KAO DATA

How Kao Data is supporting InstaDeep to become a global leader in AI powered decision-making.

In just a short space of time, InstaDeep has become a global leader in artificial intelligence (AI) powered decision-making. As experts in the field of reinforcement learning, the AI scale-up recently deployed GPU accelerated hardware at the NVIDIA DGX-Ready, Kao Data campus, which would support its research and development resources hosted in the cloud.

InstaDeep™

Kao Data's reputation as an industrial-scale facility for HPC and AI, and its technical expertise within intensive computing, positioned it as the perfect partner for InstaDeep, which required a customised and cost-effective solution to support development instances in the cloud.

Customer Background – A global leader in reinforcement learning

Founded in 2014, InstaDeep uses its expertise within GPU-powered computing, machine learning and reinforcement learning to solve some of the world's most complex challenges.

Headquartered in London, and with offices across EMEA, the company has received numerous accolades for its cutting-edge achievements in research and development. To date, InstaDeep has twice been named by CB Insights as one of the 100 most promising AI start-ups in the world, it holds a prominent position as an Elite-level service provider within NVIDIA's Partner Network (NPN) and is part of Intel's AI Builders programme.

As a high-growth, AI scale-up, the company has created several unique and market-leading solutions, including its DeepChain™ technology; a cloud-native protein design platform which enables the discovery of new protein designs, validated with molecular dynamics simulations, and without requiring any machine learning expertise.

Today, its global use cases are numerous, working with some of the largest names in its customers' sectors. The work includes a strategic collaboration with BioNTech on the discovery and development of novel immunotherapies; collaboration with Total on 3D microfossil image detection, segmentation and classification; and development of a train capacity and traffic management system for Deutsche Bahn (below).

The Situation – Testing, developing and training AI

In an environment where one in ten training and development instances lead to product developments, the requirement for a technically advanced, superfast and cost-efficient research cycle had become critical.

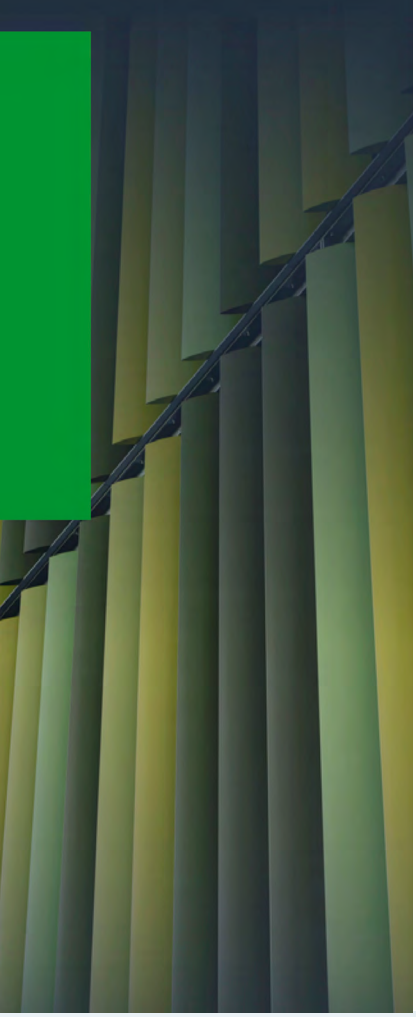
In order to use AI and intensive computing to solve complex decision-making problems, InstaDeep's engineers must work quickly to develop, test and scale their algorithms. Here, low latency connectivity, abundant compute power and application performance is essential. As such, the company quickly became an early adopter of NVIDIA's GPU servers, enabling them to accelerate their deep learning application developments.

At the beginning of its journey, InstaDeep placed a small number of NVIDIA DGX servers on premise, while arranging separate hosting agreements with both a cloud and colocation provider. With increasing numbers of customer projects underway, and more DevOps personnel joining the organisation, its compute requirements quickly escalated.

Such growth created business pressures on the company to be more agile while reducing cost. InstaDeep had reached that crucial point as an AI start-up, where it needed dedicated, intensive computing resources, deployed cost-effectively within an industrial-scale data centre, to support its production, research and experimentation instances in the cloud.

"We began to realise that owning our NVIDIA DGX-Ready hardware and hosting it on-premises would be highly beneficial to support our research and experimentation instances. Availability is essential and investing in our own supercomputer would dramatically increase the outputs from our research and testing. As such, our calculations suggested it would be financially prudent to create a hybrid environment, which would help accelerate our production instances in the cloud."

