Every Megawatt Counts: The Path to Building AI Factories with DSX MaxLPS

Ajay Kamalvanshi

Senior Director, NVIDIA



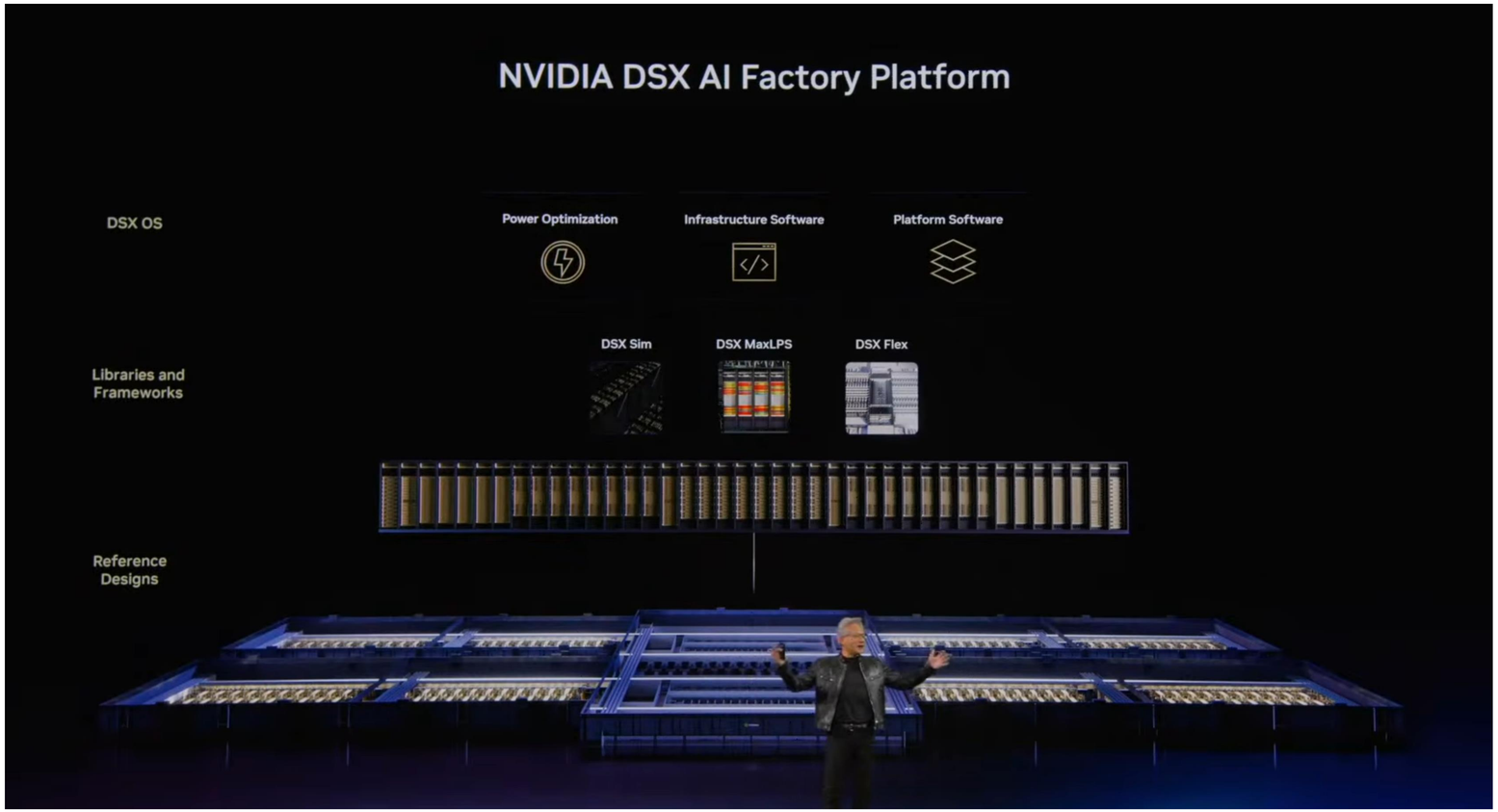
Agenda

- Background and Problem Statement
- Architecture and Theory of Operation
- DSX Stack Integration
- Ecosystem Partners
- Challenges and Collaboration

