Kalkitech

Bridging the OT/IT Gap to Unlock Capacity, Revenue, and Flexibility in Modern Data Centers

Prasanth Gopalakrishnan

White Paper

Executive Summary

The global Artificial Intelligence (AI) data center and Utility industries face a massive efficiency challenge. While the capital expenditure (capex) on data center systems and Utility investments in transmission, generation and distribution to power data centers exceed trillions of dollars over the next decade, a significant portion of this investment remains underutilized or not able to be fully deployed due to current grid interconnection guidelines, building for the worst-case scenario across the value chain. These over investments could potentially result in significant rate payer tariff increase which would increase opposition to data center projects or reduce ROI due to delays in data center projects.

This inefficiency stems from four major issues:

1. The lack of infrastructure available immediately to interconnect large data center loads.
2. The lack of visibility in real-time to the large data center loads and their impact on the grid, and lack of effective modeling of such loads in planning, and short and long term forecasts.
3. The ability to utilize the existing data center computing capacity to their fullest extent ("stranded capacity").
4. The ability to flexibly orchestrate the IT and OT systems, participate in markets and have correct financial models to orchestrate the compute and backup generation.

While hyperscaler Data Centers have tried to solve this through bespoke engineering, the wider market still struggles with the technical chasm between Operational Technology (OT) and Information Technology (IT) and act as a critical bottleneck to achieve the significant growth in Data Centers that the AI Industry needs. This whitepaper outlines a modular, software-defined architecture that bridges this gap, enabling enterprises to safely oversubscribe power, participate in grid services, and enhance grid infrastructure resilience.

However, what all these discussions do not fully comprehend, and address is how to utilize the OT equipment and infrastructure capabilities already used in MV Substations, Flexible Interconnection solutions deployed in DER Interconnections, Real-time power market Integration and integrate these into Kubernetes schedulers to achieve a more optimal real time flexible architecture that achieves the curtailment performance requirements of Utility and Power Grids, while achieving the ROI and SLA requirements of a Data Center. This whitepaper tries to address this gap and proposes an integrated and well-established approach to solving this problem.

1. The Power Dilemma: Stranded Capacity, Interconnecting Large Loads, the Load Curve and related work.

The Utility planners were focused on managing the Duck Curve when the data center load growth driven by AI Data centers added a new dimension to the problem. While the investment thesis for Data Centers is to maximize its Return on Investment (ROI) by running the IT servers continuously 24x7, the actual load pattern of these Graphics Processing Unit (GPU), Tensor Processing Unit (TPU) and Central Processing Unit (CPU) loads are not exactly base loads like for eg: an Aluminum Smelter Plant. They vary by whether it is an AI Training Load or an Inference Load. Further, there is stranded load in IT data centers and approaches like Oversubscription with Medium Voltage Power Plane[1] tries to address the same. The fundamental inefficiency arises from the difference between theoretical maximum power draw and actual consumption. Spikes in power usage are rarely correlated across thousands of servers, resulting in a large delta between the capacity built and the capacity used.

Strategy	Description	Impact on Training (KPI: Throughput / Time-to-Complete)	Impact on Inference (KPI: Latency / Reliability)	Impact on Load and Type
Job & Pod Throttling	Halting execution of a job or a pod during grid stress (e.g., peak hours or overloaded circuit due to partial outage) and restarting it when power is available	**High Impact**: Increases total Time to Completion (TTC). **Mitigation**: Checkpointing ensures no progress is lost, only delayed. Kubernetes scheduler extender distinguishes between a "pause-able" training pod and a "critical" inference pod using the Power Score.	**Not Applicable**: You cannot pause a live user request (inference) for hours.	Demand Response / Load Reduction
Voltage and Frequency Scaling	Lowering the GPU/ TPU voltage / clock speed to run slower but more efficiently (undervolting / underclocking)	**Medium Impact**: Iterations per second (throughput) drop proportionately to clock-speed reduction.	**High Impact**: Time to first token (latency) increases. **Mitigation**: Apply only if current latency is well below SLA limits or job can be scheduled in a different location.	Demand Response / Load Reduction
Workload Consolidation	Migrating active jobs to fewer GPUs to run them at high utilization, while putting empty GPUs into "sleep"	**Low Impact**: Training usually occupies full clusters anyway; less relevant for single-node training.	**Medium Impact**: Increases tail latency. Risk of cold-start delays if traffic spikes and sleeping GPUs need to wake up.	Demand Response / Load Reduction
Spatial / Geographic Shifting	Routing workload to a data-center region with better grid conditions or intermittent generation (e.g., moving load from equatorial data centers to solar-heavy regions during solar peak)	**Low Impact**: Only startup delay (transfer model/ data). Once running, throughput is normal. Large hyperscalers with equatorial data centers benefit from solar achieving 24-hour training during peak.	**High Impact**: Increased network latency due to physical distance from the user. **Cost Risk**: Higher data egress fees.	Generation Following Load
Temporal Shifting	Scheduling a job to start only when grid signals are favorable (e.g., waiting for solar peak)	**High Impact**: Queue pending time increases; the job sits in the queue longer before starting.	**Not Applicable**: Real-time inference cannot be scheduled for later; it must be served immediately.	Demand Response / Load Reduction
Approximate Computing / Model Switching	Switching to smaller, less power-hungry AI models during power constraints	**Not Applicable**: You train a specific model architecture; switching architectures mid-training is not feasible.	**High Impact**: Response quality may drop slightly. **Benefit**: Massive energy reduction per query.	Demand Response / Load Reduction

