TECHNICAL PAPER

PRACTICAL AND EMERGING APPROACHES TO A MORE EFFICIENT ELECTRIC GRID

Emerging solutions for Grid Efficiency and Speed to Power

Document number	PR-CH-2608-0001
Revision	Rev A.1
Date	17 August 2026
Status	Complete draft — Rev 1 incorporates Internal Review
Subject document	International Rethinking Electricity Forum (IREF) — "Practical and emerging approaches to a more efficient electric grid"
Scope	An engineering-led review of practical and emerging pathways to deliver firm, reliable and scalable power to dense, time-critical loads.
Prepared for	International Rethinking Electricity Forum (IREF), Le Méridien, Paris — 25 August 2026
Prepared by	Andre Canelhas — Director System Engineering

Revision history

Rev	Date	Description	By
A.1	16 August 2026	First external issue — technical paper prepared for IREF session 2026 Paris 25 August 2026	AC

Contents

1. Introduction and Context
2. From Deterministic Machine to Triple-Stochastic System
    2.1 The machine we built
    2.2 The first destabilization — the renewable transition
3. Quantifying the Challenge
    3.1 The second shock: a new kind of load
    3.2 The numbers that define the crisis
    3.3 The consequences
4. The Anatomy of the Long Queue
    4.1 What the queue is, and why it exists
    4.2 Why it has broken down
    4.3 The large-load twist
    4.4 Reform, and why it is not enough
5. Three Simultaneous Perspectives
    5.1 The Electric Power Industry (Utilities and grid operators)
    5.2 The Hyperscalers (data-centre developers)
    5.3 The OEMs (equipment manufacturers)
    5.4 The misalignment in one view
6. Why No Single Actor Can Solve It Alone
    6.1 The unilateral dead-ends
    6.2 The circular dependency
    6.3 The shared problem is the allocation of roles, responsibilities and risk
    6.4 The implication
7. Engineering-Led AND Solutions
    7.1 Co-location of generation, storage and load
    7.2 Grid-forming inverters and modular BESS
    7.3 Strategic HVDC and FACTS
    7.4 Flexible large loads as active grid assets
    7.5 Wide-area coordination and multi-timescale control
8. Practical Pathways: Bridge-to-Grid Architectures and the Speed-to-Power Playbook
    8.1 The bridge-to-grid principle
    8.2 The Speed-to-Power playbook
    8.3 The commercial and risk architecture
    8.4 From pathway to method
9. From Speed to Power to a Low-Entropy Industrial Future
    9.1 From scarcity-managed to abundance-engineered
    9.2 The compute–power symbiosis
    9.3 Low-cost electricity as the industrial foundation
    9.4 The AND future realised
10. Conclusions and Call to Action for IREF
11. References

Appendix 1 — The Power–Compute Interface

A1.1 The interface is the missing design layer
A1.2 A DC hub from standard equipment
A1.3 The 800 VDC path from facility to rack
A1.4 Storage at every timescale
A1.5 The bus itself is a component: connecting PV and storage at ultra-low inductance
A1.6 Shaping the load at its source
A1.7 Safety and reliability by defence in depth
A1.8 What the interface obliges
A1.9 The spectrum of the load, and why it matters beyond the fence

1. Introduction and Context

Scope and limitations of this paper. This paper is an engineering concept paper. It is not a feasibility study, a network development plan, a grid-code proposal, or a design specification for any particular site. The architectures it describes are candidate pathways whose applicability depends on local resource, network strength, market, permitting and reliability conditions; each requires site-specific RMS, electromagnetic-transient, harmonic, protection and availability studies, and a project-specific commercial and risk assessment, before it is adopted. Where a quantity is stated, the boundary of that quantity is given with it. Where a direction is recommended, the study that would confirm or refute it is named. The architectures described are open and multi-vendor: nothing in this paper is proprietary to, or dependent on equipment from, any single supplier, including the author's own firm; named products appear as illustrative examples only. The propositions here are offered to be tested, not to be taken on assertion.

The International Rethinking Electricity Forum (IREF) is convened under a single, unambiguous imperative: Speed to Power — the elapsed time required to deliver firm, reliable, scalable energised megawatts to a specific site.

In this context, Grid efficiency is not classical technical efficiency measured in reduced losses or higher conversion rates. It is a system-level concept: the ability of the electric power system to deliver firm, reliable, scalable megawatts to dense, time-critical loads — especially GIGA-scale AI and compute facilities — faster, at lower societal cost, with stronger reliability, and with sustainability, simultaneously.

The event's operating logic is therefore clear:

Metric: Speed to Power

Economic driver: Lowest-cost energy

Demand shape: GIGA-scale compute

Operating frame: AND solutions — approaches that achieve speed, affordability, resilience and sustainability together rather than forcing trade-offs between them.

A more efficient electric grid, under this definition, is one that can hyper-scale the delivery of dependable power without the traditional bottlenecks of multi-year interconnection queues, socialised upgrade costs, progressive loss of system strength, or false choices between speed, cost and reliability.

This rethinking of the electricity industry did not arise in isolation. It is the direct engineering response to a conclusion reached by the North American Electric Reliability Corporation's Reliability Issues Steering Committee in its 2025 ERO Reliability Risk Priorities Report [1]:

> *"Substantive simultaneous change in all system dimensions (resource, grid, load), along with increasing complexity in the interactions between them, requires rethinking traditional system planning and operating approaches."*

Grid Transformation is identified as the overarching risk profile. Traditional capacity-based reliability constructs, planning models, operating paradigms and criteria are no longer sufficient. What is now required is:

- More system margin to accommodate uncertainty
- More diverse resource, grid and operating options

- More comprehensive studies and assessments
- More coordination among all parties
- More speed in implementing necessary measures

I approach this subject after four decades of practical work in grid design, power electronics, High-Voltage Direct Current (HVDC) and Flexible AC Transmission Systems (FACTS) at Hitachi Energy (formerly ABB), Siemens and GE Vernova Alstom. That experience has taught me that the physical realities of the grid — system strength, inertia, voltage stability and the interaction of power-electronic devices — cannot be wished away by policy statements or commercial ambition. They must be engineered.

This report therefore examines the tension between explosive AI and data-centre demand and the grid's present inability to deliver firm capacity at the required speed and with adequate stability margins. It analyses the problem simultaneously from three perspectives that must be reconciled:

1. The Electric Power Industry (Utilities), charged with maintaining reliability for all consumers while absorbing load growth of unprecedented scale and density.
2. The Hyperscalers, for whom Speed to Power is now a competitive necessity measured in months rather than years.
3. The Original Equipment Manufacturers, constrained by multi-year lead times and the need for demand certainty before manufacturing capacity can be expanded.

The solutions proposed are deliberately engineering-led. They centre on the co-location of generation, storage and load; the deployment of grid-forming inverters and modular battery energy storage systems as providers of synthetic inertia and system strength; and the strategic use of HVDC and FACTS to isolate disturbances and restore stability margins. These are the practical AND solutions that convert the NERC diagnosis into the Grid efficiency that Speed to Power demands.

The remainder of this paper develops each of these elements in turn and concludes with a clear call to action for the industry leaders gathered at IREF.

2. From Deterministic Machine to Triple-Stochastic System

2.1 The machine we built

The power system is the largest and most tightly coordinated machine humanity has ever engineered. Unlike gas, coal or water, electricity cannot be economically stockpiled at grid scale; it is generated, delivered and consumed in the same instant, balanced in real time across an entire continent. For the better part of a century it rested on a single, deterministic premise: a small number of large, dispatchable central stations serving a broad, diversified and gradually growing demand, with transmission and distribution planned around predictable flows from the centre outward. For well over a century the large loads at the receiving end were themselves motor-driven, connected directly to the alternating-current grid with no power-electronic conversion stage, and the behaviour of the whole system was shaped by rotating machines.

Because there was no storage, supply and demand had to meet instantaneously, and the physics that held them together was rotational. Every synchronous generator on the system stores kinetic energy in its spinning mass; the aggregate of that stored energy — expressed as an inertia constant H, typically in the range of about two to four seconds for large machines — is what resists sudden change. When a generator trips or a large load

switches, that inertia arrests the rate of change of frequency in the critical first instants, buying the seconds the control systems need to respond. Alongside inertia, the same synchronous machines silently provide the services the grid depends on: short-circuit strength, fault current to operate protection, voltage regulation and damping. None of this was ever itemised on a bill, but all of it was essential.

A machine of continental scale is inevitably exposed — to the elements, and to the electrical and mechanical faults that will occur. The industry's answer was redundancy and fast-acting protection: detect the fault, remove the faulted element from service within a few cycles, and restore it at high speed, so that the disturbance is contained and the rest of the system rides through. This philosophy has delivered remarkable availability for more than a century. But it is, at its core, a trip-and-restore philosophy — one that accepts brief interruptions as the reasonable price of protecting the whole. As we will see in Section 3, that assumption is precisely the one the newest loads will not accept.

2.2 The first destabilization — the renewable transition

The deterministic machine has already been disturbed once, by the energy transition. As coal, oil, gas and the older nuclear stations retire, they are being replaced by intermittent solar photovoltaic and wind generation. Crucially, these resources do not connect through synchronous rotating machines; they are interfaced to the grid through power-electronic inverters — the inverter-based resources, or IBRs. As their share grows, the inertia and system strength that the synchronous fleet provided as a by-product begin to decline, and the grid becomes progressively harder to hold stable. This is the substance of what the North American Electric Reliability Corporation now ranks as the first of its critical risk profiles, Grid Transformation: substantive, simultaneous change in resource, grid and load.

To see why this matters, it helps to recall how well-behaved the old system was. Generation was deterministic and controllable. Demand was stochastic — no one can predict exactly when a household, a shop or a factory will draw power — but in a familiar, aggregated, slowly varying way that tracked economic activity — a single dominant source of randomness, on the load side, that planners had decades of experience managing.

The transition removed that comfort. Generation from wind and solar is now itself stochastic, driven by weather rather than dispatch. Add to this a large and growing volume of distributed generation embedded throughout the network, and the power flows themselves — their magnitude and, increasingly, their direction — become stochastic as well. The system that once had one well-understood variable now has three that move at once: generation, load and flow. This is the triple-stochastic grid, and the industry is still, understandably, learning to operate it.

It is into this already-unsettled system that a second shock is now arriving — larger, denser and faster than anything the transition has yet produced. That is the subject of the next section.

3. Quantifying the Challenge

3.1 The second shock: a new kind of load

The load now arriving is unlike anything the grid has absorbed before. GIGA-scale artificial-intelligence and compute campuses concentrate hundreds of megawatts — soon gigawatts — at a single node, run around the clock, and are commissioned on technology cycles measured in months rather than the multi-year horizons on which the system is planned. This is the second shock, and it lands on a grid still adapting to the first.

It is tempting to treat these facilities as simply very large inverter-based resources. That is a mistake. A hyperscale data centre is power-electronics-interfaced, but it differs from an IBR in its technologies, configurations, controls, performance and behaviour. There is no single “data-centre technology”: the sector is a fast-moving collection of architectures, and the pace of change is arguably several times faster than the industry experienced with wind and solar, with important new technologies emerging over periods as short as twelve months. The practical consequence for the power engineer is that these loads must be modelled explicitly — increasingly with detailed electromagnetic-transient (EMT) models that are more complex than those built for IBRs — and cannot be represented by borrowing IBR assumptions [2].

These loads also demand a level of continuity the classical grid was never designed to guarantee. Hyperscale operators require availability of the order of 99.999% — “five-nines” — which allows a total interruption budget of roughly 5.26 minutes across an entire year. The reason is at once economic and physical: an AI training run is a single synchronised computation spread across tens of thousands of processors, an enormous block of capital committed to a task that a momentary disturbance can corrupt or reset. Worse, these loads do not wait to be acted upon; NERC has warned that large computational loads can shed or oscillate within seconds of a disturbance, faster than conventional real-time control can respond [1]. The trip-and-restore philosophy of Section 2 — remove the faulted element, restore it at high speed, accept the brief interruption — is fundamentally incompatible with a customer that cannot tolerate the interruption in the first place.

