GPGPU Comparison for Local AI & Scientific Computing

NVIDIA Blackwell (RTX PRO 6000), Grace Blackwell (DGX Spark), and NVIDIA Ampere (RTX A6000/ RTX A5000)

Reference Budget:
$150,000 (GPU or appliance procurement) + Off-Grid Energy Infrastructure

Onri Jay Benally

December 17, 2025

Abstract

This document covers a procurement-oriented comparison of four NVIDIA GPU pathways under a fixed $150,000 reference budget, while explicitly separating (i) peak theoretical accelerator metrics, (ii) deployable system-level constraints, and (iii) off-grid energy sizing. It includes device specifications, derived cost-and-power efficiency metrics, interconnect and scaling implications, and battery-sizing sensitivity to facility power usage effectiveness (Power Usage Effectiveness, PUE), auxiliary "host" overhead, and storage efficiency assumptions. All quantitative claims are sourced from vendor documentation, technical literature, or publicly visible pricing references.

Contents

List of Tables

1 Scope, definitions, and assumptions

1.1 Intuitive explanation

A graphics processing unit (GPU) is, in practice, a massively parallel chip originally designed to draw images quickly, but now used for general-purpose computing. General-Purpose computing on Graphics Processing Units (GPGPU) means taking problems that can be split into many similar calculations and running them across thousands of small compute units. In machine learning, those calculations are usually large matrix multiplications; in scientific computing, they are often stencil updates, particle interactions, or batched linear algebra.

1.2 Framing: performance claims are typed, not scalar

Vendor "throughput" numbers are typed by *precision format*, *operation mix*, and sometimes by *sparsity assumptions*. A peak value like "4 PFLOP/s FP4 (with sparsity)" is not interchangeable with "120 TFLOP/s FP32," because the former assumes (i) a low-precision format and (ii) a 2:4 structured sparsity pattern that can, in theory, double Tensor Core throughput relative to dense execution [1, 2, 3]. In other words, if you cannot or do not enforce 2:4 sparsity in weights (or activations where supported), the headline sparse peak is not physically reachable, and you should model your expected throughput as closer to the dense case, typically by dividing the sparse peak by roughly two.

1.3 Reference budget and what it does *not* include

The $150,000 reference budget is treated as accelerator or appliance procurement only, to maintain comparability with the earlier draft. In real deployments, a complete system also needs:

- CPU platforms (and their memory bandwidth), storage, networking, racks, and power distribution.
- Cooling infrastructure and facility overhead (captured approximately by PUE).
- Software and operational overhead (driver qualification, imaging, monitoring).

These omissions are addressed explicitly in Sections 6 and 7 as sensitivity factors, rather than being hidden.

1.4 Pricing, power, and energy-storage assumptions

- **Prices:** Reference prices are sourced from vendor listings or widely cited launch/market context. The RTX PRO 6000 Blackwell Workstation Edition uses an observed retail listing price; DGX Spark uses NVIDIA Marketplace pricing; Ampere cards use widely cited launch-era reference pricing. Prices are time-varying by nature. See Table 2.
- **Nameplate power:** For safety margins, maximum power (or power supply rating, in the case of DGX Spark) is used. The RTX PRO 6000 Blackwell Server Edition is listed up to 600 W [4]; RTX A6000 is 300 W [5]; RTX A5000 is 230 W [6]; DGX Spark power supply is 240 W [7, 8].
- **Battery:** Tesla Powerwall 3 nominal energy is 13.5 kW h, and it can deliver up to 11.5 kW continuous AC power per unit (configuration-dependent) [9]. The earlier draft assumed $8,200 installed per Powerwall; real installed pricing varies widely, and published breakdowns often exceed the battery-only line item [10]. To preserve the earlier framing, $8,200 per unit is retained as a comparability knob.
- **Duty cycle:** An 8-hour continuous-load off-grid run defines the energy pool requirement.

2 Device-level specification snapshots

Table 1: Device specifications (representative, vendor-specified peak metrics).

