PARADYNE AI

SOVEREIGN AI END-TO-END SOLUTIONS · INFRASTRUCTURE SERIES · CAPACITY-PLANNING WHITE PAPER

Sizing Sovereign AI Capacity

Open-weight models, memory, throughput, and concurrent users across three deployment classes: a capacity-planning white paper for organizations that own their AI infrastructure

Reviewed by Alam Jamal, Chief Architect, Paradyne AI

Planning edition · November 2026 · Las Vegas, Nevada

Contents

1	The Capacity Question in a Sovereign Deployment	6	Demand: Tokens per User, per Task, and per Day
2	The Three Deployment Classes	7	Matching Demand to a System
3	The Four Quantities That Determine Capacity	8	The Economics of Owned Capacity
4	The Open-Weight Model Landscape and Memory Fit	9	Paradyne AI's Sizing Method
5	Throughput by Model and System	10	Limitations and Scope

Executive Summary

An organization that owns its AI infrastructure must answer a question that a hosted provider answers on its behalf and never shows it: how much work can the system do, how fast will it feel to each person using it, and how many people and automated processes can it serve at once. This paper sets out the method Paradyne AI uses to answer that question for the three deployment classes it delivers — the Sovereign Workstation (one to four NVIDIA RTX PRO 6000 Blackwell GPUs), the Deskside Station (NVIDIA DGX Station GB300), and the Sovereign Appliance (a 5U system with two to eight RTX PRO 6000 GPUs) — against the mainstream open-weight language models available in November 2026.

Four quantities determine capacity. Memory decides which models fit, and is the first and hardest constraint. Memory bandwidth decides how fast a single response is generated. Batching across concurrent requests decides aggregate throughput, which rises many-fold above the single-stream rate before each user's speed begins to fall. Compute decides how fast input is read, which matters most for retrieval-augmented and agentic workloads whose input tokens outnumber their output tokens by one hundred to one or more. The paper gives the arithmetic for each, the published measurements that anchor it, and planning ranges for representative models on each system.

The central finding is that user count is the wrong sizing input. Twenty people using a conversational assistant consume a few million tokens a day, a fraction of what a single-GPU workstation produces; what moves an organization from one class to the next is the model class the mission requires, the number of models that must run at once — a Paradyne platform such as Sentinel is a set of models, not one — and the share of demand that comes from agents and document processing rather than conversation. Paradyne AI's practice is to size on that demand profile, select the class with headroom, characterize the delivered unit in its own laboratory, and replace every planning figure in this paper with a measured one before any service level is committed.

KEY FINDINGS

1. **Memory fit is decided at 4-bit precision.** At 4-bit precision, every mainstream open-weight model below 230 GB — including the 120B class and the 284B- to 397B-parameter mixture-of-experts models — fits a four-GPU workstation or a four-GPU appliance, and all but the largest of them fit the DGX Station's HBM3e. The 670B class runs on the eight-GPU appliance at multi-user scale, and on the DGX Station from its 748 GB coherent memory at single-user to small-team speed. The trillion-parameter class at full precision requires rack-scale systems outside the three classes.
2. **Single-stream speed follows bandwidth.** A single RTX PRO 6000 generates 160–210 tokens per second for one user on a 120B-class model and 30–35 on a dense 70B model; the DGX Station GB300 measures 390 and an estimated 55–110 respectively. The difference is memory bandwidth: 1.8 TB/s against 7.1 TB/s.
3. **Concurrency multiplies, up to a ceiling.** Published measurements show aggregate throughput of 8 to 25 times the single-stream rate before per-user speed falls below 50 tokens per second: roughly 1,500–1,800 tokens per second on one workstation GPU, 4,900 at 32 users and 9,300 at 128 users on the DGX Station, and 7,000–10,000 on an eight-GPU appliance for the 120B class.
4. **Demand is tokens per task, not users.** A conversational turn costs 1,500–3,000 tokens; a document question with retrieval 5,000–20,000; an agentic task 50,000 to several million, with input tokens dominating. Daily demand is the sum of those products, not the head count, and a demand profile dominated by agents or ingestion is prefill-bound, favoring the systems with the most compute per model.
5. **The ceiling is computed; the fraction is measured.** Realized throughput is 20–80 percent of the bandwidth ceiling depending on the inference engine, the model architecture, and the maturity of kernels for the GPU generation. The ceiling is arithmetic; the realized fraction is measured — which is why Paradyne characterizes the delivered unit rather than quoting a specification.
6. **A Paradyne capability is a set of models, not one.** Sentinel, Herytek, and SICE each run several models at once. The resident set, not the largest model, decides the memory class, and the appliance's independent GPU lanes suit it directly.

1. The Capacity Question in a Sovereign Deployment

1.1 Definitions

A token is the unit in which a language model reads and writes text; in English, roughly three-quarters of a word. Inference is the use of a trained model to produce output, and it has two phases. Prefill reads the input — the question, the retrieved documents, the conversation so far — and is bound by compute. Decode generates the response one token at a time, and is bound by memory bandwidth, because every generated token requires the active weights of the model to be read from memory once. The time to first token is the prefill time the user waits before the response begins; tokens per second per stream is the speed at which the response then appears. Aggregate throughput is the total tokens per second a system produces across all concurrent streams.

An open-weight model is one whose parameters are published for download and local operation; the models discussed here are released under the MIT, Apache 2.0, or similar licenses. A mixture-of-experts (MoE) model holds many specialized parameter groups but activates only a few for each token; its total parameters determine memory footprint, and its active parameters determine per-token speed. Precision is the number of bits used to store each parameter: 8-bit (FP8) and 4-bit (NVFP4, MXFP4) formats reduce footprint and increase speed with small, measurable effects on quality, and the Blackwell GPU generation executes 4-bit arithmetic natively. The key-value (KV) cache is the memory a model uses to hold the context of each active conversation; it grows with context length and with the number of concurrent streams, and it competes with the weights for the same memory.

1.2 Why the question belongs to the owner

A hosted provider performs this capacity planning internally and presents the result as a rate limit and a bill. Sovereign deployment transfers the planning to the owner, who buys capacity once and then allocates it. The question an executive asks — how many tokens per second, for how many users, at what cost per question — is therefore not an enterprise question wrongly applied to a sovereign system; it is the right question, with one term that does not translate. There is no cost per token on owned hardware. The marginal cost of a token is electricity, and the planning unit is the system's capacity per working day and what the organization chooses to spend it on.

The answer has three parts: what the system produces in aggregate, what each person experiences, and what the organization's work demands. Sections 2 through 5 address the first two; Section 6 the third; Section 7 brings them together.

2. The Three Deployment Classes

Table 1 sets out the three classes against the properties that govern capacity. Figures are manufacturer specifications; memory bandwidth and GPU memory are the two that decide most sizing outcomes.

