CASE STUDY: BOSON AI

Accelerating LLM Innovation with Arc

ARC COMPUTE

BosonAI

NVIDIA SUPERMICRO DELL AIVRES

CASE STUDY (1/10)

Accelerating LLM Innovation with Arc

OVERVIEW

Boson AI, an emerging leader in large language model (LLM) development, sought to build a high-performance compute cluster that could handle the rigorous demands of training and running next-generation AI models. Having previously relied exclusively on cloud GPUs, Boson AI needed a robust on-prem solution to control costs, accelerate innovation, and maintain complete visibility into their infrastructure. Arc Compute stepped in to architect, procure, and deliver a 65-node NVIDIA HGX H100 cluster—complete with InfiniBand networking—in record time, enabling Boson AI to reduce operating costs and expedite model development.

COMPANY BACKGROUND

About Boson AI

- **Industry Focus:** AI research, specifically large language models (LLMs) for a range of verticals.
- **Business Goal:** Develop novel LLMs that can be tailored to various industries, ensuring faster time to market and significant cost savings compared to cloud-based training.

Pre-Arc Infrastructure

- **Deployment Model:** Fully reliant on cloud GPUs for testing and initial development.
- **Key Pain Points:**
  - **High Costs:** Running large-scale model training in the cloud proved increasingly expensive.
  - **Limited Control:** Boson AI needed deeper insight into and control over hardware configurations to optimize training workflows.
  - **Scalability Concerns:** Cloud GPU availability and procurement complexity posed hurdles for large-scale expansions.

CASE STUDY (2/10)

Accelerating LLM Innovation with Arc

BosonAI

THE CHALLENGE

Key Business & Technical Requirements

- **Lower TCO:** Boson AI aimed to significantly reduce the total cost of ownership by shifting to on-prem servers without sacrificing performance.
- **Infrastructure Control:** Full visibility into hardware configurations and the ability to customize networking to meet HPC standards.
- **Performance & Scalability:** Deploy a cluster architecture capable of handling large-scale LLM training and inference.
- **Tight Timelines:** Boson AI had aggressive production deadlines, requiring rapid deployment and minimal downtime.

Unique Constraints & Compliance

- **Supply Chain Constraints:** Many vendors quoted lead times exceeding 12 weeks for high-demand H100 GPUs.
- **Regulatory & Reference Architecture:** Ensuring compliance with US regulations and adherence to NVIDIA reference architectures for optimal performance.

CASE STUDY (3/10)

Accelerating LLM Innovation with Arc

BosonAI

THE RIGHT CHOICE

Why Boson AI Chose Arc Compute

Originally, Boson AI planned to purchase 65 NVIDIA HGX H100 8-GPU systems through a well-established vendor. However, extended delays in delivery timelines jeopardized Boson AI's production schedules. In contrast, Arc Compute provided:

- **Accelerated Timelines:** A fast-tracked procurement process that delivered hardware in under four weeks.
- **Hands-On Concierge Service:** Arc Compute acted as an infrastructure concierge, overseeing every step, from vendor negotiations to manual component installations.
- **Cost Optimization:** Negotiated pricing on GPUs, high-speed networking, and server configurations to ensure the best ROI.
- **Technical Expertise:** Proficiency in HPC and AI infrastructure design, ensuring compliance and performance requirements were met.

CASE STUDY (4/10)

Accelerating LLM Innovation with Arc

BosonAI

TECHNICAL CONSIDERATIONS

Recommended Hardware Architecture

- **Core Cluster:**
  - 65× NVIDIA HGX H100 8-GPU systems from Supermicro (64 primary nodes + 1 redundancy node).
  - Quantim-2 400G InifiBand switches and CX-7 NICs.

- **Power & Cooling:**
  - A data center for colocation was selected that was capable of supporting the higher power draw and advanced cooling needs of dense GPU nodes.
  - Supermicro chassis was chosen for its energy-efficient design.

InfiniBand Challenges

- **NIC Availability:**
  - Boson AI required eight CX-7 NICs per system, totaling 520 NICs across the cluster.
  - Standard lead times from server OEMs were incompatible with Boson AI's deadlines.

- **Solution:**
  - Arc Compute sourced NICs from an alternative supplier at a lower cost and faster turnaround.
  - Arc engineers manually installed all NICs post-delivery at the data center.

CASE STUDY (5/10)

Accelerating LLM Innovation with Arc

BosonAI

PROCUREMENT & DEPLOYMENT PROCESS

Needs Assessment & Design

- **Requirement Gathering:** Arc Compute collaborated closely with Boson AI's executive and technical teams and NVIDIA to finalize performance metrics and cluster design.
- **Vendor Selection:** Arc leveraged industry partnerships to secure discounted pricing and confirm stock availability on short timelines.