Device	AI/ Tensor Peak	FP32 Peak (dense)	Memory (type, capacity)	Power + interconnect notes
RTX PRO 6000 Blackwell	4 PFLOP/s FP4 (with sparsity) [4]	120 TFLOP/s [4]	96 GB GDDR7 ECC; 1,597 GB/s (server listing) [4]	Up to 600 W; PCIe Gen 5 x16; supports Multi-Instance GPU (MIG) up to 4 instances @ 24GB [4]
DGX Spark (GB10 Grace Blackwell)	1 PFLOP/s FP4 (theoretical, with sparsity) [11, 7]	—	128 GB LPDDR5x coherent unified system memory; 273 GB/s [11, 7]	240 W PSU; SoC TDP 140 W [8, 12]; integrated CPU+GPU; includes ConnectX-7 NIC [11]
RTX A6000 (Ampere)	309.7 TFLOP/s Tensor (peak) [13]	38.7 TFLOP/s [13]	48 GB GDDR6 ECC; 768 GB/s [13]	Max 300 W; PCIe Gen 4 x16; 2-way NVLink bridge option [5]
RTX A5000 (Ampere)	222.2 TFLOP/s Tensor (peak) [14, 15]	27.8 TFLOP/s [14, 15]	24 GB GDDR6 ECC; 768 GB/s [15]	Max 230 W; PCIe Gen 4 x16; 2-way NVLink bridge option [6]

3 Price assumptions and budget arithmetic

Table 2: Reference price assumptions (time-varying by nature).

Device	Reference Unit Price	Source Type	Reference
RTX PRO 6000 Blackwell (Workstation)	$8,999.99	Retail listing	[16]
DGX Spark (appliance)	$3,999.00	NVIDIA Marketplace	[11]
RTX A6000 (Ampere)	$4,650.00	Launch/market context	[17]
RTX A5000 (Ampere)	$2,500.00	Conservative “street” knob	[18]

Note on “MSRP” vs “list” vs “street.” Workstation-class accelerators are frequently sold through OEM and distribution channels, where prices drift with availability, warranty terms, and regional procurement. For this reason, Tables 3–4 should be treated as a *reference envelope*, not a price guarantee.

4 Aggregate compute, memory, and bandwidth under a $150,000 reference spend

Table 3: Aggregates under $150,000 procurement (accelerator/appliance-only).

Solution	Units	Leftover Budget	Total AI Peak	Total FP32 Peak	Total Memory Pool	Total Memory BW
RTX PRO 6000 Blackwell	16	$6,000.16	64.0 PFLOP/s (FP4, sparse)	1,920 TFLOP/s	1,536 GB (GDDR7)	25.552 TB/s
DGX Spark	37	$2,037.00	37.0 PFLOP/s (FP4, sparse)	—	4,736 GB (unified system memory)	10.101 TB/s
RTX A6000	32	$1,200.00	9.910,4 PFLOP/s (Tensor peak)	1,238.4 TFLOP/s	1,536 GB (GDDR6)	24.576 TB/s
RTX A5000	60	$0.00	13.332 PFLOP/s (Tensor peak)	1,668 TFLOP/s	1,440 GB (GDDR6)	46.080 TB/s

Interpretation guardrails. The DGX Spark "memory pool" is distributed across 37 separate systems; it is not a single shared coherent pool across all units. Conversely, the PRO 6000 and Ampere cards are discrete accelerators whose effective pool for a *single* model shard depends on per-node GPU count, NVLink availability, and distributed training strategy (data parallel, tensor parallel, pipeline parallel).

5 Derived efficiency metrics (per-dollar, per-watt, and per-GB)

Table 4: Per-unit derived metrics (using reference prices and nameplate power).