Table 1. Paradyne AI deployment classes and the properties that govern capacity

PROPERTY	SOVEREIGN WORKSTATION	DESKSIDE STATION	SOVEREIGN APPLIANCE 5U
Configuration	Tower chassis; one to four RTX PRO 6000 Blackwell (Workstation or Max-Q edition) [1]	NVIDIA DGX Station GB300: one Blackwell Ultra GPU and one 72-core Grace CPU [3]	Dual AMD EPYC 9005; two to eight RTX PRO 6000 Blackwell Server Edition (eight is the chassis maximum); NVMe array [2]
GPU memory	96 GB GDDR7 per GPU; 96–384 GB in separate pools	252 GB HBM3e, plus 496 GB LPDDR5X in one coherent 748 GB space that the GPU addresses directly [3, 26, 27]	96 GB GDDR7 per GPU; 192–768 GB in separate pools
Memory bandwidth	1.79 TB/s per GPU [1]	7.1 TB/s (HBM3e); 396 GB/s (CPU memory); 900 GB/s CPU–GPU link [3]	1.6 TB/s per GPU; 12.8 TB/s in aggregate at eight GPUs [2]

PROPERTY	SOVEREIGN WORKSTATION	DESKSIDE STATION	SOVEREIGN APPLIANCE 5U
Peak 4-bit compute	4.0 PFLOPS per GPU (Max-Q: 3.5) [1]	20 PFLOPS [5]	3.7 PFLOPS per GPU; ~30 PFLOPS at eight [2]
GPU interconnect	PCIe Gen 5	Single GPU; no multi-GPU synchronization	PCIe Gen 5; no NVLink
Power and facility	0.3–0.6 kW per GPU plus host. One Workstation Edition GPU or two Max-Q GPUs on a standard 120 V, 20 A circuit; anything beyond that needs a dedicated circuit, and three or four GPUs a 208–240 V circuit	1.6 kW system; office environment, dedicated circuit [5]	2–6 kW by GPU count; 208 V; 42–48 in. cabinet; air-cooled
Serves	One to a few users; several isolated model lanes	A team on one frontier-class model; analyst-facing	A team to a department; several models and data engineering together

Peak compute figures are sparse FP4 tensor-core ratings and are theoretical ceilings. The DGX Station figures are the shipping configuration common to NVIDIA's system partners [26, 27]; early units and some published material cite 288 GB and 8 TB/s [4]. Power: a single Workstation Edition GPU or two Max-Q GPUs operate on a standard office circuit; every larger workstation configuration, and every appliance, is specified with the client's facility team before order. A fourth, rack-scale class (HGX B300 or NVLink-connected systems) lies beyond the scope of this paper and is noted where the model landscape requires it.

2.1 Memory as the first constraint, and the Station's coherent pool

On the workstation and the appliance, a model that does not fit in GPU memory does not run usefully: memory is held in separate 96 GB pools, a model larger than one pool is split across GPUs by tensor parallelism, which works but synchronizes the GPUs over PCIe at every token and reduces the scaling that additional GPUs provide [14], and weights beyond the GPUs altogether fall to host memory over PCIe at a fraction of the speed.

The DGX Station is different in kind. Its GPU addresses the 496 GB of CPU memory directly over the 900 GB/s chip-to-chip link and executes against it, so the HBM3e serves as the high-speed working set and the LPDDR5X as a large extension for weights, KV cache, or mixture-of-experts parameters — frequently used experts resident in HBM, rarely used experts in system memory — without falling back to CPU execution [26]. The CPU tier is read at 396 GB/s rather than 7.1 TB/s, so the fraction of each token read from it sets the speed, and the practical planning boundary is this: multi-user service lives in the 252 GB of HBM3e, and the full 748 GB pool extends the Station to the 500B- to 1T-parameter class at single-user to small-team speed. Published measurements on comparable units show a 504B model at about 60 tokens per second, a 550B model at about 35 for one user scaling to four or five concurrent, and a 1.1T model at 40–50 tokens per second in total [27]. No other deskside system of the three runs that class at all.

2.2 Bandwidth and interconnect

For one user, the DGX Station's HBM3e gives it roughly four times the single-stream speed of a single RTX PRO 6000 on the same model. For many users, the appliance's eight separate GPUs can run eight independent model instances — eight lanes — whose aggregate throughput exceeds the Station's on models that fit in one GPU, or can be paired into two four-GPU replicas for a model that needs 300 GB. The two classes are therefore not ordered: the Station is the fastest single-model system of the three and the only one that reaches the 500B–1T class from a single chip; the appliance holds the most GPU memory and the most independent lanes.

3. The Four Quantities That Determine Capacity

The arithmetic below is deliberately simple. It produces ceilings and planning ranges; Section 5 anchors it to published measurements, and Section 9 explains how Paradyne replaces it with measured values on the delivered unit.

3.1 Memory fit

Weight footprint in gigabytes is approximately total parameters in billions multiplied by bytes per parameter: about 1.05 at FP8 and about 0.56 at 4-bit, including the scaling factors the formats carry. A 70B dense model therefore occupies about

74 GB at FP8 and 39 GB at 4-bit; a 284B mixture-of-experts model about 298 GB and 159 GB. The KV cache and runtime overhead must fit beside the weights: Paradyne's planning rule is that weights should occupy no more than 70 percent of the memory pool for multi-user service, and no more than 90 percent for a single user with modest context. Table 3 applies that rule.

3.2 Single-stream decode speed

Tokens per second for one stream is bounded by memory bandwidth divided by the bytes read per token — the active parameters at the stored precision, plus the portion of the KV cache read at each step. A dense 70B model at FP8 on a 1.79 TB/s GPU has a ceiling near 25 tokens per second; at 4-bit, near 45. Published single-user measurements of 32–34 tokens per second at 4-bit on the RTX PRO 6000 [10, 11] correspond to about 75 percent of that ceiling. Mixture-of-experts models read far fewer bytes per token — gpt-oss-120b reads roughly 5 GB — and so decode far faster than their total size suggests: 160–210 tokens per second on the same GPU [9, 11, 12]. The realized fraction of the ceiling, however, falls as models become more complex and engines less mature: a 320B/18B-active model on the DGX Station measures 140–145 tokens per second against a ceiling several times higher [6].

3.3 Aggregate throughput and concurrency

Because decode is bandwidth-bound, the weights read for one token can serve many streams at once: batching 32 requests reads the weights once per step and produces 32 tokens. Aggregate throughput therefore rises nearly linearly with concurrency until the system becomes compute-bound or the KV cache is exhausted. The practical ceilings in published measurements range from eight to twenty-five times the single-stream rate. The quantity a planner needs is the number of concurrent streams at which per-user speed falls to the target — 50 tokens per second is comfortable; 20 is acceptable; 10 is reading speed — and Figure 2 in Section 5 plots that curve for each class.

3.4 Prefill and the compute-bound workload

Reading input is compute-bound, and its rate scales with the GPU's tensor-core throughput rather than its bandwidth. Measured prefill rates range from roughly 4,000–6,500 tokens per second on one RTX PRO 6000 for a 120B-class model at short context [11], to 32,000–44,000 tokens per second on the DGX Station for mixture-of-experts models of the 120B to 320B class [6]. Prefill decides the time to first token, and it decides the capacity of any workload in which input dominates: retrieval over large document sets, long-document review, batch ingestion, and agentic tasks, where published analysis finds input tokens outnumbering output by more than 150 to one [23].

The planning arithmetic in one place

Footprint (GB) ≈ parameters (B) × 1.05 at FP8, or × 0.56 at 4-bit. Fit: weights ≤ 70 percent of the memory pool for multi-user service.

