sustainable Supercomputing for AI: GPU Power Capping at HPC Scale*, Oct 2023, Proceedings of the 2023 ACM Symposium on Cloud Computing

[12] See Appendix for assumptions.

the frequency of EDP power transients significantly surpasses the typical test conditions for these systems. Focusing specifically on UPS. Most, if not all, three-phase UPS can support overloads to varying degrees based on their magnitude and duration. However, they weren't designed to support repetitive overloads (multiple times per second) over weeks and months. A specialized energy storage system, mentioned above, is the appropriate solution to absorb EDP peaks. It's difficult to predict the long-term impact these overloads will have on power equipment. **It's likely that the probability of failure will increase over time compared to the same equipment operating at the rated capacity.**

For data centers with power system redundancy (e.g., 2N, N+1) designed into its switchgear, UPS, and generator, it may be tempting to use this redundancy as increased capacity. **We do not recommend this practice as it comes with inherent downtime risks and requires changes to your current facility operations when your data center loses redundancy.** For example, AI training would be stopped for scheduled UPS maintenance and during other degraded states (e.g., equipment failure). Also, production would cease while operating on generator because idle-to-EDP "step loads" may degrade voltage and frequency output, or worst case, cause generator shut down.

Given that these AI clusters are an emerging application, we recommend engaging equipment vendors regarding your application and specific power profile. Standards for addressing this challenge will develop as the industry matures. We also anticipate that accelerator vendors will provide more software control over these power profiles, specify additional capacitance in server power supplies, and make design changes to future accelerator generations.

2. Small power distribution block sizes are impractical

How this challenge impacts the power system

The *accelerator network communication* attribute discussed in the introduction implies that rack densities will continue to increase well beyond 100 kW per rack. Power distribution units (PDU) and remote power panels (RPP) are typically rated for about 300 kW. This means that each PDU or RPP could support three 100 kW racks. This is impractical because the smaller the capacities, the more units must be maintained. This also wastes space, not only with the footprint of each unit, but the service clearances required. As distribution block size decreases, both infrastructure space consumption and maintenance expenses increase. This challenge is exacerbated by redundant distribution configurations.

How to address this challenge in existing data centers

PDU, RPP, or feeder circuit breaker capacity ratings (block size) should increase to accommodate increasing rack densities. However, increasing capacities isn't as simple as increasing the breaker amperage. Design preferences and constraints tend to dictate the maximum distribution block size. A key constraint is the design of your data center's existing power system. Some others include:

- Increasing distribution capacities also increases fault current (discussed in the next challenge).
- In N+1 redundancy schemes, as distribution capacities increase, stranded power increases (the "+1" is stranded). Stranded capacity decreases with increasing "N" (2+1, 3+1, etc.) but power distribution complexity increases.
- The need for a symmetrical layout of compute and support racks (each with their own rack density) makes it challenging to maximize the capacity utilization of both feeders and rack PDUs.

- Standard, off-the-shelf, ratings for equipment like busway and rack PDUs are limited. This challenges designers to achieve goals related to cost, power efficiency, redundancy, space utilization, etc. On the other hand, the more standardization, the less opportunity for unique failures.

Given the unique characteristics of existing power systems, engage with your physical infrastructure and IT vendor(s) for advice on how to adapt to your proposed AI cluster. Validated AI cluster reference designs may also provide some ideas. For example, Reference Design 108, *7392 kW, Tier III, IEC, Chilled Water, Liquid-Cooled AI Clusters*, uses 100% rated 800A breakers, providing 575 kW. Reference designs also account for electric code rules when increasing the capacity of existing systems.

3. Increased risk of arc flash hazard complicates work practices

How this challenge impacts the power system

This challenge arises mainly through increasing the transformer capacity as proposed in the previous challenge. Unfortunately, the higher the transformer capacity, the lower the impedance to fault current. This means that if there were a fault downstream of the transformer, more fault current would flow with a higher-capacity transformer compared to a lower-capacity transformer. This is important because fault currents beyond 10 kiloamps (kA) at the rack may pose work restrictions for IT admins. In regions like Europe that don't use PDU transformers, this challenge arises mainly as a result of higher capacity circuits (larger cables).

