Ai-MicroCloud® Pro

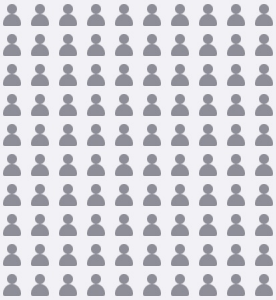
AI FACTORY
AND NEOCLOUD

TODAY · HOSTED AI

Hosted AI API
CLAUDE · GPT

Prompts leave your network.
Per-token fees.
Limits set by someone else.

YOUR ORGANISATION
OR YOUR TENANTS



500+ users
in a conversation at once,
about 142 generating

REPOINT

same API,
local URL

SECURITY PERIMETER · YOUR AI FACTORY



Fill out the rack
8–10 nodes
per 42U

Start with 3 nodes
fill out the rack as demand grows

LLM serving and API · Agents and tools · Guardrails · Notebooks and data · Teams and quotas · Studios: optional add-ons

Your models, roles, keys. No per-token bill. Air-gapped or connected.

Private AI cloud in your rack.

Three nodes, up to twelve GPUs. The AI factory for an enterprise and the platform for a neocloud provider: three nodes serve 500+ users from day one, and every node you add brings the same tenants, models, metering and controls, inside your perimeter.

500 → 2,100	people in a conversation at once, 3 to 12 GPUs[1]
2.4×	users per GPU with intelligent context offloading[1]
132B → 621B	secure tokens a year, in house[1]
100%	of prompts and documents stay on site
Multi-tenant	Pro edition: AI factory and neocloud, one-click deployment behind an OpenAI-compatible API

Zeblok Computational · © 2026 · Specifications subject to change without notice

Zeblok COMPUTATIONAL

One platform. Two ways to run it.

Start with three nodes. Fill out a rack.

Built for both.

AI factory for the enterprise

One platform for every team on shared GPUs; organisations, quotas and resource plans decide who gets what. Start with three nodes; add nodes as teams join.

Platform for neocloud providers

Sell AI capacity to your customers: isolated organisations, per-key token metering, expiring API keys and audit logs. Fill the rack as tenants sign.

Up and running in four steps.

Fully automated, one-click deployment.

1 **Pick your model**

Import any Hugging Face model. “Where can I run this?” checks it fits your GPUs.

2 **Deploy in one click**

An inference endpoint goes live on your GPUs, served by vLLM behind an OpenAI-compatible API.

3 **Get your keys**

API keys with expiry, revocation and per-key token usage.

4 **Ready to use**

Chat, agents and notebooks open to your teams; the Studios are optional add-ons.

Fully secure. Fully controlled.

Nothing leaves the site

Models, prompts, documents and answers stay on the Neuron; air-gapped install from a local registry.

Access you control

Organisations, user groups, roles and isolation policies, with email or authenticator two-factor sign-in.

Keys and quotas

Expiring API keys per user or agent, token accounting per key, resource plans per team.

Safe answers

Guardrails with PII and secret redaction, tool allow-lists, egress monitoring and audit logs.

Zeblok Computational · © 2026 · Specifications subject to change without notice

Zeblok COMPUTATIONAL

Three to twelve GPUs. One rack.

288 → 1,152 GB of GPU memory across three nodes.

Choose a configuration

8B model · 70B model · one row per node

	1 GPU per node · 3 GPUs	2 GPUs per node · 6 GPUs	3 GPUs per node · 9 GPUs	4 GPUs per node · 12 GPUs
				
	288 GB	**576 GB**	**864 GB**	**1,152 GB**
Chats at once	**142** (78 + 64)	**228** (100 + 128)	**342** (150 + 192)	**534** (150 + 384)
Tokens per hour	**15.6M**	**27.8M**	**41.7M**	**73.5M**
Knowledge-base pages per hour	1,500–3,900	3,000–7,800	4,500–11,700	9,000–23,400
Runs	70B 4-bit on two nodes; 8B on the third	70B FP8 across two cards on two nodes; two 8B on the third	70B FP8 across two cards and an 8B on every node	70B FP8 across two cards and two 8B on every node

How many people each workload serves.

