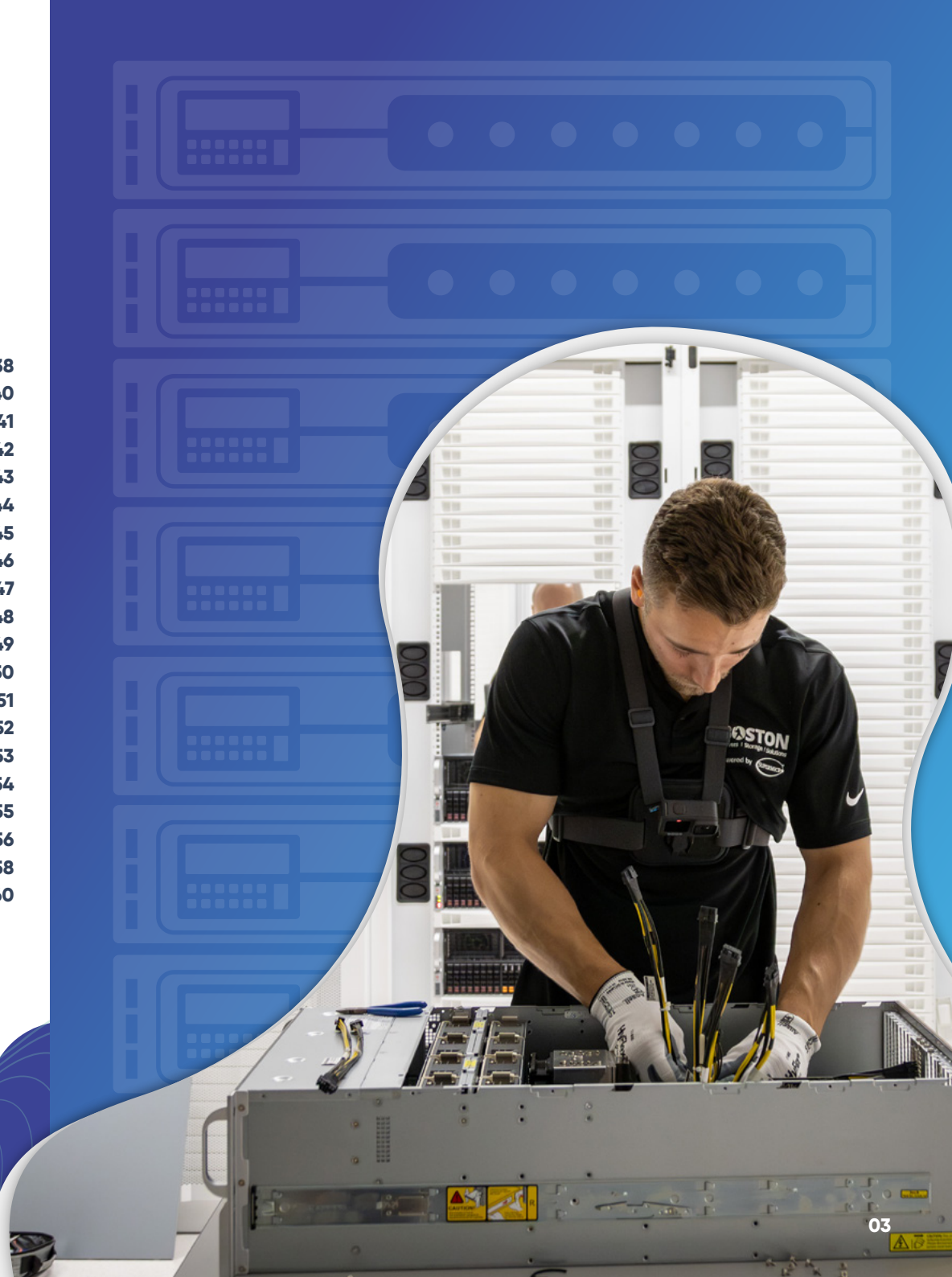
BOSTON
Servers | Storage | Solutions
A DIVISION OF SOURCECODE

HIGH PERFORMANCE COMPUTING

HPC

CONTENTS

WHAT'S INSIDE?



WHAT SETS US APART?

INNOVATIVE SOLUTIONS

We innovate tech solutions for diverse sectors, adapting to evolving business needs

BESPOKE CUSTOMISATION

Tailored, bespoke solutions harmonise with customer objectives, defining our excellence in technological delivery

GLOBAL REACH

Our global presence assures swift support, regardless of location, backed by a history of successful collaborations

TRACK RECORD

A proven track record with diverse projects instils confidence in us as a dependable partner

MANUFACTURING AND TESTING

Our in-house design, manufacturing and testing in Boston Labs guarantees quality and quick turnaround

RESEARCH AND DEVELOPMENT

Ongoing R&D equips us with expertise for the latest tech advancements, providing customers a competitive edge

SUSTAINABILITY

We are environmentally conscious and prioritise eco-friendly, sustainable solutions

ONE-STOP SHOP

Our diverse tech offerings, from computing to AI, position us as a one-stop shop for all your needs

CUSTOMER-CENTRIC

Our customers are our cornerstone, understanding their challenges and exceeding expectations in the solutions we create

BOSTON: POWERING INNOVATION FOR OVER 30 YEARS

WHO ARE WE?

Boston Limited has spent over three decades redefining what's possible in high-performance computing. From our early days pioneering bespoke server solutions to becoming a global leader in AI, cloud and advanced computing, our journey has been shaped by a relentless drive to push technology forward.

We design tailor-made HPC clusters, delivering end-to-end solutions for complex HPC challenges, including computational fluid dynamics (CFD), molecular dynamics, molecular biology and particle physics amongst many others.

At Boston, we don't just build systems—we engineer solutions tailored to the demands of the most complex and performance-hungry workloads. Whether it's AI and machine learning, high-frequency trading, cloud infrastructure or large-scale storage, our platforms are designed for efficiency, reliability and uncompromising performance.

But what truly sets us apart is our approach; every solution we create is backed by expert integration, rigorous testing and deep technical insight. We offer fully branded, custom-built platforms, global lifecycle management and hands-on support that extends far beyond deployment. From training and advisory services to full-scale implementation, we help organisations harness the power of HPC, automation and data to drive real results.

With demand for high-performance technology accelerating, Boston continues to expand—bringing our expertise to more industries, more markets and more innovators who refuse to settle for off-the-shelf solutions.

If performance, precision and partnership matter to you, Boston is ready to power your success.

HPC CLUSTER ARCHITECTURE AND WORKFLOW

HPC clusters are enabling breakthroughs in AI, scientific research, financial modelling and beyond. But building a truly optimised cluster isn't just about stacking servers—it's about engineering an ecosystem where compute, networking, storage and software work seamlessly together.

At the heart of every high-performance cluster is a carefully balanced architecture. CPUs handle serial workloads and orchestrate processes, while GPUs accelerate parallel computations, delivering the raw power needed for AI training, simulations and complex modelling. High-speed networking ensures that data moves swiftly between nodes, preventing bottlenecks and enabling real-time processing. Storage solutions—whether high-capacity disk arrays or lightning-fast NVMe drives—are selected to match the needs of data-intensive applications. And on top of it all, the right software stack ensures that every component operates in perfect unison, from job scheduling to monitoring and optimisation.

Designing and deploying such an ecosystem requires expertise. Poorly planned clusters can lead to inefficiencies, underutilised resources and spiralling costs. IT teams often struggle with incompatible software, networking misconfigurations and systems that fail to scale effectively. That's where Boston comes in.

> Boston understands the end-to-end process of designing, implementing and managing clusters—ensuring your project is delivered on time and within budget.

With decades of experience, Boston takes the complexity out of cluster deployment. Our team pre-engineers, tests and optimises complete HPC solutions in our state-of-the-art Boston Labs, ensuring seamless integration of hardware and software. Whether leveraging open-source or commercial cluster management tools, we tailor every system to match our customers' environments, workflows and performance goals.

The result? Clusters that are not just powerful, but also easy to deploy, manage and scale. With Boston, you don't just get a collection of servers—you get a finely tuned, high-performance computing ecosystem, ready to drive innovation from day one.

BOSTON HPC SOLUTION STACK SELECTOR

BOSTON SERVICES	Solution Design & Consultancy	Installation	Warranty & Support	Maintenance & Managed Services
SOFTWARE	OpenFOAM	STAR-CCM+	ANSYS Fluent	LS-DYNA
INFRASTRUCTURE MANAGEMENT	NVIDIA Base Command	ClusterVision TrinityX	Supermicro DCMS	OpenHPC
JOB SCHEDULERS	Slurm	Altair Grid Engine	Altair PBS Professional	IBM Spectrum LSF
CONTAINERS	Apptainer	Kubernetes	OpenShift	Rancher
COMPUTE	AMD 5th Gen EPYC™	Intel® Xeon® 6	NVIDIA Grace	AmpereOne
ACCELERATED SOLUTIONS	AMD Instinct™ MI300 Series	Intel® Gaudi® 3	NVIDIA Blackwell Architecture	NVIDIA Ada Lovelace Architecture
NETWORKING	NVIDIA Spectrum-X	NVIDIA Quantum-2 InfiniBand	NVIDIA Spectrum-4 Ethernet	SONiC Ethernet
STORAGE	Boston Igloo AI+	VAST	WEKA	DDN
COOLING	Direct Liquid Cooling	Liquid Immersion	Rear Door Heat Exchanger (RDHx)	Air Cooling
DATA CENTRE SERVICES	Colocation	Self-Hosted	Managed Services	Modular

COMPUTE SOLUTIONS

Supermicro's compute solutions deliver high-density, multi-node server systems designed for scalability, reliability and energy efficiency. From rackmount to blade servers, they provide exceptional performance for HPC, data centres and cloud computing.

AMD's compute solutions excel in HPC, AI and enterprise workloads. EPYC processors power advanced HPC clusters, while AMD's high-density, multi-node solutions offer scalability and efficiency for demanding applications.

Ampere empowers enterprise computing with their high-performance AmpereOne processors, offering exceptional core counts, clock speeds and memory capacity. Built on ARM architecture, Ampere's solutions deliver power-efficient, scalable performance tailored for cloud-native and high-performance workloads.

The Intel® Xeon® 6 processors deliver high performance and energy efficiency, making them ideal for HPC, cloud-native applications and high-density deployments. Designed for demanding workloads, they offer exceptional compute power for AI, scientific research and enterprise applications.

The NVIDIA Grace CPU is purpose-built for AI supercomputing, HPC and large-scale data processing, delivering high memory bandwidth and energy efficiency for demanding applications. Its advanced memory integration with LPDDR5x makes it ideal for complex simulations, AI model training and real-time analytics.

SUPERMICRO COMPUTE SOLUTIONS FOR HPC

Supermicro and Boston have been delivering HPC solutions for over three decades, powering some of the world's most demanding workloads. Supermicro's high-density, multi-node platforms are engineered for maximum efficiency; a perfect fit for HPC, AI and scientific research. Their liquid cooling solutions further enhance performance by reducing energy consumption and improving thermal management—critical for sustainable, large-scale computing.

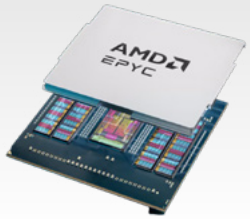
Their reference designs for enterprise HPC, AI and deep learning provide a strong foundation for customised solutions, allowing organisations to fine-tune their infrastructure to the minutest of detail. Boston trusts on Supermicro as a primary building block for our HPC solutions, ensuring customers benefit from optimised performance, energy efficiency and seamless scalability.

Supermicro systems have been deployed in HPC clusters worldwide, including CERN, where they contribute to groundbreaking scientific discoveries. A significant portion of data centres globally also leverage Supermicro's technology due to its reliability, density and competitive performance-to-price ratio.