Strategy	Description	Impact on Training (KPI: Throughput / Time-to-Complete)	Impact on Inference (KPI: Latency / Reliability)	Impact on Load and Type
HVAC and Liquid Cooling Optimization	Throttling HVAC and CPU cooling aligned with CPU throttling using IT/OT integration	**Moderate Impact**: HVAC throttling reduces compute usage during summer.	**Moderate Impact**: HVAC throttling reduces compute usage during summer.	Seasonal DR / Load Reduction
BESS and DG Exporting to Market	Use backup generation and battery for grid support	**High Impact**: Allows full operation with higher risk due to reduced SOC and DG backup usage to support data-center operations.	**High Impact**: Allows full operation with higher risk due to reduced SOC and DG backup usage to support data-center operations.	Generation Support / Ancillary Services

Table 1: Scheduling strategies optimizing power usage and being responsive to load limits.

While the AI GPU's and CPU's largely provided ability to monitor power usage and throttle the clock cycle and performance, they were usually used to manage heat and life cycle of the platform, and not so much as a load management and job and pod management engine for quite some time. In[6] the ability of processor load management and optimization of power usage and capacity utilization is discussed based on CPUs. The work at Google in improving the performance of Kubernetes are discussed in[4] and[5]. The Software defined load management can provide significant advantages as shown by[3], however involves customized algorithm tuning and would become obsolete as efficient processors and training and inference methods evolve.

The NERC Large Load Task Force Whitepaper[8] lists the following High Priority Stability Risks to Bulk Power System (BPS) from Large Data Centers: Ride-through, Voltage Stability, Angular Stability and Oscillations. The whitepaper further lists Resource Adequacy, Balancing, and Reserves as high priority risks from Data Centers. The whitepaper states the following: *"Large loads pose risks in both the planning and operations horizons. Their quick interconnection timelines and large peak demands drive generation and transmission adequacy risks. The fast ramp rates and variability of the loads could exhaust reserves for balancing and contribute to voltage and frequency instability. If system planners and operators lack accurate dynamic models, they may be unable to predict ride-through and system behaviors during events. Much of the PEL load can contribute to harmonics and voltage fluctuations. Large loads like data centers can also be susceptible to cyber-attacks that could trigger load ramps. Additionally, the rapid pace of load integration with large loads' magnitude could negatively impact system resilience. Large loads must be considered in load-shed obligations, UFLS program design, and black start restoration."*

The Large Data Center Loads in 100's of MW or GW cannot be considered Demand Response assets and need to be integrated to the market and the grid such that they are part of the solution and not seen as the problem from a reliability and resiliency standpoint. The Flexibility that the Data Centers provide the grid is for reliability, while the Flexibility the Grid provides the Data Center is for faster time to market and better ROI. These need to match and that is where better OT integration with IT systems play a pivotal role.