Device	AI Peak/ $	AI Peak/ kW	FP32/ kW	GB/ $	GB/s / W	Comment
RTX PRO 6000 Blackwell	4/8999.99 = 4.44×10^{-4} PFLOP/s USD^{-1}	4/0.6 = 6.67 PFLOP/s kW^{-1}	120/0.6 = 200 TFLOP/s kW^{-1}	96/8999.99 = 0.0107 GB USD^{-1}	1597/600 = 2.66 GB s^{-1} W^{-1}	Peak AI assumes 2:4 sparsity [4, 1]
DGX Spark	1/3999 = 2.50×10^{-4} PFLOP/s USD^{-1}	1/0.24 = 4.17 PFLOP/s kW^{-1}	—	128/3999 = 0.0320 GB USD^{-1}	273/240 = 1.14 GB s^{-1} W^{-1}	Includes CPU+GPU+NIC; power is PSU rating [7, 8]
RTX A6000	0.3097/4650 = 6.66×10^{-5} PFLOP/s USD^{-1}	0.3097/0.3 = 1.03 PFLOP/s kW^{-1}	38.7/0.3 = 129 TFLOP/s kW^{-1}	48/4650 = 0.0103 GB USD^{-1}	768/300 = 2.56 GB s^{-1} W^{-1}	Useful when per-GPU memory size matters more than raw GPU count [13]
RTX A5000	0.2222/2500 = 8.89×10^{-5} PFLOP/s USD^{-1}	0.2222/0.23 = 0.97 PFLOP/s kW^{-1}	27.8/0.23 = 121 TFLOP/s kW^{-1}	24/2500 = 0.0096 GB USD^{-1}	768/230 = 3.34 GB s^{-1} W^{-1}	Often bandwidth-rich per-dollar, but increases management and network complexity [15]

6 Topology, scaling, and communication overhead

6.1 Why topology dominates beyond "how many FLOPs"

When you scale from 1 GPU to many GPUs, the system increasingly becomes a communication machine. Gradient synchronization, tensor-parallel all-reduces, and activation checkpointing traffic can dominate wall-clock time. NVIDIA Collective Communication Library (NCCL) implements multiple collective algorithms, and its scaling behavior depends on the interconnect (PCIe vs NVLink vs network) and on GPU count [19]. In other words, if your bottleneck is all-reduce bandwidth, buying more GPUs without improving interconnect can increase cost while reducing realized performance.

6.2 PCI Express (PCIe) generations and why Gen 5 matters here

PCI Express 5.0 supports 32.0 GT/s per lane per direction of raw bandwidth [20]. Increasing generation primarily changes achievable host-to-device and peer-to-peer transfer capacity. In practice, this matters when:

- Your workload streams data from host memory or storage (data pipelines, out-of-core training).
- Your model is too large for single-GPU memory and relies on frequent transfers.
- You are operating without NVLink and therefore depend on PCIe peer-to-peer.

6.3 "Minimum viable cluster" sketches (not budgeted, but operationally decisive)

To avoid hiding practical deployment burden, Table 5 sketches plausible packaging. These sketches do not add cost into the $150,000 procurement number; rather, they highlight that the operational footprint differs radically.

Table 5: Cluster packaging sketches (illustrative).

Solution	Likely packaging	Scaling constraints	Operational notes
RTX PRO 6000 Blackwell (16 GPUs)	2 × 8-GPU PCIe Gen 5 servers, or 4 × 4-GPU workstations	Without NVSwitch-class fabrics, cross-node collectives are network-limited; within-node depends on PCIe/NVLink availability	High power density (600 W class GPUs) pushes liquid cooling sooner; fewer GPUs reduces management overhead
DGX Spark (37 units)	37 mini-systems plus top-of-rack switching (and power distribution)	Distributed training becomes "small-node" and network-heavy; local per-box memory is large but fragmented	Excellent for prototyping and inference farms; less ideal for tightly coupled multi-GPU training
RTX A6000 (32 GPUs)	4 × 8-GPU PCIe Gen 4 servers, or 8 × 4-GPU workstations	2-way NVLink helps only within paired GPUs; scaling beyond 2 GPUs requires PCIe/network collectives	Large per-GPU VRAM improves shard size, reducing communication volume for some parallelization schemes
RTX A5000 (60 GPUs)	7 × 8-GPU servers plus 1 × 4-GPU node (or many smaller nodes)	Many-GPU management overhead, and network becomes the bottleneck quickly	Attractive on aggregate bandwidth and FP32, but complexity is a first-order cost

7 Energy infrastructure, off-grid feasibility, and sustainability

7.1 Power Usage Effectiveness (PUE) and why it belongs in off-grid math

PUE is defined as total facility power divided by IT equipment power. For many facilities, average PUE has historically hovered around roughly 1.5–1.6, while highly optimized facilities can achieve near 1.2 or lower [21, 22]. In off-grid contexts, PUE is not just a "data center" concept; it captures everything you must power beyond the chips (cooling, power distribution losses, pumps, fans, controllers).