From computational fluid dynamics (CFD) and finite element analysis (FEA) to weather prediction, oil and gas research, automotive, aerospace and naval engineering, Supermicro's solutions are trusted across industries to accelerate workloads while minimising operational costs.

HPC requires immense computing power, but efficiency is key. Supermicro champions greener computing with power-efficient configurations and liquid cooling technology, helping businesses reduce total cost of ownership while advancing their sustainability goals.

AMD EPYC™ 9005 SERIES PROCESSORS



AMD has taken you on a journey through Naples, Rome, Milan and Genoa. Now, the next stop is Turin—codename for AMD EPYC™ 9005 Series processors. The EPYC family has consistently delivered exceptional performance, efficiency and security, and the EPYC 9005 Series processors continues this legacy. With architectural advancements and cutting-edge memory and I/O capabilities, it redefines compute power for modern workloads.

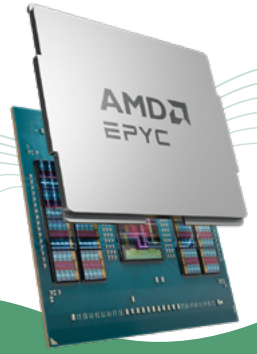
AMD EPYC 9005 Series processors features up to 192 “Zen 5” Cores, delivering industry-leading core density. With DDR5 memory support and 12 channels, it provides immense bandwidth and capacity for data-intensive applications. PCIe® Gen 5 ensures high-speed connectivity, making it the foundation of next-generation enterprise computing. All this translates to lower total cost of ownership (TCO) and improved energy efficiency, making it a smart investment for data centres.

KEY FEATURES:

- Up to 192 “Zen 5” Cores
- Up to 6TB of DDR5-6000 in 12 memory channels
- 128 PCIe® Gen 5 I/O lanes (up to 160 in 2-socket servers)
- CXL 2.0 support extends memory capacity and improving efficiency for large-scale simulations

BENEFITS:

- Maximum performance per socket
- Lower TCO and energy consumption
- Exceptional performance per watt, driving higher efficiency and faster time-to-insight



AMPEREONE

The AmpereOne™ family delivers high-efficiency compute with a scalable architecture optimised for cloud, HPC and edge workloads. Designed for energy efficiency and performance per watt, AmpereOne™ processors provide a balanced memory and compute structure, ensuring seamless execution of demanding applications.

With high core counts, PCIe® Gen 5 connectivity and efficient power management, AmpereOne™ processors are ideal for Telco Edge, Enterprise and HPC environments, offering a sustainable solution without compromising performance.

KEY FEATURES:

- Optimised for enterprise, HPC and telco edge workloads, delivering scalable performance across diverse environments
- Up to 192 Cores, enabling high-performance parallel computing for data-intensive applications
- Clock speeds up to 3.7 GHz for swift and responsive processing

BENEFITS:

- Configurable TDP (250W to 400W) balancing power efficiency and performance
- PCIe® Gen 5 support, providing high-speed connectivity for next-generation accelerators and storage solutions
- Based on Arm8.6+ architecture delivering highest performance per watt.

THE INTEL® XEON® 6 PROCESSOR FAMILY

The Intel® Xeon® 6 processor family is designed to power HPC, cloud and enterprise workloads with a scalable and efficient architecture. Featuring Performance Cores (P-Cores) for maximum compute power and Efficient Cores (E-Cores) for energy-optimised scaling, the Xeon 6 family delivers tailored solutions for demanding workloads.

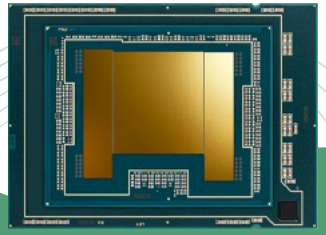
For HPC environments, the Xeon 6 6900P (Advanced Platform - AP) and 6700P (Scalable Platform - SP) processors provide the highest performing cores, I/O capabilities and memory bandwidth to accelerate complex computations. The Xeon 6 6900E and 6700E processors prioritise energy efficiency, making them ideal for high-scale cloud workloads where power consumption is a key factor.

KEY FEATURES:

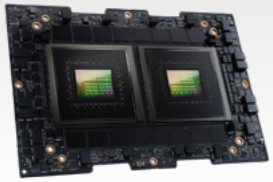
- **High Core Count:** Offers a large number of Cores per socket, allowing for parallel processing of complex tasks
- **P-Cores (Performance Cores):** Dedicated high-performance Cores for computationally intensive workloads
- **Advanced Matrix and Vector Engines:** Optimised for efficient matrix operations, crucial for AI and scientific calculations

BENEFITS:

- **High Memory Bandwidth:** Enables fast data transfer between the processor and memory, important for large datasets
- **AI Acceleration:** Integrated AI engines to accelerate machine learning and deep learning tasks
- **Scalability:** Designed to be easily scaled to meet the demands of large HPC clusters



NVIDIA GRACE



Designed to tackle the growing demands of AI, HPC and large-scale data analytics, NVIDIA Grace delivers exceptional speed, efficiency and scalability. With a purpose-built architecture, Grace unlocks new levels of performance while reducing power consumption—making it ideal for next-generation data centres.

Grace is available as a standalone CPU or as part of a hybrid CPU-GPU solution for AI-driven applications. The Grace Hopper Superchip pairs the Grace CPU with an NVIDIA Hopper GPU, optimised for AI training and inference, while Grace Blackwell integrates a Grace CPU with NVIDIA's Blackwell GPU architecture, delivering high-performance computing for large-scale workloads. Engineered for seamless integration, Grace leverages Arm Neoverse V2 Cores and NVIDIA's high-bandwidth interconnect technology, ensuring streamlined data transfer and accelerated processing for even the most complex workloads.

KEY FEATURES:

- **144 Arm Neoverse V2 Cores:** Ensures high single-threaded and multi-threaded performance for demanding workloads
- **Up to 1024 GB/s Memory Bandwidth:** Utilises LPDDR5X memory with ECC to provide fast, reliable data access
- **NVLink-C2C Interconnect:** Delivers high-speed, low-latency connectivity between Grace CPUs and NVIDIA GPUs for unified performance

BENEFITS:

- **Optimised AI and HPC Workloads:** Powers complex AI models, data analysis and simulations with industry-leading speed
- **Energy-Efficient Design:** Offers up to 2x the energy efficiency of traditional x86 systems, reducing overall operational costs
- **Scalability for Data Centres:** Supports flexible configurations to accommodate evolving data centre needs

ACCELERATED SOLUTIONS

AMD

AMD's MI300A is a versatile APU that integrates Zen 4 CPU Cores and CDNA3 GPU architecture on a single package, designed for HPC and AI workloads with a focus on energy efficiency and reduced latency.

AMD

AMD's MI325X is a datacentre GPU designed for high-performance AI and machine learning workloads, offering advanced memory bandwidth and compute capabilities optimised for LLMs and AI.

intel

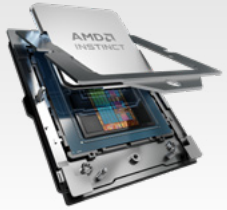
The Intel® Gaudi® 3 AI Accelerator is engineered to deliver high-performance, cost-effective scalability for complex AI deployments, supporting a wide range of industry AI models and frameworks.

NVIDIA

The NVIDIA B200 GPU is built to provide powerful AI and deep learning performance for edge and data centre environments, enabling real-time processing and low-latency inferencing. Its efficient design makes it a versatile solution for industries demanding rapid, scalable AI deployment.

AMD

AMD INSTINCT™ MI300A SERIES PLATFORM



The AMD Instinct™ MI300A is a powerful accelerated processing unit (APU) designed to handle demanding server and supercomputer workloads, including cutting-edge AI, machine learning and large-scale simulations.

Built with advanced CDNA3 architecture and Zen 4 CPU Cores, and leveraging 5nm process technology, the MI300A provides exceptional performance for a variety of intensive computing tasks. Equipped with 128 GB of HBM3 memory and Infinity Fabric technology, the MI300A ensures reduced latency and faster data access for memory-intensive applications. This makes it a versatile solution for AI inferencing, HPC and hybrid workloads where both general-purpose and accelerated computing are essential.

KEY FEATURES:

- High memory capacity of 128 GB HBM3 and Infinity Fabric Link technology minimise latency, enabling quicker insights and faster decision-making
- Hybrid architecture with integrated Zen 4 CPUs and CDNA3 GPUs allows for high-efficiency mixed workloads in environments such as scientific simulations and AI inferencing
- Built-in enhanced security features provide additional safeguards for sensitive data and AI models, offering peace of mind for mission-critical applications

BENEFITS:

- Combines Zen 4 CPU Cores and CDNA3 GPU architecture to handle a wide variety of workloads, from AI and machine learning to traditional HPC tasks, providing flexibility in deployment
- Features robust security measures, including Secure Memory Encryption and AMD Infinity Guard, to safeguard intellectual property and sensitive data—critical for industries requiring high levels of data protection
- Delivers superior performance per watt, reducing operational costs and environmental impact, making it ideal for sustainable datacentre operations

AMD INSTINCT™ MI325X SERIES PLATFORM

The AMD Instinct™ MI325X is a cutting-edge data centre GPU engineered for the most demanding AI, deep learning and high-performance computing (HPC) applications. Built upon the advanced CDNA™ 3 architecture and fabricated using a state-of-the-art 5nm process, the MI325X sets new standards in GPU performance and efficiency.

The AMD Instinct™ MI325X accelerator features 155,648 stream processors and 2432 compute units, delivering up to 2.61 petaFLOPs of FP16 performance with sparsity, making it exceptionally well-suited for deep learning training and AI model inference. Additionally, it offers up to 5.22 petaFLOPS of FP8 performance with sparsity, providing versatility and power for a wide array of computational workloads.

KEY FEATURES:

- **Exceptional Performance:** With 19,456 stream processors and 304 compute units, the MI325X is designed for extreme performance, suitable for exascale computing and large AI models
- **Extensive Software Ecosystem:** The MI325X supports AMD ROCm™ and open software frameworks, enabling seamless integration with existing AI, HPC and deep learning workflows while optimising performance across a range of applications
- **Advanced Fabrication Process:** Utilising a 5nm lithography process, the MI325X offers higher transistor density, translating to improved performance and efficiency for intensive computational tasks

BENEFITS:

- **Massive Memory and Bandwidth:** With 256 GB of HBM3E memory and 6 TB/s bandwidth, the MI325X handles trillion-parameter AI models efficiently, minimising partitioning
- **Powerful Compute Performance:** Built on CDNA™ 3, it delivers up to FP8 5.22 PFLOPs and FP16 2.61 PFLOPS with sparsity for accelerated AI training and inference
- **Scalability and Seamless Integration:** Seven Infinity Fabric™ links ensure efficient multi-GPU connectivity

INTEL® GAUDI® 3 AI ACCELERATOR

Building upon the proven performance and efficiency of its predecessors, the Intel® Gaudi® 3 AI Accelerator is engineered to meet the escalating demands of large-scale AI applications. It delivers a significant leap in performance and productivity for AI training and inference on popular LLMs and multimodal models.

The Gaudi® 3 architecture is optimised for deep learning workloads, offering enhanced performance and scalability. This makes it an ideal choice for enterprises aiming to accelerate their AI initiatives.