This points to a different way of thinking about resilience. The compute load is not a passive block of demand; it already carries redundancy inside it. A modern AI rack is fed by multiple power units in parallel, with enough headroom that the loss of an individual unit — even a whole shelf — is absorbed within the rack, which continues running while the faulty part is replaced. Only the loss of the entire rack, through a total power or cooling failure, actually interrupts the work; at that point the job is recovered on other healthy racks. Redundancy at the small scale, recovery-by-restart at the large scale — that is the nature of the load.

One productive way forward is therefore to treat the Power Component and the Compute Component as parts of a single system, matching the form of redundancy to the level at which failures actually occur, and keeping the different timescales clearly in view: on-rack energy storage smooths sub-second power swings, while facility UPS and batteries ride through seconds to minutes until on-site generation takes over.

Five-nines availability itself may usefully be reframed. Rather than an annual downtime budget tallied after the year has passed, it can be approached as a target held continuously, cycle by cycle. An optimisation loop running every five to fifteen minutes that looks ahead — at solar output and cloud forecast, battery state of charge, exchange with the grid, rack demand, coolant-outlet temperature and the parasitic draw of cooling — and works to preserve the required redundancy and operating margin on every cycle offers one concrete path. Reliability then becomes something actively managed over a rolling horizon.

On this footing, a natural next step — and a priority for further work — is to explore closer cooperation between the layers: the possibility of acting on an early indication of local power depletion so that work can be moved pre-emptively, before an interruption occurs, rather than only recovering after it, as the power industry has been practising with loads historically. That direction points toward a more integrated form of Power–Compute resilience.

3.2 The numbers that define the crisis

The scale is best seen first at the global level. The International Energy Agency projects worldwide data-centre electricity consumption to roughly double to around 945 TWh by 2030, with AI servers driving almost half of the net increase [3]. The United States is the leading edge of that curve: national data-centre demand is on track to more than double by 2030, and EPRI estimates data centres could consume between 9% and 17% of all US electricity by the end of the decade [4]. CIGRE frames the same phenomenon in capacity terms — global data-centre capacity roughly doubling from about 103 GW in 2025 to 200 GW by 2030, with around half of that workload dedicated to AI [2].

Two features make these totals more dangerous than they appear. The first is density and concentration: AI racks can reach power densities up to ten times those of conventional facilities, and the load piles up in a handful of locations rather than spreading across the network. In Virginia, data centres already draw roughly a quarter of the state's electricity and could reach between 39% and 57% by 2030 — so a benign-sounding national figure of “ten percent” can translate into local grid areas where data centres are the dominant load class [4]. The second is that this is not a US-only story. The same avalanche is now building in Europe, and fast-growing compute demand is following in India, Brazil, Canada and elsewhere. The United States is simply first and largest; the engineering problem it exposes is global.

The most important number, however, is the one that is almost never announced: how much of this capacity is actually deliverable, at a specific site, by a specific date. Announced megawatts are not usable megawatts.

Indicator	Headline figure	What it actually delivers
US interconnection queue (end 2024)	~1,400 GW of generation plus ~890 GW of storage seeking connection	Only ~13% of requests from 2000–2019 had reached commercial operation by end-2024; the median wait has doubled to more than four years.
Planned US capacity additions to 2030 (RAND, projected)	~300 GW nameplate (151 GW front-of-meter + 149 GW behind-meter)	Only ~82 GW of net available capacity after completion rates, retirements and reliability contribution are applied.
The metric that matters	Nameplate / announced pipeline	Accredited, firm, deliverable capacity at the load node by the required energisation date — i.e. Speed to Power.

Sources: Lawrence Berkeley National Laboratory [5] (interconnection-queue figures, as of end-2024); RAND [6] (net available capacity, projected to 2030).

3.3 The consequences

If firm capacity cannot be delivered through the grid at the required speed, it has to be built where the load is. Co-location of generation, storage and load therefore moves from a preference to a leading candidate, and in the most constrained cases to the only route that meets the schedule. Whether it is also the least-cost or lowest-risk route is site-specific — turning on the local resource, the duty cycle, the energy duration required, land, water, fuel, emissions and the permitting position — and it should be screened against grid-first and hybrid alternatives rather than assumed. In practice this produces partially grid-connected facilities — on-site

generation carrying the bulk of the demand, with a smaller grid connection retained for auxiliary supply, back-up and redundancy — a configuration CIGRE now treats explicitly in its work on large loads [2].

These hybrids do not fit the categories the industry is built on. A facility that generates, stores and consumes behind a single connection point is neither a generator, nor a load, nor a battery in the terms used by connection processes, energy-management systems and markets. The administrative borders between generation, transmission and distribution — and the market granularity that assumed a slow, well-behaved, purely consuming load — begin to fade. What was a clean separation of roles becomes a continuum.

This raises the obvious question. If the capacity exists on paper and the need is urgent, why can the system not simply connect it? The answer lies in the mechanism that now governs access to the grid — the interconnection queue — which is the subject of the next section.

4. The Anatomy of the Long Queue

4.1 What the queue is, and why it exists

Before any new power plant — and, increasingly, any large load — can connect to the transmission system, it must pass through the interconnection queue. The principle is sound: each new entrant is studied for the effect it will have on the shared network, the upgrades its connection will require are identified, and the cost of those upgrades is allocated before a connection agreement is signed. This is how the system has historically protected reliability and shared network costs fairly among those who cause them. The queue is not obstruction; it is the grid's admission process, and for a slower, generation following market-growth led world it worked.

4.2 Why it has broken down

That process is now overwhelmed. By the end of 2024 roughly 1,400 GW of generation and a further 890 GW of storage — close to twice the entire installed generating capacity of the United States — were waiting in interconnection queues, the median time from request to commercial operation had roughly doubled to more than four years, and only about 13% of the capacity that entered the queue between 2000 and 2019 had reached operation by the end of 2024 [5]. The cause is partly self-inflicted. Because entry was cheap and speculative, projects crowded in; because the studies are serial and interdependent, each withdrawal forces those behind it to be re-studied, and withdrawals cascade. Uncertainty over who pays for network upgrades — often assigned late, and in large, lumpy amounts — triggers still more withdrawals. The queue congests itself. Set this multi-year process against a compute campus financed and built in months, and the mismatch of Section 3 becomes concrete: the gate cannot open at the speed the load arrives.

4.3 The large-load twist

The queue compounds the problem in a way its designers never intended, because it was built to admit generators, not gigawatt-scale power-electronic loads. Those loads now enter the same congested process, and there is as yet no settled framework for studying them, nor for coordinating among the parties responsible — grid operator, developer and equipment supplier — whose timelines and incentives do not align [2]. Generation and load are still assessed through separate processes even as the co-located, partially grid-connected facilities of Section 3.3 fuse the two behind a single connection point. This is the machinery behind the plain statement made earlier: announced megawatts are not usable megawatts. The gap between the two is, quite literally, the queue.

4.4 Reform, and why it is not enough

The problem is recognised, and reform is under way. In the United States, FERC Order No. 2023 replaced the old first-come, first-served handling with a first-ready, first-served cluster study, raised the financial commitments required to enter and remain in the queue, and imposed firm study deadlines backed by penalties [7]. Operators are also experimenting with connect-and-manage approaches that admit projects sooner subject to curtailment, and with grid-enhancing technologies — dynamic line rating, advanced power-flow control and topology optimisation — that draw more capacity from existing wires. Each of these helps. None closes the gap. They make a generation-era admission process faster and fairer; they do not by themselves turn it into something that can energise a gigawatt of dense, always-on load, at a chosen site, on a compute schedule. Others have pushed the logic further, reframing speed to power as, at root, a risk-allocation problem: rather than asking the grid to underwrite every new load, the customer brings its own firming — storage, on-site generation and controllable, dispatchable demand — and receives faster, non-firm access in return, as in ERCOT's provisional controllable-load-resource model [8], [9].

This reckoning is not confined to the United States. The same wall is being met, by different instruments, across every major market:

Jurisdiction	Reform response to the large-load surge	Instrument
United States	FERC Order No. 2023 — first-ready, first-served cluster study, higher deposits and firm study deadlines [7]	Queue-process reform
Brazil	PNAST (Dec 2025) — competitive “access seasons” replace the chronological queue [10]	Access-allocation reform
European Union	ENTSO-E — national connection requirements for large loads; data centres framed as dual demand / flexibility assets [11], [12]	Connection standards + flexibility
India	CEA — projected data-centre demand mandated into state resource-adequacy plans [14]	Anticipatory planning
Canada (Alberta)	AESO — interim 1,200 MW cap allocated pro-rata (≈16 GW requested vs ≈12 GW peak) [13]	Cap-and-allocate
Singapore	Green Data Centre Roadmap (2024) — capacity released through calls for application and efficiency gains after an earlier build pause [15]	Capacity allocation + efficiency
Australia	AEMC draft rule (2026) — new connection and technical-performance standards for data centres as grid-risk assets [16]	Connection standards

The instruments differ; the limit is common — each makes grid access faster or fairer, but none, on its own, delivers firm capacity at the speed the load demands.

That will take more than a faster queue: it will take a different division of responsibility and a different engineering approach — which is why, as the next sections argue, no single actor, reforming only its own domain, can solve this alone.

5. Three Simultaneous Perspectives

The deadlock over speed to power is often blamed on obstruction, but that misreads it. Each of the three principal actors is behaving rationally within its own mandate; the difficulty is that their metrics, their timescales and the risks they fear do not align. The problem must therefore be read from all three sides at once.

5.1 The Electric Power Industry (Utilities and grid operators)

The utility's first duty is to every existing consumer: to keep the system stable and the lights on, even as the synchronous fleet that once guaranteed that stability retires. And the load is not only large but accelerating: the adoption of AI as a productivity engine, and as the foundation of a widening data economy, is driving demand beyond even the most optimistic forecasts — faster than any planning cycle was built to absorb. It carries that duty inside a framework built for a slower age — multi-year planning, interconnection and rate-case cycles, and a thick layer of grid codes and permitting that exists, legitimately, to protect reliability and to stop the cost of network upgrades being socialised onto ordinary ratepayers. But that same framework is also red tape: procedurally heavy, sequential and slow to say yes. The operator's caution is not obstruction; it is the rational response of a body that will be blamed if a single hyperscale load trips or oscillates and takes part of the system down with it [1], and that may be left with stranded upgrades if a promised campus is cancelled. Rational caution, seen from the outside, reads as delay.

5.2 The Hyperscalers (data-centre developers)

For the hyperscaler the clock runs in months. Compute revenue depends on energising firm capacity quickly, at very high availability, while the underlying technology — chips, racks, power architectures — turns over faster than the industry has ever seen. That pace is itself a problem for the grid: connection standards and grid codes need a degree of stability to be written and enforced, yet the thing they must describe changes every twelve months. The answer is not to freeze the technology but to agree stable, performance-based standards — what the load must do at the connection point — rather than device-by-device rules obsolete on arrival.

There is a deeper shift here, foreshadowed in Section 3: the data centre need not remain a passive load. With its on-site generation, storage and controllable demand it can behave as a power utility in miniature — a virtual power plant able to firm its own supply and, at times, support the grid around it, the dual role that system operators are beginning to formalise [12]. The hyperscaler's real opportunity is to embrace that role — to be dispatchable and transparent — rather than to demand firm power while refusing to share the risk of delivering it [8], [9].

5.3 The OEMs (equipment manufacturers)

The equipment makers hold the third corner. They supply the transformers, switchgear, power electronics, turbines, cable and battery systems on which everything else depends, and they face lead times measured in years. Understandably, they sell what they already build — and what their R&D road maps had predicted the market would need. Even where they had prepared for a grid moving beyond synchronous generators toward one dominated by inverter-based resources, including on the load side, the dense, gigawatt-scale, power-electronic, always-on campus was not on their radar; much of the existing product line needs adaptation. Equally understandably, they will not commit the capital to retool and expand until they see firm, certain demand — while buyers will not commit until they can see firm, deliverable power. That circular dependency is the OEM bottleneck.

