ARC COMPUTE

Your End-to-End GPU Infrastructure Partner

Empowering AI & HPC Innovation Through Consultative, End-to-End Infrastructure Solutions

Arc Compute at a Glance

ARC COMPUTE

Arc Compute powers high-performance GPU infrastructure for AI and HPC workloads. We provide the hardware, expertise, and support to help businesses deploy and scale GPU systems—whether for training large AI models, running simulations, or powering demanding applications.

We specialize in:

- Sourcing and deploying enterprise-grade GPU clusters
- Rack-level infrastructure for on-premise or colocation environments
- High-bandwidth networking and scalable storage solutions

Arc Compute helps you get the most out of your GPUs — so you can focus on performance, not procurement and setup.

Our Partners

ARC COMPUTE

NVIDIA

SUPERMICRO

Hewlett Packard Enterprise

DELL

AIVRES

CAMBRIDGE Computer

ARISTA

intel

GIGAIO

tenstorrent

The AI Infrastructure Challenge

- Growing AI workloads demand massive compute power, which can be prohibitively expensive and difficult to manage.
- Organizations face tough decisions between cloud, on-prem, or colocation. Each choice impacts performance, cost, and flexibility.
- High Performance GPU infrastucture is highly complex and difficult to design, deploy and maintain

Cost Overruns

Public cloud GPU instances can quickly become cost-prohibitive at scale.

Capital Constraints

On-premises solutions often require significant upfront investment.

Expertise Gaps

Sourcing, installing, and managing GPU clusters demands specialized knowledge.

Comparing Deployment Options

We guide you through these options, weighing factors like workload consistency, data security, and budget. The result: a deployment model that optimizes both performance and cost and allows you to focus on getting your products to market faster, not focus on your infrastructure.

Cloud Instances

- Pros: Rapid deployment, no hardware to manage.
- Cons: Potentially high operating costs at scale, poor long-term feasability, vendor lock-in.

On-Premises

- Pros: Greater control, cost savings for long-term, heavy workloads.
- Cons: Higher initial capital expenditure, may require in-house expertise.

Colocation

- Pros: Shared facility and infrastructure costs, better control than public cloud, lower upfront costs than fully on-prem.
- Cons: Requires coordination with a data center partner, though Arc Compute handles this on your behalf.

Comparing Deployment Options

5-Year Cost Example: 5x NVIDIA HGX 8-GPU Systems

Deployment	Year 1	Year 2	Year 3	Year 4	Year 5	5-Year Total
Public Cloud (40 GPUs + Minimal Extras)	$825,920	$825,920	$825,920	$825,920	$825,920	$825,920 × 5 = **$4,129,600**
On-Premises (Purchase + Full Staffing)	$1,620,000	$280,000	$280,000	$280,000	$280,000	$1,620,000 + (4 × $280,000) = **$2,740,000**
Colocation (Purchase + Partial Staffing)	$1,675,000	$335,000	$335,000	$335,000	$335,000	$1,675,000 + (4 × $335,000) = **$3,015,000**

The cost figures presented in this comparison are for illustrative purposes only. Actual expenses may vary significantly based on factors such as specific hardware requirements, market conditions, regional labor costs, and logistical complexities. Arc cannot guarantee these estimates or be held liable for variations in real-life deployment costs.

Our Consultative Approach

Customized for Your Workloads

Arc Compute begins every engagement by learning about your AI or HPC goals. We then design a tailored solution—whether you need short-term rentals for bursty workloads or a long-term GPU cluster for continuous training.

Why This Matters

Our hands-on, consultative model ensures you get the best mix of performance, cost efficiency, and scalability—without juggling multiple vendors or misaligned solutions.

Assessment & Strategy

Technical consultations to pinpoint the ideal solution (cloud, on-prem, or colocation)

Procurement & Logistics

We source and deliver NVIDIA GPU servers, handling all details with OEMs and other vendors

Deployment & Management

Installation in your data center or a colocation facility, plus ongoing managed services as needed

Our Process

Architecture Consultation:

Every GPU deployment starts with understanding your specific needs.

- We evaluate your workload types (training vs. inference), expected user demand, and growth projections.
- Our team proposes a modular architecture that supports horizontal and vertical scaling.
- We identify optimal server topology, multi-rack configurations, and tenant isolation models.
- Networking, storage, and virtualization strategies are considered early, ensuring you build infrastructure that grows without performance bottlenecks.
- The result is a blueprint that balances cost-efficiency, performance, and future growth.

Our Process

Hardware & Infrastructure Planning:

Once the architecture is defined, we design the infrastructure stack — the physical backbone of your deployment.