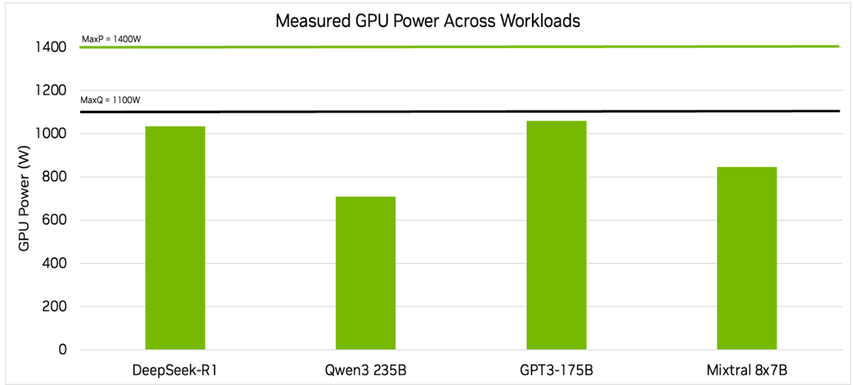
AI Factory Output Is Constrained by Power

Energy is at the foundation of the 5-layer AI cake: power constraints limit growth

There is so much power that is squandered away in AI Factories… every unused watt is revenue lost – Jensen GTC 26 keynote



Example



Measured GPU Power Across AI Workloads on GB300
Inference workloads, no power capping

Static provisioning strands power

Workload power needs vary; not all workloads need Max-P power

Problem: Power is not optimally utilized

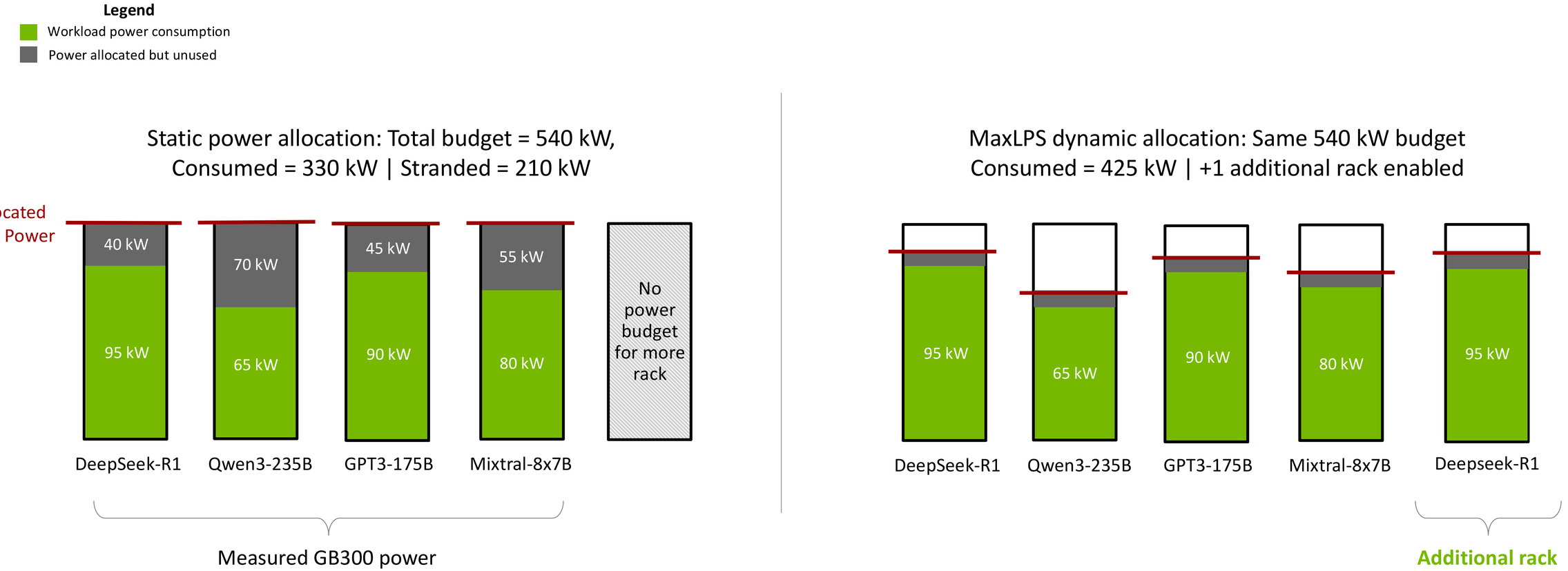
Shows 33% more GB200 racks based on Kimi-K2.5 FP4 Dynamo TRT-LLM 8K/1K