Rapid Quoting & Purchasing

- **Fast Turnaround:** The cluster architecture was designed, quoted, and approved within two weeks.
- **Cost-Effective Approach:** Arc balanced Boson AI's need for immediate availability with the best possible hardware pricing, reducing overall TCO.

CASE STUDY (6/10)

Accelerating LLM Innovation with Arc

BosonAI

PROCUREMENT & DEPLOYMENT PROCESS

Delivery & Installation

- **Shortened Lead Times:** Despite global supply constraints, Arc fulfilled orders in under four weeks, drastically faster than the 12+ weeks quoted by other vendors.
- **Manual NIC Installation:** Arc's engineering team managed the physical installation of ConnectX-7 NICs in all 65 servers at the chosen data center.
- **Data Center Integration:** Arc worked with the facility to ensure racks, power, and cooling were up to HPC standards.

Validation & Go-Live

- **Initial Testing:** While Boson AI led final benchmarking, Arc Compute's validation team was on standby to troubleshoot issues.
- **The Transition from Cloud:** With the on-prem cluster operational, Boson AI migrated LLM training workloads off cloud GPUs.

CASE STUDY (7/10)

Accelerating LLM Innovation with Arc

BosonAI

IMPACT & RESULTS

Performance Gains

- **Reduced Training Times:** HGX H100's advanced tensor core and parallel processing capabilities drastically cut model training durations compared to cloud instances.
- **Improved Throughput:** The InfiniBand networking backbone delivered high-speed interconnects, ensuring minimal latency across the cluster.

Cost Savings

- **Significant TCO Reduction:** Moving from cloud GPUs to on-prem lowered monthly expenses and eliminated unpredictable cloud billing.
- **Optimized Power Usage:** Supermicro's energy-efficient hardware and a carefully selected data center partner led to sustainable operational costs.

CASE STUDY (8/10)

Accelerating LLM Innovation with Arc

BosonAI

IMPACT & RESULTS

Faster Time to Market

- **Rapid Model Development:** With dedicated high-performance nodes, Boson AI could iterate on model architectures more frequently, accelerating innovation.
- **Revenue Realization:** Quicker model training cycles enabled Boson AI to launch new LLM-based products faster, positively impacting the bottom line.

Ongoing Collaboration & Support

- **Regular Check-Ins:** Arc Compute maintains a proactive support schedule, holding regular calls to ensure the cluster continues to meet performance SLAs.
- **Knowledge Transfer:** Dedicated Slack channels and direct collaborations with Boson AI's data scientists and IT staff foster a seamless operational handover.
- **Future Deployments:** Given the success of this initial project, Boson AI plans to engage Arc Compute for additional upgrades and expansions as its LLM workloads grow.

CASE STUDY (9/10)

Accelerating LLM Innovation with Arc

BosonAI

TAKEAWAYS

Lessons Learned & Future Outlook

- **Partner Selection is Critical**
  - Choosing an infrastructure provider who understands HPC and AI hardware can make or break a deployment. Arc Compute's concierge-style approach simplified complex networking, servers, and compliance decisions.
- **Supply Chain Flexibility**
  - During times of global GPU shortages, having an agile partner who can source critical components from alternate channels is invaluable.
- **Reference Architectures Matter**
  - Aligning closely with NVIDIA's recommended designs ensures top performance and a smoother deployment experience.
- **Scalability Planning**
  - Given the explosive growth of AI/LLM workloads, designing with future expansions in mind helps avoid expensive rework down the line.

CASE STUDY (10/10)

Accelerating LLM Innovation with Arc

BosonAI

CONCLUSION

Future Deployments

- Boson AI, encouraged by the successful HPC deployment, is already exploring further expansions. Thanks to Arc Compute's ability to handle end-to-end procurement and implementation, Boson AI confidently plans to leverage Arc's expertise again for upcoming large-scale HPC projects.

By partnering with Arc Compute, Boson AI overcame the twin challenges of high cloud costs and constrained hardware supply to deploy a 65-node H100 GPU cluster quickly. The move to on-prem infrastructure delivered both immediate and long-term benefits: significantly lowered TCO, faster development cycles, and a platform that can scale for next-generation LLM workloads. This case study underscores the importance of finding an experienced partner able to navigate the complexities of HPC deployments—from hardware sourcing to compliance—to unlock top-tier AI performance and innovation.

Contact Us

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

Website
www.arccompute.io

Email Address
info@arccompute.io

ARC COMPUTE

About Arc Compute

Arc Compute specializes in designing, procuring, and implementing state-of-the-art AI and HPC infrastructure. Through strategic industry partnerships, deep technical knowledge, and a hands-on procurement and deployment model, Arc Compute offers end-to-end solutions that accelerate enterprise AI initiatives while optimizing total cost of ownership.

NVIDIA SUPERMICRO DELL AIVRES