# Ai-MicroCloud® Lite

ZEBLOK SOFTWARE +
ACE NEURON HARDWARE




# Private AI cloud on your own workstation.

Instead of sending 100 users to a public AI cloud, serve them in house on one Ace Neuron AI Workstation. Prompts, documents and models never leave your perimeter.

| | |
|---|---|
| **100 → 500+** | users, base to fully expanded[1] |
| **2.4×** | users per GPU with intelligent context offloading[1] |
| **11.8B → 118B** | secure tokens a year, in house[1] |
| **100%** | of prompts and documents stay on site |
| **1 click** | to deploy a model behind an OpenAI-compatible API, on one workstation, on premises |



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

Zeblok COMPUTATIONAL

# Up and running in four steps.

Fully automated, one-click deployment: from any Hugging Face model to a live endpoint your teams can use.

## One click to a live endpoint.

Each step is automated by the platform.

**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, notebooks and the optional knowledge base open to your teams.

## 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

# Start with one GPU. Grow to four.

16 → 304 GB of GPU memory, in one box.

## Choose a configuration

8B model · 70B model

| | BASE | + 1 × RTX PRO 6000 | + 2 × RTX PRO 6000 | + 3 × RTX PRO 6000 |
|---|---|---|---|---|
| | **16 GB** | **112 GB** | **208 GB** | **304 GB** |
| Chats at once | **20** | **60** (40 + 20) | **70** (50 + 20) | **114** (50 + 64) |
| Tokens per hour | **1.4M** | **3.9M** | **4.7M** | **13.9M** |
| Knowledge-base pages per hour | 400–800 | 400–800 | 400–800 | 1,500–3,900 |
| Runs | 8B for chat, agents and builds | 70B 4-bit on the RTX PRO 6000; 8B on the RTX PRO 2000 | 70B FP8 on two RTX PRO 6000s; 8B on the RTX PRO 2000 | 70B FP8 on two RTX PRO 6000s; 8B on the third |

## How many people each workload serves.

| | BASE GPU | + RTX PRO 6000 | INCLUDES · MODELS IMPORTED AT DEPLOYMENT |
|---|---|---|---|
| Chat assistant and API | **20** | **40** | API with keys and token accounting, chat, General Assistant agents<br>Llama 3.1 8B / Qwen3 8B · Llama 3.3 70B |
| RAG and knowledge-base search | **21** | **64** | Optional add-on: knowledge base in ChromaDB, PDF datasets with OCR and citations, RAG agents<br>Nemotron 3 Nano 4B · Nemotron 3 Nano 30B |
| Agents and automation | **25** | **64** | Agent Hub: six agent types, MCP tools, 18 database connectors, guardrails<br>Nemotron 3 Nano 30B · Nemotron 3 Super |
| Data analysis | **7** | **57** | Data Analyst agents on PostgreSQL, MySQL, MongoDB and more; notebooks<br>Qwen3 14B / Qwen3 32B |
| Data science and ML | **40 GB + 6 vCPU**<br>free of 128 GB · 32 vCPU | | JupyterLab (PySpark, RAPIDS, Hugging Face), MLflow, per-user plans<br>Any Model Hub model |

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



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

Zeblok COMPUTATIONAL

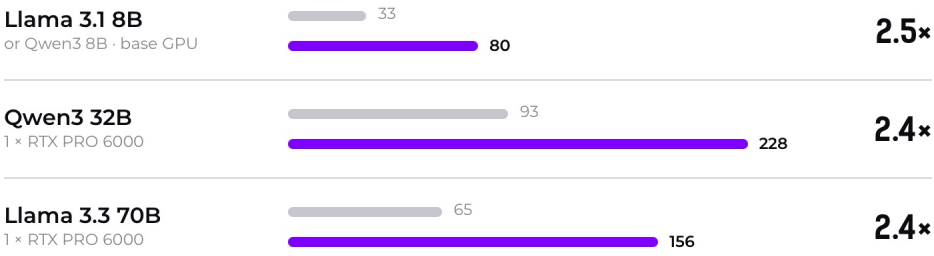
# Same people. Fewer GPUs.

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



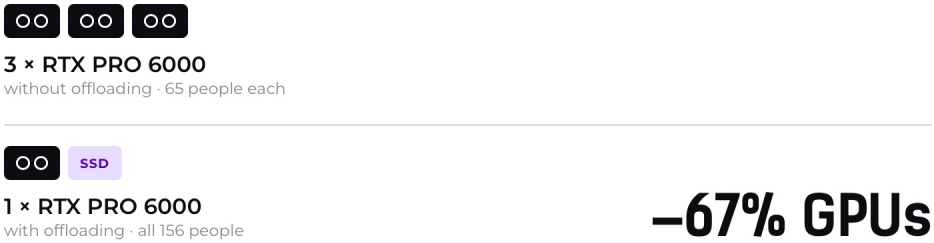

## More people per GPU.

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.



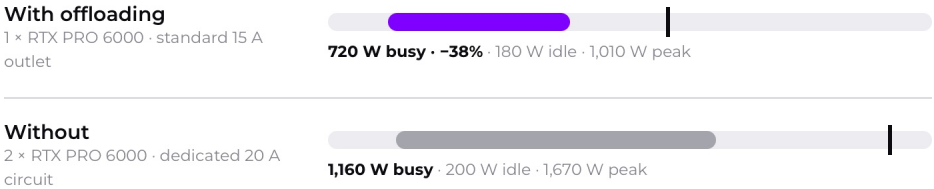

## 156 people, one card.

Llama 3.3 70B. Two cards hold 192 without offloading. Nemotron hybrid models keep little context per user and see no change. Needs the offloading image and an NVMe SSD in a free M.2 slot.