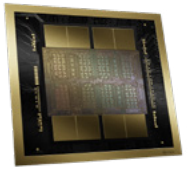
KEY FEATURES:

- **High-Efficiency Deep Learning:** Provides up to 1,835 TFLOPS of FP8 and BF16 compute performance, ensuring faster AI model training and inference
- **Massive Memory Bandwidth:** Equipped with 128 GB of High-Bandwidth Memory 2e (HBM2e) and a memory bandwidth of 3.7 TB/s to handle large, complex datasets
- **Integrated Networking:** Features 24 integrated 200-Gigabit Ethernet (GbE) ports, delivering an aggregate bandwidth of 4.8 TB/s in each direction for seamless, high-speed data transfer

BENEFITS:

- **Enhanced AI Training Speed:** The high compute power and memory bandwidth reduce training times for complex AI workloads, improving overall productivity
- **Scalable Architecture:** Supports large-scale deployments with efficient, industry-standard Ethernet networking, eliminating the need for proprietary solutions
- **Flexible Deployment Options:** Available in both custom Open Accelerator Module (OAM) and standard PCIe form factors, with power configurations of 900W and 600W, respectively

NVIDIA B200



The NVIDIA B200 GPU is engineered to deliver AI, HPC, and deep learning capabilities in a high-performance, scalable form factor. Designed for data centre and cloud AI workloads, the B200 accelerates complex computations while optimising energy efficiency. With up to 192 GB of HBM3e memory and an industry-leading 8 TB/s memory bandwidth, the B200 is built to handle the most demanding workloads, from large-scale AI training to high-precision scientific simulations.

A core component of the DGX B200 and HGX B200 platforms. The B200 enables seamless scalability and high-bandwidth interconnects for next-generation AI factories.

KEY FEATURES:

- **Massive Memory Capacity:** Equipped with up to 192 GB of HBM3e memory, the B200 ensures efficient handling of large datasets and complex models, essential for HPC and AI workloads
- **Unprecedented Memory Bandwidth:** With an industry-leading memory bandwidth of 8 TB/s, the B200 facilitates rapid data access and processing, minimising latency and enhancing overall system performance
- **Advanced Interconnectivity:** Featuring fifth-generation NVIDIA NVLink with 1.8 TB/s bandwidth and PCIe Gen6 support at 256 GB/s, the B200 enables seamless multi-GPU communication, crucial for scaling HPC applications

BENEFITS:

- **Tensor Core Performance:** TF32: 2.2 petaFLOPS for AI training and mixed-precision workloads. FP32: 80 teraFLOPS for scientific and enterprise computing. FP64: 40 teraFLOPS for double-precision HPC applications
- **Optimised Energy Efficiency:** With a maximum thermal design power (TDP) of 1,000W, the B200 balances high performance with energy efficiency, reducing operational costs in data centres
- **Scalability for Future Growth:** The B200's architecture supports the expansion of AI and HPC infrastructures, allowing organisations to scale their computational capabilities in line with evolving demands

RACK-SCALE SOLUTIONS

NVIDIA GB200 NVL72 SuperPOD is a cutting-edge, powerful computing infrastructure that harnesses the immense processing capabilities of NVIDIA GPUs to accelerate artificial intelligence and high-performance computing workloads.

The NVIDIA B200 GPU stands out as a compact powerhouse, engineered for edge computing and embedded AI applications. Its low-power architecture, makes it the ideal choice for environments where space and energy efficiency are paramount.

NVIDIA's Prestigious SuperPOD Specialization Programme

This achievement positions Boston Limited at the forefront of AI innovation, enabling customers to build their own AI Factories powered by NVIDIA DGX SuperPOD solutions.

NVIDIA Elite Partner

NVIDIA® GB200 NVL72 SUPERPOD

The NVIDIA GB200 NVL72 is a highly advanced rack-scale AI system, designed to deliver exascale computing within a single rack. It integrates 36 GB200 Grace Blackwell Superchips, interconnecting 72 Blackwell GPUs and 36 Grace CPUs using the NVIDIA NVLink Switch System. This architecture forms a 72-GPU NVLink domain that functions as a single massive GPU, achieving 30X faster real-time trillion-parameter LLM inference.

Built as a liquid-cooled, rack-scale solution, the GB200 NVL72 provides 130 terabytes per second (TB/s) of low-latency GPU communications, accelerating AI model training, inference and HPC workloads. Boston offers a range of GB200 NVL72 SuperPOD reference architectures, developed in collaboration with leading storage vendors such as DDN, IBM Spectrum Storage, VAST Data and WEKA. These fully integrated, ready-to-deploy solutions ensure best practices for scalability, efficiency and seamless AI infrastructure deployment.

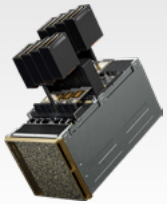
KEY FEATURES:

- **Extreme Density & Performance:** A fully liquid-cooled, rack-scale system packing 36 GB200 Superchips and 72 Blackwell GPUs, delivering up to 130 TB/s NVLink bandwidth for breakthrough AI and HPC workloads
- **Next-Gen Networking:** Equipped with Quantum-X800 InfiniBand, Spectrum-X800 Ethernet and BlueField-3 DPUs, ensuring ultra-low latency and high-bandwidth data transfer across the cluster
- **NVIDIA AI Ecosystem:** Comes with full access to NVIDIA AI Enterprise, white-glove support and NVIDIA's industry-leading software stack for seamless AI deployment and optimisation

BENEFITS:

- **Unmatched AI & HPC Acceleration:** 30X faster real-time trillion-parameter LLM inference and exascale computing power for next-gen AI training and scientific simulations
- **Seamless Scalability & Integration:** 5th Gen NVLink enables seamless GPU-to-GPU communication, creating a unified, high-performance compute fabric
- **Enterprise-Grade Reliability:** Backed by NVIDIA support, white-glove services and a robust software ecosystem, ensuring maximised uptime and streamlined deployment

NVIDIA® DGX SUPERPOD WITH DGX B200 SYSTEMS



The NVIDIA DGX SuperPOD, now featuring DGX B200 systems, offers a scalable AI infrastructure designed to meet the demands of large-scale AI and high-performance computing workloads. Each DGX B200 system is equipped with eight NVIDIA Blackwell GPUs, delivering exceptional performance and efficiency.

Built for maximum performance, the DGX B200 delivers up to 72 petaFLOPS for training and 144 petaFLOPS for inference, supported by 1,440GB of GPU memory. With 5th Gen NVLink providing 1.8 TB/s bandwidth per system and PCIe Gen6 at 256 GB/s, the DGX SuperPOD ensures rapid data transfer, optimising AI model training and inference.

KEY FEATURES:

- **High-Performance Computing:** DGX B200 systems deliver unparalleled computational power, enabling rapid training and inference for complex AI models
- **Scalable Architecture:** The SuperPOD design allows for seamless expansion, accommodating growing AI workloads and datasets
- **Integrated Networking:** High-speed interconnects ensure low-latency communication between nodes, optimising performance for distributed AI applications

BENEFITS:

- **Accelerated AI Development:** With the combined power of DGX B200 systems and the SuperPOD infrastructure, organisations can significantly reduce time-to-insight for AI projects
- **Simplified Deployment:** The turnkey nature of the DGX SuperPOD streamlines the implementation of AI infrastructure, allowing enterprises to focus on innovation rather than integration
- **Future-Proof Scalability:** Designed with growth in mind, the SuperPOD enables organisations to effortlessly scale their AI capabilities in response to evolving business needs

BOSTON JOINS NVIDIA'S SUPERPOD SPECIALIZATION PROGRAMME

1 of 3 SuperPOD Specialised Partners in EMEA

Access to the latest NVIDIA advanced technology

Expert guidance and tailored solutions

As our customer's trusted AI advisor, Boston's not only has a deep expertise in the design, consultancy, and delivery of SuperPOD systems but also in the identification and implementation of AI business solutions. With our understanding of identifying business use cases for and applying generative AI we are uniquely positioned to support your journey every step of your journey to achieve your goals.

This NVIDIA accredited certification means you can be assured that our team are fully equipped to guide you through the complexities of purchasing and implementing SuperPOD solutions. Ensuring that your infrastructure meets your specific needs and sets the stage for AI success.

AT THE HEART OF EVERY AI FACTORY: NVIDIA SUPERPOD

The NVIDIA SuperPOD serves as the powerhouse behind AI Factories, combining **massive computational power with cutting-edge technology** to handle the most demanding workloads. Key features include:

- **Next-generation NVIDIA Tensor Core GPUs**—scalable configurations delivering up to exascale-class AI performance.
- 400Gb/s NVIDIA Quantum-X InfiniBand & Spectrum-X Ethernet networking, ensuring **ultra-low latency and high-throughput AI training.**
- NVIDIA GPU Direct Storage, enabling rapid data ingestion, **seamless GPU-to-GPU communication and real-time analytics.**

This fully integrated solution accelerates every stage of the AI lifecycle, from data analysis and model training to deployment, empowering businesses to innovate faster and scale smarter.

START BUILDING YOUR AI FACTORY TODAY

Contact us to learn how NVIDIA SuperPOD solutions can transform your business.

E: sales@boston.co.uk **T:** +44 (0) 1727 876 100

STORAGE SOLUTIONS

BOSTON Servers | Storage | Solutions

Purpose-built for AI workloads, the Boston Igloo AI+ delivers high-performance storage at an accessible price point, making AI projects viable without compromise. With ultra-low latency, scalable architecture and seamless GPU integration, it ensures fast, efficient data access tailored for AI-driven enterprises.

ddn

Designed for large-scale AI and HPC environments, DDN Exascaler delivers high-bandwidth, low-latency storage with parallel file system technology. Trusted by leading research institutions and enterprises, it enables rapid AI model training, real-time analytics and seamless data movement across hybrid cloud infrastructures.

VAST

A revolutionary unified storage and compute platform, VAST Data eliminates storage silos by integrating high-performance flash storage with database and containerised computing. Ideal for multi-modal AI training and generative AI workloads, it provides limitless scalability with enterprise-grade efficiency.

WEKA

Engineered for data-heavy AI and ML applications, WEKA®FS™ delivers unprecedented storage throughput and IOPS, ensuring massive-scale AI training, deep learning and GPU-accelerated workloads operate without bottlenecks. With a cloud-native, software-defined architecture, WEKA maximises performance from edge to core to cloud.

BOSTON
Servers | Storage | Solutions

PEAK AIO

BOSTON IGLOO AI+

Investment in accelerator technology is critical to the success of an AI project. The Boston Igloo AI+, powered by PEAK:AIO, delivers exceptional performance to support AI workloads without compromising investment in accelerators. The NFS protocol enables multiple GPU-accelerated hosts or entire HPC clusters to access the storage simultaneously, ensuring seamless scalability as AI compute demands grow.

The Boston Igloo AI+ strikes the perfect balance between cost and high performance, making it an ideal solution for storage requirements in the hundreds of terabytes—ensuring accelerators are continuously fed with data for fast model development.

For smaller environments with a single HPC host, the Boston Igloo AI+ can be connected directly, eliminating the need for a switch. This allows for greater investment in additional compute resources, ensuring long-term scalability and efficiency.

KEY FEATURES:

- Block or file access
- Designed for AI workloads
- Up to 80GB/s+ throughput
- RAID 1, 10, 5 and 6 support
- GPU direct support
- Up to 737.3TB RAW

BENEFITS:

- Exceptional value for money
- Packaged appliance - Built, configured and remote installation service from Boston
- NFS protocol allows additional HPC client systems to be added over time
- Set it and forget it. No need for intensive administration

SPECIFICATIONS

MODEL	IGLOO AI+ 1U FORM FACTOR	IGLOO AI+ 2U FORM FACTOR
Drives	6 or 12x NVMe	6, 12 or 24x NVMe
CPU	(2x) Intel Gold 6442Y	(2x) Intel Gold 6442Y
Memory	256GB	256GB
Network Adapter	NVIDIA Connect-X 7 Dual Port	NVIDIA Connect-X 7 Dual Port
Installation	Local or remote	Local or remote
Capacity	Up to 368.6TB RAW	Up to 737.3TB RAW
Throughput	40GB/s+	80GB/s+

EXAMPLE USE CASE:

CHALLENGE

Many organisations face a choice when buying storage to support NVIDIA DGX technology. They wish to invest as much as possible into DGX only to find that little of the budget remains for fast parallel storage. Enterprise All Flash does not deliver enough performance, while a traditional parallel storage system is too expensive.

