

The National Compute Grid:

Pooled Infrastructure for the Frontier Research Ecosystem

[Authors Anonymized for Submission to the 2027 Proceedings of MLSys]

grid@nationalcompute.com • nationalcompute.com

Abstract

Frontier AI progress in the United States and allied nations is increasingly constrained not just by a shortage of compute, but by its fragmentation. Single-tenant clusters in the independent ecosystem frequently exhibit node occupancy below 60%, MFU below 30% and average net FLOPs utilization below 15 percent, while individual teams struggle to forecast capacity needs accurately in advance. The result is a systemic misallocation: enormous capital is deployed into capacity that sits idle, while the researchers best positioned to advance the frontier cannot access capacity when they need it. We introduce the National Compute Grid, a shared infrastructure network that pools AI compute across laboratories, clouds, and chip architectures and clears it in real time, in the tradition of the interconnected power grid and the Interstate Highway System. The Grid's technical architecture descends from the Borg, xBorg, and Quota Marketplace scheduling lineage that raised node occupancy at Google from roughly 65 percent to 99 percent at gigawatt scale. We formalize the statistical basis for pooling gains, describe the Grid's market design, in which members transact in Grid Credits and pay only for useful completed work ("goodput") rather than GPU-hours, and publish an open request for comment to the security, research and policy communities ahead of a broader rollout. The Grid enters general availability on January 15, 2027, with more than \$1.1B in long term grid reservations accumulated during the 2026 technical preview phase, and a consortium target of 2 GW of pooled capacity by 2030. We argue that a readily available, neutral pool for compute is the enabling infrastructure for the priorities articulated in America's AI Action Plan and the Genesis Mission. We invite the research community, infrastructure providers, and public institutions to participate in the development of full stack scaling and alignment standards for the national interest.

1 Introduction

Every general-purpose technology in American history has passed through the same bottleneck: the transition from isolated capability to shared infrastructure. Electricity became transformative when isolated generators were interconnected into a grid. Transportation reshaped the economy when disconnected roads were unified into the Interstate Highway System. Finance scaled when bilateral credit gave way to exchanges and clearinghouses. In each case, the underlying capability existed for decades before the infrastructure that made it fungible, reliable, universally accessible, and secure.

Artificial intelligence has reached this transition point. The United States leads the world in the design of AI capabilities. However, the compute that powers frontier research is organized the way electricity was organized in 1900: as isolated, single-tenant installations, each sized for its owner's peak demand, each idle most of the time, and none able to share surplus with a neighbor. This paper describes the National Compute Grid, an effort by a coalition of frontier laboratories, chip manufacturers, research universities, and investors to

build the missing infrastructure layer, and the evidence and design principles behind it.

This is a systems paper, not a political manifesto. The authors of this paper compete vigorously with one another on models, products, and talent. We do not believe, however, that we should compete in the wastefulness of our infrastructure. A shared grid makes each of us stronger, makes the American and allied frontier ecosystem more resilient, and increases returns to the public.

2 The Utilization Problem

2.1 Fragmentation by default

The prevailing procurement model for frontier AI compute is the multi-year, single-tenant lease: a laboratory hast to forecast its demand years ahead, signs for dedicated capacity at that forecast, and bear the full cost of that lease, whether the nodes are optimally utilized or not. This model was inherited from an era when only a handful of organizations trained large models. It has three structural failures:

1. **Forecast error is irreducibly large.** Frontier research demand is driven by research outcomes, which are unpredictable by definition. A breakthrough triples demand overnight; a negative result strands a cluster. In our members' experience, self-reported demand forecasts have run approximately 2x below realized consumption, in both directions across time. No amount of planning sophistication removes this error, because the error is the research. The same pattern holds inside the largest fleets: as the Quota Marketplace deployment at Google documents, demand ranges from scheduled training runs to short bursts for demos and conference sprints, and the assumption that chip requirements can be predicted for a full quarter is highly questionable [6].
2. **Peak-sizing guarantees idleness.** A single-tenant cluster must be sized for peak demand, and frontier demand is extremely peaked: a large pre-training run may consume an entire fleet for weeks, followed by months of ablations, evaluation, and post-training that use a fraction of it. Across independent frontier laboratories, we observe average net FLOP utilization of single-tenant data centers below 15 percent (Figure 1). The capital simply burns.
3. **Idle capacity cannot move.** When one laboratory's cluster idles while another's researchers queue, no mechanism exists to move the surplus. Fleet state lives in spreadsheets; contracts prohibit resale; schedulers stop at the organizational boundary. The ecosystem-level loss compounds the firm-level loss.

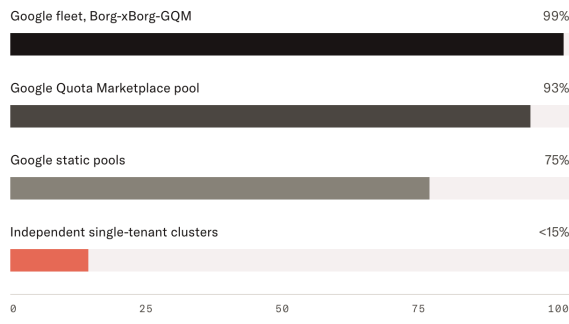


Figure 1: Net utilization across allocation regimes. Independent single-tenant clusters vs. static pools, the Quota Marketplace pool, and the pooled Google fleet under the Borg-xBorg-GQM stack [6].

Figure 1: Net utilization across allocation regimes. Independent single-tenant clusters vs. static pools[6].

2.2 The cost of fragmentation

The consequences are not confined to wasted capital, though the waste is enormous: at current buildout rates, sub-15-percent net FLOPs utilization implies that an absurdly large majority of frontier AI capital expenditure purchases idleness. The deeper cost is to the structure of the ecosystem. When compute must be bought years ahead in units only the largest organizations can afford, scale becomes a prerequisite for participation rather than a reward for results. New entrants, academic groups, and government laboratories are priced out not by the cost of compute but by the cost of committing to compute. America's AI Action Plan identifies exactly this failure, observing that companies seeking large-scale compute must often sign long-term hyperscaler contracts beyond the budgetary reach of most academics and many startups, and directs the government to accelerate a healthy unified pool for compute on the model of spot and forward commodity markets [2]. The National Compute Grid is the private-sector construction of that pool.

3 The Statistical Basis for Pooling

3.1 Multiplexing gains at ecosystem scale

The case for pooling is not novel mathematics; it is the same statistics that underlie the telephone network [7], the interconnected power grid, and portfolio theory [8]. Its application and measurement at AI-fleet scale, however, is new, and the magnitudes involved are unusually large because frontier AI demand is unusually bursty.

Consider N participants whose instantaneous compute demands have mean μ , standard deviation σ (coefficient of variation $c_v = \sigma/\mu$), and average pairwise correlation ρ . A participant provisioning alone must carry headroom proportional to c_v above mean demand to meet a given service level. A pool serving all N participants requires headroom, per unit of mean demand, proportional to:

$$c_v \sqrt{\frac{1 - \rho}{N}} + \rho$$

Equation 1

Two properties of Equation 1 guide our scaling decisions (Figure 2). First, required headroom, equivalently structural idleness, falls as $1/\sqrt{N}$ while demand streams are imperfectly correlated: pooling ten independent laboratories cuts the waste per FLOP by a factor of roughly three, and pooling a hundred by a factor of ten. Second, the gains saturate at a correlation floor $c_v \sqrt{\rho}$: when everyone's demand spikes together, as during synchronized scaling races or the release of a new chip generation, no amount of pooling helps, and the residual risk must be carried by physical reserve capacity. The Grid's design takes both properties seriously: it maximizes N and demand diversity (across laboratories, modalities, workload types, and time zones) to exploit the $1/\sqrt{N}$ regime, and it argues for explicitly ring fenced

experimental/reserve capacity (Section 5) against the correlation floor.