## Less power, too.

100 people in a conversation at once · Llama 3.3 70B · 600 W cards. Bar: idle to busy; tick: peak. Max-Q cards draw about 540 W and 810 W. Estimated from component ratings, not measured.






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

Zeblok COMPUTATIONAL

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

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

BASE CONFIGURATION

## 11.8B

**secure tokens a year**

3.0B generated + 8.9B prompt, all processed on site

| | |
|---|---|
| **Claude Opus 5.5**<br>Anthropic | **$95K / year**<br>$7,885 a month |
| **GPT-6 Astra**<br>OpenAI flagship | **$237K / year**<br>$19,712 a month |
| **GPT-6.1 Sol**<br>OpenAI | **$47K / year**<br>$3,942 a month |

FULLY EXPANDED · 3 × RTX PRO 6000

## 118B

**secure tokens a year**

29.6B generated + 88.7B prompt, all processed on site

| | |
|---|---|
| **Claude Opus 5.5**<br>Anthropic | **$946K / year**<br>$78,848 a month |
| **GPT-6 Astra**<br>OpenAI flagship | **$2.37M / year**<br>$197,120 a month |
| **GPT-6.1 Sol**<br>OpenAI | **$473K / year**<br>$39,424 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 workload. Leave with a GPU plan for the box, from the base unit to four cards.

**Ask about AI-MicroCloud® PRO.**

Private cloud in a rack: the same platform across three or more nodes.

[2] Anthropic and OpenAI retail list prices, standard tier, October 2026. Token volume is the workstation's planning capacity (1.4M and 14M 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

# Ai-MicroCloud® Lite specifications.

## Hardware under management.

Ace Neuron LC-F3XS2

### CPU

**16 cores**

— AMD EPYC 9115 · 32 threads
— Platform, vector store, OCR, builds

### Memory

**128 GB**

— DDR5 ECC RDIMM · 8 slots
— Offloaded context, databases, notebooks

### Storage

**2 TB**

— M.2 NVMe + 2 free PCIe 5.0 slots
— Models and documents; a free slot takes the offload SSD

### Power

**265 W**

— when serving · 850 W Gold PSU
— Fully expanded 1.1–1.6 kW: larger PSU, 20 A or 240 V circuit

### GPU

**16 GB**

— RTX PRO 2000 Blackwell · 70 W
— One model up to 8B (14B at 4-bit)

### GPU expansion

**+3 × 96 GB**

— RTX PRO 6000 Blackwell
— 70B to 120B models

### Network

**2 × 1 GbE**

— RJ-45 + fibre
— Air-gapped or connected

## Platform features.

What comes with the Lite edition on one Ace Neuron. 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.

### Batch and distributed compute

Autoscaling Ray clusters, batch pipelines as Kubernetes jobs, tensor-parallel inference across the box's GPUs.

### Teams and tenants

Organisations with their own 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.

### One box today, a rack when you need it

Console, keys and metering carry over as nodes are added. Ai-MicroCloud® Pro adds multi-site and provider features.

### Resilient 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

Zeblok COMPUTATIONAL

# Every model, sized to your GPUs.

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.

| | BASE GPU RTX PRO 2000 · 16 GB | 1 × RTX PRO 6000 96 GB | 2 × RTX PRO 6000 192 GB | BEST FOR |
|---|---|---|---|---|
| **Llama 3.1 8B / Qwen3 8B**<br>Small · 4-bit · 4.7 GB | **20 users**<br>38 tok/s | **64 users**<br>237 tok/s | **64 users**<br>379 tok/s | Team chat and knowledge base |
| **Nemotron 3 Nano 4B**<br>Nemotron Nano · FP8 · 5.3 GB | **25 users**<br>34 tok/s | **64 users**<br>210 tok/s | **128 users**<br>2 copies, one per GPU | Fast chat, RAG and agents on the base GPU |
| **Qwen3 32B**<br>Mid-size · FP8 · 34.4 GB | Does not fit | **57 users**<br>33 tok/s | **64 users**<br>53 tok/s | High-quality answers and data analysis |
| **Llama 3.3 70B**<br>Large · 4-bit · 40.9 GB | Does not fit | **39 users**<br>28 tok/s | **64 users**<br>45 tok/s | Best quality on one card |
| **Nemotron 3 Super 120B**<br>Large · MoE · FP4 · 84.9 GB · 12B active | Does not fit | **24 users**<br>48 tok/s | **64 users**<br>70 tok/s | Agentic reasoning, 1M-token context |

64 is the presets' concurrency limit.

## Nemotron 3 Super NVFP4 on one RTX PRO 6000.

Fits 79 GiB of NVFP4 weights at the platform's 95% GPU-memory setting, with an FP8 KV cache and FP16 Mamba cache; pages an hour use 4 to 16 build workers. Offloading doesn't change generation speed: it parks idle users' context (4.2 GB for a full 1M-token context) in DRAM and on the SSD and restores it in under a second instead of recomputing it (about 60 s for 1M tokens). Every user keeps a full context between turns. NVIDIA lists one B200 or DGX Spark as the minimum: validate on the unit; vLLM 0.20.

**24**
users at 15+ tokens/s

**48**
tokens/s for one user

**500–1,100**
knowledge-base pages an hour (add-on)[1]

**1M**
tokens of context for one user

**96**
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, at most 64 sequences, 30% derate; chats hold 2,000 tokens of context. Benchmark the delivered unit before committing to service levels. Sources: acecomputers.com (LC-F3XS2), nvidia.com and Ai-MicroCloud® Lite, 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 box 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

The box is discovered with its GPUs and SSD; resource plans set GPU, CPU, memory and storage per team.

### Teams and policies

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

### Watch everything

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

### Grow when ready

Add a card to the box or a node beside it; 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® Lite. 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