SOLUTION

The Boston Igloo AI+ meets the need by providing blistering storage performance at a price point that can support the limitations of a small budget. The NFS file system means that DGX systems can be added to storage over time while not compromising on performance needed to support the AI workloads.

ADVANTAGE

The Boston Igloo AI+ has been designed especially for this common use case. The low cost of this fast storage solution means that on many occasions budget can be directed into more GPUs.

DDN EXASCALER

For decades, DDN has designed, developed, deployed and optimised systems, software and storage solutions that enable enterprises, service providers, universities and government agencies to generate more value and to accelerate time to insight from their data and information on-premise and in the cloud.

Developed and optimised using the latest advances in file system software technology, the DDN EXAScaler storage appliance delivers extreme performance, scalability, capability, reliability and simplicity.

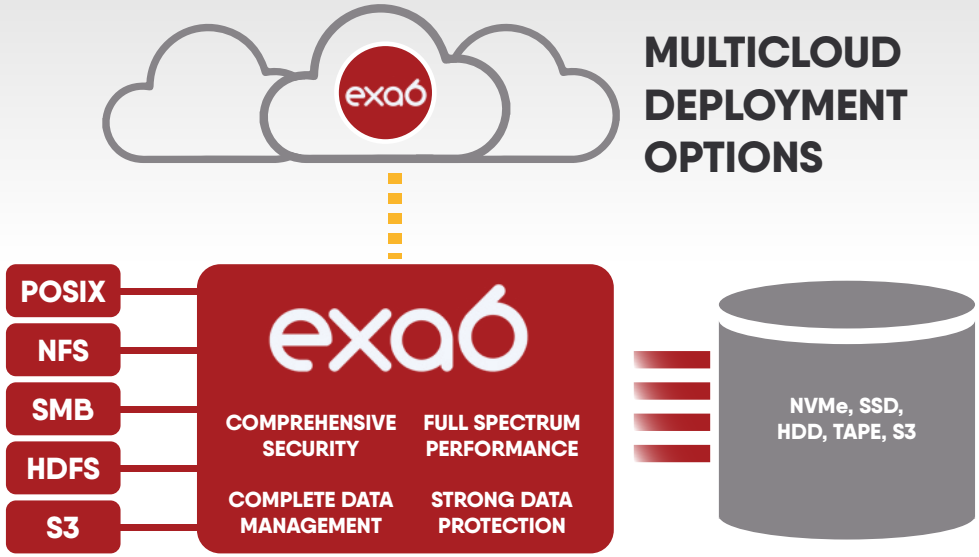
EXAScaler introduces several new data management and integrity file system features developed by DDN and is only available in its appliances and cloud offerings. Stratagem is a powerful data orchestration engine that gives users comprehensive data residency controls using policy-based placement.

KEY FEATURES:

- Enterprise storage features in an HPC capable array
- Simplify data management and orchestration with an API driven data integration tool
- Snapshots, Encryption, Data Integrity, Quotas, Cloud Workload analytics (SIA), MultiCloud readiness & Multitenancy

BENEFITS:

- A feature-rich and secure platform designed to support a wide range of use cases right out of the box
- File (NFS & SMB), Object (S3) and Container support for absolute flexibility
- Full spectrum performance using the right media for distinct data profiles



USE CASE:

CHALLENGE

A leading life sciences university required shared computational resources to accelerate genomic medicine research while staying within strict budget constraints.

SOLUTION

Enhanced existing infrastructure with a cost-effective, high-performance computing solution, integrating 200G HDR InfiniBand, 40/10 Gbps Ethernet and a DDN Storage system to improve simulations, data analysis and collaboration.

IMPACT

Delivered up to 90 GB/sec performance per appliance, significantly boosting throughput, scalability and research efficiency, enabling faster data processing and improved collaboration.

THE VAST DATA PLATFORM

The VAST Data Platform is a breakthrough approach to data-intensive computing that serves as the comprehensive software infrastructure required to capture, catalogue, refine, enrich, and preserve data through real-time deep data analysis and deep learning. This system is designed to provide seamless and universal data access and computing from edge-to-cloud, all from a platform that is designed for enterprises and cloud service providers to deploy on the infrastructure of their choosing.

The system takes a new approach to marrying unstructured data with structured data with declarative functions and can store, process and distribute data as a global data-defined computing platform. The VAST Data Platform also takes a first-principles approach to simplifying the data experience, by introducing several new architecture conventions that break long-standing trade-offs in data-intensive computing.

KEY FEATURES:

- Scale to exabytes of low-cost flash. Shared NVMe accessible by every container
- High performance storage fabric. Support for datacentre scale ethernet or InfiniBand
- Flexible client connectivity. NFT, SMB, S3, GPUDirect and K8s CSI supported

BENEFITS:

- Multi-Protocol design enables businesses to eliminate islands of storage with a single platform
- Enable mission-critical and data-intensive enterprise production environments to consolidate workflows, unlocking fast access to all data
- VAST innovations combine to deliver compounded levels of infrastructure efficiencies that make it possible to build an always-on, ultra resilient flash cloud

USE CASE:

CHALLENGE

Having developed as a lead provider of biological information to biopharma companies, a premier research organisation sought to build, host and manage an AI platform for its customers and partners to drive biological scientific discovery. This was at a time when the need for performance never stood still due to the need to keep pace with the evolving complex calculations that the platform needed to deliver.

SOLUTION

Having evolved various options, the institution chose to deploy the VAST Data Platform. The decision was made due to the value of an all-flash feature rich VAST storage to complement the Boston designed managed service solution and to deliver the performance needed to support scientific discovery.

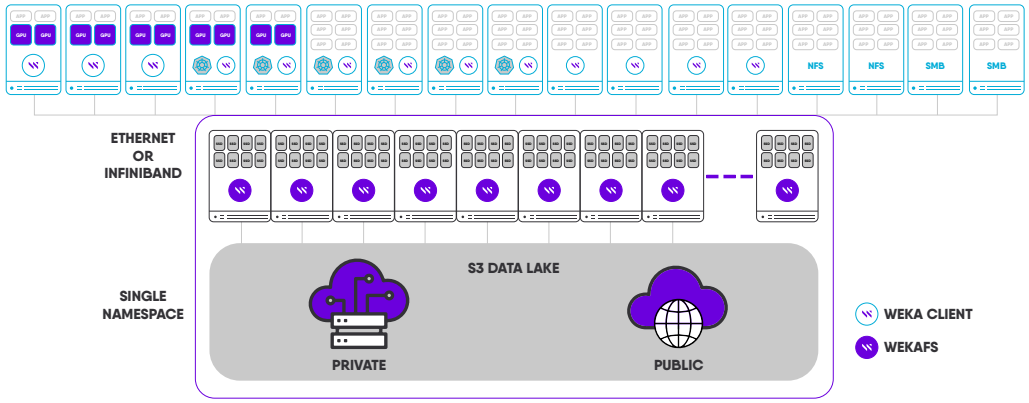
ADVANTAGE

The solution supports the need to address many use cases. The all-flash architecture will contribute to energy savings over the lifetime of the solution and the researchers and partners will be supported by a storage platform that will enable researchers to deliver results rapidly in a world where the storage will also need to support the complexity associated with evolution of more complex calculations.

WEKA IO

Built for NVMe flash and cloud-native, WEKA is a parallel file system designed to enable organisations to maximise the full value of their high-powered IT investments - compute, networking and storage. By leveraging existing technologies in new ways and augmenting them with engineering innovations, WEKA's software delivers a more powerful and simpler solution that would have traditionally required several disparate storage systems. The resulting software solution delivers high performance for all workloads (big and small files, reads and writes, random, sequential and metadata heavy).

WEKA's Data platform was built from the ground up to shatter the limits imposed by legacy storage architectures by taking a unique software-only approach to address the needs of modern workloads.



KEY FEATURES:

- Manage all your file and object data on NVMe, SSD flash or disk in the system within a single global namespace
- Management is easy with a simple GUI. Rapidly provision, manage policies, change data protection levels and monitor system health intuitively
- No tuning is required once set up. WEKAFS™ is ideal for the challenges of mixed workloads

BENEFITS:

- Manage all your data from one, centralised location
- Eliminate guesswork
- User friendly management

USE CASE:

CHALLENGE

Genomics England had previously implemented a scale-out NAS solution from a leading vendor to support the 100,000 genome project; however, it had already hit its limit on storage node scaling and performance suffered when the system was near capacity.

SOLUTION

A two-tier architecture that takes commodity flash and disk-based technologies, presenting it as a single hybrid storage solution. The primary tier consists of 1.3 Petabytes of high performing NVMe flash storage which supports the working data sets.

ADVANTAGE

Genomics England was able to realise no limit on capacity scaling, over 10x improvement in performance, 75% reduction in storage cost per genome, as well as now embedding a full disaster recovery strategy and offering integration with public cloud for compute elasticity.

LIQUID COOLING

Liquid cooling is rapidly becoming an inevitable approach to HPC cooling. Why is that? The short answer is that air cooling will not be able to keep pace with the wattage of components. The long answer is that on top of meeting temperature requirements, it also works towards sustainability goals as liquid cooling requires less energy to maintain, and generally there is less maintenance with a liquid cooled setup.

Using liquids to cool a system can come in many shapes and sizes. There are two categories which then also break into two types; single-phase and dual-phase liquid cooling. Each approach can be applied directly on the chip or by immersing components, or even the entire system.

DIFFERENT TYPES OF LIQUID COOLING

DIRECT LIQUID COOLING (DLC)

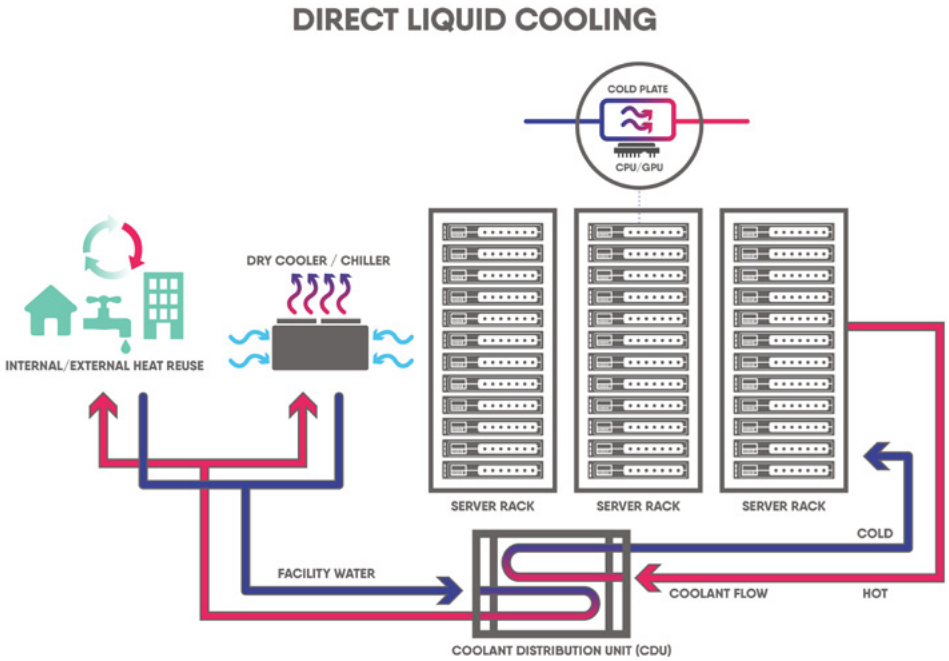
In the realm of liquid cooling, single-phase DLC maintains a consistent liquid state, efficiently absorbing heat from components and subsequently dissipating it through cooling mechanisms before being recirculated, finding applications in both direct-on-chip and immersion cooling setups.