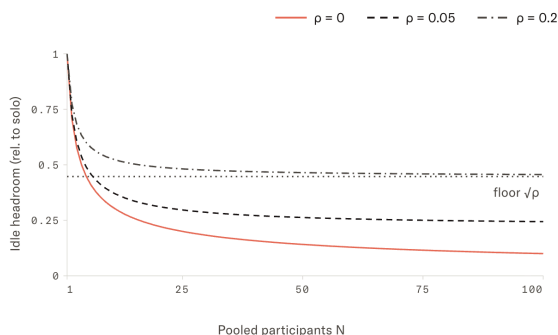


Figure 2: Idle headroom per unit of demand as a function of pool size N (Equation 1). Gains decay as $1/\sqrt{N}$ toward a correlation floor $\sqrt{\rho}$.

Figure 2: Idle headroom per unit of demand as a function of pool size N (Equation 1). Gains decay as $1/\sqrt{N}$ toward a correlation floor $\sqrt{\rho}$.

3.2 The gigawatt-scale natural experiment

The claim that pooling can approach full utilization at frontier scale is not without precedent. It is an observed result from at least experiment at scale i.e. Google. Internally, Google operates its fleet as a single pooled system, scheduled by the Borg lineage of cluster managers [5] and, in its current generation, by the Quota Marketplace (QM), a deployed market mechanism that allocates several hundreds of thousands of ML accelerators across business units with dynamic pricing that clears roughly every minute [6]. Over the past five years, as these systems were extended across AI accelerators at gigawatt scale, fleet occupancy rose from approximately 65 percent to 99 percent.

The QM deployment also provides the cleanest controlled comparison available. For the same popular resource type, in two pools of comparable size with hundreds of thousands of chips each, the conventional static pool ran at 75 percent overall occupancy while the Quota Marketplace pool ran at 93 percent, with opportunistic (unprioritized) occupancy collapsing from 33.5 percent to 5.67 percent (Appendix, Figure A1) [6]. Occupancy in the market pool tracked supply within days of new hardware landing, driven by price signals alone, while the static pool lagged for weeks. The mechanism is also proven theoretically: the market guarantees Pareto efficiency and approximate max-min fairness even when workloads have heterogeneous values, properties that static allocation and chip-hour mechanisms provably cannot deliver [6].

Members of the National Compute Grid founding team built and operated these systems at Google, including the Quota Marketplace itself. The Grid's core engineering thesis is that the same architecture, supplemented with the VCG mechanism discussed in Section 5 of [6], operated by a neutral party, can deliver the same result across organizational boundaries rather than within one

company, and that the gains available to the fragmented independent ecosystem, starting well below 60 percent rather than at 65, are correspondingly larger.

4 System Architecture

The National Compute Grid is an interconnected network of AI compute capacity spanning multiple chip architectures (NVIDIA, AMD, TPU, and emerging silicon), multiple supply sources (member-owned fleets, specialized clouds, and hyperscalers), and multiple physical sites, presented to members as a single pool with real-time allocation (Appendix, Figure A2). Four subsystems matter most.

1. **The scheduler.** The Grid scheduler descends directly from the Borg-xBorg-QM lineage: a fleet-wide control plane that tracks every node's state, packs workloads to maximize occupancy, preempts and migrates according to published priority rules, and clears supply against demand continuously rather than through static leases, with sub-minute market cycles as demonstrated in the QM deployment [6]. Its design goal is the Google result, near-total occupancy, delivered across trust boundaries with cryptographic isolation between tenants, and with a core-selecting VCG mechanism to prevent competitors from depleting other's credits.
2. **Bring-your-own-capacity (BYOC) and fleet management.** Today each laboratory tracks its fleet across providers in manual spreadsheets. The Grid replaces this with a fleet-management layer that gives members real-time tracking and utilization telemetry for capacity they already own or lease, at no charge. Connected fleets can then opt idle capacity into the pool, where it earns income at the clearing price. BYOC is the on-ramp that converts the ecosystem's existing stranded capacity into Grid supply, mirroring QM's finding that transiently available holding-pool and buffer resources, on the order of total committed capacity, are the largest untapped source of supply [6].
3. **Spike reservations and just-in-time compute.** Members reserve clusters in bulk for 3 to 6 month windows for large training runs, replacing 24x7 multi-year leases with reservations shaped like the workload. Beneath reservations sits a continuous pool, accessible through a command-line interface and AI agent (Marshall), through which any participant can buy or sell capacity in real time. Idle reserved capacity is automatically backfilled with spot workloads, and the backfill income accrues to the reserving member, so reservations stop being a bet against your own forecast.