CAISO Demand, AI Chatbot, and YouTube Streaming on Sep 1, 2025

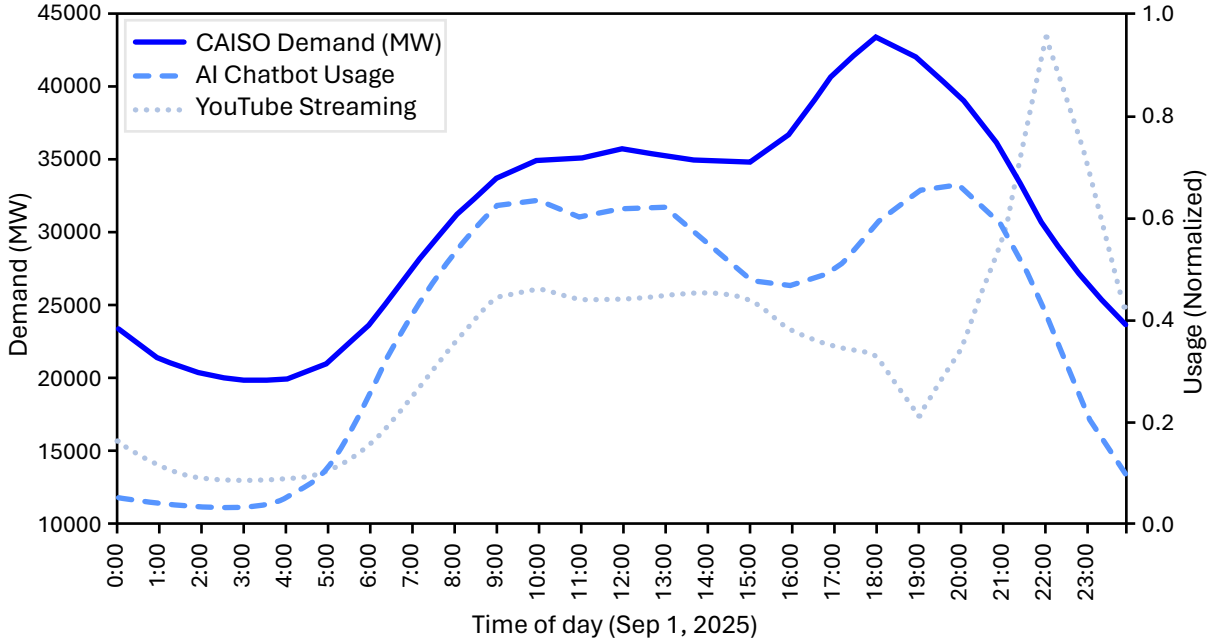

Figure 1: CAISO demand duck curve for September 1, 2025, superimposed on an illustrative normalized average Streaming and AI Chatbot usage

Figure 1 shows the CAISO demand duck curve for September 1, 2025, superimposed on normalized average Streaming and AI Chatbot usage. The Streaming and AI Chatbot usage is normalized and averaged from available approximations and is for illustration only. As the business usage of AI increases, one can assume the usage to peak during office hours for inference, while the current search load will morph into AI search loads during evening peaks.

This allows us to put the problem in the context of the shifting usage patterns and the load distribution as the usage of AI expands into the corporate and might end up having a different load profile compared to conventional IT loads.

2. Unlocking Stranded Capacity – Flexible Interconnection, Flexible IT / OT Orchestration

Flexible Interconnection[2] is a way to Interconnect a load, with the ability for the Utility to curtail / limit load during times of grid stress or certain times or days of the year when there is peak load. This service is offered in CA by some IOU's to Distribution connected loads like EV Fleet Charging stations, BESS and similar flexible loads that want to get connected to the grid quickly and start operations, without waiting for grid upgrades to be complete. The approach adopted is for the Utility to offer full limit for most of the year and during periods when the network is stressed, the loads are curtailed over IEEE 2030.5 and the loads sends data every 30 seconds to the Utility to provide visibility and fail-safe operations. Data Center loads are larger loads connected at transmission or sub-transmission level and can also achieve flexible interconnection if they can be modeled and the impact studied with the ability to control the load during times of extreme grid stress. Further these loads are integrated with the Utility EMS / DMS systems and can be monitored and controlled over DNP3 or ICCP or IEEE 2030.5.

Internal Flexible IT Orchestration[1] like the ones explored by Google implements Medium Voltage Power Plane (MVPP) and Priority Aware IT load capping to achieve oversubscription and utilize the allowed load to its maximum capacity and avoid stranded loads. The solution uses meter reading over Modbus Protocol with a latency of 2 seconds and can throttle IT loads over Remote Procedure Call (RPC) mechanism with dedicated agents. However, these approaches do not fully utilize the real-time OT capabilities already deployed in Utility Substations connected to the Data center. The other work referenced in[4], [5], [6] and the strategies listed in Table 1, all try to monitor power or control power usage by tasks and optimize the load limits to be within utility limits assuming compute based algorithmic limits are adequate. However, Utilities require more certainty than that, if these large data centers are to be much more than a demand response actor.