- We recommend the right GPUs (H100, H200, B100, MI300X, etc.) based on your performance and density needs.
- Our engineers design storage layers to match throughput demands.
- We spec InfiniBand or Ethernet for low-latency, high-bandwidth interconnects.
- Power, cooling, and rack planning is handled to align with data center constraints or colocation specs.
- Everything is optimized to ensure maximum GPU utilization and minimal data bottlenecks.

Our Process

GPU Procurement & Deployment:

Arc Compute handles the full lifecycle of sourcing and deploying hardware — saving you months of delays and reducing risk.

- We procure systems from leading OEMs (e.g., Supermicro, HPE, Dell, Aivres), with access to early allocation and preferred pricing.
- We manage vendor negotiations, compatibility validation, and lead time coordination.
- Hardware is delivered directly to your chosen data center or colocation partner.
- Our team oversees rack & stack, network cabling, BIOS/firmware tuning, and system validation.
- You get ready-to-run infrastructure configured for your exact use case.

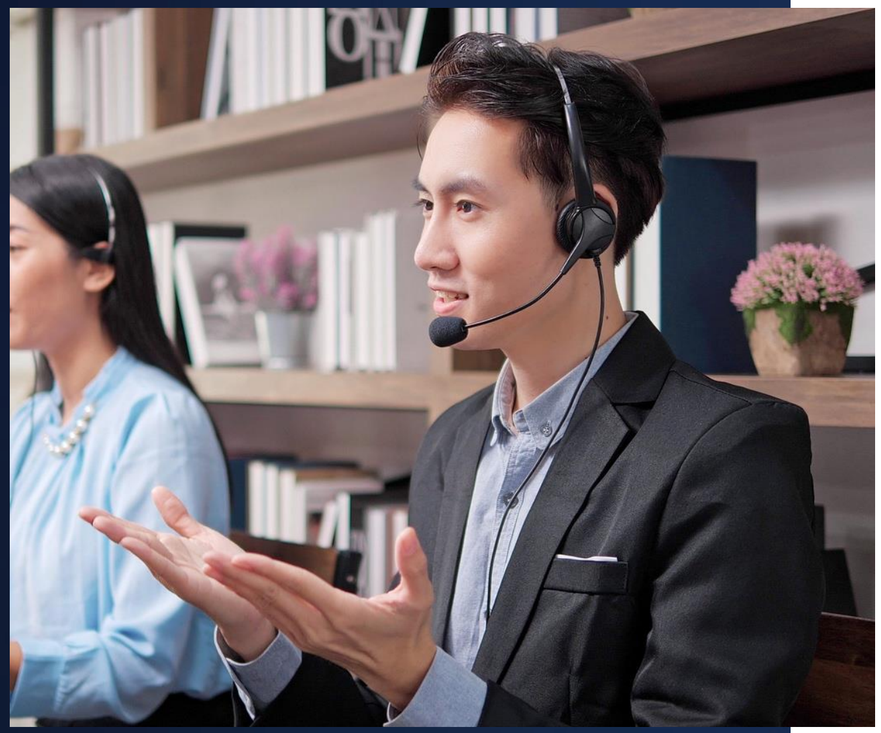
Our Process

Scaling & Optimization:

Your GPU infrastructure isn’t static — it evolves as your business grows. Arc Compute helps you scale smoothly while maintaining performance.

- We assist in capacity planning, helping forecast when to expand and how much.
- Add new nodes or racks without disrupting existing users or workflows.
- Performance tuning includes NUMA configuration, PCIe load balancing, GPU driver stack optimization, and network QoS policies.
- We can monitor cluster efficiency and thermal performance, identifying optimization opportunities.
- This means lower operational costs and a better user experience as you grow.

Our Process



Ongoing Support & Expansion:

Post-deployment, Arc Compute doesn’t disappear — we remain your infrastructure partner.