Single-stream ceiling (tokens/s) ≈ memory bandwidth (GB/s) ÷ bytes read per token (GB); realized fraction 0.2–0.8, measured per engine and model.

Aggregate throughput (tokens/s) ≈ single-stream × concurrency, up to a ceiling of 8–25 × single-stream; users served ≈ aggregate ceiling ÷ per-user target.

Prefill rate (tokens/s) ≈ effective compute ÷ (2 × active parameters); time to first token ≈ input tokens ÷ prefill rate.

Daily demand (tokens) = Σ users × interactions × tokens per interaction + Σ agent tasks × tokens per task + pages ingested × ~600. Daily capacity ≈ aggregate throughput × seconds in service × duty cycle.

4. The Open-Weight Model Landscape and Memory Fit

Table 2 lists the mainstream open-weight models an organization would consider in November 2026, with their total and active parameters, license, and estimated weight footprints at the two precisions used in sovereign deployment. Footprints are computed with the factors in Section 3.1; actual checkpoints vary by a few percent, and DeepSeek-V4.1-Flash carries roughly 190 GB of additional memory tables beyond its parameters [21, 22].

Table 2. Mainstream open-weight models, November 2026

MODEL	DEVELOPER	TOTAL / ACTIVE (B)	LICENSE	FP8 (GB)	4-BIT (GB)	ARCHITECTURE
gpt-oss-20b	OpenAI	21 / 3.6	Apache 2.0	22	12	MoE, native MXFP4
Qwen3.8-27B	Alibaba	27 / 27	Apache 2.0	28	15	Dense
Llama 3.3 70B	Meta	70 / 70	Llama 3.3	74	39	Dense
gpt-oss-120b	OpenAI	117 / 5.1	Apache 2.0	123	66	MoE, native MXFP4
Qwen3.5-122B-A10B	Alibaba	122 / 10	Apache 2.0	128	68	MoE
DeepSeek-V4-Flash	DeepSeek	284 / 13	MIT	298	159	MoE, 1M context
GLM-5.3-Flash	Z.ai	320 / 18	MIT	336	179	MoE, 1M context
Qwen3.5-397B-A17B	Alibaba	397 / 17	Apache 2.0	417	222	MoE
DeepSeek-V3.2 / R1	DeepSeek	671 / 37	MIT	705	376	MoE, prior generation
DeepSeek-V4.1-Flash	DeepSeek	552 / 16	MIT	770	404	MoE + memory tables; ~190 GB extra
GLM-5.3	Z.ai	753 / 40	GLM-5.3	791	422	MoE
Kimi K2.6 / K2.7	Moonshot AI	1,000 / 32	Modified MIT	1,050	560	MoE
DeepSeek-V4-Pro	DeepSeek	1,600 / 49	MIT	1,680	896	MoE, 1M context
Kimi K3	Moonshot AI	2,800 / 104	Kimi K3	2,940	1,568	MoE

Sources: developer model cards as summarized in [21, 22]; gpt-oss models are distributed in MXFP4 and their FP8 figures are nominal. Parameter counts for GLM-5.3 (active) and Kimi K3 are as reported by third parties.

4.1 Dense and mixture-of-experts models

The landscape has moved decisively to mixture-of-experts designs: of the fourteen models in Table 2, only two are dense. The consequence for sizing is that footprint and speed have separated. A 397B-parameter model with 17B active parameters needs 222 GB of memory at 4-bit but reads only about 10 GB per generated token, so it decodes faster than a dense 70B model on the same hardware while requiring more than five times the memory. Memory capacity, not bandwidth, is therefore the quantity that determines which class a modern frontier-class open model requires.

4.2 Fit across the three classes

Figure 1 places each model's 4-bit footprint against the memory pools of the three classes, and Table 3 applies the 70-percent planning rule.

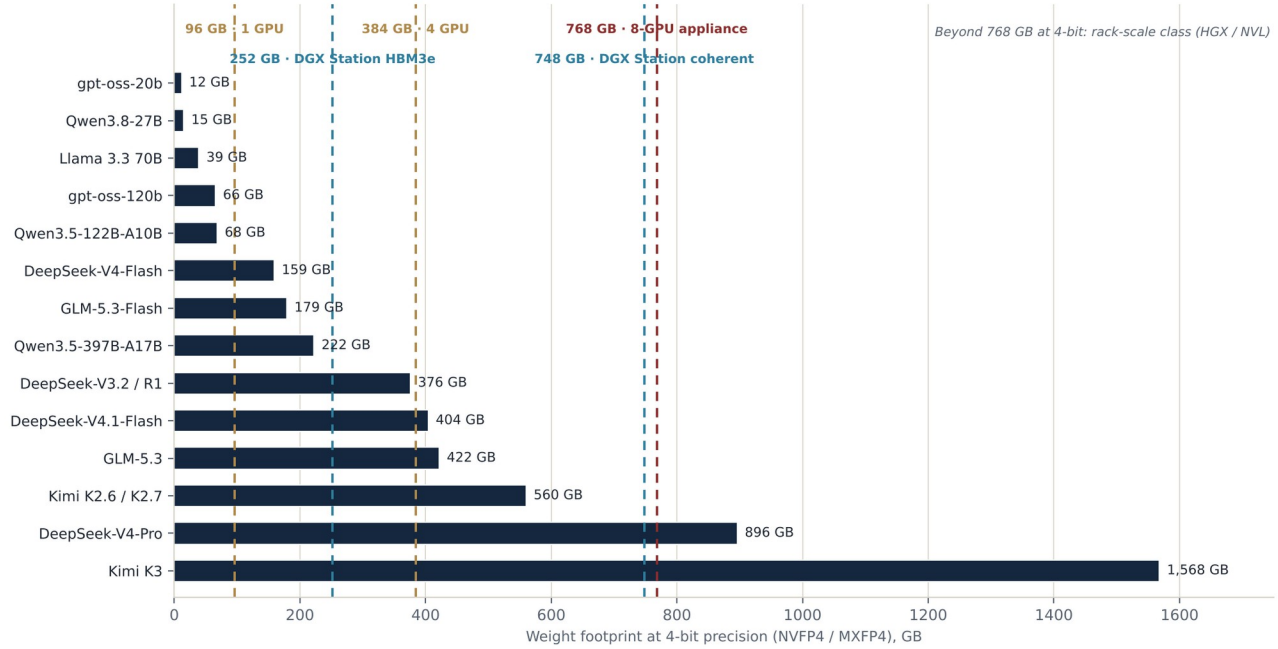

Figure 1. Weight footprint at 4-bit precision against the GPU memory pools of the three deployment classes. A model to the left of a line fits that pool; the 70-percent rule for multi-user service is applied in Table 3.