	3 GPUs	12 GPUs	Includes · models imported at deployment
Chat assistant and API	**142**	**534**	API with keys and token accounting, chat, General Assistant agents Llama 3.1 8B / Qwen3 8B · Llama 3.3 70B
RAG and knowledge-base search	**192**	**768**	Knowledge Studio add-on: cited knowledge base, PDF datasets with OCR, RAG agents Nemotron 3 Nano 4B · Nemotron 3 Super
Agents and automation	**72**	**384**	Agent Hub: six agent types, MCP tools, 18 database connectors, guardrails Nemotron 3 Nano 4B · Nemotron 3 Super
Data analysis	**171**	**684**	Data Analyst agents on PostgreSQL, MySQL, MongoDB and more; notebooks Qwen3 32B
Data science and ML	**3**	**12**	GPU notebooks: JupyterLab (PySpark, RAPIDS, Hugging Face), MLflow, per-team resource plans Any Model Hub model

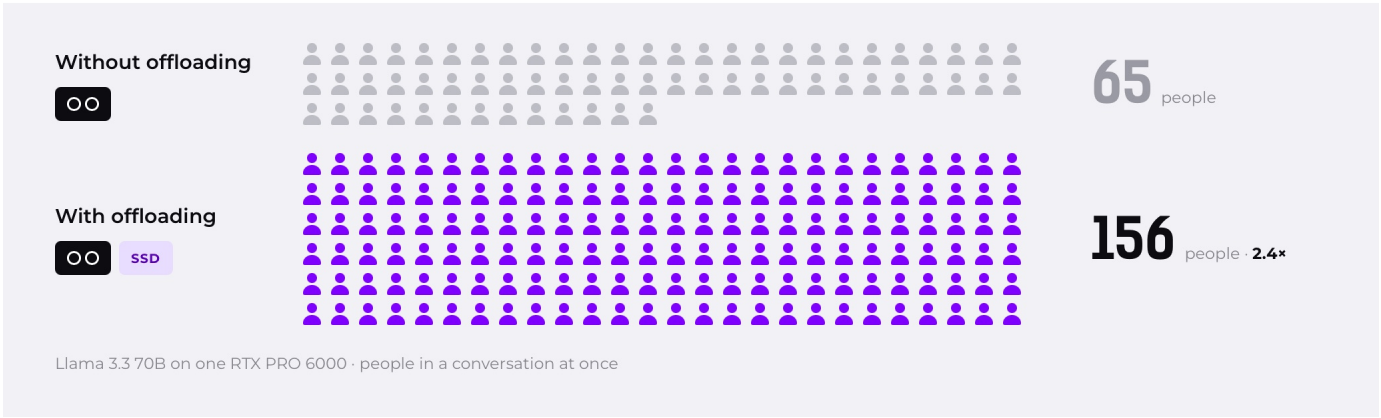
Simultaneous sessions at 15+ tokens/s each.[1]

Zeblok Computational · © 2026 · Specifications subject to change without notice

Zeblok COMPUTATIONAL

Same people. Half the GPUs.

GPUs are usually bought to hold context, not to compute. Ai-MicroCloud® moves idle conversations' context to system memory and an NVMe SSD in each node, and brings it back the moment the user replies.



More people per GPU.

People in a conversation at once; grey without offloading, violet with. Each person generates about a quarter of the time, with a 2,000-token context.

Qwen3 32B 1 × RTX PRO 6000	93 228		**2.4×**
Llama 3.3 70B 1 × RTX PRO 6000	65 156		**2.4×**
Nemotron 3 Super 2 × RTX PRO 6000	256 256		**same**

1,000 people, fewer cards.

Llama 3.3 70B: a card holds 65 conversations without offloading and 156 with it. Needs the offloading image and an NVMe SSD in every node.