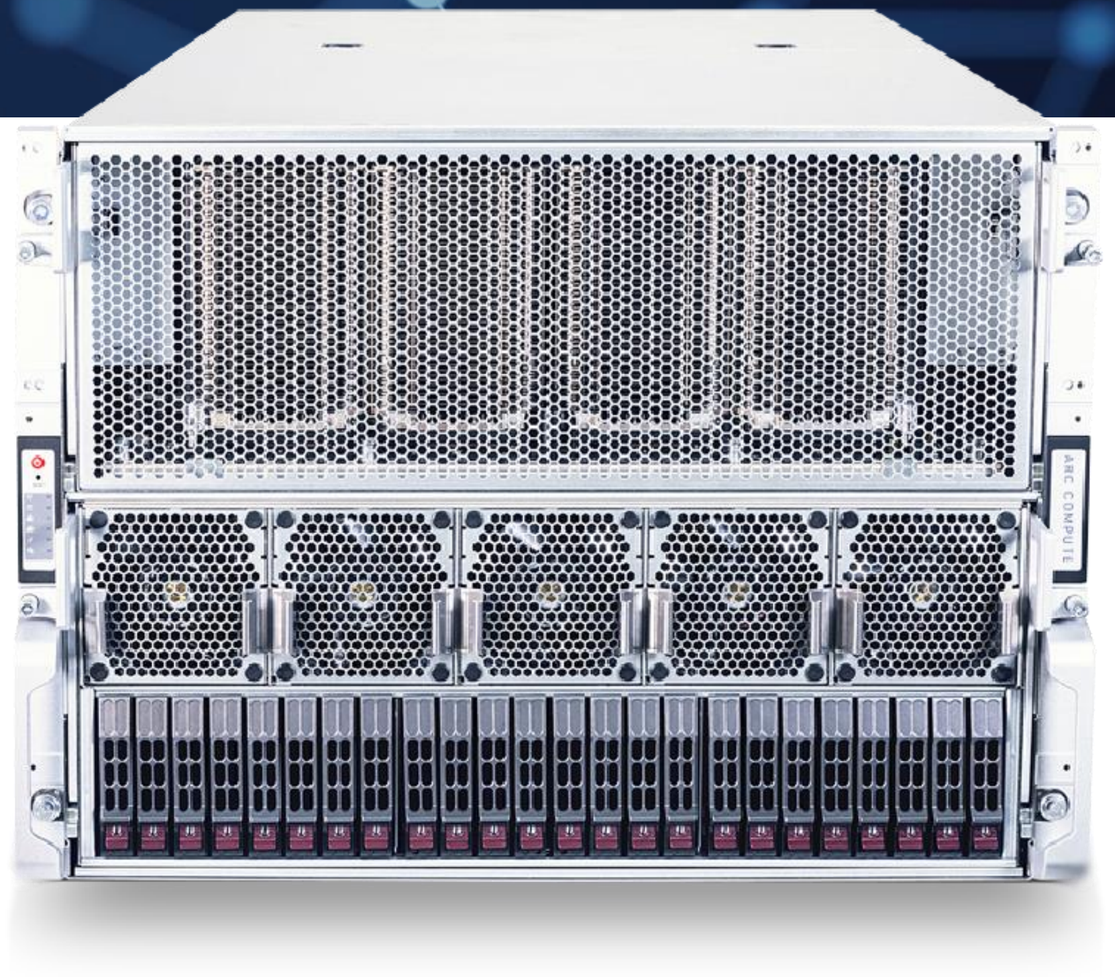
- 24/7 support options for hardware issues, system health, and uptime monitoring.
- Direct line to OEMs for fast RMA, replacement parts, and firmware updates.
- Assistance with rolling upgrades to newer GPUs, memory, or storage as models evolve.
- Help desk integration and optional remote hands support at your colocation facility.
- Guidance on new workloads (e.g., LLM fine-tuning, inference serving) as they arise.
- You get long-term stability, peace of mind, and an expert team that grows with your infrastructure.

Our Core Solutions

NVIDIA HGX Systems

Arc Compute specializes in deploying NVIDIA HGX systems, featuring the latest H100, H200, or B200 GPUs, along with the new GB200 NVL72.

These platforms deliver industry-leading performance for training advanced neural networks, running complex simulations, and processing large datasets.