But rigidity among the incumbents opens the door to fast, productised entrants. Tesla's Megablock, launched in 2025 as a factory-built, rapidly deployable grid-scale battery block with grid-forming capability, is exactly the kind of standardised, speed-to-power product — one of a broader emerging class of factory-built, grid-forming blocks — that a slow supply chain invites in [17]. The lesson is not that incumbents will be swept away, but that the market now rewards whoever can deliver adaptable, firm capacity fast — and that competitive pressure is itself part of the solution.

5.4 The misalignment overview

Actor	Metric that matters	Timescale	Binding constraint	Risk it fears
Utilities / grid operators	System reliability for all consumers	Years (planning and rate cycles)	Grid codes and permitting; no cost-socialisation; declining system strength	A hyperscale trip destabilising the system; stranded upgrades
Hyperscalers	Speed to Power at 99.999%	Months	Capital intensity; need for stable, performance-based standards	Losing the AI race by waiting for power
OEMs	A firm order book	Years (equipment lead times)	Demand certainty before retooling and expanding	Building for a bubble — or becoming the bottleneck

Three rational actors, three clocks that do not synchronise. No amount of good faith inside any one of them closes the gap — which is precisely why it cannot be closed by any one of them alone, the argument of the next section.

6. No Single Actor Can Solve It Alone

Section 5 described three rational actors whose clocks and risks do not align. The deeper point is that this misalignment is structural. It is not a failure of goodwill that better intentions inside any one domain could repair; it is a property of how roles, responsibilities and risk are currently distributed. No party, however diligent, and however far it reforms its own processes, can break the deadlock alone.

6.1 The unilateral dead-ends

Consider each in turn. The utility, acting alone, cannot simply grant speed. To do so it would have to relax the reliability obligations it owes every other consumer, or socialise the cost of network upgrades onto ratepayers who did not cause them. It can pre-build ahead of need, but then it carries the stranded-asset risk if the load never arrives. The queue reforms of Section 4 ease the process at the margin; they do not close the gap.

The hyperscaler, acting alone, can co-locate generation and storage, and can even island itself from the grid entirely. But full islanding forgoes the resilience a grid connection provides, concentrates every risk on a single balance sheet, and does nothing for the stability of the wider system — and no data-centre developer can build the transmission network it ultimately relies on.

The OEM, acting alone, will not commit the capital to retool and expand until it sees firm, certain demand — and it cannot conjure that demand into being. It can only respond to it.

6.2 The circular dependency

The result is a closed loop. The hyperscaler waits for firm, deliverable power before it commits; the utility waits for demand certainty — and for equipment — before it invests; the equipment maker waits for firm orders before it expands. Each actor's precondition is another actor's commitment, so each, quite reasonably, waits for one of the others to move first.

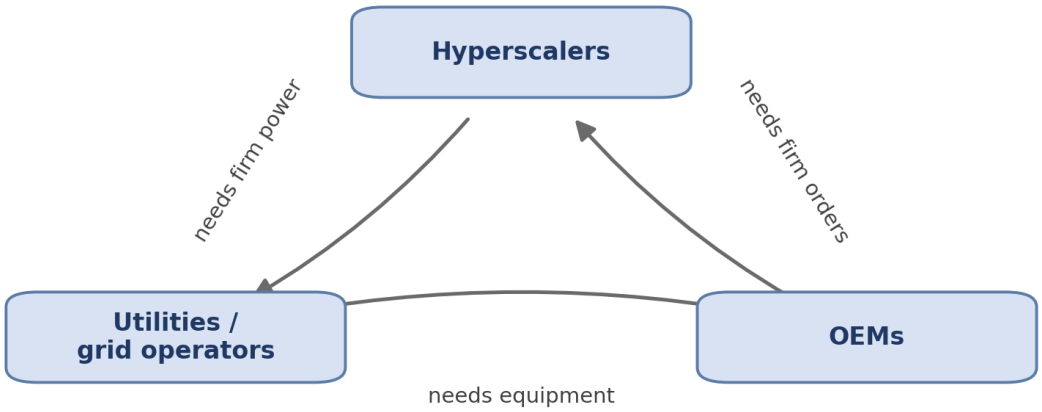

Figure 1 — The circular dependency: each actor's precondition is another actor's commitment.

6.3 The shared problem is the allocation of roles, responsibilities and risk

Beneath the loop lies a single question: who carries the risk of new capacity, and on what terms. Framed this way — as fundamentally a risk-allocation problem [8], [9] — the impasse becomes tractable, because risk can be re-allocated only by the parties acting together. What the situation demands is a new distribution of roles, responsibilities and risks among utility, developer and manufacturer — and, decisively, one that does not socialise the cost of serving these loads onto ordinary consumers. That constraint is not incidental; it is a founding principle of the Rethink Electricity frame set out in Section 1. The objective is the lowest possible cost of electricity, achieved through AND solutions — speed, affordability, resilience and sustainability together — not by forcing trade-offs between them, and not by quietly shifting the bill. It is also what the reliability and system-operator communities are themselves calling for: NERC's prescription is “more coordination among all parties” [1], while CIGRE and ENTSO-E identify the absence of a shared framework as the core gap [2], [12].

6.4 The implication

If no single actor can solve it — and if the answer is neither pure regulation, nor pure market, nor pure self-supply — then the resolution must be an integrated, shared-risk, engineering-led design: one that assigns each party the role it is best placed to carry, and prices risk where it truly sits. Suggesting an engineered solution is the task of the rest of this paper. The next section sets out the engineering-led AND solutions through which it appears to becomes real.

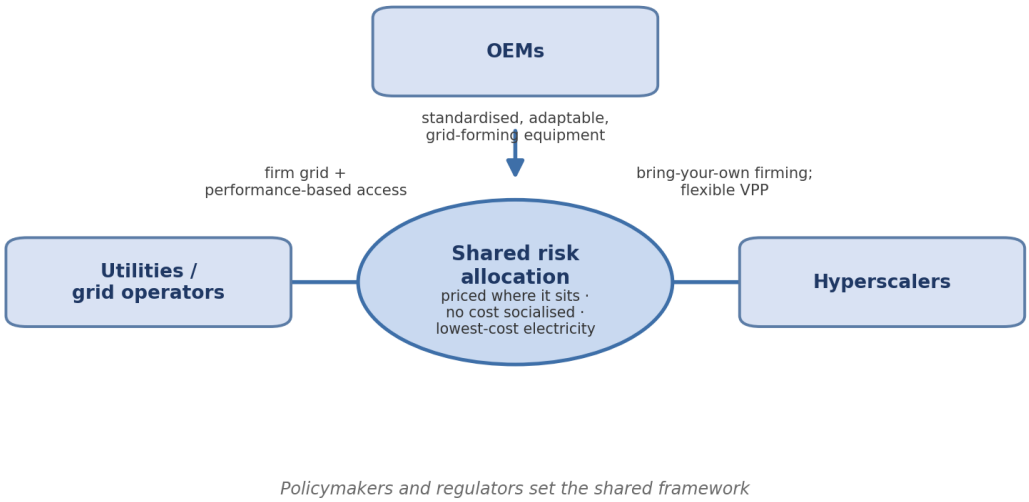

Figure 2 — Sharing responsibility and risk: each actor underwrites its best-placed role around a shared framework, with no cost socialised onto consumers.

7. Engineering-Led AND Solutions

The preceding sections make a demand of the solution: it must be engineering-led, it must reallocate roles and risk rather than socialise cost, and it must deliver speed, affordability, resilience and sustainability together. The individual elements below are established rather than speculative: each is mature, commercially available, or emerging at the scale intended here, and none waits on a technical breakthrough. Their integration is a separate question. Component maturity does not by itself establish an integrated architecture at gigawatt scale, where the controls, the protection philosophy, the operational interfaces, the permitting position and the commercial structure all have to be shown to work together; each element accordingly carries its own operating envelope and its own study requirements, noted as it arises. Their power lies in combination — recombined into a single integrated design, they offer a repeatable way of addressing the deadlock of Section 6, subject to that validation.

7.1 Co-location of generation, storage and load

The first move is to collapse the distance between generation and load — and it runs in both directions. In some cases the right answer is to build the generation and storage where the load is; in others it is exactly the opposite — to build the load where the generation can be, and so avoid transmission altogether. A megaload need not sit in a congested metropolitan grid. It can be placed where premium solar and wind resources exist without significant environmental, social or economic impact, and a siting decision made well resolves a whole cluster of problems at once: access to low-cost renewable energy, cooling, physical security of the assets, positive regional and economic development, and the build-out of the telecommunications infrastructure the compute itself requires. Understood this way, “sharing the same site” is an electrical statement rather than a literal one: generation and load need only sit within the same wide-area electrical system, close enough that power need not be wheeled across a continent to meet. Either way the campus is supplied primarily by on-site or nearby photovoltaics and battery storage, with firm backup and a minimal grid connection retained for auxiliary supply and redundancy — the partially grid-connected facility of Section 3.3, and the practical form of the bring-your-own-firming principle [8], [9]. The developer underwrites the risk of its own supply, at the lowest-

cost energy, without asking the wider system to carry it; and a well-chosen location turns the load's density from a liability into a regional asset.

7.2 Grid-forming inverters and modular BESS

Co-located supply is only firm if it is stable, and stability is precisely what the transition eroded. Grid-forming (GFM) inverters, paired with modular battery storage, can contribute much of it. Unlike the grid-following inverters that dominated the first wave of renewables, GFM inverters establish voltage and frequency themselves, providing synthetic inertia, fast frequency response, voltage regulation and a defined contribution to system strength — services the retiring synchronous fleet used to give away for free (Section 2). What they deliver is bounded, and the bounds matter. The services available depend on converter current capability, control design, protection philosophy and the energy standing behind the converter; grid-forming plant is not a general substitute for synchronous short-circuit contribution, nor for the fault-current levels on which existing protection was set and graded. Which services can be relied upon, over what voltage and frequency envelope and with what current limits, has to be established by electromagnetic-transient and protection-coordination study for the network in question. Delivered as factory-built, rapidly deployable blocks — Tesla's grid-forming Megablock is one example [17] — they also match the speed the load demands. Grid-forming inverters plus storage are, on present evidence, the principal means by which a power-electronic grid feeding power-electronic loads can be held together — provided the protection and stability case is made for each system [18].

7.3 Strategic HVDC and FACTS

Where power must still be moved — from remote, low-cost generation to a load centre, or between systems that cannot be synchronised — the answer is to use AC and DC transmission each for what it does best, and to modernise the network in the process. Much of the existing grid rests on ageing, in some cases obsolete, AC transmission. Some of those corridors can be rebuilt with modern designs at higher voltage — up to 765 kV double-circuit — which introduces strong, high-capacity connectivity where it is needed; other rows of pylons can be decommissioned and dismantled altogether. In their place, embedded or inter-area voltage-source-converter (VSC) HVDC interconnectors, with fully controllable power flow and no need for synchronism, can limit the propagation of selected disturbances between the systems they join. The extent of that decoupling is a property of the design rather than of the technology: it depends on converter topology, DC-side protection, control blocking behaviour, the strength of the AC network on each side, and the contingencies actually studied. Such a link decouples frequency and controls power flow; it does not by itself isolate every AC fault. Used together in this way, each technology plays to its strengths to provide generation ties that combine strong grid support with the steady, firm delivery of electricity the megaloads require. Flexible AC transmission systems (FACTS) — STATCOMs and series compensation — do the complementary work within the stronger, reinforced AC network, holding voltage and damping oscillations in tandem with the multiple VSC connections. Together they give planners controllable ties: the means to balance dense loads across areas that use them as an anchor, rather than stranding them behind a single congested corridor. Where this paper calls such a tie firm, it means firm in a defined sense — a stated transfer capability under stated N-1 conditions, for a stated duration, with grid-code performance, black-start or islanding behaviour and energy availability all specified — and not merely a physical connection of nominal rating. This is mature technology, deployed at scale for decades; the task now is to apply it deliberately, as part of the integrated design.

