AI infrastructure for your enterprise.

We supply and install the servers, build private AI systems on your data, and sell tokens from capacity in your region. Start with one and add the others as you grow.

What we offer

Hardware

Sizing, sourcing and installation of servers, workstations, networking and storage, from a single node to a multi-rack pod.

Sovereign AI

Open models adapted to your data and languages, deployed on premises, in a sovereign cloud or air-gapped. You own the weights and pipelines.

Token services

Metered access to open-source models through an OpenAI-compatible API, served from GPU capacity in the region you choose.

Ways to buy

Option	How you pay	Best when
Token services, shared	Per million tokens used	Pilots, internal tools, variable workloads. Fastest to start
Token services, dedicated	Reserved capacity, recurring term	Steady volume, predictable latency, no hardware to own
Own hardware	Purchase, phased delivery where useful	High, steady volume and the tightest data control
Hybrid	Tokens now, hardware later	Prove value before committing capital

All options use the same open-source models and API, so you can change course without rebuilding applications. We model power, space, support and staffing in every proposal.

GPU models we quote

GPU	Memory	Best for
NVIDIA B300	288 GB HBM3e	The largest models, long-context and reasoning workloads
NVIDIA B200	180 GB HBM3e (HGX B200)	High-throughput serving and training of large models
NVIDIA H200	141 GB HBM3e, 4.8 TB/s	Large-model inference and fine-tuning, long context
NVIDIA H100	80 GB HBM3 (SXM), 94 GB (NVL)	Proven training and inference
RTX PRO 6000 Blackwell	96 GB GDDR7 with ECC	Private inference, large-scene rendering, mixed workloads
RTX PRO 5000 Blackwell	48 GB GDDR7	Developer and studio workstations
NVIDIA L40S	48 GB GDDR6	Cost-efficient inference, render farms, virtual workstations
GeForce RTX 5090	32 GB GDDR7	Individual artists and developers

Figures are from the manufacturer's published specifications and vary by configuration. Availability and pricing are confirmed per order. GeForce products are intended for desktop and workstation use and are not licensed for data centre deployment.

Open-source models we deploy

DeepSeek V4, GLM-5.2, Qwen 3, gpt-oss and Mistral families, released under Apache 2.0 or MIT licences, shortlisted by language and size, then evaluated on your own data before hardware is committed. Details at avanticompute.com/models.

Planning your facility

Power and cooling decide what you can install and when. Indicative maximums for planning, confirmed against the vendor datasheet for your configuration:

System	Indicative maximum power	Cooling
8-GPU Hopper server (H100, H200)	Up to about 10 kW per server	Air-cooled
8-GPU Blackwell server (B200)	Up to about 14 kW per server	Air-cooled or direct liquid
RTX PRO 6000 Server Edition	Up to 600 W per GPU, plus CPUs and fans	Air-cooled; passive cards need server airflow

How an engagement works

1. Brief

You tell us the workload, site and timeline. We reply within two business days.

2. Proposal

A configuration, lead times and a price, with assumptions written down.

3. Delivery

Procurement, installation, burn-in and benchmarks on your own models.

4. Support

Warranty handling, updates and a plan to expand.

What every quote includes

Bill of materials with manufacturer part numbers and licences.
Itemised price and its validity period.
Confirmed lead time, delivery terms, freight and duties.
Warranty terms and support options for your site.
Power, cooling, weight, space and network requirements.
Installation plan and acceptance tests.

Who we serve

Banks and insurers, healthcare, telecom, energy and industry, design and media, and data centre operators: organizations whose data, rules or workloads need AI infrastructure under their own control.

Contact

Calgary

Suite 5000, Suncor Energy Centre West Tower
150 6 Avenue SW
Calgary, AB T2P 3Y7

Montréal

4055 Rue Sainte-Catherine Ouest, Unit 152
Montréal, QC H3Z 3J8

contact@avanticompute.com · avanticompute.com

NVIDIA, GeForce, RTX and related product names are trademarks of NVIDIA Corporation. Avanti Compute is an independent company.