16 × RTX PRO 6000 · 4 nodes
without offloading · 65 people a card

7 × RTX PRO 6000 · 2 nodes SSD
with offloading · 156 people a card

–56% GPUs

Less power, too.

1,000 people · Llama 3.3 70B · 600 W cards and a reference node of about 350 W under load. Bar: idle to busy; tick: peak. Estimated from component ratings, not measured.

With offloading
7 × RTX PRO 6000 · two nodes busy
3,820 W busy · –55% · 630 W idle · 5,560 W peak

Without
16 × RTX PRO 6000 · four nodes, needs a fourth
8,480 W busy · 1,270 W idle · 12,390 W peak

Zeblok Computational · © 2026 · Specifications subject to change without notice

What the public cloud would charge for the same tokens.[2]

Retail list prices, 8 hours × 22 days a month.

STARTING POINT · 3 NODES, ONE GPU EACH

132B

secure tokens a year

32.9B generated + 98.8B prompt, all processed on site

Claude Opus 5.5 Anthropic	**$1.05M / year** $87,859 a month
GPT-6 Astra OpenAI flagship	**$2.64M / year** $219,648 a month
GPT-6.1 Sol OpenAI	**$527K / year** $43,930 a month

RACK FILLED · 3 NODES, FOUR GPUS EACH

621B

secure tokens a year

155.2B generated + 465.7B prompt, all processed on site

Claude Opus 5.5 Anthropic	**$4.97M / year** $413,952 a month
GPT-6 Astra OpenAI flagship	**$12.42M / year** $1,034,880 a month
GPT-6.1 Sol OpenAI	**$2.48M / year** $206,976 a month

Public cloud tokens are shared-infrastructure tokens.

Retention

Prompts and answers cross shared infrastructure and are kept up to 30 days for abuse monitoring; zero retention is by agreement only.

Residency

Keeping inference in one country costs 10% more (Anthropic US-only inference; Bedrock and Vertex regional endpoints).

Reserved capacity

Dedicated private capacity is billed by the hour, used or not, with a 15-unit minimum (Azure OpenAI provisioned throughput).

Contact sales

sales@zeblok.com
+1 (631) 223-8233
1500 Stony Brook Road
Stony Brook, NY 11794

Book a complimentary 30-minute capacity planning session.

Bring your multi-model workload. Leave with a node and GPU plan, from three nodes to a full rack.

Ask about the Studio add-ons.

Knowledge Studio and Document Studio are licensed separately; ask for a demo on your documents.

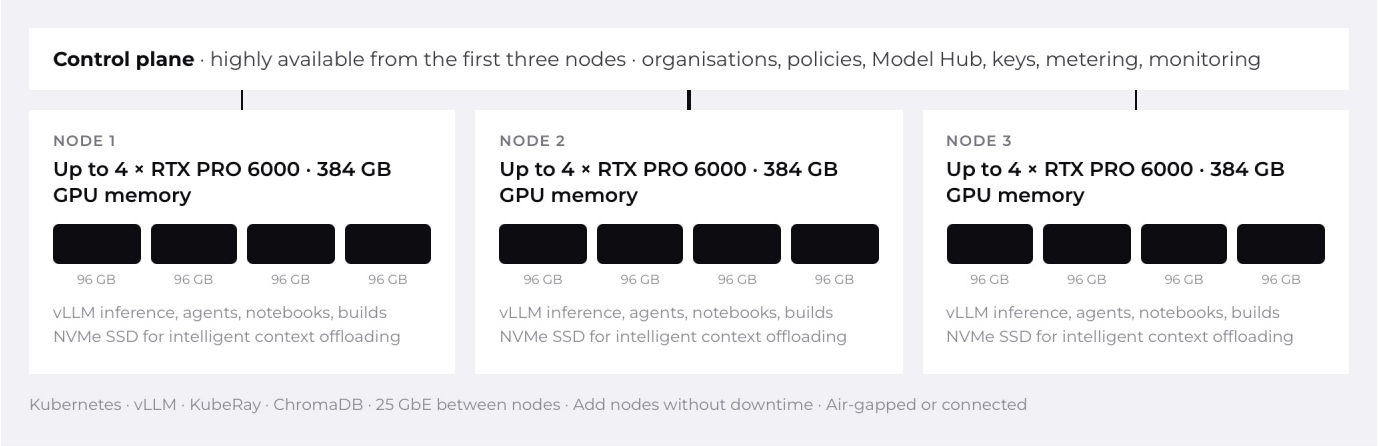
[2] Anthropic and OpenAI retail list prices, standard tier, October 2026. Token volume is the workstation's planning capacity (15.6M and 73.5M generated tokens an hour, 3 prompt tokens per generated token) over 2,112 hours a year. Retention, residency and provisioned-capacity terms from the providers' published policies and price lists, October 2026. Network security and egress for a cloud service come on top. The comparison is by token volume, not model capability.