Table 3. Memory fit at 4-bit precision (Yes: ≤ 70 percent of the pool; Tight: 70–90 percent, single user or limited context; Coherent: beyond HBM3e but within the Station's 748 GB pool, single-user to small-team speed; —: does not fit)

MODEL	4-BIT (GB)	WORKSTATION 1 GPU · 96 GB	WORKSTATION 4 GPU · 384 GB	DGX STATION · 252 GB HBM	APPLIANCE 4 GPU · 384 GB	APPLIANCE 8 GPU · 768 GB
gpt-oss-20b	12	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
Qwen3.8-27B	15	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
Llama 3.3 70B	39	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
gpt-oss-120b	66	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
Qwen3.5-122B-A10B	68	**Tight**	**Yes**	**Yes**	**Yes**	**Yes**
DeepSeek-V4-Flash	159	—	**Yes**	**Yes**	**Yes**	**Yes**
GLM-5.3-Flash	179	—	**Yes**	**Tight**	**Yes**	**Yes**
Qwen3.5-397B-A17B	222	—	**Yes**	**Tight**	**Yes**	**Yes**
DeepSeek-V3.2 / R1	376	—	—	**Coherent**	—	**Yes**
DeepSeek-V4.1-Flash	404	—	—	**Coherent**	—	**Yes**
GLM-5.3	422	—	—	**Coherent**	—	**Yes**
Kimi K2.6 / K2.7	560	—	—	**Coherent**	—	**Tight**
DeepSeek-V4-Pro	896	—	—	—	—	—
Kimi K3	1,568	—	—	—	—	—

Coherent fits are served partly from the Station's CPU-memory tier (Section 2.1); trillion-parameter models have been run on the Station at sub-4-bit quantization [27] and are outside the 4-bit rule applied here. DeepSeek-V4.1-Flash has been operated on eight RTX PRO 6000 GPUs at mixed 8-bit precision

using about 91 GB per GPU [16]. Multi-GPU fits assume tensor parallelism across the pool. The appliance is configured by GPU count per engagement and grows in place; eight is the chassis maximum.

4.3 From a model to a capability

A Paradyne platform is not one model. Sentinel, Herytek, SICE, and the company's other capabilities each run several models at once — a primary model for analysis and drafting alongside the retrieval, classification, and document-reading models the work requires — under supervision and behind the Marquis layer. For sizing, the unit is therefore the set of models resident at the same time, plus the KV cache for every concurrent stream, rather than the platform's largest model. Table 4 gives the shape of the load for each platform; the composition of the set is engagement-specific and is settled in Discover.

Table 4. Paradyne platforms: shape of the load and typical class

PLATFORM	FUNCTION	SHAPE OF THE LOAD	TYPICAL CLASS
Sentinel Core	Sovereign intelligence platform: entity graph, dossier assembly, network analysis, AI-driven ETL	Analysts interactive; continuous ingestion; supervised agents	Workstation 4 GPU; DGX Station; Appliance
Sentinel Pulse	Information-environment analytics and substantiated messaging support	High-volume ingestion; batch analytics; interactive review	DGX Station; Appliance
Sentinel Screening	Screening of persons and entities against watchlists and internal records	Event-driven and batch; low latency per check	Workstation 1–2 GPU; one appliance lane
Herytek	Anti-money-laundering and know-your-customer intelligence for gaming and hospitality	Nightly batch scoring; case-analyst interactive	Workstation 2 GPU; Appliance
SICE	Subpoena Intelligence & Compliance Engine: legal-process response for any organization that receives process	Window-defined bursts, input-dominant; concurrent matters	DGX Station; Appliance 4–8 GPU
Smart Display	Sovereign operations dashboard: telemetry, SCADA, and sensor fusion with bottom-line-up-front summaries	Continuous low-latency streams; few users	Workstation; one appliance lane

Omnichannel (camera and telemetry fusion) and Ancile Core (agentic cyber defense) follow the same pattern with heavier vision and telemetry components.

Two consequences follow for anyone who will use these platforms rather than a model. First, because the sizing unit is the resident set, a Sentinel deployment for a team lands on the four-GPU workstation, the DGX Station, or the appliance even though its largest single model would fit one GPU. Second, the appliance's separate GPU pools suit a set — one model per GPU, with no tensor-parallel synchronization — while the DGX Station runs the whole set inside one HBM3e pool at the highest per-model speed. The throughput figures in Section 5 apply per model; a platform's throughput is that of its busiest model at the concurrency it is allowed.

5. Throughput by Model and System

Table 5 gives planning ranges for representative models on each class, expressed as single-stream tokens per second and aggregate tokens per second at high concurrency. Entries marked M are published third-party measurements on the stated hardware; entries marked E are Paradyne estimates from the arithmetic of Section 3, anchored to the nearest measurement. Measurements vary with engine version, prompt and output length, quantization, and harness, and the sources are listed so that each can be examined.

Table 5. Planning throughput by model and class — single-stream · aggregate, tokens per second

MODEL (PRECISION)	WORKSTATION, 1 GPU	WORKSTATION OR APPLIANCE, 4 GPU	DGX STATION GB300	APPLIANCE, 8 GPU
gpt-oss-120b (MXFP4)	160–210 · 1,500–1,800 M [8, 9, 11, 12]	176 · 3,900–4,000 M [9]	390 · 4,900 at 32 users; 9,300 at 128 M [6, 7]	176 · 7,000–10,000 as two 4-GPU replicas E [9, 14]

MODEL (PRECISION)	WORKSTATION, 1 GPU	WORKSTATION OR APPLIANCE, 4 GPU	DGX STATION GB300	APPLIANCE, 8 GPU
Llama 3.3 70B (FP8 · 4-bit)	20–25 (FP8) and 32–34 (4-bit) · 1,000–1,700 M [9, 10, 11]	33 per user at batch 1 (TP4); 3,000–4,000 E [9]	55–60 FP8, ~100 4-bit · 5,000–7,000 E [20]	32–34 · 8,000–13,000 as eight lanes E [9]
DeepSeek-V4-Flash, 284B/13B (4-bit · FP8)	Does not fit	25–35 · 1,000–2,000 E [17]	148 single M [7]; 2,500–3,500 aggregate E, anchored on GLM-5.3-Flash [6]	40–50 FP8 (200 with speculative decoding) · 650–950 measured; ~2,450 cited M [15, 16, 17]
GLM-5.3-Flash, 320B/18B (4-bit)	Does not fit	20–30 · 800–1,500 E	140–145 · 1,380 at 32; 2,517 at 256 users; prefill ~32,000 M [6]	40–50 · 650–950 E, as DeepSeek-V4-Flash
Qwen3.5-397B-A17B (4-bit)	Does not fit	20–30 · 700–1,200 E	100–150 · 1,000–2,000; limited context (88 percent of HBM) E	25–40 · 800–1,500 E
DeepSeek-V3.2 / R1, 671B/37B (4-bit)	Does not fit	Does not fit	35–60 single user from the coherent pool; a few concurrent streams E, anchored on 504B–550B measurements [27]	20–40 · 500–1,500 E, PCIe-bound [14, 18]
DeepSeek-V4.1-Flash, 552B + tables (mixed 8-bit)	Does not fit	Does not fit	35–60 single user from the coherent pool E [27]	94–134 (362 with speculative decoding) · 790 at 8 streams, ~1,540 at higher; falls to ~105 at 90,000-token inputs M [16]

M: published measurement on the stated hardware; E: Paradyne estimate. Aggregate figures are at 32–128 concurrent streams unless stated. Data-center GB300 measurements [18, 19, 20] are derated for the DGX Station's memory configuration and power envelope where used as anchors. Eight-GPU appliance figures are drawn from published measurements on comparable eight-GPU PCIe systems and from Paradyne estimates.