Our Core Solutions

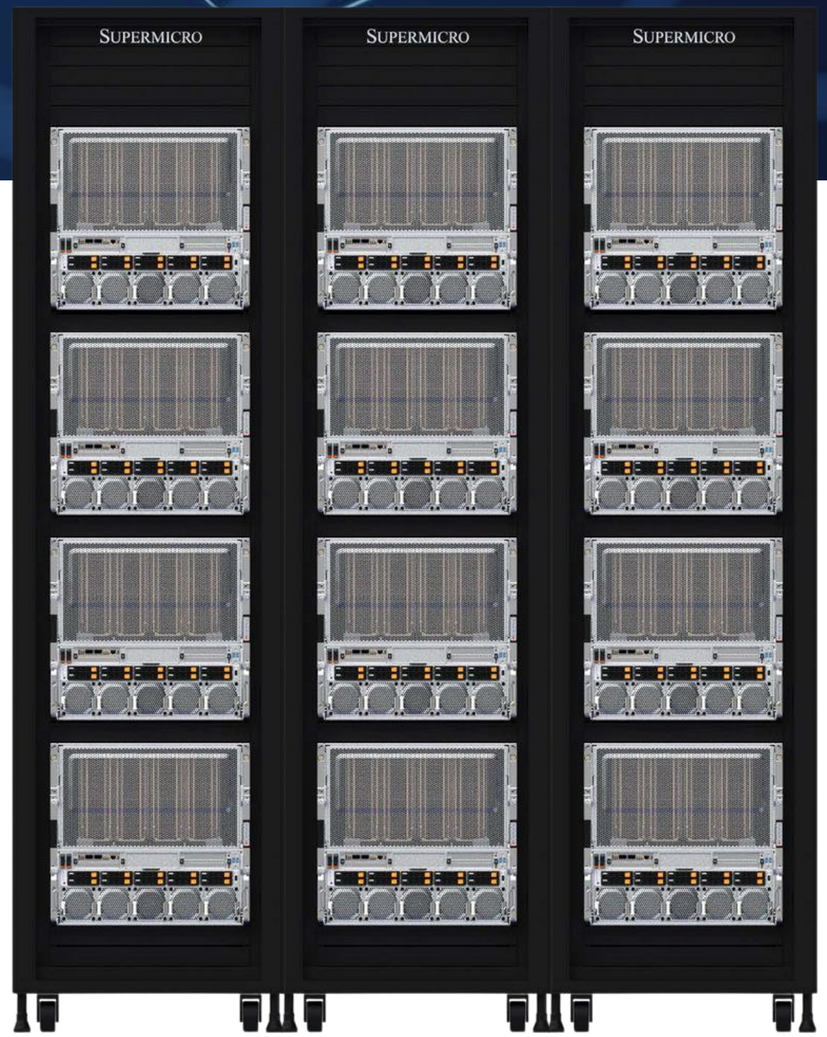
Key Benefits of HGX

- **Scalability:** Each 4/8-GPU node scales linearly for data-intensive tasks.
- **Efficiency:** High-bandwidth interconnects and large GPU memory ensure maximum throughput.
- **Future-Proofing:** Designed to meet the growing demands of large language models (LLMs) and next-gen AI research.

Our Core Solutions

Streamlined Training & Inference Clusters

- Arc Compute offers pre-configured GPU clusters built on NVIDIA-validated reference architectures. These turnkey designs simplify deployment and ensure optimal performance for AI workloads—from quick POCs to enterprise-scale production.



Our Core Solutions

Streamlined Training & Inference Clusters

- **Rapid Deployment:** Clusters with tested networking, storage, and software stacks to reduce complexity.
- **Scalable Performance:** Start with a single 8-GPU node and seamlessly scale to multi-rack clusters as demands grow.
- **Customizable Configurations:** Custom-built for training or inference workloads.

Our Customers

Trusted Across Industries

We've helped startups and enterprises deploy their own GPU infrastructure—quickly, cost-effectively, and with full control, and our clients stay with us because we deliver — from first rack to full-scale rollout.

"Arc Compute helped us stand up our infrastructure faster than we thought possible and support has been outstanding from day one."

Alex Smola (CEO of Boson AI)

Boson AI

OpenCall.ai

QIM
QUANTITATIVE
INVESTMENT
MANAGEMENT

xingtera™

LeddarTech®

Next Steps with Arc Compute

- **Initial Consultation**
  - Schedule a brief call to discuss your specific hardware requirements, current infrastructure, and deployment objectives.

- **Customized Procurement & Deployment Strategy**
  - Our team will outline a clear, step-by-step plan covering vendor selection, logistics, and on-site or remote deployment options.

- **Agreement & Coordination**
  - Once you approve the strategy, we'll finalize the necessary agreements, coordinate with suppliers, and manage the procurement process on your behalf.

- **Deployment & Ongoing Support**
  - We'll oversee the procurement and delivery of your new hardware, ensuring a seamless transition. Post-deployment, Arc remains available for maintenance, troubleshooting, and future enhancements as your needs evolve.

Arc Compute is your long-term partner in building scalable, high-performance GPU infrastructure.

Contact Us

Address

31 Scarsdale Rd, Unit 4, North York, ON, Canada

Website

www.arccompute.io

Email Address

info@arccompute.io

ARC COMPUTE

NVIDIA Elite Partner | SUPERMICRO | Hewlett Packard Enterprise | DELL | AIVRES | Authorized Partner