How to address this challenge in existing data centers

White Paper 194, *Arc Flash Considerations for Data Center IT Space*, states: "The term "arc flash" describes what happens when electrical short circuit current flows through the air. A fault (the common term for short circuit) usually occurs between a live conductor (e.g., wire, bus) and another live conductor(s) or grounded metal. In many cases, a single-phase fault quickly evolves into a three-phase fault. In an arc flash, the current literally travels through the air from one point to the other, releasing a large amount of energy, known as *incident energy*, in less than a second. This energy is released in the form of heat, sound, light, and explosive pressure - all of which can cause harm. Some specific injuries can include burns, blindness, electric shock, hearing loss, and fractures."

The two most important factors[13] that determine the amount of incident energy (measured in calories/cm^2) are:

- Available fault current – measured in kiloamps (kA), is the maximum amount of current available (at the location of a fault) to "feed" a fault and is dependent on the electrical system design.
- Arc duration – measured in milliseconds (ms), is the amount of time it takes for a fuse or circuit breaker to open and clear a fault.

A data center's electrical design controls both factors. A short circuit analysis determines how much fault current is available at the rack PDU input. This fault current should typically be limited to 10,000 amps (10 kA) or less. If the fault current is greater than 10 kA at the rack PDU cord cap (i.e., connector), implement one or a combination of the following solutions (listed below in order of most preferred to least preferred). Consult with your physical infrastructure vendor as some cord caps or distribution equipment may have ratings greater than 10 kA.

[13] White Paper 194, *Arc Flash Considerations for Data Center IT Space*, p. 2

- PDU transformer impedance – transformers should be specified with 5-9% impedance to help limit fault current at the rack. Increasing impedance must be balanced with the resultant voltage drop. Though atypical in IEC countries,[14] PDU transformers are an effective means for limiting fault current.
- Current-limiting breaker – these breakers usually protect a group of branch breakers, and each branch breaker protects a rack PDU. If a fault occurs at the rack PDU, the main breaker limits the fault current seen by the branch breaker. It does this by partially opening its contacts thereby causing an arc which increases the impedance, limiting the amount of fault current it lets through.[15] If the branch breaker doesn't open, the main breaker ultimately will.

  When pairs of breakers are "series rated", current-limiting breakers are usually used. The breaker pair is tested together despite the downstream breaker having a lower fault current rating than what is present on that circuit. Furthermore, when used with a current-limiting upstream breaker, series rating may provide selectivity which means that the branch breaker trips before the main breaker. Note that this solution mostly applies for branch breakers rated for 63 amps or less.
- Conductor length – the longer a circuit's electrical wires (i.e., conductors) the more impedance and therefore the lower the fault current available at the end of the circuit. Sometimes increasing the length of circuits, beyond what's required, provides just enough impedance to meet a specification.
- Short-circuit current limiter block – these devices limit current similar to current-limiting breakers except that they are connected to the circuit breaker.
- Line reactor – these devices add impedance to a circuit much like transformers do. They are made of wire wound around a metal core and are sometimes referred to as a choke since it resists the flow of fast changes in current.
- Redundant current-limiting breakers upstream of the branch breaker – placing two current-limiting breakers in series complicates selectivity but reduces the fault current seen by the branch breaker.
- Fuse – in general fuses interrupt fault current faster than circuit breakers given the same voltage and current rating. However, in data center applications, fuses are generally a last resort since they must be replaced after a fault, increasing mean time to repair. A supply of spare fuses must also be available.

4. 120/208 V distribution is impractical to deploy

Onsite power sources	Main SG	UPS systems	Critical dist.
			⚠

How this challenge impacts the power system

This challenge applies *only* to data centers in non-IEC countries that use PDU transformers to distribute 120 volts (single-phase) and 208 volts (three-phase) in data centers. The lower the voltage, the more current you need for the same power. Consequently, the wire must be larger to provide greater current. At 120/208 V, it would take five 60-amp circuits to power an 80 kW rack (each circuit equals 120 V x 3 phases x 60 A x 80% derating = 17.3 kW) at 1N redundancy. Most racks can accommodate 6 vertical rack PDUs with a rack extension kit (**Figure 4**) using left and right rack channels. Therefore, 5 circuits per rack is feasible at 1N but not at 2N.

A PDU transformer is not typically used in countries with 230V distribution. This is because the data center *input* voltage (230V) is already compatible with IT equipment. The data center *input* voltage in most North American (NAM) data centers is 277V which is too high for most IT equipment. This is why PDU transformers are required to step down 277V to usable IT voltage.

[14] The International Electrotechnical Commission oversees the standards for electrical equipment.