7.2 Battery sizing equations

Let P_{IT} be the nameplate IT load (kW), f_{host} be a multiplier for non-GPU overhead (CPU, RAM, storage, networking), and PUE capture facility overhead. The 8-hour energy is:

$$E_{8\mathrm{h}} = P_{\mathrm{IT}}\ f_{\mathrm{host}}\ \mathrm{PUE}\ t$$

To map this to required battery energy, include an overall discharge-path efficiency η and a planning depth-of-discharge DoD:

$$E_{\mathrm{battery}} = \frac{E_{8\mathrm{h}}}{\eta\ \mathrm{DoD}}$$

Powerwall count is then $\lceil E_{\mathrm{battery}}/13.5\ \mathrm{kWh} \rceil$ [9].

7.3 Results: optimistic vs engineering vs "legacy-room"

Table 6 shows three scenarios. The earlier draft roughly corresponds to the optimistic case (nameplate-only, no losses). The engineering and legacy-room cases are intended as conservative electrical engineering planning baselines, not as worst-case fear.

Table 6: Off-grid feasibility (8-hour run) under multiple facility assumptions.

Build	Scenario	IT Power	Facility Power	8h Energy	Battery Energy Req.	Powerwalls
RTX PRO 6000 Blackwell	Optimistic (no losses)	9.60 kW	9.60 kW	76.8 kW h	76.8 kW h	6
RTX PRO 6000 Blackwell	Engineering ($f_{\mathrm{host}} = 1.25$, PUE=1.2, $\eta = 0.90$, DoD=0.90)	9.60 kW	14.40 kW	115.2 kW h	142.22 kW h	11
RTX PRO 6000 Blackwell	Legacy-room ($f_{\mathrm{host}} = 1.25$, PUE=1.56, $\eta = 0.90$, DoD=0.90)	9.60 kW	18.72 kW	149.76 kW h	184.89 kW h	14
DGX Spark	Optimistic (nameplate only)	8.88 kW	8.88 kW	71.04 kW h	71.04 kW h	6
DGX Spark	Engineering (PUE=1.2, $\eta = 0.90$, DoD=0.90; host included)	8.88 kW	10.66 kW	85.25 kW h	105.24 kW h	8
DGX Spark	Legacy-room (PUE=1.56, $\eta = 0.90$, DoD=0.90; host included)	8.88 kW	13.85 kW	110.85 kW h	136.82 kW h	11
RTX A6000	Optimistic (no losses)	9.60 kW	9.60 kW	76.8 kW h	76.8 kW h	6
RTX A6000	Engineering ($f_{\mathrm{host}} = 1.25$, PUE=1.2, $\eta = 0.90$, DoD=0.90)	9.60 kW	14.40 kW	115.2 kW h	142.22 kW h	11
RTX A6000	Legacy-room ($f_{\mathrm{host}} = 1.25$, PUE=1.56, $\eta = 0.90$, DoD=0.90)	9.60 kW	18.72 kW	149.76 kW h	184.89 kW h	14
RTX A5000	Optimistic (no losses)	13.80 kW	13.80 kW	110.4 kW h	110.4 kW h	9

Table 6: Off-grid feasibility (8-hour run) under multiple facility assumptions. (continued)

Build	Scenario	IT Power	Facility Power	8h Energy	Battery Energy Req.	Powerwalls
RTX A5000	Engineering ($f_{\text{host}} = 1.25$, PUE=1.2, $\eta = 0.90$, DoD=0.90)	13.80 kW	20.70 kW	165.6 kW h	204.44 kW h	16
RTX A5000	Legacy-room ($f_{\text{host}} = 1.25$, PUE=1.56, $\eta = 0.90$, DoD=0.90)	13.80 kW	26.91 kW	215.28 kW h	265.78 kW h	20

7.4 Total CapEx: hardware + storage (using $8,200 per Powerwall as a knob)

Table 7: Total CapEx with Tesla Powerwall 3 storage (illustrative; assumes $8,200 per unit).