3. The Flexibility Gap: The OT vs. IT Divide

For most organizations, replicating hyperscale Data Centers vertically integrated success is a decade-long, multi-million-dollar effort. The primary barrier is the "Flexibility Gap" between the two distinct worlds managing the data center:

- **Operational Technology (OT)**: The world of power systems (switchgear, generators, UPS, protection relays). It prioritizes reliability, safety, and physical assets, operating on microseconds, milliseconds, or seconds, but focused on ensuring real-time performance and control capabilities of grid assets, using protocols like IEC 61850, DNP3, IEEE 2030.5 and Modbus.

- **Information Technology (IT)**: The world of virtualized workloads. It prioritizes performance, scalability, and software logic, operating in millisecond timescales, but offering best case performance without hard real-time characteristics.

These domains speak different languages and are managed by different teams, creating a fundamental disconnect between them that results in underutilized / stranded capacity and operational risk for the Data centers. The efforts so far in addressing these issues have revolved around addressing these problems in software or using AI but purely from a siloed approach[3]. [1] tries to integrate the OT into their oversubscription calculations but focuses purely on meter reading at 2 second latency and custom integrations with the Generator controllers over API's.

These efforts have moved the needle forward, but the chasm between OT and IT remains, and addressing these as an integrated problem and the Data Center as an integral grid asset would address the problem more holistically.

4. The Solution: An Integrated IT/OT Architecture for Flexibility in Data Centers

To achieve hyperscale-level efficiency data centers require a standards-based hard real-time platform that translates between OT and IT. This solution integrates deep utility protocol expertise with modern cloud-native architecture.

The foundation of this architecture is a smart automation controller (e.g., SYNC SAC) designed for the grid edge. It is ruggedized to IEC 61850-3 standards for harsh industrial environments and provides precision time synchronization (IEEE 1588 PTP). Running intelligent edge control software (Kalki.io Edge KIOE), this controller ingests data from facility assets using native protocols (IEC61850, DNP3 and Modbus) and acts as a single point of monitoring and control.

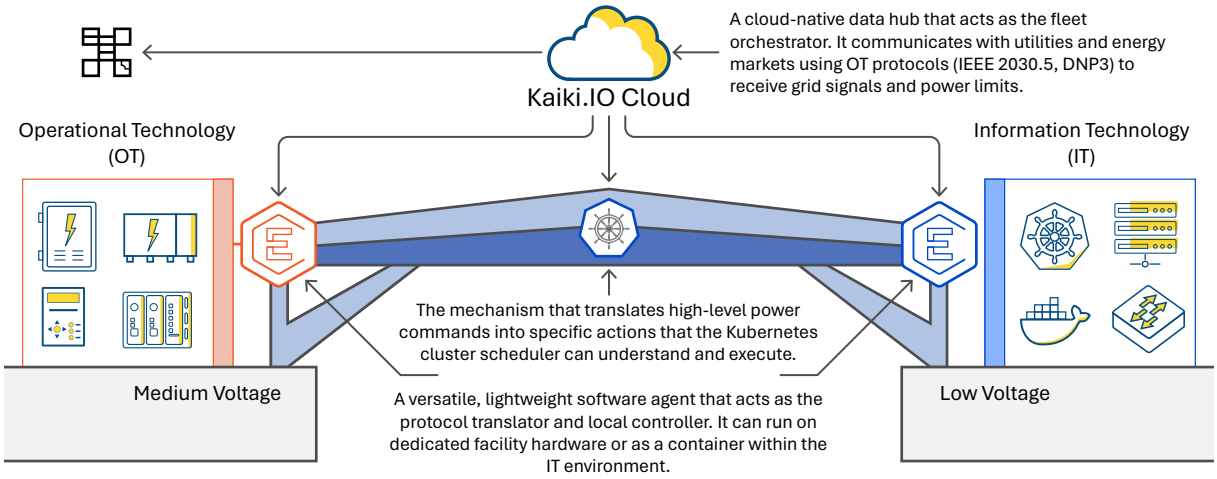

Figure 2: OT and IT Integration and orchestration to unlock efficiency and grid flexibility

Step 1: Real-time monitoring and control at the Facility Edge

The SYNC Automation Controllers (SAC) at the site level are connected to the dedicated IEC 61850 Process Bus and monitors Sample Values from the Merging Units providing real-time values from the Medium Voltage Network. Further it also receives power consumption data from site and rack meters over Modbus or DNP3 or IEC. To address sub-synchronous oscillation and mitigation, PMU over IEEE C37.118 or IEC6185090-2 / Sampled values are also supported by the SAC and the SYNC application stack with Oscillation Monitoring System (OMS). The KIOE edge software is deployed in the AI Racks and Servers monitoring the Processor Utilization, Clock Cycle, Temperature and Memory and CPU loading. These data are sent to the SAC over IEC 61850 GOOSE every 4-20ms. The SAC Controller provides an IEC 61131 real-time programming environment that uses all these input data to compute power score and ROI score and other optimization scores.