4. **Goodput as the unit of value.** The Grid does not sell GPU-hours. It prices completed work via Grid Credits: for training, uninterrupted contiguous runtime (an interrupted run is free); for inference, availability of declared replica counts; for batch, throughput on work completed. GPUs are an input; goodput is the product. This pricing shape moves reliability risk from the researcher, who cannot control it, to the network operator, who can, and it aligns the operator's incentives with exactly the utilization and reliability engineering the ecosystem needs.

5 Security and Governance: A Request for Comment

A pooled network is only as valuable as it is secure and reliable, and trust in shared infrastructure is both a technical and governance problem. The coalition has aligned on the destination: deep, continuous availability under published rules that no participant, including the Grid's operator, can bend. We have deliberately not finalized the rules themselves. This section states the open design questions and the candidate principles under discussion, and we publish them as an invitation. Open questions span two areas.

Security governance: In light of recent frontier cybersecurity incidents, [10] Coalition members are considering how frontier models should be trained, tested and released on the grid in a secure way at scale.

A shared grid gathers the weights, checkpoints, and training data of many labs onto hardware run by many operators. That makes the Grid one of the most attractive targets in the western AI ecosystem for well-resourced state actors. Its security will be set by the weakest connected site and by the integrity of the scheduler that decides where every workload runs. We therefore propose that security be enforced by the Grid's placement mechanism itself, not left to each member's contracts. Every workload would carry a declared security tier. Tiers would run from open-data batch and spot jobs, through regulated data such as clinical and government workloads, up to frontier training runs that hold unreleased weights. Every site would be certified to a published tier, using established reference points: the RAND security levels for model weights, FedRAMP and DoD Impact Levels for government work, and HIPAA for medical data. The scheduler would be technically unable to place a workload on capacity below its tier. The Grid Index would price tiers separately, so that members pay for higher assurance rather than assume it.

The candidate controls under discussion follow from that principle. Preemption, migration, and backfill make pooling efficient, but they are also the moments when

one tenant's data is most likely to be left behind for the next. So nodes are wiped and re-attested before every reuse, and the highest tiers require hardware confidential computing where the silicon supports it. The control plane, including the Marshall agent, would be protected like a certificate authority. Policy changes would require multi-party approval, every placement and access decision would be written to an append-only log that members can audit, and agents would act with narrowly scoped authority. Site operators would accept common physical and personnel standards, hardware chain of custody, rights of audit, and incident notification measured in hours, not weeks. Coalition training and the evaluation exchange would give members access to one another's models only through monitored interfaces, never by copying weights. Finally, the Grid's flexibility is itself a security asset. A pool that can quickly drain a compromised site and move its workloads elsewhere can contain incidents that would cripple a single-tenant cluster. We ask the security community to scrutinize the tier definitions, the attestation requirements, and the constraints on the operator itself before they are written into the Grid Constitution.

Economic governance: Involving the denomination of transactions in Grid Credits and the creation of a public Grid Index; anchor availability, which seeks to balance member-committed baseline capacity with spot participation; and neutrality constraints, such as capacity concentration limits, operator restrictions on model training.