7.4 Flexible large loads as active grid assets

The load itself is the most under-used asset in the system. A hyperscale campus with on-site generation, storage and controllable demand need not be a passive draw on the grid; it can act as a virtual power plant — curtailing, shifting or exporting on request, and providing demand response and ancillary services to the system around it. This is the dual role that system operators are moving to formalise [12], and the basis of the flexibility compact and of non-firm access schemes such as ERCOT's controllable-load model [8], [9]. Embracing it is what turns the data centre from the threat of Section 3 into a stabilising presence — and what earns the faster, non-firm connection the developer wants, because a load that can be dispatched is a load the grid can safely say yes to sooner.

7.5 Wide-area coordination and multi-timescale control

None of these elements delivers its full value alone; the returns come from operating them as one system. That requires coordination across every timescale at once: grid-forming inverters acting in milliseconds; storage and dispatch responding over seconds to minutes; a rolling optimisation loop re-planning every five to fifteen minutes against solar and cloud forecasts, battery state of charge, grid exchange, load and cooling; and longer-horizon planning above that. It is here that reliability is actually delivered — five-nines held continuously, cycle by cycle, as a managed target rather than an annual tally (Section 3) — and here that the Power Component and the Compute Component are designed as a single, integrated system. Wide-area coordination, underpinned by validated dynamic models of these new loads [19], is the connective tissue that turns co-located generation, grid-forming storage, HVDC ties and flexible demand into one coherent, controllable whole.

This coordination is only as good as the models beneath it — and here a clear discipline is required, one the modelling community is already converging on. Every new system, and every newly developed technology brought to the grid, should comply with an open standard and be accompanied by a high-fidelity model, shared and exercised in a model-in-the-loop environment, so that its behaviour can be fully integrated and calibrated against the rest of the system before it is energised. In practice this means that manufacturers and developers provide validated high-fidelity (or black-box) models that can be exercised in a common model-in-the-loop environment and calibrated against the wider system before energisation. Given how quickly these technologies now change, and how much more complex a data-centre model is than a conventional inverter model (Section 3.1), this open, shared modelling basis is not optional housekeeping; it is a precondition for operating the integrated design at all [19].

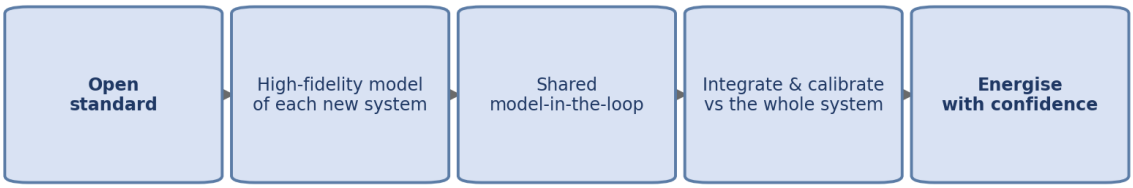

Figure 3 — A shared modelling discipline: open standards and high-fidelity models, tested in the loop, integrated and calibrated before energisation.

Taken together, these five moves are not a menu from which to choose one; they are a single design. Co-location provides the power, grid-forming storage makes it stable, HVDC and FACTS move and firm it, flexible operation turns the load into an asset, and wide-area control binds the whole into a system that can be relied upon. Each, on its own, advances more than one of the forum's four objectives at once; together they satisfy all four — which is the entire meaning of an AND solution.

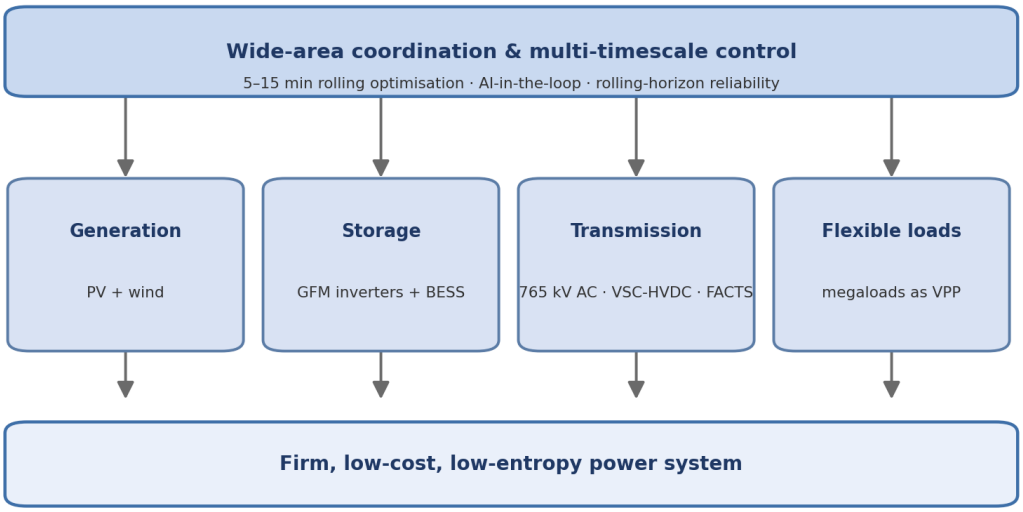

Figure 4 — The integrated system: co-located generation, storage, transmission and flexible loads operated as one, under a single wide-area control layer.

Solution	Speed	Affordability	Resilience	Sustainability
7.1 Co-location	Sited to skip transmission and the queue	Lowest-cost on-site energy	On-site firm supply + redundancy	Best renewable sites; low impact
7.2 GFM + modular BESS	Factory-built, rapid deploy	Defers network upgrades	Synthetic inertia; system strength	Enables high renewable share
7.3 HVDC & FACTS	Controllable bulk transfer	Higher asset utilisation	VSC firewall; firm ties	Connects remote clean power
7.4 Flexible loads (VPP)	Fast non-firm access	Avoids peak-sized build	Demand response; ancillary services	Absorbs surplus renewables
7.5 Wide-area control	Real-time optimisation	Least-cost dispatch each cycle	Rolling-horizon reliability	Maximises clean utilisation

Table — Each engineering solution advances several of the four objectives at once; together they deliver all four.

8. Practical Pathways: Bridge-to-Grid Architectures and the Speed-to-Power Playbook

The engineering of Section 7 becomes speed to power only when it is sequenced correctly. The organising idea is the bridge-to-grid architecture: rather than waiting for the grid to be ready before the load can run, the campus is energised first from its own bridging resources and then integrated with the grid progressively, as the

connection and the network upgrades mature. Where a viable bridging supply, the necessary consents, EPC capacity and a compliant minimal connection can be secured, first power can be brought forward to a matter of months rather than years; full integration then follows over years — and the load earns throughout. The achievable schedule is site-specific and set by that critical path, not by the architecture alone.

8.1 The bridge-to-grid principle

In its simplest form the principle inverts the usual order of events. On-site generation, storage and grid-forming inverters supply the first blocks of load in a largely islanded or minimally connected configuration — the bridge — so that compute can begin commercial operation

.while the interconnection process, network reinforcement and equipment delivery proceed in parallel rather than in series. As those mature, the facility leans progressively on the grid and, in turn, begins to support it. EPRI frames the same continuum through its Speed-to-Power strategies — islanded, grid-flexible and bridge-to-grid — and the choice between them is simply a matter of where a given site sits on the path from self-supply to full integration [20].

8.2 The Speed-to-Power playbook

Reduced to a repeatable sequence, the method has six steps.

Step 1 — Site for deliverability. Choose the location for how fast firm power can actually be delivered, in either direction: generation at the load, or the load at the generation (Section 7.1).

Step 2 — Build modular. Deploy standardised blocks that scale from tens of megawatts to a multi-gigawatt complex, energising each block the moment it is built rather than waiting for the whole to be complete.

Step 3 — Bridge power first. Stand up on-site generation, storage and grid-forming capability behind a minimal grid tie, so that the first load can run as soon as the bridging supply, the consents and the minimal connection allow.

Step 4 — Operate flexibly from day one. Run the campus as a dispatchable, transparent virtual power plant (Section 7.4); it is this flexibility that earns the faster, non-firm connection.

Step 5 — Overlay and reinforce. As the cluster grows, add the VSC-HVDC and FACTS overlay and rebuild the surrounding AC network (Section 7.3), using the load as an anchor for regional strengthening.

Step 6 — Firm progressively. Convert non-firm access to firm as upgrades complete, so resilience deepens over time without the initial build ever having stalled.

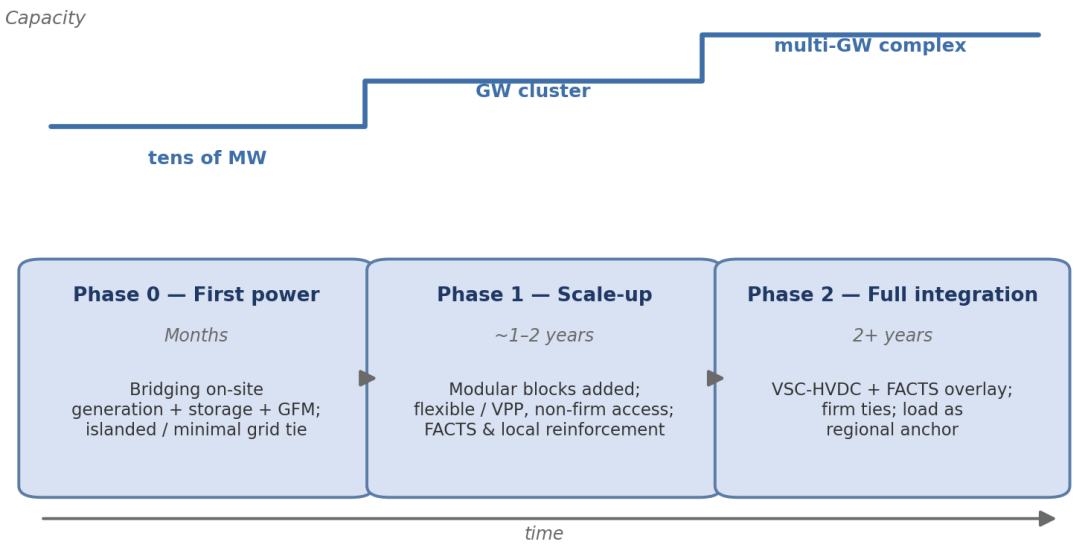

Figure 5 — The bridge-to-grid ramp: first power in months, integrated over time.

8.3 The commercial and risk architecture

A pathway is only real if the risk is bankable. The playbook therefore rests on the risk-allocation principle of Section 6: each party underwrites what it is best placed to carry, and nothing is quietly socialised onto consumers. In practice this means private firming underwritten by the developer; milestone-based and call-option co-developer structures that release capital as deliverability is proven; and standardised, productised contracts and hardware — the same logic that lets a factory-built battery block ship in volume [17] — so that each new project executes a known template rather than a bespoke negotiation [8], [9]. Speed comes not only from the engineering, but from removing the friction around it.

8.4 From pathway to method

None of this is a workaround to be abandoned once the grid catches up. It is a repeatable industrial method for delivering firm, low-cost, dependable power at the pace the digital economy now sets — and, applied at scale, it points toward something larger, which the next section sets out.

9. From Speed to Power to a Low-Entropy Industrial Future

9.1 From scarcity-managed to abundance-engineered

Speed to power solves an immediate problem, but it opens onto a larger one. The grid we have today is, in a real sense, a high-entropy system: energy is curtailed for want of a wire, capacity sits stranded in queues, upgrades are built then wasted, and disturbances propagate for want of the system strength to contain them. Much of the industry's effort goes into rationing this disorder — deciding who waits, who pays and who is turned away. The integrated design of the previous two sections does something different: it drives the entropy down. Co-located generation removes the need to move power across a congested continent; grid-forming storage and controllable HVDC contain disturbances instead of spreading them; flexible operation and rolling-horizon control keep the system ordered cycle by cycle. The result is a grid that is no longer a bottleneck to be rationed, but a platform to be built upon. The entropy language is a device, and the outcomes it stands for are measurable ones: interconnection lead time, curtailed energy, available transfer capability, transmission losses, unserved

energy, capacity factor, capital cost per firm megawatt and the standard reliability indices. Those are the quantities against which everything proposed here should finally be judged.