The Hardware Foundation (SYNC SAC)

Purpose-Built for the Grid Edge	
	Ruggedized & Reliable: Designed to IEC 61850-3 standards for substation and industrial environments.
	High-Performance: Available in X86 (Intel Core/Xeon) and ARM configurations to balance performance and efficiency.
	Precision Time Synchronization: Native support for IEEE 1588 PTP, essential for deterministic control and grid applications.

Software Intelligence (Kalki.IO Edge)	
	Runs as the primary control platform on the SYNC SAC.
	OT Protocol Mastery: Ingests data from facility assets (switchgear, meters, generators, UPS) using their native protocols (Modbus, DNP3, IEC 61850).
	Local Control: Capable of executing deterministic control logic (IEC 61131-3 via CODESYS) for high-speed local response.

Figure 3: SYNC Automation Controller

Customers can build bespoke optimization engines that meet their business needs while providing flexibility and conforming to real-time limits from the grid and participating as a responsible grid asset reacting to emergencies or participating in the market.

Step 2: Integrating with Utility and Power Markets to respond to Flexibility Limits, Grid Stress and participate in the market.

The AI Data Center is a large load (in hundreds of MWs) and being a responsive participant in the grid that can react and respond to the grid s flexibility requirement for secure and reliable operations, opens up more power capacity immediately rather than wait for grid build out to get the full load limits requested. Further with local backup generation at the data center that is dispatchable, it allows for improved ROI that optimizes for the capital investment and best monetary return for the asset, rather than assuming the only return possible is from AI / IT Load.

The Controller with IEC 61131 engine can be programmed together with Kalki.IO Edge and Cloud to manage Ramp-up / Ramp-down, Voltage and Frequency Ride throughs to support the grid during disturbances and ensure that such large loads disconnecting does not result in cascading grid failure or other reliability issues.

Kalki.IO Cloud and SYNC Controllers allow for connection with Utility EMS over ICCP / TASE.2, IEEE 2030.5, DNP3, Open ADR and REST APIs to allow for Flexible Load Limits and Schedules, as well as Participate in Power Markets through Demand Response Providers or Scheduling Coordinators.