We invite researchers, market designers, infrastructure operators, and public institutions to weigh in before these rules harden. The mechanisms of the Google Quota Marketplace took years of deployed iteration inside a single company [6]; the Grid's rules will govern shared infrastructure of national consequence, and they deserve at least as much scrutiny, applied in the open. Comments are welcome at grid@nationalcompute.com. A first Grid Constitution will be published at general availability and revised publicly against the comments received.

5.1 Coalition scaling: shared frontier research

Coalition training allows member laboratories to pool data and compute for joint training runs across organizational boundaries, with the Grid as neutral administrator of pooling, checkpoint access tiers, licensing, and billing. As frontier runs grow beyond any single independent laboratory's capacity, joint runs administered by a neutral party are the only mechanism by which the independent ecosystem can stay at the frontier. The evaluation exchange allows members to submit models for structured early-access red-teaming and benchmarking by other members before release, raising the safety and quality floor of the whole coalition while keeping evaluation independent of any single developer.

6 Alignment with National Science and Technology Policy

The Grid is a private undertaking, financed and operated by its members, at no cost to the taxpayer. It is also, deliberately, the execution of publicly articulated national priorities, and we document the correspondence here (Appendix, Figure A5).

The compute market directive. America's AI Action Plan directs the acceleration of a healthy financial market for compute, with spot and forward mechanics, so that startups and academics need not sign hyperscaler-scale long-term contracts to participate in frontier research [2]. The Grid Exchange is that market, built and capitalized by the private sector: a spot market with published clearing prices, forward capacity via reservations, and a public price index.

The Genesis Mission. Executive Order 14363 launched the Genesis Mission to double the productivity and impact of American science within a decade, integrating supercomputers, models, instruments, and datasets across the national laboratories [4]. The binding constraint on AI-for-science at national scale is flexible access to compute; the Grid's founding commitments include compute-credit grants to U.S. government laboratories, and the coalition intends the Grid to serve as interoperable civilian infrastructure for Genesis Mission workloads, on the same open terms available to every member.

National security preparedness. The Action Plan directs agreements codifying priority access to computing resources in national emergencies [2]. Among the governance proposals published for comment in Section 5 are a standing mobilization protocol and guaranteed best execution for cleared U.S. government workloads. A liquid national pool, with real-time fleet state across the western independent ecosystem, is precisely the instrument such a protocol requires; no such instrument exists today.

New market mechanisms for science. Science: A New Golden Age calls for experimentation with new market mechanisms that direct resources to problems where breakthroughs matter most [3]. A transparent price for compute is the substrate for such mechanisms: once compute clears at a public index, funders, from federal agencies to philanthropies, can denominate grants, prizes, and advanced market commitments in the actual scarce input of modern science, and researchers can spend them anywhere on the network. We regard the Grid Index as public scientific infrastructure in the same sense as a standard reference material, and we commit to keeping it open.

7 Launch, Traction, and Roadmap

The Grid entered invite-only early access in June 2026. In its first ninety days it accumulated more than \$1.1B in reservations; every major western frontier laboratory has joined the network or is in active pipeline. The Grid Exchange enters general availability on September 15, 2026, launching as the largest spot market for AI compute available to startups, universities, and government laboratories, with the on-demand command-line interface reaching general availability in October. Coalition members have set a pooled-capacity target of 2 GW by 2030 (Appendix, Figure A3), with a diversified silicon base.

A key learning during this phase was that frontier AI research is increasingly agentic. To meet the changing needs of frontier researchers, an agentic interface was added (Fig A.6) to improve the ergonomics of the Grid. This agent, codenamed “Marshall”, has been designed to recursively self improve on performance and security capabilities based on reinforcement learning from grid member feedback. We aim to publish quantitative results in a follow up paper.

The roadmap beyond launch follows the architecture of this paper: deepen availability (more members, more connected fleets, more supply integrations across specialized clouds and hyperscalers); harden the Index into a benchmark the ecosystem can write contracts against; stand up coalition training and the evaluation exchange as production services in 2027; and publish, with the research community, the utilization and allocation data that will let the pooling gains of Section 3 be measured in the open, as a matter of scientific record rather than internal telemetry.

