GPCN™

GLOBAL PRIVATE CLOUD NETWORK

WHITE PAPER

Running AI Workloads with Predictable Cost and Control

A private GPU and CPU infrastructure model for production inference, data pipelines and continuous AI operations

ONE GPU	SSD INCLUDED	PRIVATE	NO EGRESS
Right-sized private accelerator capacity	Durable storage in the complete GPU service	Allocated capacity for production workloads	None charged by GPCN; external fees may apply

Public Edition | July 2026

Executive Summary

Production AI changes the infrastructure equation

As artificial intelligence moves from experimentation into core production services, infrastructure priorities change. The challenge is no longer simply gaining access to a GPU. Organizations must secure dependable capacity, control data movement, automate repeatable environments and forecast the complete cost of operating AI continuously.

GPCN™ provides private GPU and CPU infrastructure through a controlled cloud-consumption model. A complete GPCN GPU service can start with one right-sized accelerator, includes durable SSD capacity and is privately allocated for the customer. When an AI workload needs additional CPU or system memory, customers can add it from the standard GPCN CPU product pool without oversizing the GPU layer.

This right-sizing matters because some standard public-cloud accelerator configurations—including selected H100 and H200 offerings—have a minimum eight-GPU node or VM. The correct comparison is therefore the complete capacity the customer must purchase, not an artificial per-GPU division of a required multi-GPU bundle.

THE PRODUCTION AI PRINCIPLE Use high-cost GPU resources where acceleration creates measurable value. Place data preparation, application services, orchestration and suitable inference workloads on right-sized CPU infrastructure. This separation improves utilization and stabilizes cost.

What changes when AI enters production

- Inference becomes continuous and demand follows users, applications and data growth.
- Model checkpoints, embeddings, prompts and output data create sustained storage and transfer requirements.
- GPU availability becomes an operational dependency rather than an experimental convenience.
- Security, residency, availability and change control must meet enterprise standards.
- Engineering and finance teams need a common, forecastable operating model.

A complementary AI infrastructure strategy

Public cloud and managed AI services remain useful for rapid prototyping and provider-native capabilities. GPCN™ adds a private production layer for sustained workloads that benefit from allocated capacity, explicit infrastructure control and reduced exposure to variable data-transfer costs.

The Production AI Cost Equation

GPU hourly rates are only one part of the operating model

AI infrastructure is frequently compared using the accelerator rate alone. A useful production comparison must also include the provider's minimum purchasable GPU bundle, supporting CPU services, durable storage, checkpoint movement, inter-service traffic, availability design, engineering effort and the cost of unused or unavailable capacity.

COMPLETE COST Minimum purchasable accelerator capacity + CPU services + durable storage + data movement + resilience + operations. Comparing only a divided per-GPU rate can conceal the cost of a required multi-GPU node.

Match each workload stage to the right resource

AI workload stage	Primary infrastructure	Why it fits
Data ingestion and preparation	CPU + storage	ETL, cleansing, indexing and application logic often do not require premium accelerators.
Training and fine-tuning	GPU	Parallel matrix operations and high-memory acceleration provide the greatest performance benefit.
Inference and model serving	CPU or GPU	Selection depends on model size, latency, concurrency and throughput requirements.
RAG, embeddings and vector workflows	CPU + selective GPU	Use acceleration for embedding or model execution while keeping retrieval and orchestration right-sized.
Control plane and operations	CPU	API services, CI/CD, monitoring and scheduling should not consume high-cost GPU capacity.

Why separation matters

Many production AI platforms keep GPUs occupied with tasks that can run efficiently on CPU infrastructure. Separating the application, data and control layers from the acceleration layer increases GPU utilization and makes capacity planning easier. It also allows teams to scale the expensive resource independently from the rest of the platform.

Data movement is an infrastructure decision

Training data, model artifacts, vector stores and inference outputs can move repeatedly between compute and storage. GPCN™ does not charge egress fees for data leaving or moving within GPCN. If a GPCN workload exchanges data with AWS, Azure, GCP or another external service, that provider may charge for data leaving its environment. Those third-party charges are separate and are not imposed by GPCN. Placement should still account for latency, residency and dependent-system location.

The GPCN™ AI Infrastructure Model

Private capacity, open automation and customer-directed operations

Enterprise NVIDIA platforms

GPCN™ provides access to NVIDIA A100, H100, H200, RTX A6000 and RTX PRO 6000 Blackwell accelerator platforms. Supporting compute, memory and storage are aligned to the deployment design rather than presented here as permanent public specifications. Availability remains subject to location, deployment timing and applicable commercial terms.

Allocated for production use

GPU capacity is planned and privately allocated for the agreed deployment rather than treated as best-effort excess capacity. The complete service can start with one GPU and includes durable SSD capacity. Additional CPU and RAM can be supplied from the standard GPCN CPU product pool without purchasing another GPU, keeping the accelerator layer aligned to actual model demand.

1 Provision	2 Operate	3 Scale intentionally
Select the GPU or CPU profile, geography, operating model and connectivity required by the workload.	Integrate through the REST API and documented Terraform workflows; retain control of the OS, security posture and runtime stack.	Add capacity, reproduce environments or expand regions through explicit customer-directed changes.