9.2 The compute–power symbiosis

The most striking feature of this future is that the load which strains the grid also carries the means to run it. A dense compute campus is not only a consumer of power; it is a source of order. Its waste heat can be captured and reused; its demand can be shifted, curtailed and shaped; and, most powerfully, the same artificial intelligence that drives its growth can be turned on the power system itself — forecasting, optimising and balancing generation, storage, exchange and load in the rolling loops described in Sections 3 and 7. The data centre thus becomes both the largest consumer on the system and part of its intelligence: a participant that helps hold the low-entropy state it depends upon. The threat of Section 3, fully realised, becomes an instrument of stability.

9.3 Low-cost electricity as the industrial foundation

Beneath Speed to Power lies an older truth: abundant, low-cost electricity is the foundation of industrial prosperity. It lowers the marginal cost of production, attracts capital-intensive industry and raises real incomes. What the AI era changes is the urgency — low-cost power that arrives too late has no value — and the geography. Because a megaload can be sited where premium renewable resources are strongest (Section 7.1), it becomes an anchor: it pulls new generation, a reinforced grid, telecommunications and skilled employment into regions that would otherwise have waited decades for them. The lowest-cost, firm, clean electron, delivered fast, is not merely an input to the digital economy; it is the seed of a new industrial geography — and it is delivered without socialising its cost onto ordinary consumers.

9.4 The AND future realised

Set against the false choices that opened this report, the change is fundamental. Speed, affordability, resilience and sustainability cease to be traded against one another and begin to compound: co-location is fast and cheap and clean; grid-forming storage is fast and resilient and enabling of renewables; flexibility is inexpensive and stabilising at once. This is the whole meaning of an AND solution, carried from principle to practice. The low-entropy industrial future is not a utopia to be waited for; it is the cumulative result of applying, deliberately and together, engineering that already exists.

Dimension	Today — high-entropy grid	Integrated design — low-entropy system
Access to power	Multi-year queue; ration by waiting	Bridge-to-grid; first power in months
Power flow	Congested corridors; curtailment	Controllable HVDC and FACTS; minimal waste
The load	Passive threat to stability	Dispatchable asset and system intelligence
Reliability	Trip-and-restore; annual availability tally	Rolling-horizon, continuously managed
Cost	Socialised upgrades; stranded assets	Risk priced where it sits; lowest-cost energy
Geography	Power chases distant load	Load anchored at the best clean resource

Table — From a high-entropy grid to a low-entropy, integrated system.

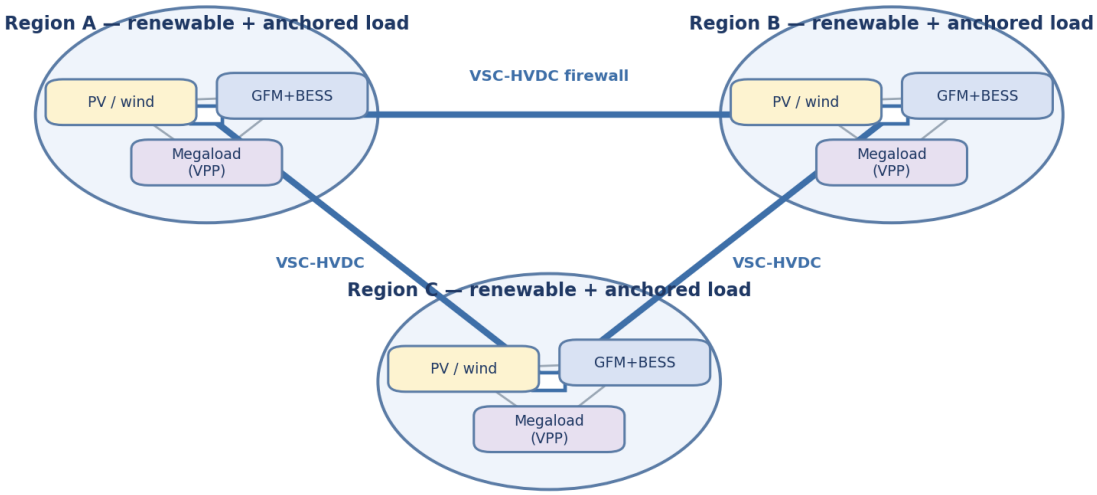

Figure 6 — The efficient final arrangement: renewable regions with anchored megaloads and a reinforced 765 kV AC backbone, linked by controllable VSC-HVDC ties between asynchronous regions.

10. Conclusions and Call to Action for IREF

This paper began with a single imperative — speed to power — and has tried to show what it truly demands. The grid we inherited was a deterministic, synchronous machine; the energy transition made it triple-stochastic; and the arrival of dense, gigawatt-scale, always-on compute has exposed the limits of the processes, the markets and the equipment built for an earlier age. The interconnection queue cannot open at the speed the load arrives; no single actor — utility, hyperscaler or manufacturer — can break the resulting deadlock alone; and the way out is not more rationing, but a different allocation of roles, responsibilities and risk, delivered through engineering that already exists.

That engineering is not speculative. Co-location, grid-forming inverters and modular storage, strategic HVDC and FACTS, flexible loads operating as virtual power plants, and wide-area multi-timescale control are proven technologies. Sequenced through the bridge-to-grid playbook, they deliver first power in months and firm, integrated, low-cost capacity over time — and they do so as AND solutions, achieving speed, affordability, resilience and sustainability together rather than trading one against another.

The case is no longer confined to engineering. Firm, low-cost, clean electricity delivered fast is increasingly understood as the decisive industrial advantage of the coming decade rather than a constraint upon it — speed to power expressed as strategy, and available, in different forms, to every region willing to engineer for it.

The forum's task is to turn that recognition into practice. Concretely:

- **Grid operators and utilities** should move from rationing to enabling — adopting readiness-based, performance-based connection standards, offering non-firm and flexible access, and treating the large load as a partner in stability rather than only a risk to be managed.

- **Hyperscalers** should earn the speed they need by bringing their own firming and operating transparently as dispatchable, grid-supporting assets — accepting the responsibilities of the power-utility role they are, in effect, assuming.
- **OEMs** should productise and standardise — shipping adaptable, grid-forming, rapidly deployable equipment against the demand certainty that co-developed, milestone-based structures can now provide.
- **Policymakers and regulators** should write the shared framework that is still missing — one that reallocates risk to those best able to carry it, shields ordinary consumers from socialised costs, and rewards the lowest-cost, firm, clean electron delivered fastest.
- **All parties** should commit to the common modelling discipline without which none of this can be coordinated: open standards and shared, high-fidelity models, tested in the loop before energisation.

None of this requires waiting for a breakthrough. It requires deciding, together, to build. The electricity industry does not need to be rebuilt from first principles; it needs to be re-coordinated around a new metric and a new allocation of risk. Do that, and the dense loads that today threaten the grid become the anchors of a low-entropy industrial future — abundant, dependable, low-cost power at the speed the age demands.

Let there be power.

11. References

[1] North American Electric Reliability Corporation (NERC), 2025 ERO Reliability Risk Priorities Report, Reliability Issues Steering Committee, July 2025.

[2] CIGRE Large Load Joint Task Force (Study Committees C4, C2 and C1), “Integration of Large Power-Electronics-Interfaced Loads in Power Systems — Challenges and Opportunities, Part I: Technical Foundations,” CSE N°41, June 2026.

[3] International Energy Agency (IEA), Energy and AI, 2025.

[4] Electric Power Research Institute (EPRI), Powering Intelligence 2026, 2026.

[5] Lawrence Berkeley National Laboratory, Queued Up: 2025 Edition — Characteristics of Power Plants Seeking Transmission Interconnection as of the End of 2024, 2025.

[6] RAND Corporation, Assessing the United States’ Additional AI Power Capacity by 2030 (RR-A3845-1), 2025.

[7] Federal Energy Regulatory Commission (FERC), Order No. 2023, “Improvements to Generator Interconnection Procedures and Agreements,” July 2023.

[8] A. Sharma Frank, “From Bottlenecks to Breakthroughs: Rewriting the Grid Planning Playbook in the Southwest Power Pool and Texas,” Center for Strategic and International Studies (CSIS), July 2025.

[9] A. Sharma Frank, “Power Access for AI: The Flexibility Compact,” Center for Strategic and International Studies (CSIS), July 2025.

[10] Brazil, Ministério de Minas e Energia, Política Nacional de Acesso ao Sistema de Transmissão (PNAST), December 2025.

[11] ENTSO-E, Position on the Need for National Connection Requirements to Ensure EU Power System Stability, December 2025.

[12] ENTSO-E, Data Centres and the Power System: Expected Trends, Challenges and Opportunities, May 2026.

[13] Alberta Electric System Operator (AESO), Interim Approach to Large Load Connections, 2025.

[14] India, Central Electricity Authority (CEA), directive to incorporate projected data-centre demand in state resource-adequacy planning, 2025.

[15] Singapore, Infocomm Media Development Authority and Energy Market Authority, Green Data Centre Roadmap, 2024.

[16] Australian Energy Market Commission (AEMC), Draft Rule — Connection Standards for Data Centres (large loads), 2026.

[17] Tesla, Megablock and Megapack 3 — factory-built grid-scale battery systems, launched September 2025.

[18] CIGRE, Technical Brochure 943 — Evaluation of Voltage Stability Assessment Methodologies in Modern Power Systems with High Penetration of Inverter-Based Resources, 2024.

[19] B. Badrzadeh and Z. Emin (eds.), CIGRE Green Book — Power System Dynamic Modelling and Analysis in Evolving Networks, 2025.

[20] Electric Power Research Institute (EPRI), Speed to Power (Data Centers) initiative, speed2power.epri.com, 2026.

[21] J. Huntington and M. Tu, 800 VDC Architecture for Next-Generation AI Infrastructure — The Architectural Imperative of 800 VDC and Integrated Energy Storage, NVIDIA Corporation, 2025.

[22] NVIDIA, “How New GB300 NVL72 Features Provide Steady Power for AI,” technical blog, 2025.

[23] E. Choukse et al., Power Stabilization for AI Training Datacenters, arXiv:2508.14318 (Microsoft and NVIDIA), August 2025.

[24] P. Patel, E. Choukse, C. Zhang, Í. Goiri, B. Warrier, N. Mahalingam and R. Bianchini, “Characterizing Power Management Opportunities for LLMs in the Cloud,” ASPLOS ’24, doi:10.1145/3620666.3651329, April 2024.

[25] Electric Power Research Institute (EPRI), Torsional Interaction Between Electrical Network Phenomena and Turbine-Generator Shafts: Plant Vulnerability, Technical Report 1013460, November 2006.

[26] B. A. Ross and J. Follum, Electromagnetic Transient Modeling of Large Data Centers for Grid-Level Studies (Alpha Release), Pacific Northwest National Laboratory, PNNL-38817, December 2025.

[27] M. Bahrman, E. V. Larsen, R. J. Piwko and H. S. Patel, “Experience with HVDC — Turbine-Generator Torsional Interaction at Square Butte,” IEEE Transactions on Power Apparatus and Systems, vol. PAS-99, no. 3, pp. 966–975, May 1980.

[28] R. J. Piwko and E. V. Larsen, HVDC System Control for Damping of Subsynchronous Oscillations, EPRI EL-2708, Research Project 1425-1, Final Report, 1982 (origin of the Unit Interaction Factor).

[29] K. Chatterjee, M. Ramesh, S. Biswas, B. A. Ross, A. C. Varghese, S. Nekkalapu and S. Kincic, “Assessing Risks of Hydro-Generator Shaft Fatigue from Data Center Load Oscillations,” arXiv:2607.14412, July 2026.

[30] F. Hossain, N. Ray Chaudhuri, A. Sinha, S. G. Vennelaganti and M. E. Nassar, “Limiting the Impact of AI Data Centers on Fatigue Life of Thermal Turbine Generators in the Grid: A Frequency-Domain Approach,” arXiv:2605.01173, May 2026.

[31] North American Electric Reliability Corporation (NERC), Large Loads Task Force, Characteristics and Risks of Emerging Large Loads, white paper, July 2025.