[15] Discrimination, cascading, and enhanced discrimination by cascading

How to address this challenge in existing data centers

Replace 120/208V PDUs with 240/415V PDUs. This simplifies the distribution (less circuits needed) and allows more space for network cable trays. As suggested in the previous “arc flash” challenges, increasing the PDU transformer % impedance will help limit the amount of fault current at the IT rack.

5. Standard 60/63A rack PDU impractical to deploy

Onsite power sources	Main SG	UPS systems	Critical dist.
			⚠

How this challenge impacts the power system

Space in AI racks is limited due to deeper servers and network cable density. Space becomes even more constrained with liquid-cooled servers because the liquid manifold occupies one side of the accessory channel in the rear of the rack. In this case, a typical rack can accommodate a maximum of two rack PDUs. At 230 V 63 A (IEC) and 240 V 60 A (non-IEC), the highest-capacity standard rack PDUs provide 43.6 kW or 34.5 kW respectively. Two rack PDUs will support rack densities up 87.2 kW or 69.0 kW. This is still not enough for rack densities over 100 kW and doesn’t account for redundant power paths.

How to address this challenge in existing data centers

While not a standard offering for all vendors, there are higher rated rack PDUs available. For example, 240 V 125 A (non-IEC) rPDU provides 71.9 kW. Using two of these can support 143.8 kW. For rack densities greater than 143.8 kW, it may be possible to add a third rPDU if the rack vendor offers a 1400 mm deep rack, or a rack extension kit. See yellow highlight in **Figure 4**. This will allow up to three rack PDUs at 1N or 6 rack PDUs at 2N or N+1 redundancy if both rack channels are available. For higher rack densities and redundancy options, we recommend specifying custom rack PDUs, as shown in **Table 2**. Note that IEC capacities in **Table 2** may be lower with busway distribution, due to the circuit breaker’s thermal derating within tap-off boxes.

Figure 4

Rack extension kits provide space for an additional rack PDU

	Standard			Custom		
Non-IEC	40 A	60 A	100 A	125 A	150 A	175 A
240/415 V	23.0 kW	34.5 kW	57.5 kW	71.9 kW	86.3 kW	100.6 kW

IEC	32 A	63 A	125 A	150 A	160 A
230/400 V	22.2 kW	43.6 kW	86.6 kW	103.9 kW	110.9 kW

Table 2

Usable 3-phase power density per rPDU based on circuit breaker amp rating and voltage (line-to-neutral)

Another possible solution to increase the number of feeds to the rack is to use Open Compute Project (OCP) racks. These racks have integrated power supply shelves and power busbars that distribute power to rack-mounted OCP-type IT equipment. In essence, the power shelves save space inside the rack by consolidating the individual power supply units (PSU) in each server.

Since this architecture integrates the power distribution into the rack, it doesn't require traditional rack PDUs. Instead, the power whip from the branch breaker feeds the connector at the bottom or top of the OCP rack using an IEC or NEMA connector as illustrated in Figure 5. This distribution approach alleviates a significant amount of space in the rack's accessory channel. The power shelves also allow multiple power redundancy options (e.g., N+1, N+2, 2N, etc.).

Figure 5

OCP Open Rack V3 power distribution to integrated power shelf

Source: Open Rack V3 Meta High Power AC Whip Power Cable specification

6. CDU pump VFD harmonics risk causing equipment malfunctions

Onsite power sources	Main SG	UPS systems	Critical dist.
		⚠	⚠

How this challenge impacts the power system

Harmonics are not new to data centers and their negative consequences have largely been kept in check over the last few decades through the adoption of power factor corrected (PFC) power supplies. However, there have been a few recent instances where the deployment of coolant distribution units (CDU) has coincided with an increase in equipment malfunction.

The thermal design power (TDP) of some chips necessitates direct-to-chip liquid cooling. The water in these systems is supplied with CDUs that may include pumps with a variable frequency drive (VFD). These drives may generate harmonics. Harmonics could negatively affect electrical devices such as air conditioning compressors. Harmonics can alter the shape of the voltage and current sine wave supplied to other equipment. Depending on the equipment and the severity of the distortion, the equipment may malfunction.