5.1 What the figures show

Three patterns hold across the table. First, for the 120B-class mixture-of-experts models that most organizations will deploy first, a single workstation GPU already serves a small team at comfortable speed, the DGX Station serves a department, and the eight-GPU appliance serves the same department with capacity to spare for additional models. Figure 2 plots the per-user speed each class sustains as concurrent streams increase.

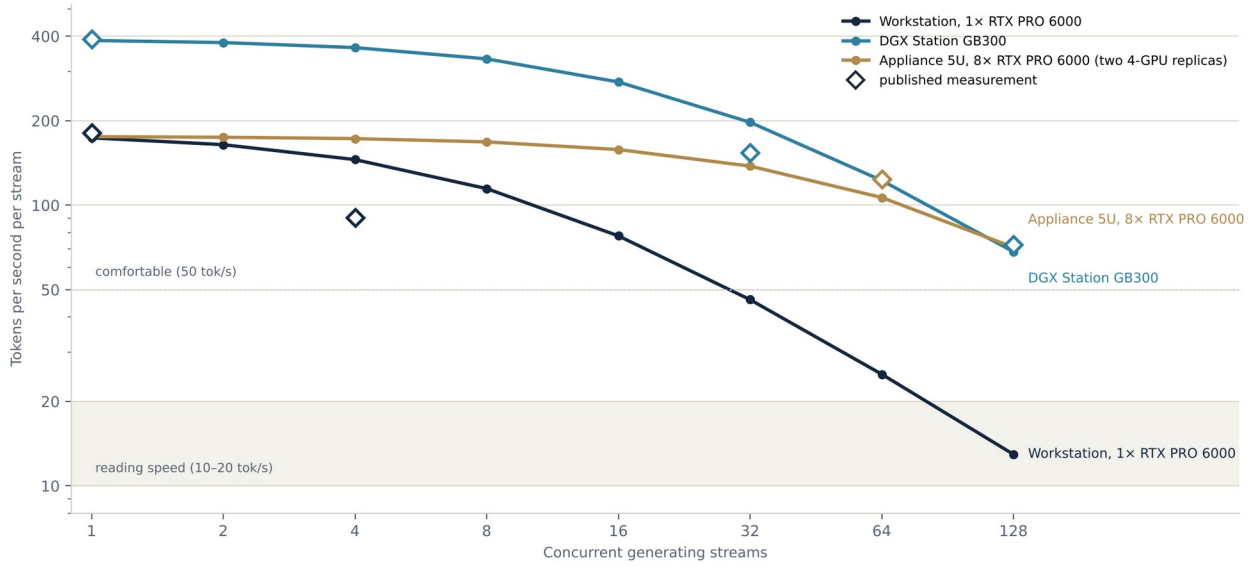

Figure 2. Per-stream generation speed against concurrent streams for gpt-oss-120b, modeled from the arithmetic in Section 3 and anchored to published measurements (diamonds) [6, 7, 9, 12]. The shaded band is reading speed. The appliance curve assumes two four-GPU replicas; eight single-GPU lanes would sustain the workstation's per-stream speed at eight times its concurrency.

Second, dense models are slow per stream everywhere except on HBM. A dense 70B model is the case where the DGX Station's bandwidth is decisive for a single analyst; for many users, batching recovers the aggregate on any class. Third, the 300B-class mixture-of-experts models — the current frontier-class open models — are where the classes diverge most. On the DGX Station they run in HBM at 140–200 tokens per second for one user with a prefill rate above 30,000 tokens per second; on the eight-GPU appliance they run at 40–50 tokens per second per user from PCIe-connected GDDR7, rising to about 200 with speculative decoding, and their aggregate is bounded by PCIe synchronization [14, 15]. The appliance's advantage at this model class is GPU memory and lanes: it serves the 670B class to many users, and it can hold a 300B model and a 120B model at once. The Station's answer is its coherent pool, which runs the 500B–1T class for an analyst or a small team from one chip.

5.2 The realized fraction

The gap between the bandwidth ceiling and measured throughput is not constant. For gpt-oss-120b on one RTX PRO 6000, measurements reach 45–60 percent of the ceiling, and a dense 70B model about 75 percent; on the DGX Station's far higher bandwidth, single-stream decode of a small-active-parameter model is limited by kernel and compute overhead rather than bandwidth, and the fraction falls to 25–30 percent; for GLM-5.3-Flash and DeepSeek-V4-Flash on the newest GPU generation, 20–40 percent, with published field reports documenting order-of-magnitude swings between engine versions on identical hardware [15, 16]. Two consequences follow. A planning table dated November 2026 understates what the same hardware will do in mid-2027, because engines improve faster than hardware changes. And a figure quoted from a specification sheet is not a capacity; only a measurement on the delivered configuration, with the engine and model that will run in production, is.

6. Demand: Tokens per User, per Task, and per Day

Capacity is meaningful only against demand, and demand is measured in tokens per unit of work, not in head count. Table 6 gives the planning units Paradyne uses, with the published figures that support them.

Table 6. Demand units for capacity planning

WORK PATTERN	TOKENS PER UNIT	TYPICAL DAILY VOLUME PER USER	SHAPE OF THE LOAD
Conversational assistant turn	1,500–3,000 (input and output) [25]	30–60 turns: 50,000–150,000	Balanced; decode speed is what the user sees
Document question with retrieval	5,000–20,000; mostly retrieved input	5–20 questions: 50,000–300,000	Input-heavy; prefill sets the wait
Long-document review	20,000–100,000 per document	1–5 documents: 100,000–500,000	Input-heavy; KV cache per stream is large
Agentic task (multi-step, tool-using)	50,000–500,000 typical; coding agents average ~4.2 million, with 30× variance between runs [23]; 5–30× a chat exchange [24]	2–10 tasks: 0.5–10 million or more	Input-dominant (about 150:1); bursty; long-running
Batch ingestion and indexing	~600 per page	1,000–50,000 pages per night: 0.6–30 million (organization-wide)	Prefill only; schedulable off-peak
Model adaptation (fine-tuning)	GPU-hours, not tokens	Reserved windows	Compute-bound; exclusive use of the GPUs

6.1 Three demand profiles

Profile A, a twenty-person professional office. Each person holds forty assistant turns and ten retrieval questions a day: 20 × 40 × 2,500 + 20 × 10 × 12,000 ≈ 4.4 million tokens, of which about 80 percent is input. Peak concurrency is three to five

streams. A single-GPU workstation running a 120B-class model sustains more than 100 tokens per second per stream at that concurrency and produces about 24 million tokens in a working day at 50 percent duty; the office uses under a fifth of it. If this office selects a DGX Station, it does so for the model class — a 300B-class model in HBM — or for growth, not for the load.

Profile B, a 150-person department with automation. Conversation and retrieval as above for 150 people ≈ 29 million tokens; 200 agentic tasks a day at 250,000 tokens ≈ 50 million; a nightly ingestion of 20,000 pages ≈ 12 million; total ≈ 91 million tokens a day, about 85 percent input. Peak concurrency is 15–25 interactive streams plus 10–20 agent streams holding long contexts. The profile is prefill-dominant and benefits from two model lanes — a 120B-class model for interaction and a 300B-class model for agents — which places it on the eight-GPU appliance, or on a DGX Station paired with a workstation.