[32] North American Electric Reliability Corporation (NERC), Large Loads Working Group, Reliability Guideline: Risk Mitigation for Emerging Large Loads, May 2026.

[33] Electric Power Research Institute (EPRI), Assessing Subsynchronous Torsional Interactions with Respect to AI Training Loads, Product ID 000000003002034099 (DCFlex Workstream 3).

[34] Electric Power Research Institute (EPRI), Managing Oscillating Load Impacts from Data Centers on Synchronous Machines: A Framework for Estimating Acceptable Load Oscillation Limits to Minimize Turbine-Generator Shaft Impacts, Product ID 3002036374.

Appendix 1 — The Power–Compute Interface

Native DC, 800 VDC delivery, and the shaping of the load at its source

A1.1 The interface is the missing design layer

Section 7 ends at the campus boundary. Everything downstream of the converter — the distribution inside the fence, the rack, the compute node itself — has so far been treated as an opaque load: a number in megawatts, with a power factor and a ramp rate attached to it. That treatment is now one of the most consistently under-represented design layers in large-load planning. A modern AI campus is not an opaque load; it is a power-electronic system in its own right, with its own conversion chain, its own storage, its own protection philosophy and its own control hierarchy. A great deal of the efficiency the system needs, and almost all of the flexibility it needs, is created or destroyed inside that boundary.

This appendix therefore opens the box. It describes the interface between the Power Component and the Compute Component as a single engineered layer: how the facility is built so that conversion is not paid for twice; how 800 VDC delivery moves conversion out of the rack and density into it; how storage is layered across every timescale from microseconds to minutes; how the load is shaped at its source so that the grid sees a controlled envelope rather than raw compute pulses; and how safety and reliability are engineered into all of it. The elements described here are not speculative. They are drawn from the reference architecture the dominant supplier of AI compute has published [21], from the operating practice of industries that have run high-current DC continuously for decades, and from the design work behind the arrangement set out in Section 7. Their integration at the scale proposed is a different matter, and remains subject to project-specific validation: what follows is a synthesis of published reference designs and established practice, not a design specification.

A1.2 A DC hub from standard equipment

The first move is to stop converting the same energy twice. On a conventional campus, photovoltaic generation is inverted to AC, battery storage is inverted to AC, the AC is transformed and distributed, and the compute load then rectifies it back to DC — three conversions to move energy between two devices that are both natively DC.

The alternative is a native DC hub. Solar PV and the battery energy storage system feed a common DC bus through MPPT and DC/DC stages; the IT load draws from that bus directly; and a single bidirectional voltage-source converter links the hub to the AC grid. The critical point is the rating of that converter: it carries only the *net* exchange between local generation, storage and load — a fraction of facility power, not the whole of it. Heavy auxiliaries, principally cooling and the mechanical plant, remain on AC and are never converted twice; there is no efficiency case for rectifying a motor load.

Nothing in this arrangement is novel equipment. MPPT trackers, DC/DC converters, voltage-source converters, transformers and gas-insulated switchgear are field-proven, volume-manufactured products with mature supply chains — which is precisely why the arrangement can be delivered at the speed the load demands rather than waiting on a development programme. On the loss side, a hub built this way achieves an end-to-end figure from generation and storage to the IT load of approximately 1 to 1.5 per cent at rated load — a typical-loss estimate for this arrangement developed by HVDC Technologies Ltd, rather than a published vendor figure. That figure should be read as an illustrative preliminary loss budget rather than a normative benchmark. Its boundary is the path from generation and storage to the IT load at rated output, and it excludes the AC auxiliaries; it will move with loading, redundancy state, cable lengths, cooling, harmonics and component selection. Establishing it as a

benchmark would require a published loss-boundary diagram, a component-by-component budget across part-load and N+1 states, and comparison with a conventional AC architecture under identical boundary conditions. The narrower claim made here is that a figure of this order is achievable with equipment available today, and that it is the right quantity to compare.

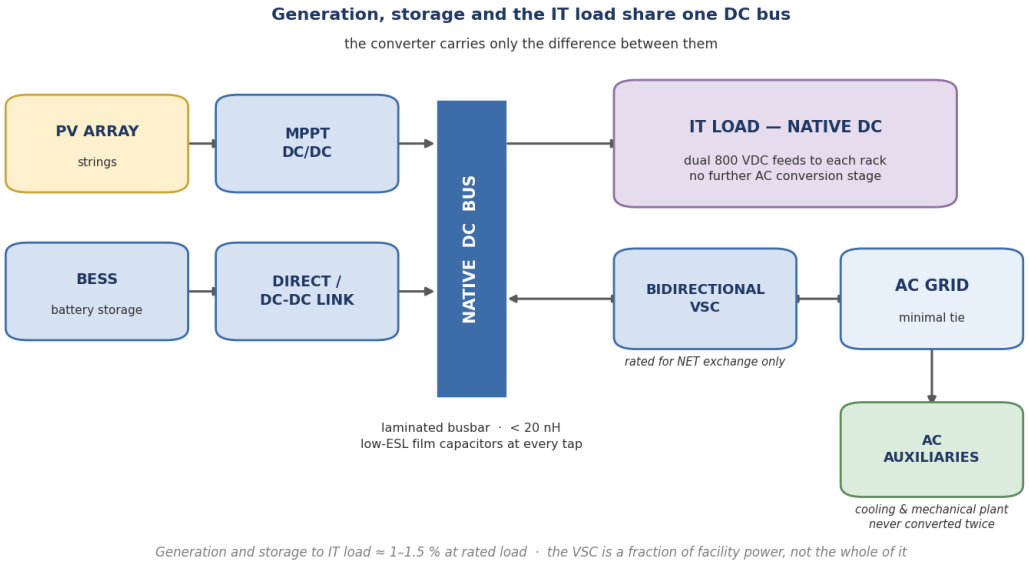

Figure 7 — The native DC hub: generation, storage and the IT load on one bus, with the converter rated only for net exchange and heavy AC auxiliaries left on AC.

A1.3 The 800 VDC path from facility to rack

Inside the fence, the constraint that now governs is copper. Traditional 415 V or 480 V three-phase distribution has supported data-centre growth for thirty years, but as rack density approaches and exceeds the megawatt scale it reaches a practical limit: whips are constrained by thermal ratings and connector standards, higher rack power demands more and larger inlets, and each additional AC feed consumes rack space and adds protection complexity. For a fixed conductor cross-section rated at 48 A continuous, moving from 415 VAC to 800 VDC transmits 157 per cent more power through the same copper; moving instead to 480 VAC — the common North American workaround — yields only 16 per cent [21].

The 800 VDC architecture takes that result to its conclusion. Alternating-current rectification is moved upstream, out of the rack and into the facility, where it can be done once at scale: by facility-level rectifier strings of the order of 1.5 MVA paralleled onto a common DC bus at around 99 per cent efficiency, or by medium-voltage rectifiers and solid-state transformers converting directly from medium voltage — of the order of 35 kV AC — to 800 VDC in units as large as 7.5 MVA at 98.5 per cent efficiency or better, eliminating the low-voltage AC layer altogether [21]. A protected 800 VDC bus is then formed, with facility battery storage connected to it, and dual independent 800 VDC feeds are taken into each compute rack.

Within the rack, conversion is pushed as close to the GPU as it can physically go. A 64:1 LLC converter with a matrix transformer takes 800 VDC directly to 12 VDC beside the compute node, eliminating the separate 400 V to 50 V and 50 V to 12 V stages of the present standard power tree; the reported gain is a further one per cent in efficiency and a 26 per cent reduction in converter area, in the most space-constrained location in the building [21]. Only a short 12 V path then remains, feeding the voltage regulators and the GPUs themselves.

The choice of a single-ended 800 VDC bus rather than a bipolar ±400 V arrangement is worth recording, because it is a good example of engineering for deliverability rather than for theoretical optimum. Bipolar configurations offer real advantages in grounding and fault isolation, but they require three-pole breakers and protection devices that the industry does not currently manufacture at volume; single-ended 800 VDC aligns with existing two-pole breaker designs, avoids load-balancing between positive and return, and can therefore be deployed now [21]. The same logic runs through this entire appendix, and through Section 8 — though it is a balance rather than a rule. Technology selection has to weigh availability and maturity against lifecycle performance, serviceability and standards compliance, and the schedule advantage of what can be built this year is only an advantage if the resulting plant can be protected, maintained and qualified. It should also be said plainly that what is described here is a supplier reference architecture under active evaluation, not a settled cross-industry standard. The standards applicable to DC distribution at this voltage are still being written, and fault interruption, insulation coordination, connector qualification, serviceability and corrosion behaviour all require project-specific treatment and type testing before selection. Conventional AC distribution, in-rack 48 or 54 V, bipolar ±400 V and medium-voltage DC all remain live alternatives that a given site may have good reason to prefer.

The result is a single 800 VDC backbone that spans the whole present and foreseeable range of rack power, from the hundreds of kilowatts of today's deployments to beyond one megawatt per rack.

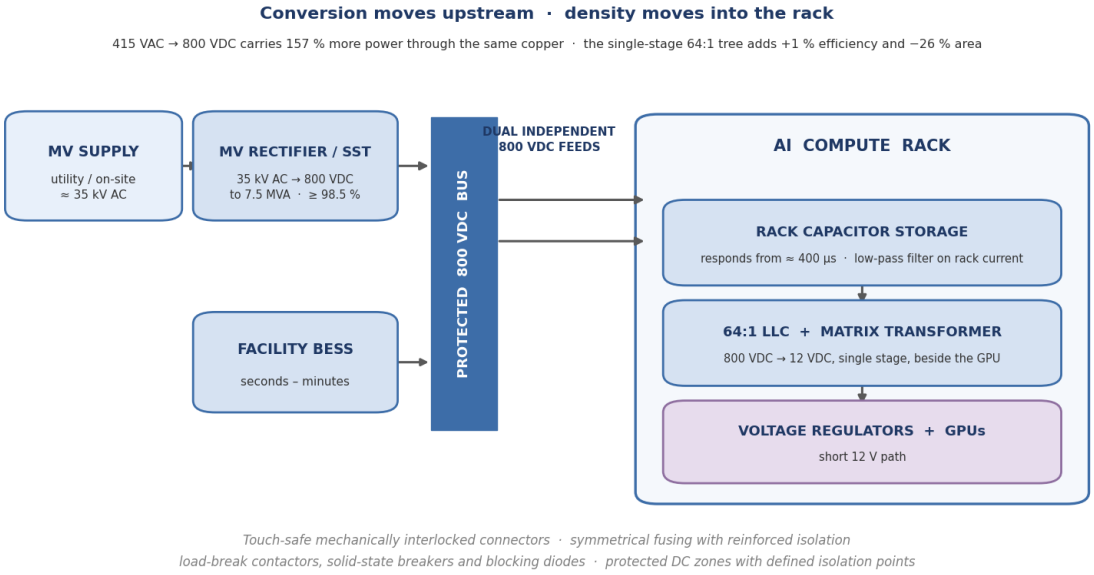

Figure 8 — The 800 VDC path from facility to rack: rectification moved upstream, protected DC bus, dual independent feeds, and single-stage conversion beside the GPU.

A1.4 Storage at every timescale

A dense compute load does not present a single dynamic problem; it presents a family of them, separated by orders of magnitude in time. During a large language-model workload, intervals of intense matrix computation alternate with intervals of data exchange, and because the GPUs in a cluster are synchronised, so are their power profiles. Unmitigated, this produces rapid swings between roughly 30 per cent of rack power at idle and 100 per cent utilisation — a problem for the rack, then for the facility, and at sufficient cluster size for the grid itself [21].

The characteristic durations are known. Overshoot and peak-power events, and typical workload idle periods, extend up to about 100 milliseconds; checkpointing takes one to five seconds; workload ramp-up and ramp-down occur over several minutes. Energy storage must span the whole of that range, and no single technology does. Volumetrically, electrolytic capacitors are the right answer below about 100 milliseconds, a mixture of technologies serves the window from 100 milliseconds to 10 seconds, and batteries are the better solution beyond it [21].