Rack Power	MaxP	MaxLPS
GPU TGP setting	1,200W	**800W**
Provisioned Rack Power	120kW	**90kW**
Throughput	1x	**Same**
Racks in Same Power Envelope	1x	**33% More Racks***

* 120kW / 90kW = 1.33x

MaxLPS recovers stranded power

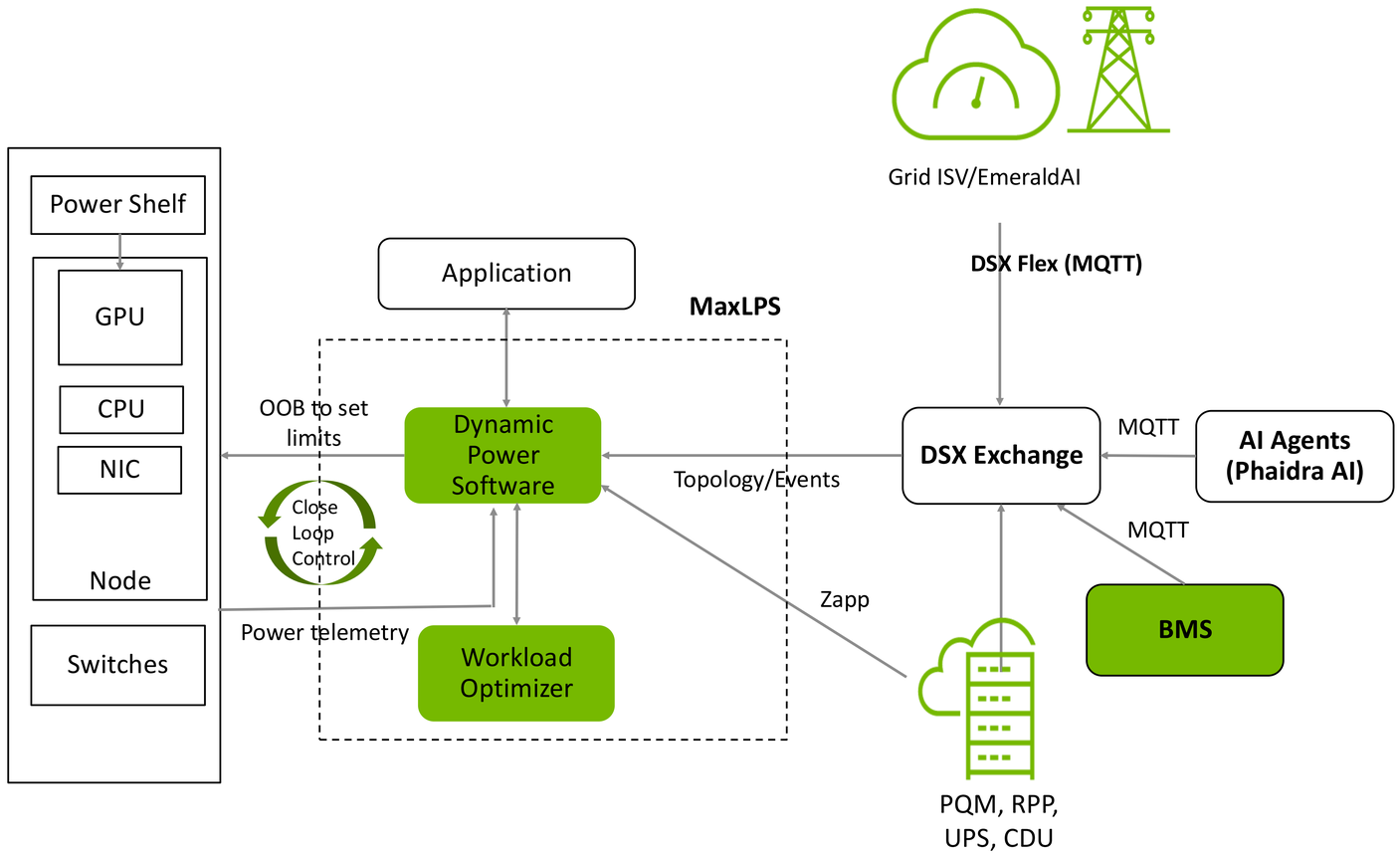
Deploying additional GPUs within the same power budget