How to address this challenge in existing data centers

Considering the recent emergence of this problem, it merits further study to better understand the root cause. In the meantime, we provide the following guidance. First, given the cooling criticality of CDU pumps, they must be placed on critical UPS power in case of a power outage. Your data center's existing UPS system may not have spare capacity for the additional pump load. Therefore, install a separate UPS system to support CDU pumps. Verify that the UPS can support the pump inrush current. Alternatively, you may specify pumps with frequency converters that avoid high inrush current, sometimes referred to as "soft start".

Second, any harmonics generated by drives may be addressed by specifying a voltage and frequency independent (VFI) UPS. Alternatively, active filters (Figure 6) are capable of addressing a wide range of harmonics. For more information on this topic, see White Paper 510, *Impacts of Variable Speed Drives on a Building's Power Quality*.

Figure 6

Examples of active harmonic filters

Next steps

For organizations that will deploy AI workloads in a colocation or an existing on-prem data center, there are power infrastructure challenges they must overcome. The following next steps will help address these challenges:

Leverage validated reference designs. These tools incorporate the latest knowledge from both IT and physical infrastructure vendors. They should include a set of engineering documents such as electrical one-line diagrams, piping diagrams, floor plans, and equipment lists.

Perform a load study. Prior to adding the AI cluster, perform a load study for the up-stream infrastructure. This should include a detailed assessment of the anticipated AI load profile and the available electrical capacity.

Determine if the power peaks from the new AI cluster will overload your power system. If an AI cluster's average power consumption approaches a data center's IT capacity, the data center will likely experience consistent overloads. If this is the case, review the list of solutions most feasible for your data center.

Use power quality metering and monitoring software. This will allow you to monitor critical infrastructure loads and set threshold alerts when critical levels are approached. For example, IT admins will have time to assess which loads to temporarily throttle to avoid exceeding infrastructure thresholds.

Add PDU transformers to distribute power to the AI cluster. Though mainly used in North American data centers, PDU transformers are an effective fault current limiter for any region. Higher-capacity isolation transformers will support more racks compared to traditional PDUs. The increased fault current from larger capacity transformers will need to be counterbalanced with available fault current mitigation solutions listed in this paper. Performing a short-circuit analysis determines how much fault current is available at the rack PDU input. This fault current should typically be limited to 10,000 amps (10 kA) or less.

Implement the highest-capacity rack PDUs available for your region. AI racks can easily reach over 100 kW with the latest accelerator-based servers. Most racks have space for 2 rack PDUs at 1N redundancy. 1400 mm deep racks or racks with an optional extension kit, will allow up to three rack PDUs at 1N or 6 rack PDUs at 2N or N+1 redundancy, if both rack channels are available. If unable to support the required rack kW and redundancy with standard rack PDUs, consider higher-capacity custom rack PDUs.

Specify a UPS that can mitigate the harmonics from variable speed coolant distribution unit (CDU) pumps. If your AI servers are liquid-cooled, they will require a CDU to distribute coolant to the server's components. These variable speed pumps create harmonics that may interfere with other equipment. Placing these pumps on a UPS will mitigate these harmonics and maintain cooling during a power outage.

About the authors

Stuart Sheehan is a Lead Systems Engineer at Schneider Electric. He works to explore new technologies and incubate new solutions and architecture for the data center, focusing on the increasing union of sustainability, digitization, and power system and energy storage innovation. Stuart holds a Master's degree in Mechanical Engineering from Duke University and a Bachelor's degree in Physics from Bowdoin College.

Allegia (Gia) Wiryawan is a Senior Systems Design Engineer at Schneider Electric, where she plays a critical role in advancing solutions for data centers. Gia specializes in evaluating and analyzing emerging trends and technologies, focusing on power system architectures and energy storage. Her work includes developing reference design packages that integrate thought leadership and showcase our latest innovations, providing actionable strategies for optimizing data center operations.

Gia's expertise is demonstrated by practical and forward-thinking solutions for customers. Her contributions align with Schneider's commitment to sustainable and efficient energy management.

She holds a Bachelor's degree in Electrical Engineering with a minor in Computer Science from Tufts University. With a strong foundation in technical knowledge and analytical capabilities, she drives progress in the design and implementation of advanced data center technologies.

Acknowledgements

Special thanks to Victor Avelar for his original contributions to this paper.