Build	Scenario	Powerwalls	Total CapEx
RTX PRO 6000 Blackwell	Optimistic	6	$199,200
RTX PRO 6000 Blackwell	Engineering	11	$240,200
RTX PRO 6000 Blackwell	Legacy-room	14	$264,800
DGX Spark	Optimistic	6	$199,200
DGX Spark	Engineering	8	$215,600
DGX Spark	Legacy-room	11	$240,200
RTX A6000	Optimistic	6	$199,200
RTX A6000	Engineering	11	$240,200
RTX A6000	Legacy-room	14	$264,800
RTX A5000	Optimistic	9	$223,800
RTX A5000	Engineering	16	$281,200
RTX A5000	Legacy-room	20	$314,000

Sustainability remark (practical, not rhetorical). If you are serious about "off-grid" for sustained compute, you should treat battery sizing, charging source sizing, and heat rejection as a coupled design: any thermal throttling that reduces compute also reduces energy efficiency, and any poor facility efficiency (high PUE) increases both battery and generation needs [22, 21].

8 Workload suitability and operational fit

8.1 Workload mapping: what dominates (compute, memory, bandwidth, or comms)?

Table 8: Workload-to-platform fit.

Workload	**Primary bottleneck**	**Best fit here**	**Why**	**Failure mode to watch**
Large language model (LLM) training (multi-GPU)	Interconnect + memory + Tensor throughput	RTX PRO 6000 Blackwell (dense GPU nodes)	High peak low-precision throughput; PCIe Gen 5; fewer GPUs eases orchestration [4, 20]	If you cannot realize 2:4 sparsity or FP4 stability, the effective advantage shrinks; network-limited scaling [1]
LLM inference (many requests)	Memory capacity + batch scheduling	DGX Spark farm *or* PRO 6000	Spark provides large per-box unified memory and easy replication; PRO 6000 gives high per-GPU memory with high tensor peak [7, 4]	Spark performance can be power- and thermally limited if sustained loads cannot hold near-peak [12]
Memory-bound scientific computing (out-of-core, streaming)	Bandwidth + memory footprint	RTX A6000 (pair NVLink if useful)	Large per-GPU VRAM improves shard sizes; stable FP32/TF32 tooling [5, 23]	Without NVLink and with small nodes, PCIe and host memory become bottlenecks [20]
Throughput FP32 simulations (embarrassingly parallel)	FP32 compute + many workers	RTX A5000 cluster	High aggregate FP32, and strong aggregate memory bandwidth [15]	Management overhead and network overhead dominate quickly; more failure points

9 Feasibility scoring

9.1 Scoring rubric

The earlier star ratings are retained, but the rationale is tightened: the "AI throughput" dimension is split between (i) peak low-precision AI throughput and (ii) realism of achieving it (sparsity/quantization readiness). This is why DGX Spark can score well on deployment ease while still carrying uncertainty about sustained peak behavior for long-duration runs [12].

Table 9: Feasibility scoring matrix (5 stars best).

Metric	**PRO 6000**	**DGX Spark**	**A6000**	**A5000**
Peak AI throughput per $	★★★★★	★★★★	★	★★
Sustained-performance confidence	★★★★	★★★	★★★★	★★★★
Power efficiency	★★★★★	★★★★	★★	★★
Deployment ease	★★	★★★★★	★★★	★★
Per-device memory capacity	★★★★★	★★★★	★★★	★★
Scale-out complexity (ops)	★★★	★★	★★	★