The engineering discipline that follows is to mitigate each swing as close to its source as possible. Capacitor storage integrated at the rack acts as a low-pass filter on rack current, responding to events as fast as 400 microseconds; facility battery storage handles seconds to minutes, providing load averaging, grid-forming support and transition power during a transfer from utility supply to on-site generation; and the campus storage of Section 7.2 addresses the minutes-to-hours horizon that the grid actually trades in. Filtering close to the GPU is not merely elegant: every piece of equipment upstream must otherwise be sized for the peak current and for the additional RMS heating that pulsed loading imposes. A square wave at 50 per cent duty cycle with a 50 per cent peak above average produces a 25 per cent increase in RMS losses [21] — a penalty paid, in a conventional design, by every conductor, breaker and transformer between the rack and the grid.

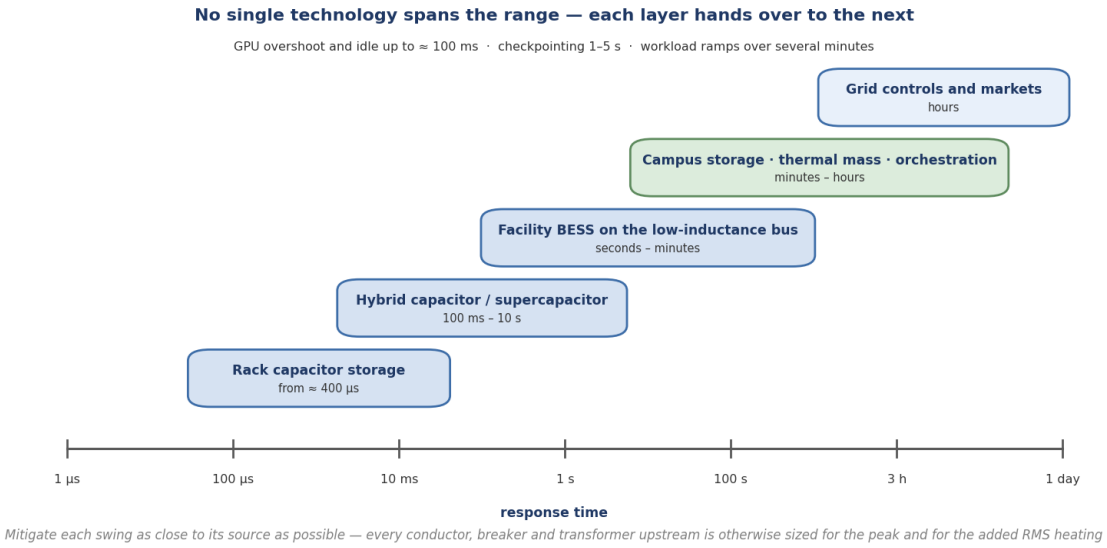

Figure 9 — Storage and control across timescales: no single technology spans the range from the microsecond event to the market hour.

A1.5 The bus itself is a component: connecting PV and storage at ultra-low inductance

The published rack architecture places its fastest storage near the compute, and leaves the facility battery to the slow side of the problem — load averaging, grid-forming support, transition power. That division is a consequence of how the facility is usually built, not a law of nature. If the battery is connected to the DC bus through a path of sufficiently low series inductance, it becomes a fast element in its own right, and a large part of the spike attenuation can be performed once, at facility level, instead of only at the rack.

The topology is the hub of Section A1.2, with every element sharing one protected, low-inductance bus: PV strings through their MPPT DC/DC converters, paralleled directly onto the bus; the battery connected directly, or as near to directly as voltage matching allows; the compute racks taking their dual feeds from the same bus; and the bidirectional converter connected to it for net grid exchange only.

True zero inductance is unobtainable, but practical designs reduce the series path to the order of tens of nanohenries — low enough that the battery's own internal impedance and the control bandwidth of its converter, rather than the bus, become the limit on how fast it can respond. The methods are established practice in high-power converters, electric vehicles, traction and industrial DC. Laminated busbars — positive and return plates stacked with a thin high-voltage dielectric between them — cancel the magnetic field of the opposing currents and typically fall below 20 nH, reaching single figures in short, optimised links. Connections from every MPPT output cabinet and every battery rack terminal to the main distribution boards are made short, wide and closely spaced, so that loop area is minimised. Low-ESL film capacitors, with bulk or hybrid capacitors behind them, are mounted directly on the laminated bus at each connection point — battery, PV feed, converter, rack feeder — to hold the high-frequency impedance down. The battery racks, the MPPT output cabinets and the main DC switchgear are physically co-located, so that interconnections are measured in centimetres or a few metres rather than tens of metres; where cable is unavoidable it is closely paired or coaxial, and kept short. The design intent is to insert no series filter choke between the battery and the bus, since inductance deliberately added there works directly against the objective. That intent is a starting point for design rather than a rule to be applied without analysis. The target impedance profile, the damping needed for stable operation and the permissible series inductance have to be established together, through frequency-domain, electromagnetic-transient, protection and electromagnetic-compatibility study; the aim is to minimise parasitic inductance while preserving stable converter control and selective fault isolation. In some arrangements those last two requirements will set a floor below which the inductance should not be driven.

The consequence is that this part of the system can sink or source large currents on a microsecond-to-millisecond timescale. The division of labour within it is worth stating precisely, because the four elements are not interchangeable: the fastest content is taken by the distributed capacitance and by the converter control loops acting on it, while the battery converter extends that behaviour into the milliseconds and beyond, within the current and thermal limits the cells and the battery management system allow. It is the coupling and the control that make the storage fast, not the electrochemistry. Compute pulses and rapid photovoltaic fluctuation — a cloud edge crossing the array — are absorbed locally, at the bus, and the bidirectional converter sees only the residual slow component: the controlled envelope of the following section. Its current reference stays smooth, its stress and harmonic emission fall, and the risk of a grid-code violation falls with them.

The control hierarchy of Section A1.7 is unchanged by this; the fastest layer is simply made as effective as it can be. Battery current control, or hybrid battery-and-supercapacitor control, forms the innermost and highest-bandwidth loop, prioritised for bus-voltage regulation and spike cancellation. The MPPT loops continue to maximise harvest, with any residual high-frequency content absorbed by the bus and the battery. The converter's power control and the campus energy-management and dispatch layers then operate on power that has already been smoothed.

The protection philosophy of Section A1.7 applies unchanged, with one emphasis: a stiff, low-inductance bus permits high rates of rise of fault current, so every MPPT feeder and every battery string requires its own high-speed protection — semiconductor fuses or solid-state DC breakers — rated for the di/dt the bus now allows. Reverse-current protection for the photovoltaic feeders, isolation coordinated with the battery management system, ground-fault detection, arc-flash mitigation and insulation coordination all follow the industrial-DC practice already described. Modular paralleling of MPPT outputs and battery modules provides N+1 continuity, and over- and under-voltage protection on the common bus is coordinated across all sources. Because the

battery is tightly coupled and high in bandwidth, it also improves the ride-through and transient stability of the bus as a whole.

A1.6 Shaping the load at its source

Layered storage is the passive half of the answer. The active half is that the load itself can be commanded, and this is the point at which the compute engineer and the power engineer must design together rather than exchange specifications.

Four mechanisms are available, and they are complementary. Idle periods can be minimised in software, which is the ideal outcome because it reduces the facility's power requirement outright. Energy storage charges in the troughs and discharges into the peaks, as described above. Power capping raises permitted GPU draw progressively during workload ramp-up, at a rate the facility and the grid can tolerate, rather than as a step. And where storage is exhausted — during idle periods longer than the local storage can bridge — the GPU can deliberately burn power to hold a floor, with performance throttling as a further fallback [21], [22]. Both are backstops for corner cases and should be treated as such: burning energy in order to stabilise a power profile is a poor trade in cost and in carbon, and its use belongs in a quantified and bounded corner of the operating regime rather than in normal practice.

Used together, these mechanisms convert a raw compute profile into a controlled envelope. The grid sees a predictable power request with a bounded ramp rate; the compute sees the highly dynamic source it requires; and the storage sits between the two, reconciling requirements that are otherwise in direct conflict. This is the single most important fact in this appendix, and it deserves to be stated plainly: **the AI load is shapeable at source, by design, using features that already exist in the silicon.** The volatility that system planners currently treat as an inherent property of the load can, to a material degree, be shaped instead — within the service levels, thermal limits and software controls the operator has agreed to work inside. How much can be shaped, and at what cost to compute performance, is a question to be answered for each facility rather than a general figure.

Beyond the rack, the same principle extends across the campus. Compute is schedulable, curtailable and migratable between clusters; the compute mix itself can be designed for elasticity, since inference and training have very different tolerance for interruption; and the thermal system is a storage resource in its own right, capable of being over-run and then released. Taken together with generation and storage behind the single converter of Section A1.2, these make the campus a virtual power plant in the full sense of Section 7.4 — not a load with a demand-response contract attached, but a dispatchable resource whose internal state is designed for dispatch.

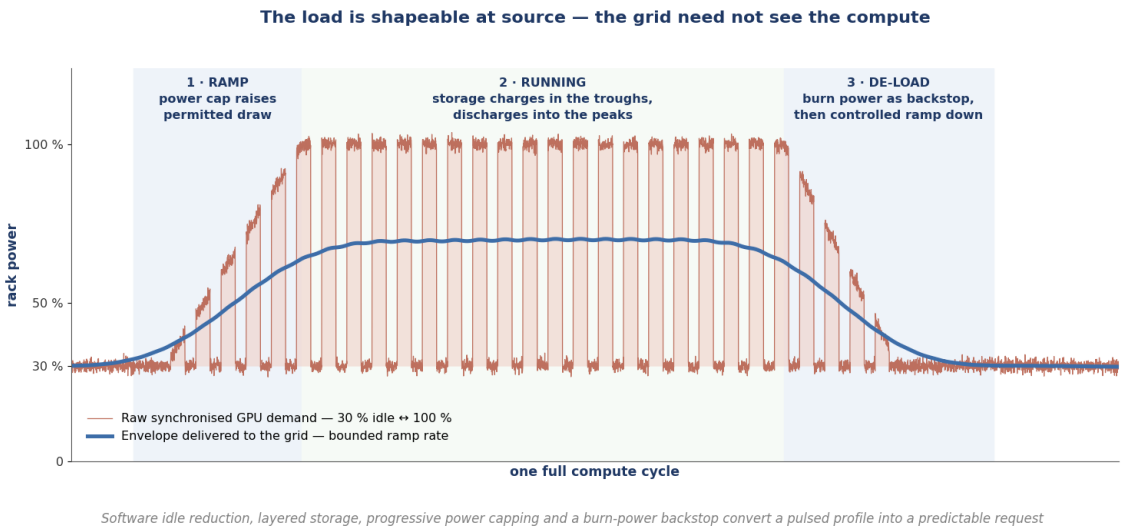

Figure 10 — Shaping the load at source: a pulsed compute profile becomes a controlled envelope with a bounded ramp rate before it ever reaches the grid.

A1.7 Safety and reliability by defence in depth

An architecture of this density is only admissible if it is demonstrably safe and fail-operational, and that has to be demonstrated installation by installation — through hazard analysis, protection and arc-flash studies, equipment qualification and type testing, maintainability analysis and written operating procedures. What can be said in advance is that the means are conventional: the design borrows from industries that have operated high-current DC continuously for decades — aluminium smelting with its rectiformers and busbars, railway traction with its DC switchgear and selective protection, and industrial DC drives — and adapts that practice to an indoor, high-density environment. Medium-voltage rectification itself is proven at scale in mining, electrolysis, rail and grid-scale storage [21]; what is new is the application, not the technology.