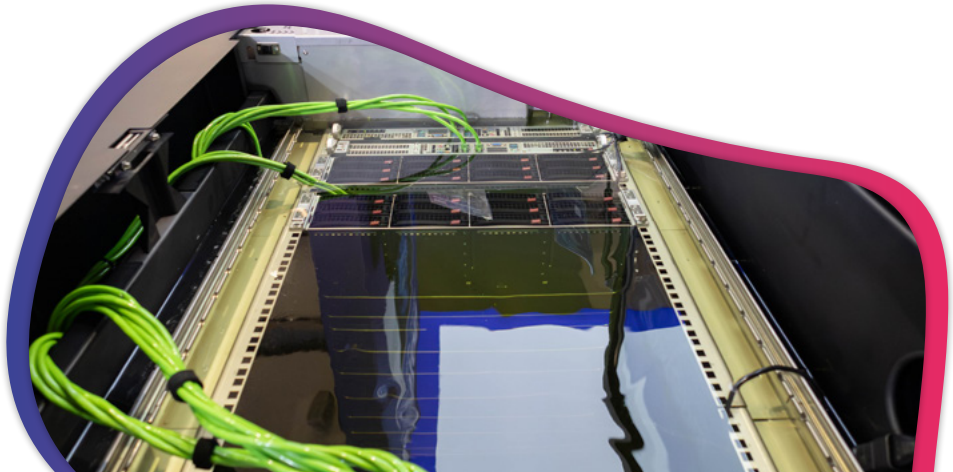
DUAL-PHASE DIRECT LIQUID COOLING (2-PHASE DLC)

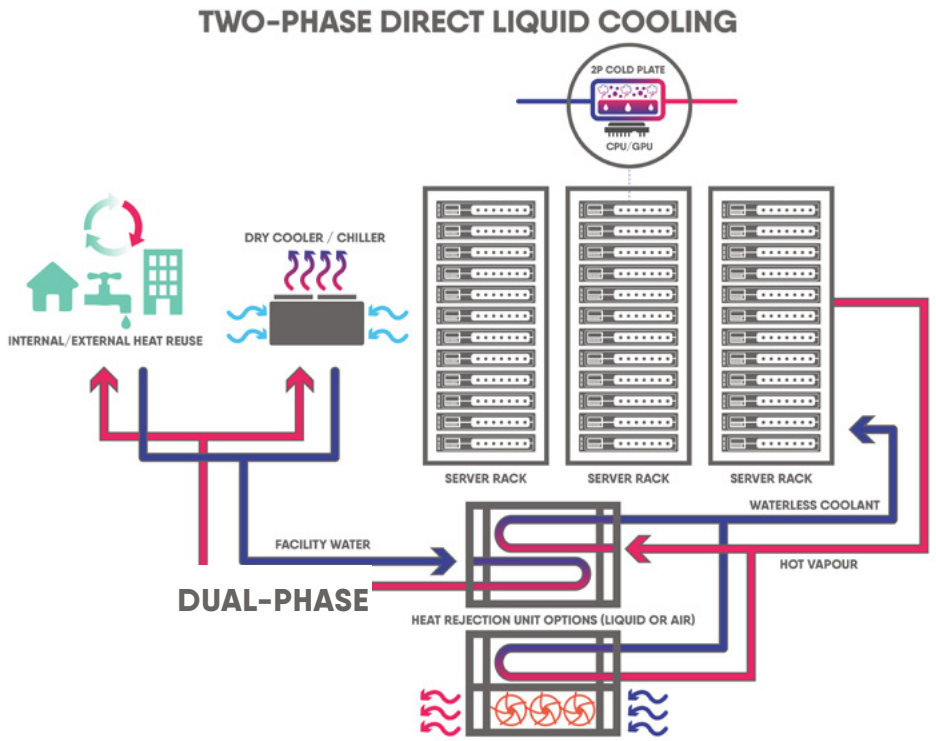
Dual-phase DLC involves an intricate process where a liquid undergoes a phase change into vapour upon contact with heat and, after vapour cooling, it condenses back to its liquid form, a method often employed in immersion cooling; it relies on a precise choice of a fluorocarbon-based liquid that transitions states around 50°C, demanding careful handling.

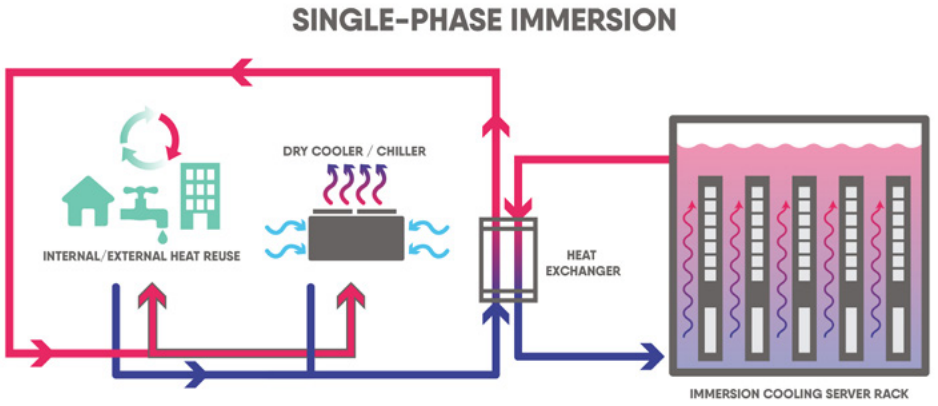
IMMERSION COOLING

Immersion cooling submerges components within a dielectric fluid, offering the choice of both single and dual-phase cooling loops, its dielectric properties ensuring safety as it efficiently absorbs and dissipates heat from electronic parts; with a secondary heat exchanger in play, the captured heat can be redirected for dissipation or repurposing, making it particularly well-suited for high-power applications such as data centres and industrial processes.









LIQUID COOLING

BOSTON Servers | Storage | Solutions

The Boston GlacierCore series is a range of liquid cooling-ready servers, designed for businesses that require robust, reliable performance with fully validated and warranted designs.

Castrol ON

Boston Limited has partnered with Castrol, a global leader in lubricants and thermal management, to bring advanced cooling solutions to the data centre industry.

submer

Submer specialises in single-phase liquid immersion cooling, offering solutions that scale from edge data centres to hyperscale deployments, addressing the evolving challenges of high-density computing.

SUPERMICRO

Supermicro provides a comprehensive end-to-end liquid cooling ecosystem, from direct-to-chip cooling solutions to rack-level infrastructure and cooling tower integration, reducing data centre electricity costs by up to 51%.

ZUTACORE

ZutaCore is a two-phase direct-to-chip liquid cooling specialist, delivering high-performance, energy-efficient solutions that maximise compute density, system longevity and operational efficiency.

THE GLACIERCORE SERIES

PURPOSE-BUILT FOR LIQUID COOLING

Our GlacierCore range of servers is designed from the ground up for immersion cooling, ensuring maximum efficiency and performance. Built to handle the demands of modern data centres, these servers leverage Castrol's cooling fluids to optimise performance, reduce thermal stress and support sustainable operations.

GLACIERCORE MODELS OVERVIEW:

GlacierCore 1100
A compact and cost-effective solution ideal for entry-level deployments, offering efficient immersion cooling for essential tasks.

GlacierCore 1110
A powerful choice for high-density applications, designed for businesses requiring robust computing performance with reliable cooling.

GlacierCore 1200
Built for advanced workloads, this model supports intensive applications such as AI training, benefiting from high-performance cooling.

GlacierCore 1210
Designed to support data-intensive and mission-critical environments, this model ensures stability even under heavy processing demands.

GlacierCore 1220
The top-tier choice for extreme processing needs, this model is optimised for AI, ML and HPC tasks, delivering sustained performance with unparalleled cooling efficiency.

APPLICATIONS

- AI and Machine Learning Training
- High-Performance Computing
- Data Analytics and Financial Modelling

Each model can be tailored to specific use cases, from high-density tasks to intensive AI applications, ensuring that you find the ideal solution for your data centre's needs.

CASTROL

CASTROL OIL & IMMERSION COOLING TECHNOLOGY

At the core of our liquid cooling solutions is Castrol's pioneering immersion cooling technology. Castrol's specially engineered cooling fluids provide superior thermal management, ensuring efficient heat transfer even during high loads. This results in a lower energy footprint, helping organisations minimise operational costs while extending the lifespan of their equipment.

Castrol's immersion cooling fluids are engineered for superior heat transfer, optimising the cooling process and enabling data centres to operate more efficiently, this also minimises the need for traditional air conditioning which reduces energy consumption, supporting Boston's mission to deliver sustainable technology options to our customers.

KEY FEATURES:

- The liquid submersion fluid is purpose-designed to ensure optimal thermal performance and efficiency
- It is environmentally friendly and safe, making it a sustainable choice for modern data centres
- The system provides 100% heat encapsulation, ensuring maximum thermal management and efficiency

BENEFITS:

- This technology enables significant reductions in cooling and operational costs, enhancing long-term savings
- Components benefit from improved reliability and an extended lifespan, reducing maintenance and replacement needs
- With reduced or even zero reliance on air cooling, facilities can achieve greater energy efficiency and sustainability

LIQUID COOLING WITH BOSTON LIMITED

As data demands grow, so does the need for innovative cooling solutions that support high performance, efficiency, and sustainability. At Boston Limited, we're leading this evolution through our partnership with Castrol, a global leader in thermal management and immersion cooling fluids.

By combining Castrol's advanced liquid submersion technology with Boston's expertise in high-performance computing, we deliver cutting-edge cooling solutions that encapsulate 100% of the system's heat output. This approach significantly reduces or eliminates the need for air conditioning, lowering energy consumption and operational costs.

Additionally, cooler operating components experience improved reliability and a longer lifespan, leading to fewer failures, reduced maintenance requirements, and a lower total cost of ownership (TCO). With Boston and Castrol, businesses can achieve unparalleled efficiency while meeting their environmental sustainability goals.

WHY LIQUID COOLING?

Liquid cooling dissipates heat more efficiently than traditional air cooling, making it ideal for data centres running intensive workloads. By immersing components directly in a specialised cooling fluid, heat is removed at the source, ensuring consistent thermal management even under peak loads.

SUSTAINABILITY & COST-EFFICIENCY

With reduced reliance on air conditioning, liquid cooling significantly lowers energy consumption, helping businesses meet sustainability targets. Additionally, cooler operating temperatures improve hardware reliability and lifespan, leading to fewer failures, reduced maintenance needs and a lower total cost of ownership (TCO).

Our partnership with Castrol gives Boston customers access to customised, high-efficiency cooling solutions that go beyond traditional methods. Designed for demanding environments, these systems enhance reliability, reduce operational costs and support sustainable business growth.

SUBMER

Submer is a leading provider of single-phase liquid immersion cooling tanks, designed to make data centre operation and construction more efficient and sustainable. Their technology allows hyperscalers and large industries to improve energy efficiency, reduce costs and optimise space utilisation. Submer's approach goes beyond just providing cooling solutions—they develop and refine their products, platforms, APIs and installation processes to create future-proof infrastructures. Their solutions tackle key challenges such as heat reuse, net-zero water consumption and site selection, helping businesses meet both current and future sustainability goals.

Their single-phase immersion cooling technology is highly scalable, allowing data centres to increase density without compromising efficiency. Whether for edge, enterprise or hyperscale deployments, Submer's technology offers a flexible, space- and energy-efficient approach to high-performance computing.