10 Decision tree

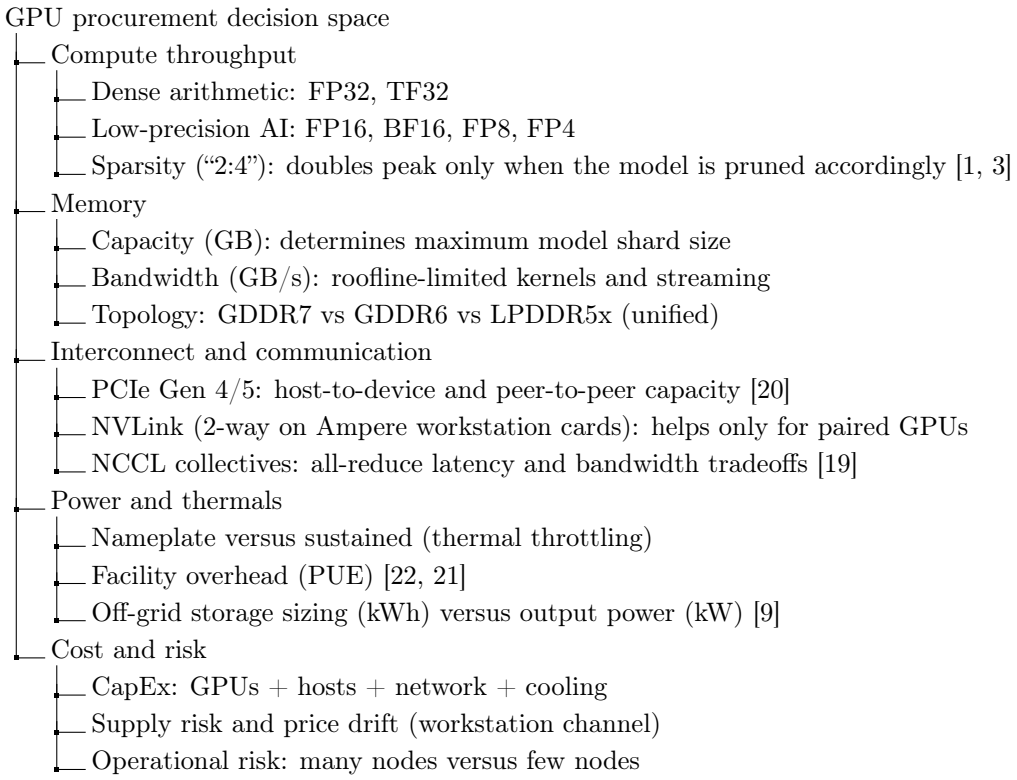

Figure 1: Decision mind-map: metrics, constraints, and terminology connections.

11 Portmanteaus, etymologies, and naming notes

- **GPGPU** is a portmanteau of “general-purpose” and “GPU,” reflecting the migration of parallel compute from graphics to scientific and AI workloads.
- **Tensor** derives from Latin *tendere* (to stretch), and in mathematics it generalizes scalars, vectors, and matrices into multi-index objects. In GPU marketing, “Tensor Cores” are specialized matrix-multiply units.
- **Sparsity** derives from Latin *sparsus* (scattered), and in modern deep learning it often means many weights are zero. NVIDIA’s 2:4 structured sparsity is a constrained pattern that hardware can exploit efficiently [1, 3].

12 Acronym glossary

Table 10: Acronym glossary.

Acronym	Meaning (and why it matters here)
AI	Artificial intelligence; here it mainly implies neural-network training and inference workloads.
BF16	Brain Floating Point 16; a 16-bit format common in training, balancing range and precision.
CUDA	Compute Unified Device Architecture; NVIDIA's programming and runtime stack for GPGPU.
ECC	Error-correcting code; reduces silent memory errors, relevant for long scientific runs and training stability.
FLOP	Floating-point operation; a rough unit for arithmetic throughput.
FP4/FP8/ FP16/FP32	Floating point with 4/8/16/32-bit representations; lower precision increases throughput but can change numerical behavior.
GDDR6/GDDR7	Graphics Double Data Rate (v6/v7); high-bandwidth memory on discrete GPUs.
GPGPU	General-Purpose computing on Graphics Processing Units; using GPUs for non-graphics compute.
kW/kWh	Kilowatt (power) / kilowatt-hour (energy); off-grid design requires both, not one.
LLM	Large language model; large transformer models that are often memory- and bandwidth-intensive.
MIG	Multi-Instance GPU; hardware partitioning that allows one GPU to present multiple isolated instances [4].
NIC	Network interface card; relevant for multi-node training and inference scaling.
NVLink	NVIDIA high-speed GPU interconnect; on these workstation Ampere cards it is typically 2-way bridging.
PUE	Power Usage Effectiveness; facility power divided by IT power [22, 21].
PSU	Power supply unit; nameplate power often derives from PSU ratings in appliance systems.
TF32	TensorFloat-32; Tensor Core mode that accelerates FP32-typed training without rewriting models [23].
TDP	Thermal design power; a thermal engineering target, not always identical to real sustained draw.
VRAM	Video random-access memory; on discrete GPUs this is local device memory, distinct from host RAM.