Zeblok Computational · © 2026 · Specifications subject to change without notice

Zeblok COMPUTATIONAL

The cluster under management.

Three nodes to start, a rack to grow into.



Reference node sized to four GPUs: 32–64 cores, 256–512 GB DDR5 ECC, NVMe SSDs, 25 GbE, two 3 kW power supplies; final node specification per partner. A 42U rack takes eight to ten such nodes; the control plane and metering carry over as nodes are added.

Platform features.

What the Pro edition adds for an enterprise AI factory or a provider rack. Everything here is in the platform; the Studios are separate add-ons.

On-ramps for models, code and data

Any Hugging Face model, public or private; your own container images and notebooks; datasets versioned by upload; your Harbor registry.

HPC-style batch and distributed compute

Autoscaling Ray clusters, batch pipelines as Kubernetes jobs, tensor-parallel inference across the GPUs of a node.

Multi-tenant by design

Organisations with their own data centers and namespaces, policy-based access rules and a network policy around every workload.

Identity and governance

Roles, user groups and isolation policies. Two-factor sign-in by email or authenticator; Google sign-in. Activity and audit logs.

Capacity planning built in

Overview, capacity, latency, cost and live-infrastructure views, with Prometheus metrics and per-pod Grafana links.

Multi-site from one console

Primary and secondary clusters registered to one control plane; workloads placed on the hub, a satellite cluster or an edge node.

High availability for serving

Inference replicas autoscale on GPU utilisation. Microservices run multiple replicas. Scheduled jobs survive a backend restart.

Metering per key

Input, output and total tokens per API key. Agent turns logged. A weekly usage report to every user.

Lifecycle management

LifeCycle Manager delivers platform updates through the Launcher, with an in-console banner. One-click add-on install.

Zeblok Computational · © 2026 · Specifications subject to change without notice

Every model, sized to your rack.

Users at 15+ tokens/s each with a 2,000-token context, and tok/s: the speed one user sees.

Which model fits which GPU.

	1 × RTX PRO 6000 96 GB	2 × RTX PRO 6000 192 GB	4 × RTX PRO 6000 one node · 384 GB	BEST FOR
Llama 3.1 8B / Qwen3 8B Small · 4-bit · 4.7 GB	**64 users** 237 tok/s	**128 users** 2 copies, one per card	**256 users** 4 copies, one per card	Team chat and knowledge base
Nemotron 3 Nano 4B Nemotron Nano · FP8 · 5.3 GB	**64 users** 210 tok/s	**128 users** 2 copies, one per card	**256 users** 4 copies, one per card	Fast chat, RAG and agents
Qwen3 32B Mid-size · FP8 · 34.4 GB	**57 users** 33 tok/s	**114 users** 2 copies, one per card	**228 users** 4 copies, one per card	High-quality answers and data analysis
Llama 3.3 70B Large · 4-bit · 40.9 GB	**39 users** 28 tok/s	**78 users** 2 copies, one per card	**156 users** 4 copies, one per card	Best quality per card
Nemotron 3 Super 120B Large · MoE · FP4 · 84.9 GB · 12B active	**24 users** 48 tok/s	**64 users** 70 tok/s across both cards	**128 users** 2 copies, 2 cards each	Agentic reasoning, 1M-token context

64 is the presets' concurrency limit per model copy.