KEY FEATURES:

- Achieve an ultra-low PUE of 1.03 for improved energy efficiency
- Support up to 100kW of compute density per Pod for high-performance scalability
- Reduce deployment time with modular Pod-based building blocks for faster installation

BENEFITS:

- Extend hardware lifespan by maintaining optimal operating temperatures
- Lower hardware failure rates with consistent and efficient cooling
- Save up to 50% on CAPEX building costs with simplified infrastructure requirements
- Minimise water usage, supporting sustainable data centre operations

SUPERMICRO

Supermicro's liquid-cooled rack solution delivers superior performance and efficiency for HPC, large-scale AI and cloud-scale compute infrastructure. As modern applications demand high-performance CPUs and GPUs, power consumption and heat generation have increased significantly. To address this, Supermicro provides a complete, end-to-end liquid cooling ecosystem, covering everything from servers and Coolant Distribution Units (CDUs) to manifolds and even cooling towers.

Supermicro offers a wide range of liquid-cooled servers, including 1U, multi-node, blade and GPU-accelerated solutions, supporting the latest AMD, Intel and NVIDIA processors and GPUs. Each component is designed in-house for maximum performance, reliability and seamless integration, delivering a fully integrated, rack-level solution.

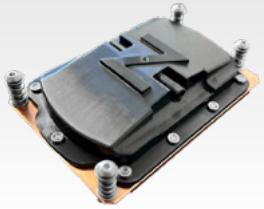
KEY FEATURES:

- **Comprehensive End-to-End Solution:** From server to cooling tower, Supermicro provides a fully integrated liquid cooling infrastructure
- **Range of CPU and GPU-Accelerated Servers:** Supporting HPC, AI and analytics workloads across 1U, multi-node, blade and GPU-accelerated architectures
- **Enterprise-Grade Reliability:** Features redundant cooling pumps, power supplies, leak-proof connectors and leak detection to ensure maximum uptime

BENEFITS:

- **Full Turn-Key Single Source Solution:** A one-stop-shop experience, with all components designed to work together seamlessly
- **Accelerated Lead Times:** Benefit from in-stock inventory, reducing procurement and deployment delays
- **Up to 51% Reduction in Data Centre Electricity Costs:** Lower cooling requirements translate to significant operational cost savings

ZUTACORE



For organisations running compute- and data-intensive applications, optimal server performance and longevity depend on efficient cooling solutions. As each new generation of processors generates more heat and requires tighter thermal tolerances, traditional cooling methods struggle to keep up.

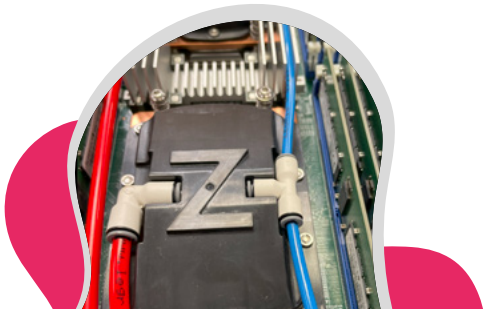
ZutaCore's two-phase dielectric liquid cooling technology offers a breakthrough solution. Unlike single-phase cooling, ZutaCore's system boils the coolant at the heat source, rapidly removing thermal energy from the processor. This phase change process enables CPUs to maintain high utilisation rates with zero throttling, improving performance and efficiency. Additionally, as a waterless system, ZutaCore eliminates the risks associated with traditional liquid cooling.

KEY FEATURES:

- **50% Reduction in Energy & Space Usage:** Optimises data centre footprint and power consumption
- **3x Processing Capacity:** Higher cooling efficiency allows for greater compute density
- **50% CapEx Reduction:** Lower infrastructure costs by reducing cooling complexity and footprint

BENEFITS:

- **Two-Phase Direct-to-Chip Cooling:** Boiling coolant extracts heat at the source, ensuring maximum thermal efficiency
- **<1000W of Cooling Per Chip:** Supports the highest-performance CPUs and GPUs without throttling
- **Waterless Dielectric Liquid Cooling:** Non-conductive coolant enhances safety and reliability



NVIDIA NETWORKING

NVIDIA Networking Ethernet products provide scalable, low-latency and efficient fabric with simplified management, reducing overall costs and power consumption compared to traditional Ethernet.

NVIDIA Networking InfiniBand switches and adapters support up to 400Gb/s, enabling scalable fabric for high-performance computing and data centres.

The NVIDIA Spectrum™-X networking platform is the first Ethernet platform designed specifically to improve the performance and efficiency of Ethernet-based AI clouds.

NVIDIA ETHERNET SPECTRUM®-4 PLATFORM

The world's first 400Gbps end-to-end networking platform, NVIDIA Spectrum®-4 provides 4x higher switching throughput than previous generations, with 51.2 terabits per second.

It consists of the NVIDIA Spectrum®-4 switch family, NVIDIA ConnectX®-7 SmartNIC, NVIDIA BlueField®-3 DPU and the DOCA™ datacentre infrastructure software to supercharge cloud-native applications at scale. Built for AI, NVIDIA Spectrum®-4 Switch arrives as data centres are growing exponentially and demanding extreme performance.

NVIDIA Spectrum®-4 features a fully-shared and monolithic packet buffer that's dynamically available to all ports. This provides excellent microburst absorption with true, port-to-port, cut-through latency.

KEY FEATURES:

- 400 Gbps end-to-end networking platform
- 4x higher switching throughput
- 51.2 Tbps of switching capacity, ensuring ultra-low latency and high-performance networking for AI, HPC and cloud data centres

BENEFITS:

- Provides extreme performance
- Enhanced efficiency
- 4x higher switching throughput than previous generations

NVIDIA INFINIBAND QUANTUM™-2 PLATFORM

The NVIDIA Quantum™-2 based QM9700 and QM9790 switch systems deliver an unprecedented 64 ports of NDR 400Gb/s InfiniBand per port in a 1U standard chassis design.

A single switch carries an aggregated bidirectional throughput of 51.2 terabits per second (Tb/s), with a landmark of more than 66.5 billion packets per second (BPPS) capacity.

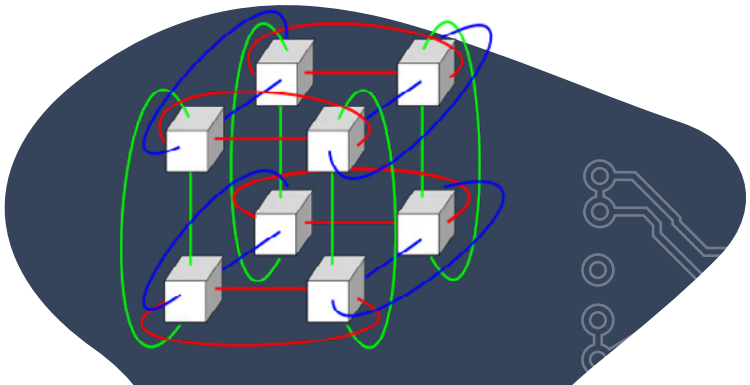
Supporting the latest NDR technology, NVIDIA Quantum™-2 brings a highspeed, extremely low-latency and scalable solution that incorporates state-of-the-art technologies.

KEY FEATURES:

- High speed, low latency and scalable
- Aggregated bidirectional throughput of 51.2 TB/s
- Delivers 64 ports of NDR 400 GB/s

BENEFITS:

- Incorporates state-of-the-art technologies
- 66.5 BPPS capacity
- 1U standard chassis design



NVIDIA SPECTRUM™-X

Spectrum™-X is a revolutionary solution created by NVIDIA for building multi-tenant, hyperscale AI clouds and AI Factories with Ethernet. With it, organisations can significantly improve the performance and power efficiency of AI clouds and gain higher predictability and consistency, resulting in faster time to market and a greater competitive edge.

NVIDIA Spectrum™-X enhances performance and power efficiency by up to 1.7X, accelerating the processing, analysis and execution of AI workloads and, in turn, the development and deployment of AI solutions. Spectrum™-X is fully standards-based Ethernet with support for open Ethernet stacks (SONiC) at cloud scale, and is tuned and validated across the full stack of NVIDIA hardware and software, creating an unmatched Ethernet solution for AI clouds.

KEY FEATURES:

- Nearly Perfect Bandwidth at Scale
- Extremely Low Latency
- End-to-End Stack Optimisation
- Advanced RoCE Extensions for Scalable AI Communications
- Deterministic Performance and Performance Isolation
- Open Network Operating System: SONiC and Cumulus

BENEFITS:

- Improved AI Cloud performance
- Reduced TCO
- Reduced time to results
- Full-stack optimisation
- Enhanced multi-tenancy features

SOFTWARE

DKube's software is an AI and machine learning platform that streamlines the end-to-end data science workflow, from data preparation and model training to deployment and monitoring.

NVIDIA Base Command Manager is a centralised GPU resource management platform that simplifies the allocation, deployment and utilisation of GPU resources across your HPC cluster.

OpenHPC is a modular, open-source framework designed for deploying and managing HPC clusters. It simplifies cluster management by providing pre-packaged, integrated software components, including tools for provisioning and resource management.

TrinityX is a cluster management solution focused on HPC and AI workloads. It enables efficient orchestration, monitoring and scaling of HPC clusters.

DKUBE

DKube is a powerful data integration and analytics platform designed to streamline how organisations manage and analyse their data. It offers a comprehensive set of tools for aggregating data from various sources, performing advanced analyses, and visualising results through customisable dashboards.

With its capability for real-time data processing and scalable architecture, DKube supports businesses in making informed decisions and staying competitive in a dynamic market.

KEY FEATURES:

- DKube integrates seamlessly with a wide range of data sources and formats, enabling efficient data aggregation and analysis
- It provides powerful analytics tools to derive actionable insights from complex data sets
- Users can create personalised dashboards to visualise data in ways that meet their specific needs
- The platform supports real-time data processing to deliver up-to-date information and facilitate faster decision-making
- It is designed to scale with the growth of data and user demands, ensuring long-term usability

BENEFITS:

- By providing comprehensive and real-time data insights, DKube helps businesses make informed decisions
- Streamlining data integration and analytics reduces the time and effort required for data management tasks
- Customisable dashboards and advanced visualisation tools enhance data accessibility and understanding
- Advanced analytics capabilities enable businesses to gain deeper insights and stay ahead of market trends

NVIDIA BASE COMMAND MANAGER

NVIDIA Base Command Manager is the cornerstone of efficient GPU resource management for your HPC cluster. This powerful platform centralises the allocation, deployment and utilisation of GPU resources, eliminating the complexities associated with GPU management. Whether you're a seasoned data scientist or a newcomer to the world of GPU computing, NVIDIA's Base Command Manager simplifies GPU workflow, making it a valuable asset for your HPC infrastructure.

The system's user-friendly interface and robust support for multiple users provide the foundation for optimising GPU-driven workloads.

KEY FEATURES:

- NVIDIA Base Command Manager offers a unified platform to manage your GPU resources across your HPC cluster
- Facilitate collaboration with support for multiple users and seamless allocation of GPU resources
- Utilise containers to streamline AI, ML and HPC workloads, ensuring efficient resource utilisation

BENEFITS:

- Streamlines model training, deployment and scaling, reducing time-to-insight
- Seamless multi-user collaboration enables secure, shared access to AI resources, improving productivity across teams
- Provides intuitive tools to orchestrate and monitor AI workloads with ease

OPENHPC

Boston offer a complete solution built around OpenHPC's comprehensive stack of management tools, Our team of experienced HPC engineers, can take you through the journey of designing and implementing a balanced, functional and cost effective cluster management solution.