8 Conclusion

Eighty-one years ago, Science: The Endless Frontier argued that the nation's scientific strength depends on infrastructure and institutions deliberately built for it [1]. The institutions Bush inspired carried American science through a century. The defining scientific instrument of this century is compute, and it is today organized as wastefully as electricity before the grid and roads before the Interstate. The undersigned believe the future of AI should be built as a community. We have committed our laboratories, our fleets, and our capital to a shared network that makes the whole American and allied frontier ecosystem stronger than the sum of its silos. The National Compute Grid opens on September 15. We invite every builder, and the public institutions that serve them, to connect.

References

- [1] V. Bush. Science: The Endless Frontier. A Report to the President. U.S. Government Printing Office, 1945.
- [2] The White House. Winning the Race: America's AI Action Plan. July 2025.
- [3] M. Kratsios, Office of Science and Technology Policy. Science: A New Golden Age. A Report to the President. July 2026.
- [4] Executive Order 14363, "Launching the Genesis Mission." Federal Register 90, November 2025.
- [5] A. Verma, L. Pedrosa, M. Korupolu, D. Oppenheimer, E. Tune, and J. Wilkes. Large-scale cluster management at Google with Borg. In Proceedings of EuroSys, 2015.
- [6] B. Sivan, R. Paes Leme, M. Tiuca, I. McFarlane, V. Gkatzelis, N. Mehta, S. Hassas Yeganeh, V. Mirrokni, and A. Vahdat. Quota Marketplace: Dynamic Pricing for Efficient Allocation of ML Training Resources. In Proceedings of OSDI, 2026.
- [7] A. K. Erlang. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. Elektroteknikeren, 13:5-13, 1917.
- [8] H. Markowitz. Portfolio selection. The Journal of Finance, 7(1):77-91, 1952.
- [9] M. Vuppapapati, G. Fikioris, R. Agarwal, A. Cidon, A. Khandelwal, and E. Tardos. Karma: Resource allocation for dynamic demands. In Proceedings of OSDI, 2023.
- [10] OpenAI. Hugging Face Incident and the Road Ahead. OpenAI Blog, 2026.

Appendix A: Supplementary Figures

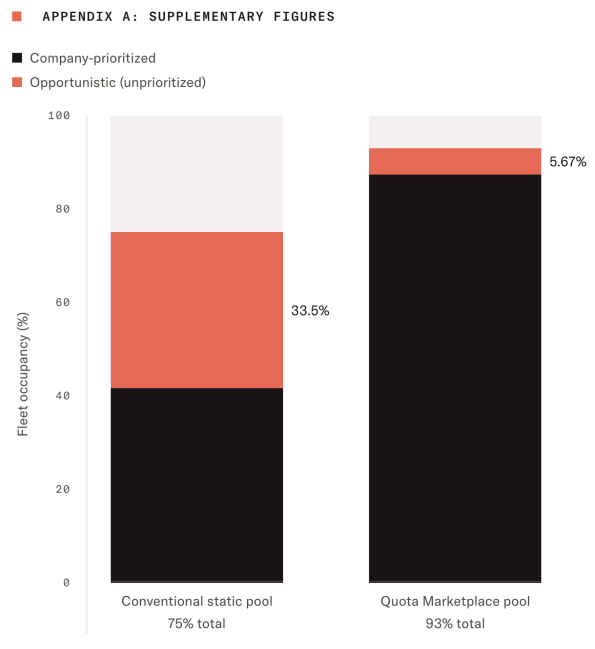


Figure A1: Controlled comparison from the Quota Marketplace deployment at Google [6]. For the same popular resource type in two pools of comparable size (hundreds of thousands of chips each), overall occupancy rises from 75 to 93 percent under the market mechanism, and unprioritized opportunistic occupancy collapses from 33.5 to 5.67 percent.