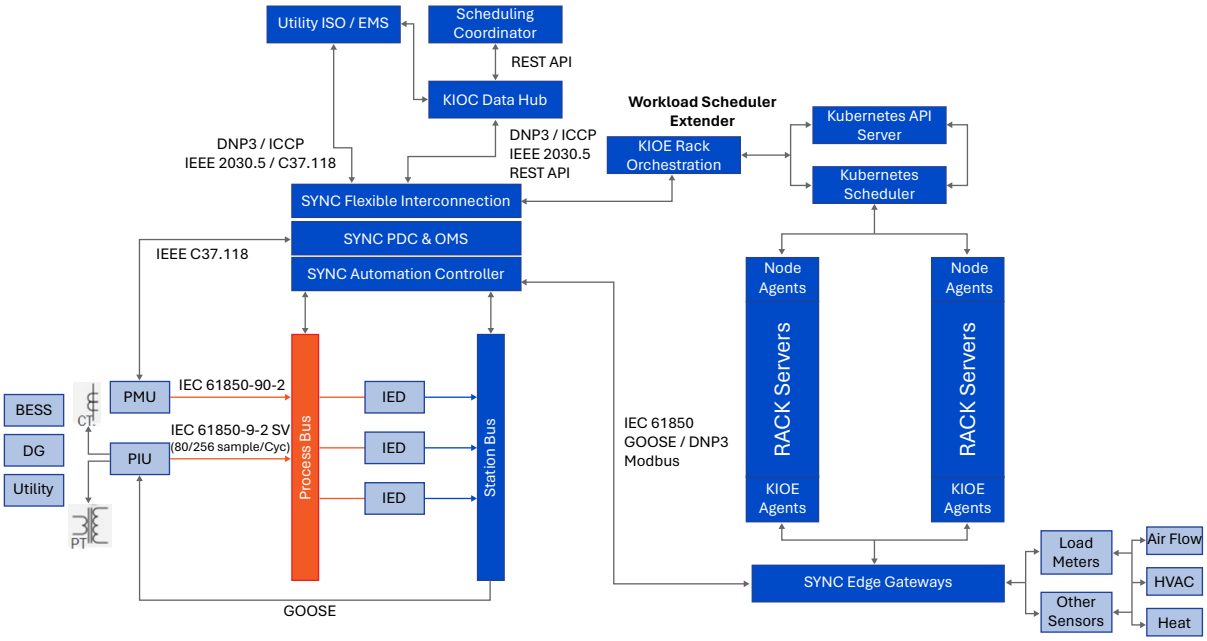

Figure 4: Integration architecture using IEC 61850 and Kubernetes Scheduler Extender for MV Power Plane

Flexible Interconnection allows the data center to operate at near full load most of the time, as long as it responds to load limits during peak seasons or grid stress. With Kalki.IO and the Flex Controller, you can implement and integrate with Utility and ISO EMS systems to receive and respond to limits.

Power markets allow the Data Center to participate in day-ahead and real-time power markets and utilize its backup generation and battery as grid assets and create arbitrage and better ROI for the investments. It averts additional investment in Peaker plants for resource adequacy planning.

These limits and market signals are used by the local site controller to decide on the best course of action and compute the real-time power limits and ROI values, that the Data Center Kubernetes scheduler can use to implement physical limits (Power) and Financial Limits (ROI) in its decision making to curtail loads. The data center operator can decide to what extent the financial limits are to be included in the control algorithm or whether it has to be managed separately and optimized separately. This allows the data center operator to look at actual ROI in a holistic manner and not purely from the load availability and compute scheduling alone, as defined below and detailed in Appendix.

$$ROI(t) = \frac{1}{CapEx} \sum_{\tau=1}^{T} \Big[\underbrace{U(\tau) \cdot \Delta_{token}(\tau)}_{\text{Effective Revenue}} - \underbrace{\big(U(\tau) \cdot P_{ppa} - \Theta(\Delta_{power})\big)}_{\text{Net Power Cost}} \Big] \, d\tau$$

Step 3: Translating Power Limits and Controls to Workload Actions

The key is not just monitoring power, but making the IT workload scheduler aware of real-time physical power constraints.

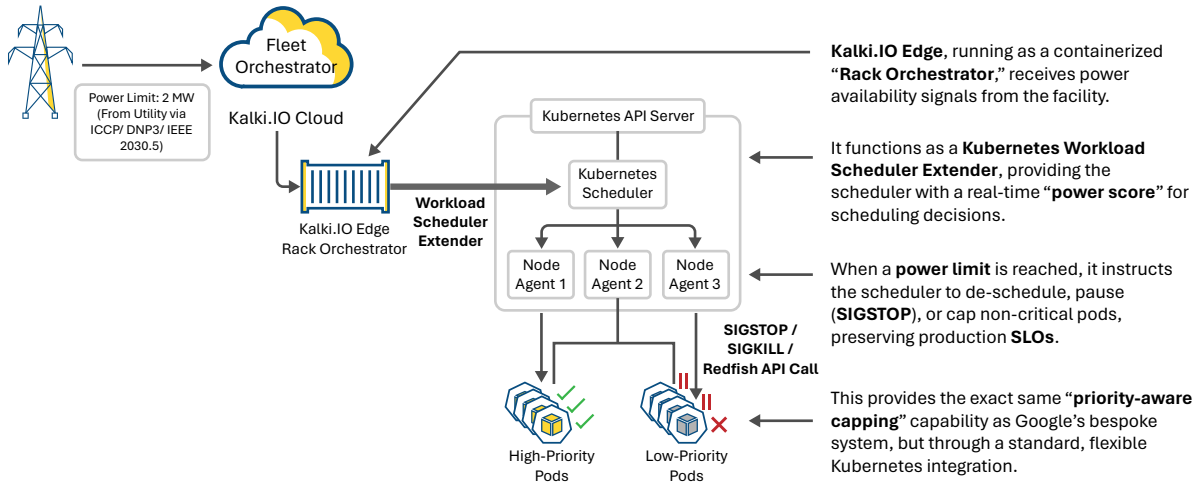

Figure 5: Kalki.IO Edge Rack Orchestration with Workload Scheduler Extender for Kubernetes

The critical innovation is making the IT workload scheduler "aware" of physical power constraints and the ROI optimization values on a continuous basis. The edge software functions as a **Kubernetes Workload Scheduler Extender**.