OpenHPC is a collaborative, community-driven project that provides a comprehensive collection of open-source software components designed to simplify the deployment and management of HPC Linux clusters. It offers a modular framework that includes provisioning tools, resource management, I/O clients, development tools and a variety of scientific libraries, all pre-built and tested for HPC environments.

KEY FEATURES:

- **Comprehensive Software Stack:** OpenHPC aggregates a wide range of pre-packaged software components essential for HPC systems, facilitating easier installation and management
- **Modular Architecture:** Its modular design allows users to select and integrate only the components that meet their specific needs, providing flexibility in cluster configuration
- **Community Collaboration:** As a community-driven initiative, OpenHPC benefits from collective expertise, continuous updates and shared best practices, ensuring a robust and evolving ecosystem

BENEFITS:

- **Simplified Deployment:** By offering pre-built packages and standardised configurations, OpenHPC reduces the complexity involved in setting up HPC clusters, enabling faster deployment
- **Flexibility:** Users can tailor their HPC environments by selecting components that align with their specific requirements, allowing for customised solutions
- **Scalability:** OpenHPC's framework supports the scaling of resources to meet the demands of growing computational workloads

TRINITYX

ClusterVision's TrinityX management software provides 24/7 oversight for your HPC & AI environments, offering a support system that is both dependable and SLA-compliant. This setup allows you to focus entirely on your research, leaving the complexities such as Linux, SLURM, CUDA, InfiniBand, Lustre, and Open OnDemand to us.

AI and High-Performance clusters demand efficiency and effectiveness. TrinityX streamlines the management of the cluster infrastructure, allowing you to add additional features and applications such as Kubernetes and Singularity in a user-friendly environment for the deployment.

KEY FEATURES:

- **Secure and Ultra-fast Cluster Deployment:** TrinityX streamlines cluster deployment through its intuitive interface, offering a straightforward setup process
- **Centralized Cluster Monitoring & Alarms:** Proactively address issues before they impact your business. Monitoring and Alarm triggers are crucial for ensuring smooth cluster operations
- **Efficient Power Resource Management:** Enabling dynamic resource scaling based on workload demands, ensuring your clusters always operate at peak efficiency

BENEFITS:

- **Enhanced Agility:** The open-source nature and modular architecture of TrinityX provide the flexibility to adapt to evolving computational needs
- **Reliability:** Designed for dependability, TrinityX ensures that HPC and AI systems operate smoothly, allowing users to focus on their research without worrying about system stability
- **Managed Services:** An inclusive Management Support framework enables, SLA-driven, Remote System Administration, which enables you to focus on research

BOSTON SERVICES

At Boston, we bring together a diverse range of expertise and services under one roof to meet all your technology needs. Whether you're looking for cutting-edge solutions, in-depth training, or reliable support, we offer a complete suite of services to support your organisation's success. From our global team of technical and HPC experts who benchmark and optimise new technologies, to our Boston Training Academy that nurtures talent and education and our Boston Labs that provide custom tech solutions and testing facilities—we cover every aspect of your technology journey.

BOSTON TECHNOLOGY CONSULTING

Our technical and HPC experts, based around the world, are constantly benchmarking new technologies, the results of which we share with our customers, to deliver fully optimised solutions.

BOSTON TRAINING ACADEMY

The mission of the Boston Training Academy (BTA) is to become a renowned developmental ground for talent engagement, education and solutions across a variety of disciplines.

BOSTON LABS

Boston provides tech facilities to test emerging technologies and enterprise-grade systems from leading manufacturers. Boston Labs is customer-centric, customised solutions with off-the-shelf components to enhance business outcomes.

PROFESSIONAL SERVICES

Our senior engineers understand the technical dependencies and requirements of your organisation, we will ensure a well thought out installation is managed and completed on schedule and with the utmost professionalism.

SUPPORT AND ONSITE WARRANTY

From tailor-made SLA solutions, warranty support and spares packages – Boston offers customised services level work packages for ongoing support.

BOSTON TECHNOLOGY CONSULTING

Our technology consultancy service helps organisations to effectively create, optimise and grow their business in today's data and IT focused market. Our consultants mastermind solutions for the most complex business needs in an effort to achieve their goals, while balancing costs and performance. Our core values are leadership, integrity, collaboration, accountability and simplicity.

Our global reach enables us to provide agile and scalable solutions to everyone from the smallest players to industry heavy hitters. Our main objective is to solve current problems while minimising risks of future issues. Such all-encompassing approach to problem solving has been utilised in the banking, financial services and insurance (BFSI), government, manufacturing, oil and gas, logistics, telecoms, pharma and retail.

Our consultancy service is aimed at enterprises striving to utilise specialised services to optimise their solutions and enhance efficiency. With our deep industry knowledge, we provide valuable guidance to unlock the full potential of your Infrastructure.

- **Cross-domain expertise:** We stand out through our wide-ranging expertise across multiple technical domains, allowing us to provide holistic solutions that address complex challenges spanning various technologies.
- **Innovative problem solving:** We excel in innovative problem-solving, employing a team of seasoned consultants who approach challenges with creativity and resourcefulness, leading to unique solutions that drive customer success.
- **End-to-end project management:** We offer comprehensive project management services, overseeing projects from inception to completion. This end-to-end approach streamlines processes, reduces customer workload and ensures successful project outcomes.

BOSTON TRAINING ACADEMY

Boston Training Academy provides structured training by world-class trainers who deliver tailored courses based on the current knowledge of the attendees. Our program content is designed by industry experts and covers data science skills, exposure to analytical tools and corporate learning and consulting on AI projects, to name a few.

The AI Centre of Excellence (CoE-AI) at our Training Academy brings together experts from industry academia and Government to provide the latest infrastructure, smart tools, skilled resources and leadership for AI exploration. The CoE-AI aims to create the most sophisticated ecosystem for AI, Data Sciences, Machine Learning and Deep Learning to enable breakthrough innovations.

We also offer online starter courses for complete beginners to gain working knowledge on the fundamentals of NLP, computer vision, deep learning and Python for data science. Boston also host NVIDIA Deep Learning Institute (DLI) workshops. Participants receive hands-on training sessions, with help from developers, data scientists and researchers. Attendees learn how to approach challenges using deep learning techniques like building transformer-based NLP, building conversational AI applications, building intelligent recommender systems, among other approaches.

We provide practical training to universities and businesses on algorithm creation, hardware development and the latest technologies. Our expert instructors empower organisations to stay competitive in a rapidly evolving digital world.

- **Expert-led training:** We offer expert-led training programs delivered by industry professionals, ensuring participants gain practical insights and skills that are immediately applicable in their professional roles.
- **Tailored learning paths:** We excel in providing personalised learning paths, allowing participants to choose courses that match their skill levels and career aspirations. This tailored approach ensures that individuals receive relevant and effective training.
- **Hands-on lab experience:** With a strong emphasis on practical learning, we offer hands-on lab experiences that enable participants to engage with real-world scenarios, fostering better understanding and confidence in their abilities.

BOSTON LABS

DISCOVER THE FUTURE WITH BOSTON LABS: YOUR TESTING PARTNER

Are you navigating the ever-evolving landscape of new and enhanced technologies? We understand the challenge of making informed decisions for your upcoming projects. The task becomes even more daunting when you can't experience and assess the hardware firsthand prior to making a commitment.

Through our Boston Labs, we bridge this gap for you. Our state-of-the-art, remote accessible, on-site test facilities offer the perfect environment to explore and evaluate emerging technologies. We recognise that sometimes, the true fit of a product or solution can only be gauged through hands-on experience.

Our dedicated technical team is poised to assemble comprehensive solution stacks using cloud-based lab systems or even loaned hardware that can be seamlessly integrated into your environment. This ensures that your investment aligns perfectly with your requirements. Immerse yourself in our incredible R&D lab, Boston Labs, where an array of the latest hardware and software solutions await your evaluation. Whether you choose remote assessment or prefer to collaborate closely with our seasoned field application engineers, the choice is yours.

Our veteran team excels in guiding you through the setup and optimisation of both hardware and software components, crafting solutions that harmonise performance, power efficiency and cost-effectiveness. When you reserve a session at Boston Labs, you gain access to expert advice on product selection and fine-tuning, helping you achieve unparalleled outcomes.

EXPERTS IN AI AND DATA

Boston Training Academy and Boston Technical Consulting are our in-house initiatives which represent the pinnacle of technological advancement, driving forward groundbreaking solutions that harness the power of artificial intelligence.

PROFESSIONAL SERVICES

PRESALES CONSULTATION AND SERVICES

Our Presales Consultation and Services aim to provide optimal solutions for your business needs. Our team of experts is dedicated to understanding your requirements, offering personalised consultations and delivering our wide range of technology solutions. We go beyond product knowledge, bringing a deep understanding of industry trends and emerging technologies to help you stay ahead of the competition.

WE CAN PROVIDE:

- Consultancy and Solution Design
- Inventory and Site Survey
- Technical Account Management
- Boston Labs and POC Services
- Bespoke System Design and OEM Manufacturing
- Product and Technology updates
- Environmental and Regulatory Testing

POST SALES AND IMPLEMENTATION

Our Post Sales and Implementation team ensures a smooth and successful deployment of your technology solutions. With skilled project management expertise, thorough testing, seamless integration and comprehensive training, we prioritise your success from purchase to implementation.

WE CAN PROVIDE:

- System Integration and Burn in Testing
- Racking and Installation
- Project Management
- Global delivery
- Turnkey Solutions and Rack and Roll
- Training and Education

DATACENTRE AND HOSTING SERVICES

Your business relies on a secure, reliable and high-performance infrastructure to thrive in the digital landscape. Boston offers state-of-the-art data centre and hosting solutions designed to meet your specific needs. Our industry expertise delivers robust and scalable data centre solutions, including dedicated servers, virtual private servers (VPS), cloud hosting and colocation services. With enterprise-grade infrastructure, customised solutions, 24/7 support and scalability, we empower your digital infrastructure to reach its full potential. Choose Boston's Datacentre and Hosting Services for exceptional performance and reliability.

WE CAN PROVIDE:

- Hosting
- Leasing
- Managed Services and Cloud
- Datacentre and Rack Design
- Rack and Roll

END OF LIFE AND EXTENDED LIFE CARE

As technology rapidly evolves, effectively managing end-of-life products and systems becomes crucial. Boston Limited's End of Life and Extended Life Care services provide comprehensive solutions to navigate this critical phase of your technology lifecycle. Our expert team understands the challenges associated with retiring outdated systems and mitigating risks related to unsupported or vulnerable hardware. With our services, we ensure a seamless transition to newer technology while maximising the value of your existing investments.

WE CAN PROVIDE:

- Decommissioning, Recycling and Disposal
- Out of Warranty Repair and Refurbishment

Please contact us if you are interested in discussing any of our additional service offerings.

SUPPORT AND ONSITE WARRANTY

At Boston, exceptional customer support is the foundation of our successful partnerships. Our Support and Onsite Warranty services offer peace of mind, knowing that our dedicated team of experts is ready to assist you whenever you need help.

With prompt and reliable assistance available 24/7, our experienced professionals with deep expertise across various technologies are just a phone call away. We provide onsite support to minimise downtime and ensure smooth operations, coupled with warranty coverage for eligible hardware, ensuring swift replacement or repair.