Profile C, ten engineers running coding agents. Four tasks each at two million tokens ≈ 80 million tokens a day, almost entirely input, in bursts, with ten agents each holding contexts of 100,000 tokens or more. The binding constraints are KV cache per stream and prefill rate, which favor the DGX Station's HBM and compute; an eight-GPU appliance serves the same profile with lower per-stream speed and more total memory.

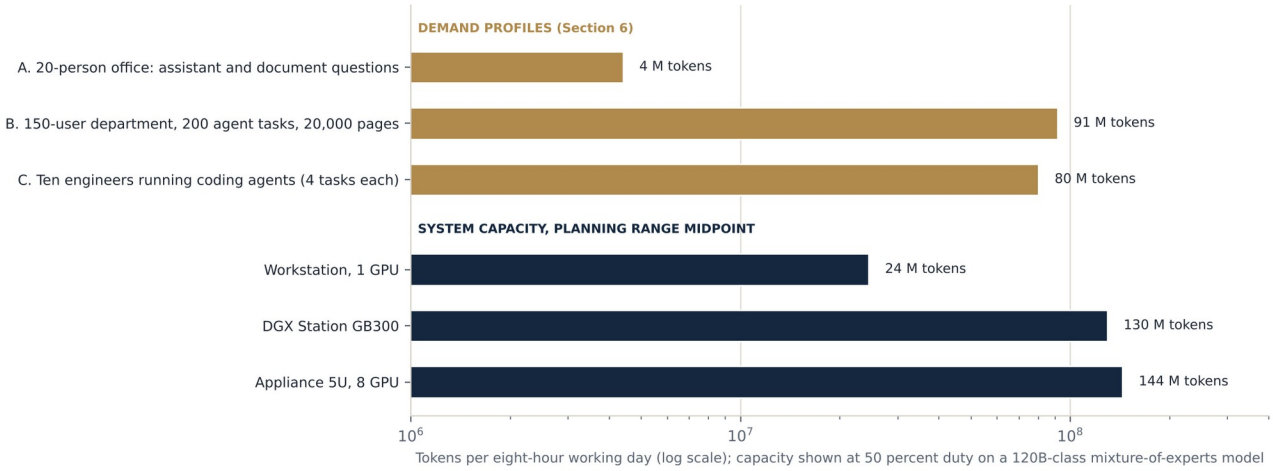

Figure 3. The three demand profiles of Section 6.1 against the planning capacity of each class in tokens per eight-hour working day, at 50 percent duty on a 120B-class model. Capacity is the midpoint of the ranges in Table 5; the logarithmic scale spans two orders of magnitude.

6.2 Burst demand and the time window

Some workloads are defined by a window rather than a day. A document-processing engine that must read 25,000 pages of records within fifteen minutes of their arrival — about 15 million tokens — requires that volume of prefill in 900 seconds: about 16,700 tokens per second. The DGX Station's measured prefill of 32,000–44,000 tokens per second on mixture-of-experts models of the 120B to 320B class meets that window with margin; a single workstation GPU at 4,000–6,500 tokens per second would need forty to sixty minutes. Window-defined demand is sized on prefill rate, and it is the case in which the DGX Station or the appliance is required regardless of user count.

6.3 How agents change the shape of demand

Agentic workloads differ from conversation in three ways that matter for sizing. Their tokens are overwhelmingly input, so they are compute-bound and consume prefill capacity rather than decode capacity. They are long-running, so each holds KV-cache memory for minutes rather than seconds, which reduces the concurrency the same memory can support. And they are variable: published analysis of coding agents finds runs on the same task differing by up to thirty times in total tokens [23], and enterprise reporting through 2026 documents agentic workloads consuming five to thirty times the tokens of a chat exchange [24]. An organization planning to automate should therefore size on the agent share of demand, with per-task token budgets enforced outside the model — a governance control as much as a capacity one.

7. Matching Demand to a System

Table 7 summarizes the selection rules that follow from Sections 2 through 6. They are stated as the conditions under which each class is the right choice; where more than one applies, the model class the mission requires decides.

Table 7. Selection rules by deployment class

CLASS	CHOOSE WHEN	HEADROOM AND GROWTH
Sovereign Workstation, 1–4 GPU	The required model is in the 120B class or below; up to roughly ten concurrent interactive streams per GPU; one to a few users, or several isolated model lanes with one model per GPU; standard office power at one GPU or two Max-Q GPUs, a dedicated circuit beyond that; a pilot that must be redeployable	Add GPUs to four; each GPU is an independent lane; 4-bit precision unlocks the 300B class across four GPUs
Deskside Station, DGX Station GB300	One frontier-class open model up to about 230 GB at 4-bit for a team; the highest single-stream speed and prefill rate of the three classes; long contexts held in HBM; window-defined demand; 1.6 kW in an office environment	The 748 GB coherent pool extends the Station to the 500B–1T class at single-user to small-team speed and carries staging and adaptation; capacity grows by adding a second unit or a workstation lane
Sovereign Appliance 5U, 2–8 GPU	Several independent models at once; concurrent matters and data engineering beside inference; the 670B class at 4-bit; a department's interactive and agentic demand together; 208 V power and a 42–48 in. cabinet are available	Grows from two to eight GPUs in place; ingestion, indexing, and retention run beside inference on the same chassis
Rack-scale (beyond this paper)	Trillion-parameter models at full precision; hundreds of concurrent users on one model; adaptation and training at scale	HGX B300 or NVLink-connected systems; liquid cooling; three-phase power

For a Paradyne platform, the sizing unit is the resident model set of Table 4 at the engagement's concurrency, not its largest model. Two rules of thumb complete the method. Size for the model class the mission requires in year two, not the load in year one; memory is the constraint that cannot be added without a chassis change. And hold interactive demand below half of planning capacity at the projected peak, so that agents, ingestion, and adaptation windows run in the remainder without degrading the experience of the people at the keyboard.

8. The Economics of Owned Capacity

Owned capacity has no price per token. The marginal cost of generation is the system's electricity: a DGX Station drawing 1.6 kW for an eight-hour day consumes about 13 kWh, and an eight-GPU appliance at 5–6 kW about 45 kWh, during which they can produce on the order of 100 million tokens on a 120B-class model. Hosted inference for enterprise use is published at a practical range of $0.10 to $15 per million tokens depending on the model [25], and token consumption in enterprises grew thirteen-fold between January 2025 and March 2026 as agentic workloads spread [25]; analysts have cautioned through 2026 that falling prices per token have not meant falling bills, because agents consume orders of magnitude more tokens per task than the chat exchanges on which the price declines were measured [24].

The comparison that matters to an executive is structural rather than arithmetic. Under a metered model, every expansion of automation is a cost increase decided by the vendor's price list; under owned capacity, it is an allocation decided by the owner, within a ceiling the owner can measure and raise. Paradyne AI sells sustainment, signed updates, and standing programs, so the distinction is not the absence of recurring cost. It is the absence of dependency: if a client stops paying, the system continues to run on the client's hardware with the client's models and data, and only the flow of new work stops.

9. Paradyne AI's Sizing Method

Sovereign AI has been Paradyne AI's vision from day one, and sizing is where that vision becomes engineering: the hardware is vectored in on the mission, calibrated with the neural-network application as one combination, and delivered as the client's property. The method follows the six-gate engagement model.