Nemotron 3 Super NVFP4 on one 4-GPU node.

Two copies per node, each across two cards; 79 GiB of NVFP4 weights per copy at 95% GPU memory with an FP8 KV cache and FP16 Mamba cache; pages an hour use 4 to 16 build workers per copy. Offloading doesn't change generation speed: it parks idle users' context (4.2 GB per 1M tokens) on the node's SSD and restores it in under a second. NVIDIA lists one B200 or DGX Spark as the minimum: validate on a node; vLLM 0.20.

128
users at 15+ tokens/s, two copies

70
tokens/s for one user

2,000+
knowledge-base pages an hour, up to 4,400 (Knowledge Studio add-on)[1]

1M
tokens of context for one user

512
concurrent users per session with intelligent context offloading

< 1 s
first token for a returning user with offloading and 1M context (60 s if recomputed)

[1] Planning estimates, not benchmarks: decode modelled as bandwidth bound (65% of peak), prompts as compute bound (40% of dense FP8); vLLM at 90% of GPU memory, 16-bit KV cache (Nemotron 3 Super: 95%, FP8 KV cache, per NVIDIA), at most 64 sequences per model copy, 30% derate; chats hold 2,000 tokens of context; capacity scales linearly across model copies. Benchmark one node before committing to service levels. Sources: nvidia.com and Ai-MicroCloud® Pro, October 2026. Agent and Gen AI workspace features are beta.

Zeblok Computational · © 2026 · Specifications subject to change without notice

Zeblok COMPUTATIONAL

Two views of the same platform.

DEVELOPER VIEW

Ship against one endpoint.

Notebook Assistant: an AI pair-programmer in JupyterLab that generates, runs and fixes cells, on an external model or a running Ai-MicroCloud® inference.

OpenAI-compatible API

Every model is served by vLLM at /v1/chat/completions; existing SDKs point at a local URL.

Keys you own

Gen AI keys per user or agent, with expiry and one-time display; token usage per key.

Agents and tools

Six agent types, MCP servers from the Agent Hub, read-only connectors to 18 database types.

Notebooks in minutes

JupyterLab images for data science, PySpark, R, RAPIDS and Hugging Face; MLflow one click away.

Your own containers

Deploy a microservice or an Ai-API from your image, an archive or a notebook.

OPERATOR VIEW

Run the rack as one system.

Per-deployment utilisation: time to first token, time per output token, and request and token counts for users and agents.

Hardware in, plans out

Nodes are discovered with their GPUs and SSDs; resource plans set GPU, CPU, memory and storage per team.

Tenants and policies

Organisations, roles and isolation policies decide who sees and runs what.

Watch everything

Prometheus metrics for nodes, workloads and inference; Grafana per pod; activity logs.

Grow without downtime

Add nodes to the cluster; inference scales out on GPU utilisation.

Keep it current

Platform updates through the Launcher, surfaced in the console before they roll.

Optional add-ons, licensed separately.

Not included with Ai-MicroCloud® Pro. Each is quoted and licensed on its own.

Knowledge Studio

Turns PDFs, Office files and web pages into a searchable, cited knowledge base, with evaluations that grade answers against it.

Document Studio

Generates marketing and sales documents (datasheets, proposals, case studies) from your own material, on your own GPUs.

Domain adaptation

Fine-tunes open models on your own data, on your own GPUs, into domain-specific models.

Ai-MicroCloud® is a registered trademark of Zeblok Computational.

Zeblok Computational · © 2026 · Specifications subject to change without notice

Zeblok COMPUTATIONAL