WARRANTY SERVICES	RETURN TO BASE (RMA)	ADVANCED WARRANTY	ONSITE WARRANTY			
			PEWTER	BRONZE	SILVER	GOLD
TECHNICAL SUPPORT DESK	YES	YES	YES	YES	YES	YES
ADVANCED PARTS SHIPPING	NO	YES	NO	YES	YES	YES
REMOTE ISSUE TRIAGE	YES	YES	YES	YES	YES	YES
ONSITE SUPPORT/ DIAGNOSIS	NO	NO	NO	NO	NO	NO
ONSITE BREAK/FIX LABOUR	NO	NO	YES	YES	YES	YES
ENGINEER TO SITE TARGET (POST DIAGNOSIS)	N/A	N/A	NBD	NBD	SBD	4 HOURS
ACCESS TO SERVICES 24 HOURS A DAY, WEEKENDS & UK BANK HOLIDAYS	NO	NO	NO	NO	NO	YES

CO-DESIGN IN HPC: NAVIGATING A NEW ERA MARKED BY POST MOORE'S LAW, EDGE-TO-EXASCALE WORKFLOWS AND AI

Three significant drivers are reshaping how organisations approach their computing infrastructure and bringing HPC towards the centre of IT strategies: (1) the proliferation of technology choices, (2) the need for a computing "fabric" that spans from device to data centre and (3) the emergence of artificial intelligence (AI) as a new ecosystem.

These forces are compelling a shift toward custom, co-designed hardware — a collaborative approach to building systems that reduces cost and complexity. Co-design unites customers with the right coalition of vendors to architect and build an optimal, integrated solution tailored to specific requirements.

THE COMPLEXITY OF CHOICE

One of the most challenging aspects of modern HPC and AI is the overwhelming number of technology options available. CPUs from Intel, AMD, Ampere and others, alongside GPUs and accelerators from NVIDIA, AMD, Intel, Cerebras, Untether, Groq, SambaNova and others. This is further complicated by interconnect and networking choices and a variety of storage options, all supporting a growing roster of software. Selecting and integrating the right component now and into the future is a difficult task.

These advances present opportunities, challenges and tradeoffs. For example, while low-precision hardware allows for faster computations, not all applications are suited to it. Many HPC workloads, such as weather modelling or molecular simulations, may still require 64-bit precision, while AI-driven applications can often operate effectively with significantly lower-precision arithmetic. But HPC applications can come out ahead if practitioners can reformulate the algorithms to harness the strengths of newer technologies without compromising the precision or accuracy the results.

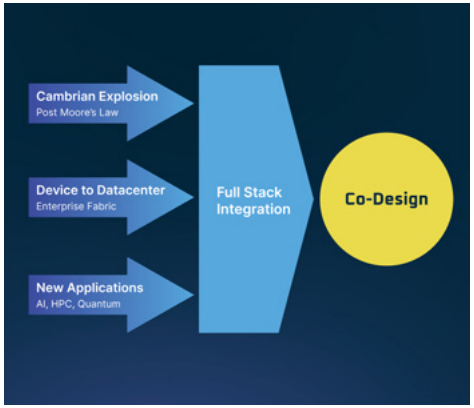
This is an example of where co-design comes into play: the ability to canvas a wide range of hardware and software options to make the best use of available technology. In such a case, it's no longer a question of selecting a pre-built platform; it's about customising and fine-tuning the platform to serve specific needs. Co-design becomes even more essential when considering the platform's evolution over time — ensuring it adapts seamlessly as new technologies emerge while maintaining the same functionality.

BUILDING A COMPUTING FABRIC: FROM DEVICE TO DATA CENTRE

In a world where everything is an AI-enabled computer, all IT solutions blend into a "fabric" – a unified web of flexible infrastructure that spans from the edge (where much of the data is generated and eventually consumed) to data centres of varying size. This distributed computing model aligns with the modern workflow, where managing data flow is as important as computational power.

Consider the landscape of a modern HPC workflow: data streams in from sensors, devices, scientific instruments or other edge environments and must flow seamlessly through various computing layers, from edge nodes to data centre supercomputers and cloud resources. The ability to manage this flow, this meta system, is critical. The co-design model ensures that the computing stack is optimised both in parts and in whole, and that it interacts and integrates with the rest of its extended ecosystem.

This is especially true when speed is of the essence. In applications where real-time data analysis is critical — such as autonomous devices or fleets, or digital twins — the ability to process data close to where it is generated while also having access to centralised, powerful computing resources can be a game-changer.



THE AI REVOLUTION AND THE DEMAND FOR MORE POWER

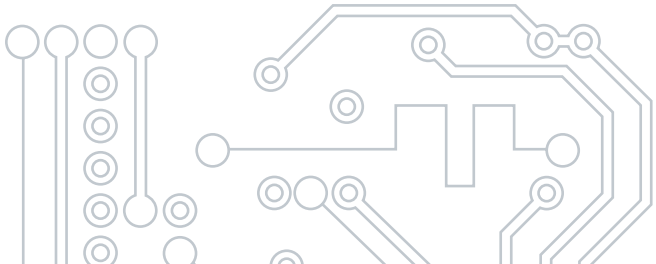
Finally, AI is creating a clean slate, removing and replacing the boundaries of what is possible, and driving the need for even more powerful machines. AI workloads differ from traditional HPC applications in that they can readily take advantage of low-precision data types and arithmetic, and they benefit from a favorable funding environment that enables massively scaled infrastructure. This shift has profound implications for the design of HPC systems because wherever there is speed, HPC is looking to exploit it.

Co-design allows HPC systems to meet these demands head-on. By aligning hardware innovation with the specific requirements of AI workloads, organisations can build systems that are not only powerful but also efficient and scalable.

THE TIME FOR CO-DESIGN IS NOW

As HPC enters this new era of technological complexity, distributed computing and AI-driven workloads, the need for a co-design approach has never been greater. The proliferation of technology choices, the demand for a seamless computing fabric from device to data centre, and the new AI ecosystem make it clear that only a broad, vendor-neutral, engineering-led, mission-focused approach can meet all needs. Co-design enables organisations to build HPC systems that are greater than the sum of their parts, allowing them to harness the full potential of the modern computational landscape.

Additionally, power consumption and cooling are becoming more critical as systems grow increasingly powerful. A thoughtful co-design approach can help optimise energy efficiency, ensuring that new HPC infrastructures remain sustainable while meeting performance goals.



HOW SOURCECODE CAN HELP WITH YOUR CO-DESIGN NEEDS

SourceCode's history and business model are rooted in the co-design approach. Partnering with SourceCode for your HPC and AI needs offers distinct advantages:

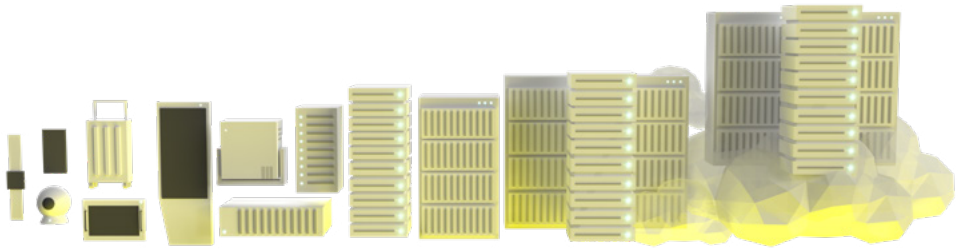
- Accelerates time to market via proven co-design methodology that brings the right expertise at the right time
- Reduces technical risk via access to best-in-class components and building blocks
- Balances design constraints: cost, performance, power, cooling
- Maximises the potential of rapidly evolving technologies, complementing in-house expertise

SourceCode offers expertise across a broad spectrum of leading-edge technologies that span from device to data centre, including exascale-class systems from Eviden.

As a vendor-agnostic partner, SourceCode provides the flexibility to select the best components for optimal performance. Our streamlined co-design process enhances efficiency, saving both time and cost.

Additionally, SourceCode's in-house, U.S.-based environmental testing lab ensures that your hardware meets top-quality standards and achieves the necessary certifications, providing unmatched reliability for your most critical systems.

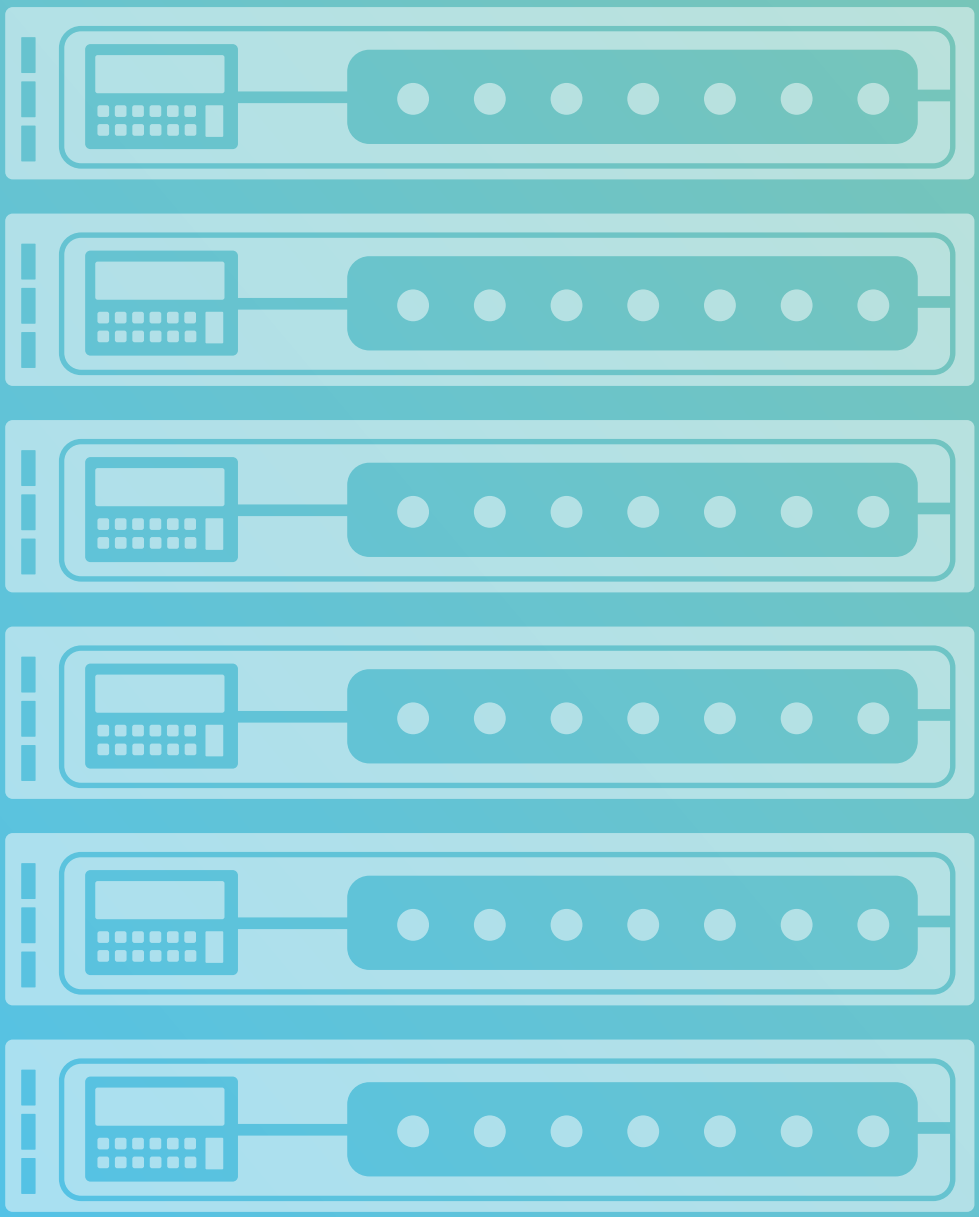
Learn more at www.sourcecode.com or reach out to us at info@sourcecode.com.



BOSTON

Servers | Storage | Solutions

A DIVISION OF SOURCECODE



sales@boston.co.uk | www.boston.co.uk | +44 (0) 1727 876 100 | Boston Limited, Unit 5 Curo Park, Frogmore, St Albans, AL2 2DD, UK

Austria | Australia | France | Germany | India | Taiwan | South Africa | Switzerland | **UK** | USA

March - 2025