- **Discover.** The engagement begins with the demand profile, not the hardware: users and their interaction patterns; the agentic tasks to be automated and their token budgets; ingestion volumes and windows; adaptation and fine-tuning requirements; the model class the mission requires; and the facility's power, cooling, and cabinet constraints. Discover is a working conversation conducted over weeks to months, because the demand profile is rarely known at the outset.
- **Architect.** The deployment class is selected with headroom: the model class required in year two, interactive demand below half of planning capacity, and the lanes needed for independent models. Compute is one line item among several; network, storage, power, and cabinet are specified with it and matched to the class.
- **Fine-Tune.** Open-weight models are adapted to the client's domain and data, and the adapted models are retained as the client's property. Precision is chosen per model on measured quality, not assumed.
- **Assemble.** The configuration is built and then characterized in Paradyne's Las Vegas laboratory on the specific unit to be delivered: single-stream speed, aggregate throughput at 1, 8, 32, and 128 streams, prefill rate, time to first token at the client's context lengths, long-context behavior, and power and thermal performance under sustained load. Laboratory characterization replaces every planning figure in this paper with a measured one.
- **Validate.** Acceptance is measured against the client's key performance indicators — accuracy on client data, p95 latency at projected peak concurrency, and cost to serve — with red-team testing of the boundary.
- **Deliver and sustain.** The system is delivered with its measured capacity record, and the record is updated at each signed engine or model update, so that the realized fraction of the ceiling is tracked over the system's life rather than assumed at delivery.

10. Limitations and Scope

This paper is a planning instrument. Its throughput figures are drawn from third-party measurements published between October 2025 and November 2026 on the stated hardware and engines, and from Paradyne estimates derived from the arithmetic of Section 3; they are not service commitments and should not be read as the performance of any delivered Paradyne system, which is established only by laboratory characterization and the acceptance tests of a specific engagement. Model parameters, footprints, and licenses are as published by developers and summarized by the sources cited, and the open-weight landscape changes quarterly. The DGX Station's shipping specification differs from its announced specification, and both appear in published measurements. Eight-GPU appliance figures are published third-party measurements on comparable PCIe systems and Paradyne estimates, not Paradyne laboratory results; eight is the chassis maximum. Inference engines for the Blackwell GPU generation are maturing rapidly, so the realized fraction of the bandwidth ceiling will rise; the planning ranges here are conservative for that reason. Quality differences between precisions, between models, and between adapted and unadapted models are outside the scope of this paper and are addressed in Paradyne AI Research's benchmark series. This is the planning edition; a subsequent edition will replace the planning figures for the DGX Station GB300 and the appliance with Paradyne laboratory measurements on the systems it delivers.

Sources

1. NVIDIA. "NVIDIA RTX PRO 6000 Blackwell Workstation Edition" and "Max-Q Workstation Edition" specifications. nvidia.com, 2025–2026. 96 GB GDDR7 ECC; 1,792 GB/s; 600 W (Max-Q: 300 W); 4,000 (Max-Q: 3,511) AI TOPS.
2. NVIDIA. "RTX PRO 6000 Blackwell Server Edition" datasheet, April 2025 (doc. 3850600), and partner specifications. 96 GB GDDR7; 1,597 GB/s; up to 600 W; 3.7 PFLOPS peak FP4.
3. NVIDIA. "DGX Station Technical Specifications," 2026 datasheet. 1× Blackwell Ultra GPU; 1× Grace 72-core CPU; 252 GB HBM3e at 7.1 TB/s; 496 GB LPDDR5X at 396 GB/s; NVLink-C2C 900 GB/s; ConnectX-8 up to 800 Gb/s.

4. ServeTheHome. “NVIDIA DGX Station Shipping Specifications,” GTC 2026 coverage, March 2026 — noting the change from the announced 288 GB / 8 TB/s configuration; and Tom’s Hardware, “Nvidia launches DGX Station with its bleeding-edge GB300 Grace Blackwell superchip,” 2026 (748 GB total memory).

5. TechRadar. “Nvidia GB300 DGX Station desktop PC goes on sale,” November 2026 (system power 1,600 W; up to 20 PFLOPS FP4; “as much as 1,200 tokens per second” for continuous agent workloads); corroborated by independent specification summaries.

6. Petronella Technology Group. “LLM Benchmark Results: Every Configuration Measured on Our Own GPUs,” including “gpt-oss-120b Benchmarks” and “DGX Station Price and GB300 Efficiency,” September–October 2026. DGX Station GB300, vLLM: gpt-oss-120b 390.7 tok/s single user, 4,924 tok/s at 32 users; GLM-5.3-Flash NVFP4 140–145 tok/s single user, 2,517 tok/s at 256 users, ~32,000 tok/s prefill; Qwen3.8-Flash-Next 205 / 3,007 tok/s, 43,576 tok/s prefill.

7. StorageReview, DGX Station GB300 (ASUS ExpertCenter Pro ET900N G3) measurements as compiled in “How Many Tokens Can DGX Station GB300 Actually Generate?”, November 2026. gpt-oss-120b 9,280 output tok/s at 128 concurrent 512/512 requests; 5,023 tok/s at 8,192/1,024; gpt-oss-20b 22,161 tok/s at 128 and 536 tok/s single stream.

8. Database Mart. “Pro 6000 vLLM Inference Benchmark: Pro6000 vs. H100 vs. A100-80GB,” 2026. gpt-oss-120b 1,525 output tok/s (1,779 total); 32B dense ~830 output tok/s; 8B ~2,750.

9. Eurokommerz. “NVIDIA RTX PRO 6000 Blackwell benchmarks: MLPerf, LLMs, rendering and limits,” October 2026, compiling StorageReview (October–November 2025) and NVIDIA figures: gpt-oss-120b 163 tok/s single request on one Workstation Edition; 176 tok/s per user on two Server Edition GPUs (TP=2, NVFP4); 3,956 tok/s total on four Server Edition GPUs (TP=4, batch 32); Llama 2 70B 32.9 tok/s per user at TP=4, batch 1; NVIDIA’s 1,724 output tok/s per Server Edition on Llama 3.3 70B FP4.

10. Pulsed Media. “NVIDIA RTX Pro 6000 (Blackwell) — 96GB GPU for LLM Inference,” 2026. Llama 3.3 70B Q4_K_M 34 tok/s single stream (llama.cpp); 1,031 tok/s batched (vLLM).

11. Hardware Corner. “RTX PRO 6000 Blackwell Local LLM Benchmarks, Context Scaling & Supported Models,” 2026. gpt-oss 120B MXFP4 210 tok/s generation and 4,579 tok/s prompt processing at 4K context (99.8 and 1,290 at 128K); Llama 3.3 70B Q4 32.0 tok/s at 4K.

12. Millstone AI. “gpt-oss-120b on 1× and 2× RTX Pro 6000 Blackwell Benchmark — MXFP4,” 2026. Single request 181 tok/s (1K context); 361 tok/s at 4 concurrent; two GPUs 664 tok/s at 6 concurrent.

13. Eordano. “LLM Inference on RTX PRO 6000 Blackwell,” 2026. FP8 1,581 vs BF16 1,103 tok/s on one GPU; Qwen3.5-122B FP8 102 tok/s single stream; gpt-oss-120b 2,656 tok/s at request rate 16 on two GPUs.