13 Conclusion

Under accelerator-only accounting, RTX PRO 6000 Blackwell dominates on typed low-precision AI throughput per dollar, and it now also matches RTX A6000 on aggregate discrete GPU memory pool under $150,000 because it is a 96 GB class part [4]. DGX Spark is operationally compelling for rapid prototyping and inference replication, especially when you value integrated CPU+GPU and large unified memory per box [7], but it is structurally disadvantaged for tight-coupled multi-GPU training because its compute and memory are naturally fragmented across many small nodes.

When you add off-grid engineering realism (host overhead, PUE, discharge-path efficiency, and depth-of-discharge), battery requirements can increase from single-digit Powerwalls to double digits for the 10–20 kW class builds (Table 6). That can invert "best value" decisions if energy infrastructure is genuinely part of the budget envelope, rather than an afterthought.

The practical recommendation is therefore conditional:

- If you are optimizing for *fewest nodes* and *highest typed AI throughput* with a plausible path to quantization and 2:4 sparsity, prioritize RTX PRO 6000 Blackwell.
- If you are optimizing for *fastest deployment* and *replicated local inference or prototyping*, prioritize DGX Spark.
- If you are optimizing for *per-device VRAM* and a conservative engineering profile, prioritize RTX A6000.
- If you are optimizing for *aggregate FP32 throughput and aggregate memory bandwidth* and you can operationalize many nodes, RTX A5000 remains a viable quantity-over-quality option.

References

[1] NVIDIA Developer Blog. *Structured Sparsity in the NVIDIA Ampere Architecture and Applications in Search Engines*. 2023. URL: https://developer.nvidia.com/blog/structured-sparsity-in-the-nvidia-ampere-architecture-and-applications-in-search-engines/ (visited on 12/26/2025).

[2] NVIDIA. *NVIDIA Ampere GA102 GPU Architecture Whitepaper*. 2020. URL: https://www.nvidia.com/content/PDF/nvidia-ampere-ga-102-gpu-architecture-whitepaper-v2.pdf (visited on 12/26/2025).

[3] Asit Mishra et al. "Design and Behavior of Sparse Tensor Cores". In: *arXiv* (2021). URL: https://arxiv.org/pdf/2104.08378 (visited on 12/26/2025).

[4] NVIDIA. *NVIDIA RTX PRO 6000 Blackwell Server Edition*. 2025. URL: https://www.nvidia.com/en-us/data-center/rtx-pro-6000-blackwell-server-edition/ (visited on 12/26/2025).

[5] NVIDIA. *NVIDIA RTX A6000 (Product Page)*. 2021. URL: https://www.nvidia.com/en-us/products/workstations/rtx-a6000/ (visited on 12/26/2025).

[6] NVIDIA. *NVIDIA RTX A5000 (Product Page)*. 2021. URL: https://www.nvidia.com/en-us/products/workstations/rtx-a5000/ (visited on 12/26/2025).

[7] NVIDIA. *NVIDIA DGX Spark (Product Page) – Specifications*. 2025. URL: https://www.nvidia.com/en-us/products/workstations/dgx-spark/ (visited on 12/26/2025).

[8] NVIDIA. *Hardware Overview — DGX Spark User Guide*. 2025. URL: https://docs.nvidia.com/dgx/dgx-spark/hardware.html (visited on 12/26/2025).