Open automation and repeatable workflows

The documented REST API and Terraform workflows support declarative provisioning, version control and CI/CD integration. Environments can be reproduced across supported locations using standard infrastructure-as-code practices, while implementation guides cover Kubernetes deployment and LLM-hosting patterns.

Operational control

- Real-time GPU inventory and GPU-series selection support capacity planning before deployment.
- Customer-defined runtime, security tooling, policies and model-serving stack.
- Two-factor authentication, API and SSH keys, session controls and role-based permissions support controlled operator access.
- Persistent block volumes can be attached, detached and resized as model and dataset requirements evolve.
- Location and capacity changes remain explicit and customer-directed, with SLA-backed service availability.

GPCN™ GPU Platform Options

Enterprise NVIDIA accelerators for different production workload profiles

GPU selection should start with model memory, throughput and workload behavior rather than the newest accelerator name. GPCN™ aligns the supporting VM configuration to the selected platform and deployment requirements, allowing the underlying CPU, system memory and storage design to evolve without changing the workload strategy described in this paper.

NVIDIA platform	GPU memory	Typical starting fit
NVIDIA A100	80 GB	Established training, tuning and sustained inference
NVIDIA H100	80 GB	High-throughput transformer training and inference
NVIDIA H200	141 GB	Memory-intensive models, larger context and datasets
NVIDIA RTX A6000	48 GB	Cost-conscious inference, vision and visualization
NVIDIA RTX PRO 6000 Blackwell	96 GB	Blackwell inference, RAG, simulation and media

PLATFORM SELECTION The accelerator options shown are intended as workload-planning guidance, not fixed public VM specifications. Supporting resources, region, capacity availability, operating model and final commercial terms are confirmed for each deployment.

Use the smallest profile that meets the production objective

- Confirm that GPU memory can hold the model, runtime overhead, active context and expected batching strategy.
- Benchmark target latency and throughput with the intended precision, framework and model-serving stack.
- Keep preprocessing, retrieval, application services and control-plane work on CPU capacity unless acceleration produces measurable benefit.
- When the selected GPU service needs more CPU or system memory, add capacity from the standard GPCN CPU product pool rather than moving to an unnecessarily larger GPU configuration.
- Scale GPU count only after validating utilization, interconnect requirements and the operational value of parallelism.

Production AI Workload Fit

Patterns that benefit from predictable capacity and controlled data movement

Always-on inference and model serving

Customer-facing assistants, recommendation engines and application-integrated models require stable latency and dependable capacity. Allocated GPU or right-sized CPU infrastructure can support predictable serving without relying on opportunistic availability.

Fine-tuning and scheduled training

Reserved acceleration windows or sustained private GPU capacity support repeatable fine-tuning, model adaptation and training programs. The appropriate design depends on dataset size, run frequency, model architecture and completion-time requirements.

Retrieval-augmented generation and embeddings

RAG platforms combine document processing, embedding generation, vector retrieval, application services and model inference. Separating these components across CPU, storage and selective GPU resources avoids using premium capacity for every stage.

Computer vision and media processing

Image, video and sensor workloads often require continuous ingestion and accelerated inference. Private placement can help control data locality, throughput and retention while keeping the processing environment consistent.

Data-intensive AI pipelines

Pipelines that repeatedly move large datasets between preparation, training, inference and analytics can benefit from a service model with no GPCN egress fees and infrastructure positioned close to the data source or consuming application.

Sovereign and compliance-sensitive AI

Explicit workload placement and customer-controlled runtime environments support AI programs with residency, governance or audit requirements. Compliance remains a shared responsibility and must be validated for the selected location and service design.

Business Outcomes and Deployment Decisions

Turn AI infrastructure from a moving target into a managed production platform

Expected outcomes

- More predictable monthly infrastructure spend for sustained workloads.
- Improved assurance that required GPU capacity is available for production operations.
- Higher utilization through deliberate separation of CPU and GPU workloads.
- No GPCN egress fees within the service model and more predictable data-movement economics.
- Repeatable deployment through API-driven infrastructure-as-code workflows.
- Greater control over placement, operating systems and runtime architecture.

ECONOMIC POSITION GPU economics vary by accelerator, region and the minimum public-cloud shape required. GPCN™ combines complete one-GPU right-sizing, included durable SSD, private allocated capacity, no GPCN egress fees within the service model and predictable monthly service. Some public-cloud comparisons require the customer to purchase an entire eight-GPU node. Evaluate the full purchasable configuration, and add CPU or RAM from the standard GPCN CPU product pool when the workload needs more supporting compute.

Six inputs for responsible sizing

- Model architecture and required GPU memory.
- Latency, concurrency and throughput targets.
- Training frequency, run duration and completion windows.
- Dataset size, storage performance and data-residency requirements.
- Availability, recovery and operational-support objectives.
- Expected utilization and the commercial term needed to secure capacity.

Conclusion

Production AI does not inherently require hyperscale complexity. It requires the correct accelerators, dependable supporting infrastructure, controlled data movement and an operating model that engineering and finance teams can both understand.

By combining private NVIDIA GPU platforms, right-sized CPU infrastructure, open automation, no GPCN egress fees within the service model and SLA-backed availability, GPCN™ provides a practical foundation for scaling AI with greater cost predictability and architectural control.

Available through EnTelegent Solutions | gpcn.entelegent.com | 800-975-7192 | 2006