- It provides real-time "power score" and "ROI score" to the scheduler, every 20 milliseconds.
- When a power limit is reached, it instructs the scheduler to de-schedule, pause (SIGSTOP), or cap non-critical pods and use the "ROI score" and SLA considerations in identifying the right candidates.
- This implements a "priority and ROI aware capping" through standard Kubernetes integration, preserving production Service Level Objectives (SLOs).

5. Use Cases: From Grid Asset to Disaster Recovery

This cohesive platform enables end-to-end visibility and control, transforming how data centers interact with the grid provides ancillary services support and handle internal failures.

Use Case 1: Transforming the Data Center into a Grid Asset in a Flexible Interconnection Scenario where there is Grid Stress or Load Limitation due to seasonal high load, a utility sends a signal via ICCP/TASE.2 or DNP3 or IEEE 2030.5 or OpenADR to reduce load.

1. **Signal Received**: The cloud platform receives a "Reduce load by 2 MW" signal.
2. **Command Dispatched**: The system calculates a new power cap and updates the on-site controllers. It reviews the local generation or storage availability and load limits and computes the "power score".
3. **Workloads Curtailed**: The Kubernetes extender automatically pauses enough low-priority batch computing pods to meet the target "power score".
4. **Outcome**: Critical services remain unaffected, and the data center generates revenue while supporting grid stability.

Use Case 2: Ensuring High Availability During Failures During a utility outage where a backup generator fails to start:

1. **Detection**: The edge controller detects the reduced available capacity from the generator controllers within seconds.
2. **Proactive Capping**: A new power limit is communicated to Kubernetes.
3. **Shedding**: Low-priority workloads are shed *before* the remaining generators are overloaded.
4. **Outcome**: A full data center is averted. This software-defined response replaces blunt hardware fail-safes (opening breakers) with a graceful, intelligent response.

Use Case 3: Participating in Power Markets during peak demand - Case for resource adequacy planning - During a peak summer season where the utility grid is stressed and power market prices are very high, switch to battery or generator or shed load and sell power:

1. **Detection**: The Data Center through scheduling coordinators and Kalki.IO Cloud participates in the power market and bids to sell power at very high prices or contractually agree to shed load for a good reward. If the bids are accepted, Kalki.IO sends signal to edge controller of load reduction or power. On the other hand, at low power prices, it opportunistically purchases power and fast tracks its model training requirement.
2. **Proactive Load Curtailment or Power Export**: The controller sends signal to the BESS and Backup Generator to export power and if required a new power limit is communicated to Kubernetes.
3. **Shedding**: Low-priority workloads are shed and BESS and Generators are dispatched.
4. **Outcome**: The Data Center participates in the grid as a Demand Response or / and Generation entity providing valuable service to the grid for significant returns.

Conclusion

The future of data center infrastructure is grid-aware, integrated, and software-defined. By adopting a modular architecture that bridges the OT/IT divide, operators can unlock stranded capacity, increase Data Center capacity, increase Utilization and ROI. This approach not only lowers capital expenditures and cost-per-watt but also transforms the data center into a flexible energy asset capable of generating new revenue streams and weathering facility failures with intelligent resilience. This OT/IT integration and optimization approach literally furthers data centers' vision.

Reference/Citations

[1] Data Center Power Oversubscription with a Medium Voltage Power Plane and Priority-Aware Capping, Varun Sakalkar and Vasileios Kontorinis and David Landhuis and Shaohong Li and Darren De Ronde and Thomas Blooming and Anand Ramesh and James Kennedy and Christopher Malone and Jimmy Clidaras and Parthasarathy Ranganathan, 2020, https://dl.acm.org/doi/abs/10.1145/3373376.3378533, Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems

[2] OpFlex Pilot Report, PG&E, February 28, 2025, https://www.cpuc.ca.gov/-/media/cpuc-website/divisions/energy-division/documents/rule21/smart-inverter-working-group/pge_opflex_report_2025.pdf

[3] Colangelo, P., Coskun, A.K., Megrue, J. et al. AI data centres as grid-interactive assets.Nat Energy (2025). https://doi.org/10.1038/s41560-025-01927-1

[4] Unlock the AI performance you need: Introducing managed DRANET for A4X Max on GKE, October 28, 2025, https://cloud.google.com/blog/products/networking/introducing-managed-dranet-in-google-kubernetes-engine

[5] The Kubernetes Network Driver Model: A Composable Architecture for High-Performance Networking}, Antonio Ojea, 2025, https://arxiv.org/abs/2506.23628

[6] Y. Li et al., "A Scalable Priority-Aware Approach to Managing Data Center Server Power," *2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, Washington, DC, USA, 2019, pp. 701-714, doi: 10.1109/HPCA.2019.00067.