14. CloudRift. “RTX PRO 6000 vs H100, H200, and L40S: LLM Inference,” 2025. Eight RTX PRO 6000 GPUs achieve 15,713 tok/s against 25,120 expected from linear scaling (GLM-4.5-Air AWQ); PCIe all-reduce as the limiting factor in tensor-parallel decode.

15. Helix ML. “SGLang vs DwarfStar vs vLLM+DSpark: Running DeepSeek 4 on the RTX Pro 6000,” 2026. Eight GPUs: SGLang two TP=4 instances 653 tok/s aggregate at 64 concurrent, 39.1 tok/s single stream; vLLM with DSpark speculative decoding 946.6 tok/s aggregate at 32 concurrent, ~204 tok/s single stream.

16. vLLM project, GitHub issues #56700 and #56892, September 2026. “Field report: running DeepSeek-V4.1-Flash end-to-end on 8× RTX PRO 6000”: TP=8, ~91 GB per GPU, 1M-token context verified; decode 94 tok/s prose and 134 structured, 362 tok/s single stream with speculative decoding; engine decode 791.5 tok/s at 8 concurrent and 1,539.8 at higher concurrency; aggregate falls to 105 tok/s at 90K-token inputs; ~2,452 tok/s aggregate cited for DeepSeek-V4-Flash on the same hardware.

17. Hugging Face, deepseek-ai/DeepSeek-V4-Flash discussion #28, “Running models with vLLM on the RTX Pro 6000 — SM120,” 2026. FP8 on 8× RTX PRO 6000: 40–50 tok/s decode for chat; up to 2,000 tok/s prefill; two GPUs ~100 tok/s on a smaller configuration. Four-GPU extrapolation per Codersera, “DeepSeek V4 Flash on 4× RTX Pro 6000 Blackwell,” 2026.

18. vLLM Blog. “DeepSeek-V3.2 on GB300: Performance Breakthrough,” February 2026. NVFP4 weights of the DeepSeek series require two 288 GB GPUs; 2,816 tokens per GPU per second at 2K input / 1K output; 7,360 prefill-only.

19. NVIDIA TensorRT-LLM. “DeepSeek-V4 on NVIDIA Blackwell: Model-Specific and Agentic-Workload Optimizations,” 2026. Peak output throughput on GB300 raised from 984 to 1,618 tokens/s/GPU; 57.5 concurrent users per GPU at 20 tok/s per user.

20. NVIDIA Developer Blog. “NVIDIA Blackwell Leads on New SemiAnalysis InferenceMAX Benchmarks,” 2025. Llama 3.3 70B: 10,000 tokens/s per Blackwell GPU at 50 tok/s per user (1K/1K); gpt-oss-120b per-GPU throughput 10,000 rising to 30,000 tok/s with speculative decoding at 100 tok/s per user.

21. Sysart Consulting. “Best Open-Weight LLMs 2026: Choosing a Local LLM for Enterprise,” November 2026 (model families, sizes, licenses, and memory requirements); Computing for Geeks, “Open Source LLM Comparison Table (2026),” October 2026 (DeepSeek-V4.1-Flash 552B plus 196B memory tables; GLM-5.3 753B / ~40B active).

22. Context Studios, “Best Open-Weight LLMs 2026 to Self-Host,” and ZenMux, “The Best Open-Source and Open-Weight LLMs in 2026,” 2026 (Qwen3.5-122B-A10B and Qwen3.5 397B / 17B; gpt-oss-120b on a single 80 GB GPU); Morph, “The Best Open Source LLMs (November 2026)” (Kimi K3 2.8T / 104B; DeepSeek-V4.1-Flash 8B prefill / 16B decode active).

23. Bai, L., Huang, Z., Wang, X., Sun, J., Mihalcea, R., Brynjolfsson, E., Pentland, S., and Pei, J. “How Do AI Agents Spend Your Money? Analyzing and Predicting Token Consumption in Agentic Coding Tasks.” Stanford Digital Economy Lab / arXiv:2604.22750, April 2026. Agentic coding tasks average 4.17 million tokens against 3,390 for code chat and 1,190 for code reasoning; input-to-output ratio 153.85; runs on the same task differ by up to 30×.

24. Gartner, March 2026 forecast of agentic workloads at 5–30 times the tokens per task of a chatbot exchange, as reported in HackerNoon, “The Token Price Fallacy,” 2026, and Beri, “Agentic AI Costs 12x More Than CIOs Expected,” July 2026.

26. Supermicro. “Super AI Station ARS-511GD-NB-LCC” datasheet and product page, 2026 (total coherent memory 748 GB: 496 GB LPDDR5X system, 252 GB HBM3e GPU; closed-loop liquid cooling; tower or 5U); BIZON, “AI Station with NVIDIA GB300 Grace Blackwell Ultra Desktop Superchip,” 2026 (748 GB coherent memory); Hardware Corner, “This Desktop Machine Runs 1T Parameter LLMs Locally,” 2026 (GPU executes directly against LPDDR5X over NVLink-C2C; HBM as working set, system memory as extension for weights, KV cache, and experts).

27. pi3g. “Nvidia GB300 Specs, including memory bandwidth and LLM benchmarks, based on 2026 systems,” 2026. Measured on GB300 stations: Kimi K2.5 (1.1T) 40–50 tok/s station total; Nemotron Ultra (550B) ~35 tok/s single request, scaling to 4–5 concurrent; GLM-5.2-REAP (504B) ~60 tok/s. StorageReview, as compiled in [7]: ~35 output tok/s for a very large model with one user; DeepSeek-V4-Flash 148 tok/s and MiniMax M2.7 214 tok/s single user on the DGX Station GB300.

25. SmartDev. “Token Consumption in AI,” glossary, 2026. Enterprise interactions at ~2,000 tokens each; practical enterprise price range of $0.10 to $15 per million tokens; Gartner’s March 2026 finding of thirteen-fold growth in enterprise token consumption since January 2025.

About the Reviewer

This paper was reviewed by Alam Jamal, Chief Executive Officer and Chief Architect of Paradyne AI. He spent two decades in defense intelligence and AI/ML, concluding in a senior civilian role at U.S. Special Operations Command, where he led intelligence operations and full-lifecycle programs from requirement through fielding and sustainment. He holds multiple U.S. utility patents, received the Joint Meritorious Civilian Service Medal in 2026 and the NDIA Superior Achievement Award in 2025, and completed executive education at MIT, Oxford, and Yale.

About Paradyne AI

Paradyne AI engineers sovereign AI end-to-end solutions: AI systems that operate entirely within the client’s own environment, on NVIDIA-Certified infrastructure the client owns, with open-weight models adapted to the client’s data and retained as the client’s property. Its platform is patent-protected, and its Marquis layer inspects, controls, and audits every outbound connection. The company characterizes every hardware configuration in its own laboratory before delivery and engineers for warranted sense-making: findings whose evidence, confidence, and governing priorities are visible to the decision-maker. Headquartered in Las Vegas, Nevada, Paradyne AI is a member of the NVIDIA Inception program. Sentinel Core received two 2026 Silver Davey Awards, including Best Use of Responsible & Ethical AI. paradynetech.ai