On the electrical and physical side, every rack is fed by dual independent 800 VDC feeds from independent rectifiers, so that the loss of a source or a feeder does not interrupt compute; in the published reference design a 17.5 MW block is built from five 3.5 MW medium-voltage rectifiers in a five-to-make-four configuration, and a single rectifier can sustain up to 3.3 MW of compute load under worst-case conditions [21]. All user-accessible connectors are touch-safe, and all are mechanically interlocked so that disconnection under load is impossible — techniques taken directly from electric-vehicle charging, where they are already used safely by untrained consumers, and re-optimised for the data-centre environment. Fuses are fitted on both the high and low side of the DC/DC converter input, with reinforced isolation from the secondary, allowing the same rack to be supplied from either ±400 V or 800 VDC sources. Distribution is protected by load-break contactors, solid-state circuit breakers and blocking diodes, giving selective fault clearing and containment of fault current, and protected DC zones with clearly defined isolation points allow maintenance on one section while the remainder stays energised. Heavy AC auxiliaries, kept on AC as described above, never see the DC system's fault levels or isolation requirements at all.

On the operational side, reliability comes from modularity and from control. Conversion is delivered as paralleled units with N+1 or better redundancy, so that capacity is added and faults are removed without a shutdown; components are industrial-grade and field-proven rather than novel; the layered storage of Section A1.4 provides ride-through from microseconds to minutes; and the active shaping of Section A1.6 removes the

violent step changes and high-frequency content that would otherwise stress converters, busbars, and the grid beyond them. Above all, the control hierarchy must be coherent across timescales — converter and semiconductor control in microseconds to milliseconds, campus energy management, storage and thermal control in seconds to minutes, workload orchestration and virtual-power-plant dispatch in minutes to hours, and grid controls and markets in hours — so that four control systems designed by four different disciplines cooperate rather than fight. That coherence is the subject of Section 7.5, and it is where this appendix rejoins the main argument.

A1.8 What the interface obliges

If the load is a power-electronic system with its own converters, storage and controls, then it should be held to the standards that other converter-interfaced equipment is held to. The recommendation of this paper is therefore direct: **develop proportionate performance requirements for converter-interfaced mega-loads — model provision, ramp-rate limits, power-quality limits, disturbance response and agreed flexibility services — set by reference to what is already expected of generation of comparable rating.**

The specific obligations must remain jurisdiction- and connection-specific. They have to respect the local grid code, the applicable market rules, protection coordination and the safety of the customer's own process, and they belong in a connection agreement rather than in a blanket rule. But the direction is not an external imposition on an unwilling industry. The published reference architecture already anticipates it, calling for energy storage with fast real-time compensation to control ramp rates and mitigate step changes; for compute-firmware and workload pacing to limit cycle-to-cycle ramp rates; for coordinated control to meet utility requirements on ramp rate, transient stability, harmonics and flicker, and voltage ride-through; for grid-forming and fast-response controls that actively support voltage and frequency during disturbances; for internal IT-load ride-through that exceeds the ITIC curve by reference to ERCOT's voltage-ride-through requirements for large electronic loads; and for real-time power support that delivers power when the grid is stressed and absorbs it when it is not [21]. It further calls for standardised load-behaviour profiles and response metrics, so that utilities can evaluate these projects with confidence and approve them faster.

That is the same bargain this paper has argued for throughout, arrived at from the other side of the meter. The load offers controllability, transparency and grid support; the system offers faster access. What Section 7.5 asks for as a modelling discipline, and Section 8 sequences as a delivery method, the interface described here supplies as physical capability. Native-DC efficiency, storage layered across every timescale, and 800 VDC density together make the data centre a generation-like, dispatchable, grid-supporting asset behind a single converter — one integrated Power–Compute system, in which the load that created the problem of Section 3 becomes a substantial part of its solution.

A1.9 The spectrum of the load, and why it matters beyond the fence

Everything so far has treated the load's dynamics as a facility problem — something to be filtered, buffered and shaped so that the equipment inside the fence is not oversized. There is a second reason to do it, and it is the more important one. The frequencies at which a large synchronised compute cluster varies its demand are not arbitrary, and they fall close enough to the bands where the surrounding rotating machinery is least tolerant that the question cannot be left unasked. What follows is a screening argument rather than a demonstrated generic risk. Whether a particular facility presents a real exposure depends on its measured load spectrum, on the coherence of that spectrum at the point of connection, on the network transfer function between the load

and the machine, and on the location, operating point and damping of the machines concerned. None of that can be settled in a paper. What a paper can do is set out why the screening is necessary, and what it would consist of.

Consider the envelope of Figure 10 with real durations attached. The ramp and the de-load each occupy several minutes. The running phase persists for minutes to many hours, and it is inside that phase that the fast structure lives: individual peaks and troughs lasting tens to a few hundred milliseconds, checkpointing events of one to five seconds, and training iterations that repeat on a cycle of seconds [23], [24]. Convert those durations into frequencies and the picture is immediate. A half-period of 10 milliseconds is a fundamental near 50 Hz; 100 milliseconds is 5 Hz; 200 milliseconds is 2.5 Hz. And because the underlying waveform is closer to a square wave than a sinusoid, each fundamental carries odd harmonics with it: a 5 Hz modulation puts real energy at 15, 25 and 35 Hz. It is the harmonic content, more than the fundamental, that determines the exposure.

Published measurement now allows this to be put more precisely, and the picture is more interesting than a simple coincidence of frequencies. Production traces from AI training clusters show spectral energy concentrated between roughly 0.2 and 3 Hz [23] — which places the dominant fundamental of the load in the inter-area and local electromechanical range, below the shaft torsional band, and the same source puts shaft torsional criticals from about 7 Hz upwards. Taken at the fundamental alone, then, the two do not coincide, and any argument resting on that coincidence is weaker than it first appears. What connects them is the shape of the waveform rather than its rate. A near-square modulation at 2 to 3 Hz carries odd harmonics at 6 to 9, 10 to 15 and 14 to 21 Hz, and it is that harmonic content — together with the interharmonics introduced by the facility's own conversion and control — that reaches into the torsional band. The exposure is therefore real but conditional, and the quantity to be established is the spectral content at the point of connection, not the nominal cycle time of a training job.

For the converter this is undemanding. The inner current-control loop of a modern voltage-source converter operates with a bandwidth of hundreds of hertz to a few kilohertz, so oscillations at a few hertz to a few tens of hertz sit well inside its capability; it can track or reject them without difficulty, provided the DC bus behind it is stiff — which is precisely what Section A1.5 is for. This does not make the converter irrelevant to the assessment. The converter, the AC network, the connected machines, the protection system and the interactions between their controls all require examination, and which of them dominates will be specific to the topology and the operating point. The narrower point is that at these frequencies the converter is not where the margin is thinnest. The elements with the least margin sit outside the fence, and they are mechanical.

Two distinct phenomena need to be kept apart, because they are routinely conflated. The first is electromechanical: the inter-area modes by which whole regions of a synchronous system swing against one another, typically between 0.1 and 2 Hz. The second is torsional: the natural frequencies of the turbine-generator shaft train itself, which for large steam units lie broadly between 10 and 50 Hz — the IEEE First Benchmark Model, drawn from a real 3,600 rpm machine, has modes at 15.7, 20.2, 25.6, 32.3 and 47.5 Hz [25]. These torsional modes are extremely lightly damped. Recent national-laboratory work puts the consequence plainly: a small torque oscillation at the stator terminals can produce a torque oscillation between sections of the shaft twenty to forty times larger [26]. A mechanism that looks negligible at the point of connection does not stay negligible by the time it reaches the coupling.

A periodic large load acts on such a mode as a forcing function. If its spectrum contains energy at or near a natural frequency, it excites that mode directly — a forced oscillation, sustained for as long as the workload

pattern persists, accumulating fatigue in a shaft that was designed on the assumption that nothing on the network would drive it at that frequency for hours at a time. This is not the same mechanism as classical subsynchronous torsional interaction, in which converter controls introduce *negative damping* and an oscillation grows of its own accord. That phenomenon is real and well documented — it was first encountered at the Square Butte HVDC link in North Dakota, where 1977 field tests showed the converter's constant-current control loop destabilising an 11.5 Hz torsional mode of the adjacent thermal plant, and where the remedy was a subsynchronous damping controller added to the current loop [27], [28]. But for a converter-interfaced mega-load the primary concern is the simpler and more direct one: not that the load will destabilise a mode, but that it will drive it.

The exposure is a function of scale and coherence rather than of any single facility. A few tens of megawatts of uncorrelated demand is noise. Hundreds of megawatts of tightly synchronised training, varying together because the GPUs are computing the same job in lockstep, is a coherent forcing function of a magnitude the system has not previously had to consider on the load side. Which machines are most at risk depends on the local fleet: large steam turbine-generators are the best-documented case and carry the historical precedent, hydro units have been examined more recently — with Kaplan machines identified as the more vulnerable [29] — and gas turbines are generally less exposed, because their shorter, stiffer trains place the lower modes higher, though this is a matter of degree rather than immunity. The frequency-domain methods for assessing the resulting fatigue on thermal units are now being published [30].

The industry has a screening tool for this, developed for exactly the analogous problem. The Unit Interaction Factor, introduced by EPRI in 1982 in the work that followed Square Butte, compares converter rating to generator rating, weighted by the generator's share of short-circuit capacity at the converter bus; a value above 0.1 has long been the threshold above which detailed torsional study is required rather than optional [28]. That threshold was derived for HVDC converters, however, and cannot simply be carried across to loads: an equivalent formulation for a converter-interfaced load would have to be justified on its own terms, with the interaction pathway, the governing rating and the acceptance criteria all re-derived. Work in that direction is under way. EPRI has published a screening approach for subsynchronous torsional interaction with AI training loads, based on interharmonic current gains, together with a companion framework for estimating acceptable load-oscillation limits with turbine-generator shaft impacts in view [33], [34]. What is missing is therefore not the concept of screening but its settled form for loads — and the habit of asking for it at the connection stage.

Nor is this a hypothetical concern being raised for the first time here. NERC's Large Loads Task Force documented an observed forced oscillation from a cryptocurrency-mining facility with a peak-to-peak amplitude of around 25 MW [31], and NERC's May 2026 reliability guideline now asks large-load entities to declare any cyclical demand ramps *and their associated frequencies*, with particular attention to the subsynchronous range, and triggers electromagnetic-transient screening where the load profile carries significant fluctuation between 5 and 60 Hz [32]. The regulatory expectation described in Section A1.8 is not a proposal awaiting adoption; in this respect it has already been written down. Individual utilities have gone further and set numbers. Among the limits now reported in the literature are ERCOT’s restriction on repetitive large-load excursions exceeding 10 MW within a sliding five-second window; LIPA’s cap of 3.5 MW on the sum of adjacent spectral bins between 5 and 55 Hz over a ten-second window; AESO’s limit of 16 kW per 100 milliseconds with a total permissible change of 160 kW; and ATC’s 25 MW limit on active-power oscillation within five seconds for loads above 200 MW [30]. The band those limits address — broadly 5 to 55 Hz — is the torsional band rather than the band in which the

measured fundamental of an AI workload sits, which is a further reason to treat the harmonic and interharmonic content as the quantity of interest.

Which brings the appendix back to its own architecture. Every layer described in Sections A1.4 to A1.6 is, viewed from outside the fence, a filter that removes energy from precisely these bands before it can reach the alternating-current terminals. The rack capacitors take the fastest content, from microseconds to the order of a hundred milliseconds. The facility battery, coupled to the bus through the ultra-low-inductance path of Section A1.5, takes what remains in the low hertz range — the band that reaches furthest into the system and does the most damage when it arrives. Progressive power capping stretches the slow envelope over minutes, so that even the ramp presents no step. What the converter then exchanges with the grid is a smooth, low-slew net power, and the residual energy at torsional and inter-area frequencies is reduced — by how much, and whether far enough, being a matter to be shown by study against the agreed connection-performance envelope rather than asserted. Suppressing an oscillation at the rack costs capacitors; suppressing it at the shaft is not possible at all. The engineering case for shaping the load at source is, in the end, not an efficiency argument. It is a reliability argument about machines the operator of the data centre will never see.

That is also the strongest reason for the model-provision obligation of Section A1.8 and the modelling discipline of Section 7.5. A mode interaction cannot be screened if the load cannot be modelled, and a load of this complexity cannot be modelled from a nameplate rating and a power factor. The high-fidelity model, exercised in the loop before energisation, is what turns this section's concern from a warning into a calculation.