RATE THIS PAPER ★★★★★

Resources

The Different Types of UPS Systems
White Paper 1

The AI Disruption: Challenges and Guidance for Data Center Design
White Paper 110

Arc Flash Considerations for Data Center IT Space
White Paper 194

Impacts of Variable Speed Drives on a Building's Power Quality
White Paper 510

7392 kW, Tier III, IEC, Chilled Water, Liquid-Cooled AI Clusters
Reference Design 108

Browse all white papers
whitepapers.apc.com

Browse all TradeOff Tools™
tools.apc.com

Note: Internet links can become obsolete over time. The referenced links were available at the time this paper was written but may no longer be available now.

Contact us

For feedback and comments about the content of this white paper:

Schneider Electric Data Center Research & Strategy
dcsc@schneider-electric.com

If you are a customer and have questions specific to your data center project:

Contact your Schneider Electric representative at
www.apc.com/support/contact/index.cfm

Appendix

Rightsizing to EDP peak, costs 1% to 2.5% of the AI cluster's (IT) capital cost

This analysis estimates the cost to increase a power system's capacity to match an AI cluster's peak power (i.e., EDP). We assumed two different UPS design factors (1.1 and 1.2). Power system designs are typically based on "per unit" (PU) sizing factors. For example, UPSs are typically sized to 1.2 times the full IT load. We present the incremental costs as a percentage of the AI cluster's IT cost.

Assumptions:

- $25 million per MW or $25/Watt is the capital cost of an AI data center including the physical infrastructure and AI cluster.[16] Assume this cost is per MW of rated *IT* capacity and an N+1 power system.
- $6.30/W is the estimated data center physical infrastructure capital cost at N+1 power redundancy as shown in **Figure A1**. Based on the *Data Center Capital Cost Calculator*.
- $18.70/W ($25.00 - $6.30) is the cost of only the IT (i.e., AI cluster), the total cost minus the data center physical infrastructure cost.
- 6% and 15% increase in data center power system capacity. This is the amount of extra power capacity the power system needs to support the EDP peaks at 1N redundancy. These values were estimated from an energy model that accounted for wire losses (99% efficiency) and PDU losses (varying efficiency). The 6% and 15% were derived from a UPS PU value of 1.2 and 1.1 respectively.

Findings:

- The capex premium of increasing the power system capacity by 6% and 15% is $0.18/W and $0.46/W respectively. This equates to 1% to 2.5% of the IT capex.

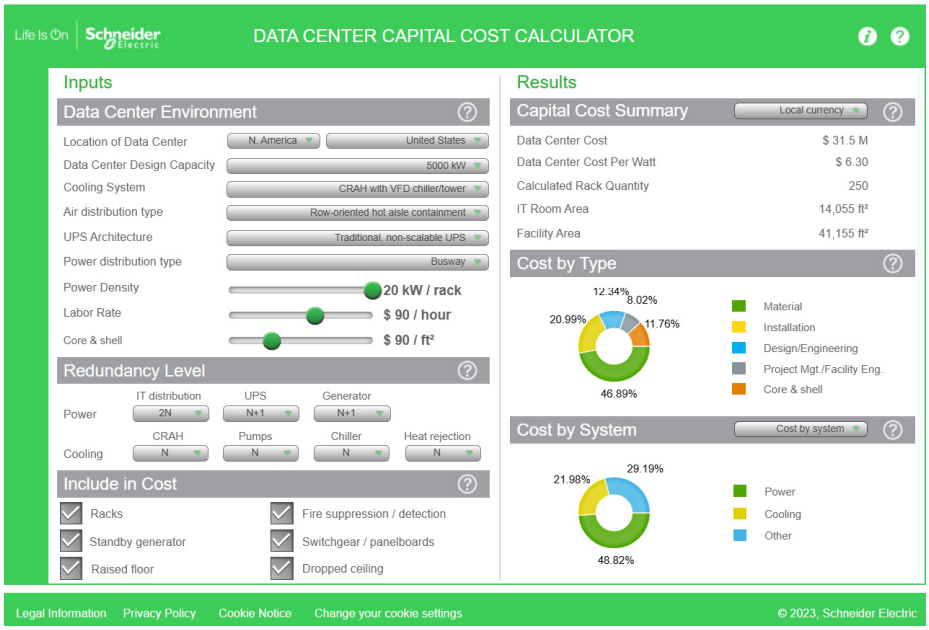

Figure A1

Data Center Capital Cost Calculator with estimated data center capital cost for N+1 power system

[16] Stephen Lacey, *Microsoft plans $80B for data centers as power constraints loom*, Latitude Media, 1/6/2026