[9] Tesla. *Powerwall 3 Datasheet*. 2025. URL: https://energylibrary.tesla.com/docs/Public/EnergyStorage/Powerwall/3/Datasheet/en-us/Powerwall-3-Datasheet.pdf (visited on 12/26/2025).

[10] SolarReviews. *Tesla Powerwall price breakdown (Powerwall 3)*. 2025. URL: https://www.solarreviews.com/blog/is-the-tesla-powerwall-the-best-solar-battery-available (visited on 12/26/2025).

[11] NVIDIA. *NVIDIA DGX Spark (Marketplace Listing)*. 2025. URL: https://marketplace.nvidia.com/en-us/enterprise/personal-ai-supercomputers/dgx-spark/ (visited on 12/26/2025).

[12] NVIDIA Developer Forums. *DGX Spark Power Clarification (240W PSU; 140W SoC TDP)*. 2025. URL: https://forums.developer.nvidia.com/t/dgx-spark-power-clarification/349668 (visited on 12/26/2025).

[13] PNY. *PNY NVIDIA RTX A6000 – Specifications*. 2021. URL: https://www.pny.com/nvidia-rtx-a6000 (visited on 12/26/2025).

[14] NVIDIA Developer Forums. *Looking for full specs on NVIDIA A5000 (includes FP32 and Tensor throughput figures)*. 2022. URL: https://forums.developer.nvidia.com/t/looking-for-full-specs-on-nvidia-a5000/217948 (visited on 12/26/2025).

[15] NVIDIA. *NVIDIA Professional Graphics Solutions Linecard (Ampere)*. 2022. URL: https://www.nvidia.com/content/dam/en-zz/Solutions/gtcs22/design-visualization/quadro-product-literature/quadro-ampere-linecard-us-nvidia-web.pdf (visited on 12/26/2025).

[16] Micro Center. *PNY NVIDIA RTX PRO 6000 Blackwell Workstation Edition Dual Fan 96GB GDDR7 PCIe 5.0 Graphics Card (Price Listing)*. 2025. URL: https://www.microcenter.com/product/694549/pny-nvidia-rtx-pro-6000-blackwell-workstation-edition-dual-fan-96gb-gddr7-pcie-50-graphics-card (visited on 12/26/2025).

[17] Tom's Hardware. *Nvidia's RTX A6000: 48GB of Memory Powers Twice The Workstation Performance of RTX 3090*. 2021. URL: https://www.tomshardware.com/news/nvidia-rtx-a6000-48gb-benchmarked (visited on 12/26/2025).

[18] AEC Magazine. *Nvidia RTX A4000 review / RTX A5000 review (street-price context)*. 2021. URL: https://aecmag.com/workstations/nvidia-rtx-a4000-review-nvidia-rtx-a5000-review-arch-viz-ray-tracing/ (visited on 12/26/2025).

[19] Xinyue Jiang et al. *Demystifying NCCL: An In-depth Analysis of GPU Communication Collectives*. 2025. URL: https://arxiv.org/html/2507.04786v1 (visited on 12/26/2025).

[20] PCI-SIG. *PCIe 5.0 FAQ (raw bandwidth per lane)*. 2025. URL: https://pcisig.com/faq?field_category_value%5B%5D=pci_express_5.0&keys= (visited on 12/26/2025).

[21] Uptime Institute. *Uptime Institute Global Data Center Survey 2024 (PUE statistic)*. 2024. URL: https://datacenter.uptimeinstitute.com/rs/711-RIA-145/images/2024.GlobalDataCenterSurvey.Report.pdf (visited on 12/26/2025).

[22] National Renewable Energy Laboratory (NREL). *High-Performance Computing Data Center Power Usage Effectiveness (PUE)*. 2025. URL: https://www.nrel.gov/computational-science/measuring-efficiency-pue (visited on 12/26/2025).

[23] NVIDIA Developer Blog. *Accelerating AI Training with NVIDIA TF32 Tensor Cores*. 2021. URL: https://developer.nvidia.com/blog/accelerating-ai-training-with-tf32-tensor-cores/ (visited on 12/26/2025).