[7] A. Alvarez de Sotomayor, D. Della Giustina, G. Massa, A. Dedè, F. Ramos, A. Barbato, IEC 61850-based adaptive protection system for the MV distribution smart grid, Sustainable Energy, Grids and Networks, Volume 15, 2018, Pages 26-33, ISSN 2352-4677, https://doi.org/10.1016/j.segan.2017.09.003.

[8] Characteristics and Risks of Emerging Large Loads, NERC Large Loads Task Force White Paper, July 2025, https://www.nerc.com/globalassets/who-we-are/standing-committees/rstc/whitepaper-characteristics-and-risks-of-emerging-large-loads.pdf

Appendix 1: ROI and Power Score Illustration

A. Physical Constraints (Φ_{phys})

L_{lim}:	Scheduled Load Limits (Maximum load that the facility should draw from the grid for a given interval or schedule, as sent over standard OT protocols like ICCP, DNP3 or IEEE 2030.5)
S_{grid}:	Grid Stress Signals that could be in the form of Demand Response signals or Emergency Curtailment Limits sent over standard OT protocols like OpenADR, IEEE 2030.5, DNP3 or ICCP, or through REST API
P_{backup}:	Backup Power Availability at the site to respond to grid stress or curtailment schedules. Linked to SOC of the battery, ramp rates and DG synchronization times and UPS limits

B. Contractual & Performance Constraints (Φ_{ops})

α_{sla}:	Service Level Agreement (Required uptime / availability %)
A_{server}:	IT Server Availability (Actual hardware uptime)
η:	Efficiency (Power Usage Effectiveness or conversion efficiency)

C. Economic Deltas (Δ)

$P_{mkt}(t)$:	Spot Market Price of Power
P_{ppa}:	Power Purchase Agreement Price
$\delta(t)$:	Hardware Depreciation rate. The rate will be impacted by new generation of processors that has higher efficiency (processing or power usage)
R_{perf} :	Performance Ratio (Current Generation vs. Latest Generation processor efficiency)

Effective Utilization Function *U(t)*

$$U(t) = \min\left(L_{lim}, \frac{A_{server} \cdot \eta}{\alpha_{sla}}\right) \times \mathcal{H}(S_{grid}, P_{backup})$$

Where H is a switching function Grid Stress S_{grid} is high, the system curtails unless P_{backup} backup power is sufficient.

We define the value drivers as "Deltas" Δ, representing the net advantage or effective yield.

Power Price Delta (Δ_{power}) This represents the arbitrage opportunity or cost variance.

$$\Delta_{power}(t) = P_{mkt}(t) - P_{ppa}$$

Token Price Delta (Δ_{token}) This represents the effective revenue yield per unit of power, adjusted for hardware obsolescence and relative performance.

$$\Delta_{token}(t) = P_{token}(t) \cdot [R_{perf} \cdot (1 - \delta(t))]$$

The Return on Investment over time ***t*** is the summation of the **Net Economic Yield** (driven by the Deltas and limited by Utilization) divided by the *CapEx*

$$ROI(t) = \frac{1}{CapEx} \cdot \sum_{\tau=1}^{T} \Big[\underbrace{\big(U(\tau) \cdot \Delta_{token}(\tau)\big)}_{\text{Effective Revenue}} - \underbrace{\big(U(\tau) \cdot P_{ppa} - \Theta(\Delta_{power})\big)}_{\text{Net Power Cost}} \Big] d\tau$$

Effective Revenue ($U \cdot \Delta_{token}$): You only generate revenue when the system is utilized (U), scaled by the depreciation-adjusted token value (Δ_{token}).

$\Theta(\Delta_{power})$ **(The Demand Response Term)**: This represents the strategic value of power.

- If Δ_{power} is positive (Market Price > PPA) and you curtail usage ($U \to 0$), you may gain "credits" or avoid costs, effectively adding to ROI.
- If Δ_{power} is negative, you simply pay the PPA rate.