MaxLPS Architecture

Dynamic Power Software (DPS), Workload Profiles, and Optimization Policies

- Maximize tokens/watt with dynamic power and performance optimizations.
- Enable dynamic power management based on workloads and optimization goals
  - Max-P: maximizing performance to meet SLA
  - MaxLPS: maximizing node count with improved perf/provisioned-watt
- Unlock stranded utility power (100GW in North America) with ISV like Emerald AI
- Workload Optimizers:
  - Workload Profile (memory clock, switch optimization, process optimization)
  - Utilization/observed power-based prediction
  - Heuristic/DL based prediction



Closed-Loop Control - Details

Guarantees compliance within total DC power budget | Simple algorithm ensures robustness

Closed-Loop Control Architecture



Power Telemetry

Closed Loop

DPS

Power Provisioning

MaxLPS Decision Algorithm (per Node)

START, for each node

Consumed power vs allocated power

much less (>20% gap)

on target (5-20% gap)

near capacity (within 5%)

Reduce allocation to node

Increase allocation to node

No action

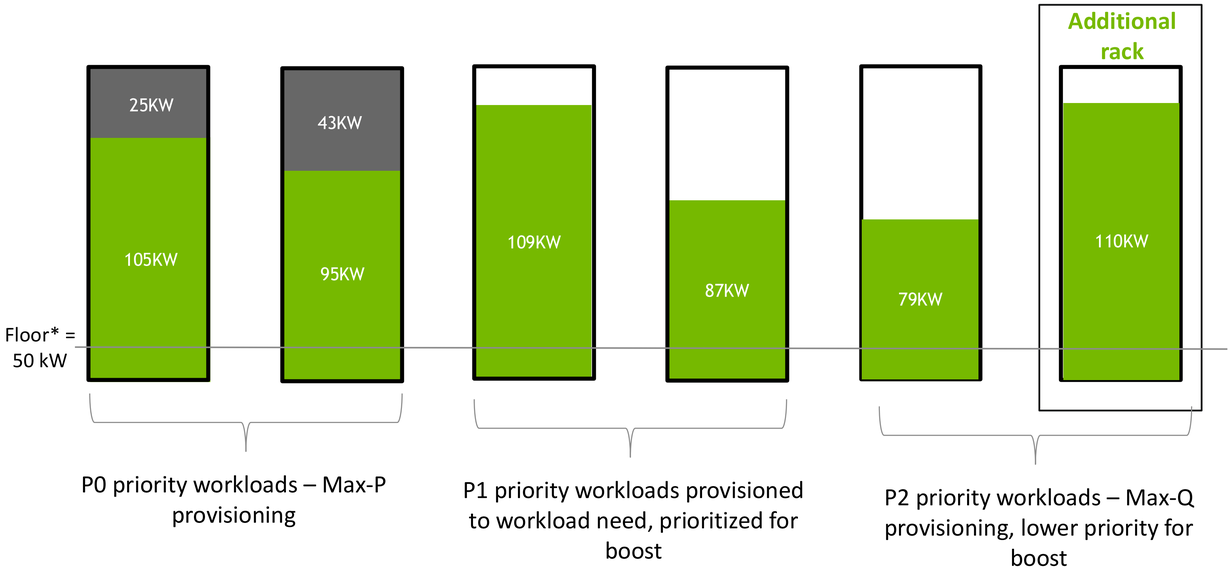
Update available DC power

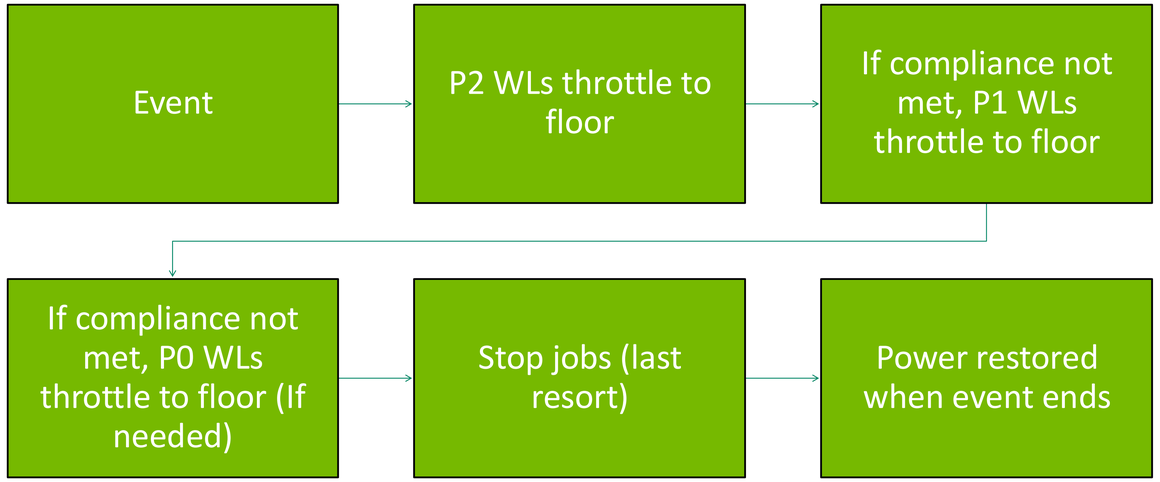
END

Continuous Loop

Power Compliance with Workload Priority

Handling SLA and power events like Load Shedding



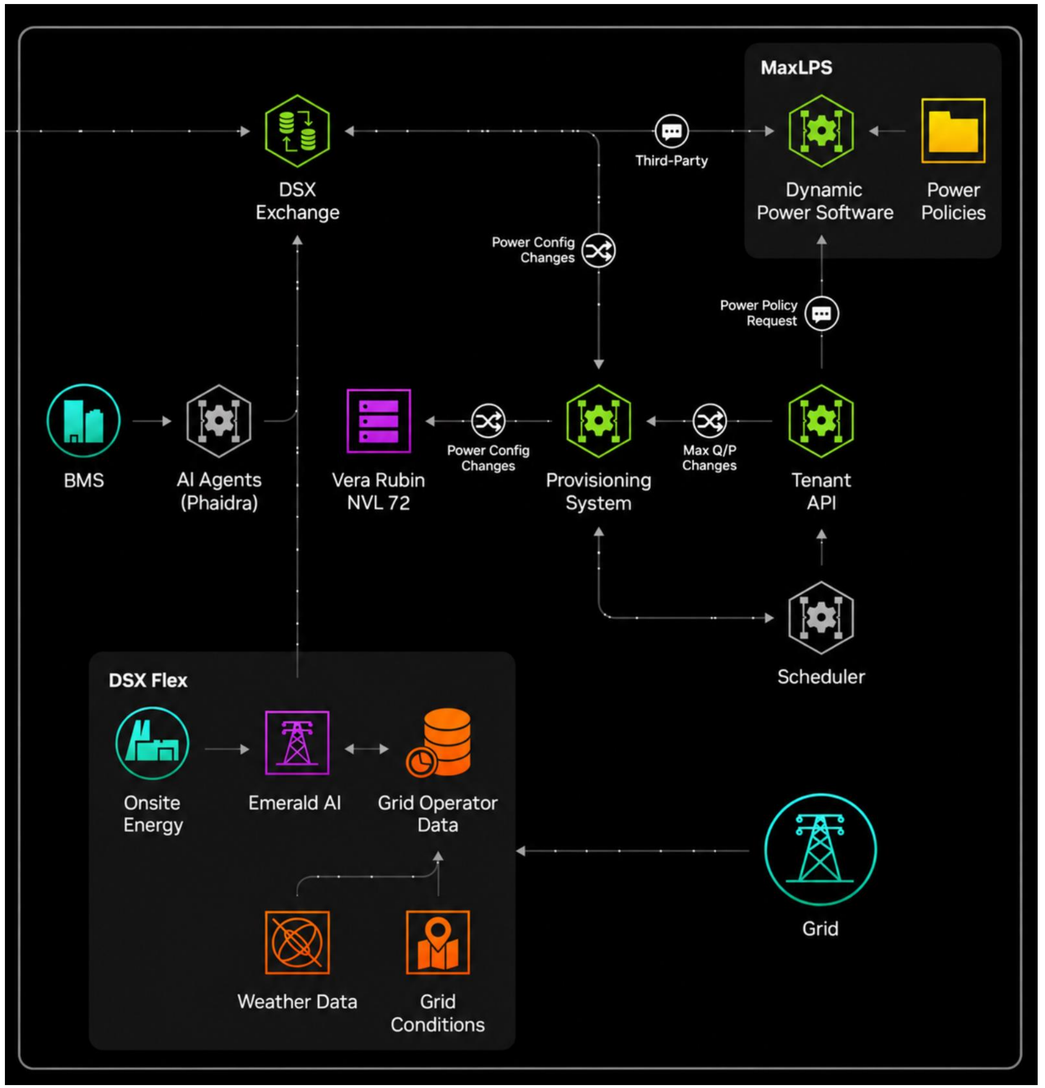


Workload priorities are set along with cap and floor power as part of power policy setting

DSX Stack Integration for Vera Rubin

Blueprint for scaling MaxLPS to 1GW Datacenter

- Scaling for large-scale clusters using DSX stack
- Integration with Provisioning systems: NICo, and Launch Layer
- Supporting multi-tenants through DSX's BGP, EVPN, and VXLAN technologies
- OT interfaces through DSX Exchange
- Include NVLink Switch and LPU MaxLPS to improve throughput for premium and ultra segments



AI Clouds Deploy NVIDIA DSX, Lambda Adopts MaxLPS

Lambda reclaiming stranded power to increase compute density within the same infrastructure footprint.



"NVIDIA DSX MaxLPS paves the way to reclaiming stranded capacity and converting it into real-world usage, with significantly more compute density in the same footprint. It's a scalable and sustainable blueprint for how we can operate and expand our infrastructure moving forward."

—***Dave Ward,*** President, Cloud Services, Lambda

CoreWeave · firmus · GMI · GMO INTERNET · indosat OOREDOO HUTCHISON · IREN · Lambda · Mistral AI · NAVER Cloud

NEBIUS · NSCALE · ResetData · SHARON AI · Visionbay.ai · YOTTA · YTL AI Cloud YTL GROUP

NVIDIA

Challenges & Collaboration

Moving towards no stranded power

- Improve latency and reliability of BMC Redfish interfaces with IBV/IFVs
- Support fast streaming telemetry to scale across the datacenter – move from Modbus/Redfish to UDP-based streaming interfaces in collaboration with the OCP GPU Management Interfaces workgroup
  - NVIDIA Zapp as a reference interface implementation from NV for v2.0.0 spec
- Incorporate fast streaming telemetry on power devices like PQMs and BESS for communication with the system management stack
- DSX Stack as a reference stack for AI factory – simulate with DSX Sim