Figure A1: Controlled comparison from the Quota Marketplace deployment at Google [6]. For the same popular resource type in two pools of comparable size (hundreds of thousands of chips each), overall occupancy rises from 75 to 93 percent under the market mechanism, and unprioritized opportunistic occupancy collapses from 33.5 to 5.67 percent.

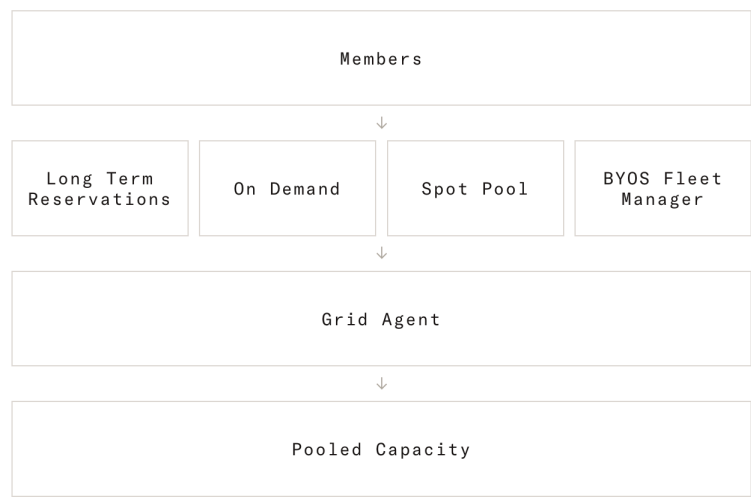


Figure A2: The National Compute Grid stack. Members access pooled capacity through Long Term Reservations, On Demand, Spot Pool, and BYOS Fleet Manager, coordinated by the Grid Agent.

Figure A2: The National Compute Grid stack. Members interact through three surfaces (the Grid Exchange, spike reservations, and BYOC fleet management), all cleared by a scheduler in the Borg-xBorg-GQM lineage against pooled multi-architecture supply and an endowment-financed Reserve.

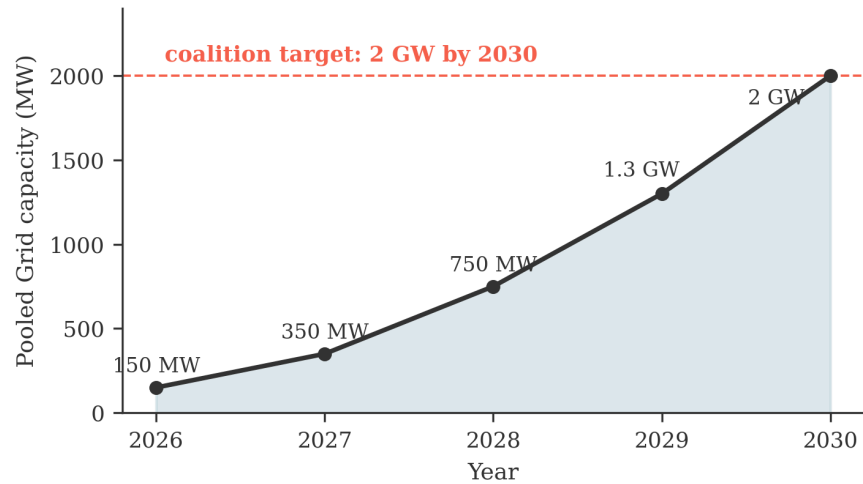


Figure A3: Illustrative trajectory of pooled Grid capacity to the coalition target of 2 GW by 2030.

Utilization Definitions

1. Goodput = <https://cloud.google.com/blog/products/ai-machine-learning/goodput-metric-as-measure-of-ml-productivity>
1. Node utilization = node occupancy i.e % of time a node has a job scheduled on it
2. MFU = <https://inference.semianalysis.com/glossary/model-flops-utilization>
3. Net FLOPs utilization (aka compound utilization) = a directional measure of overall flops that are productively consumed, estimated most frequently by 2*3