- Nacef Labidi, Lead DevOps Engineer and Project Lead at InstaDeep.



InstaDeep's DevOps team tested several alternative solutions before making the decision to deploy its first supercomputer at Kao Data. The new infrastructure would offer a dedicated, high performance, on-premise environment which would increase the speed of production and testing, with an architecture that could scale as data and demands grew.

Further, InstaDeep had reached the limits of its current colocation hosting provision, and due to its requirement for high-performance computing (HPC) capabilities, was temporarily supporting performance demands with several NVIDIA DGX 100 servers hosted at its headquarters.

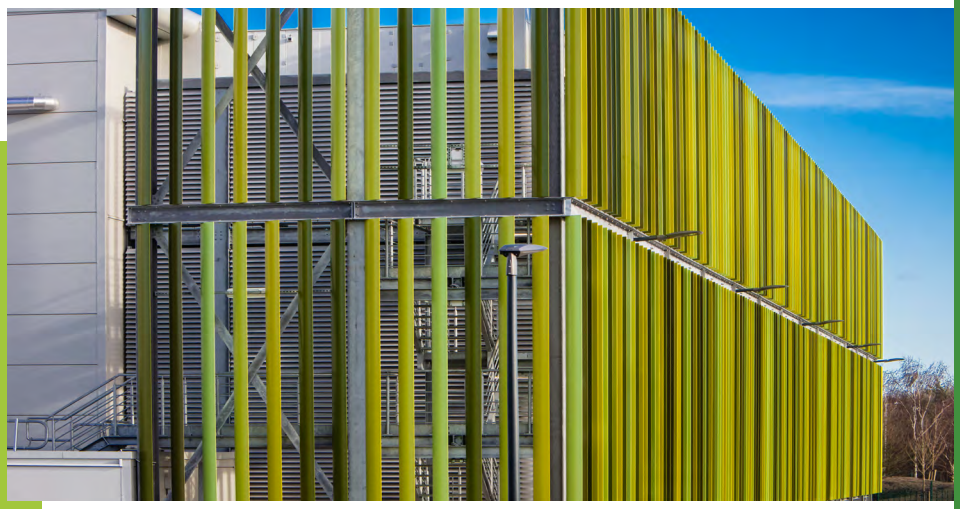
In the world of AI, trained models are just as important as the IT infrastructure on which they are hosted, and in order to progress further, InstaDeep needed to develop the perfect hybrid model; leveraging the scale offered by the cloud, and combining it with the industrial-scale performance of an on-premise, DGX-Ready, colocation data centre provider.

New Platform for Growth

InstaDeep made the decision to target its developments on 10,000 CPU cores, using the latest NVIDIA DGX 100 servers, which would provide the GPU density its engineering teams required. Once a solutions study was complete, and a specification agreed, their incumbent hosting company was contacted, together with several colocation providers, to discuss both costs and the implementation of its new infrastructure.

During the process, concerns were quickly highlighted, demonstrating that generalist and legacy colocation data centres were unable to accommodate the requirements of HPC. Issues including the need to mitigate physical, networking, and electrical concerns, coupled with challenges around power availability, rack density and cooling efficiency were common among various requests for proposals across the customer relationship. Further, in one RFP from the customer, the proposal included a spread of 10 racks, which would have increased the complexity of installation, added latency and required additional cost.

As such, it was clear that InstaDeep needed to find a facility that could accommodate the needs of AI cost-effectively. Further, it required technically excellent features, including the ability to host customised infrastructure, contiguous racks, supported by specialist technical expertise in HPC and within an environment inspired by the hyperscale cloud. Such fundamentals would be essential if it was to leverage the scale and service offered by the cloud, while benefiting from on-site processing speeds and tailored, face-to-face support.



The Solution: A data centre precision engineered for AI

Kao Data recommended that InstaDeep deploy its new supercomputer within a customised architecture, which was precision-engineered for intensive computing.

InstaDeep's new supercomputer would include AMD EPYC (Milan) powered hardware, NVIDIA DGX A100 GPUs, Mellanox switches and a Ceph distributed storage system; all of which were configured to allow all compute nodes (CPUs and GPUs) access to the storage cluster. This design would alleviate many of their challenges, allowing InstaDeep to optimise performance and accelerate the reinforcement learning process.

In addition, the cluster would be housed in a dedicated Technology Suite that provided 28kW high density adjacent and contiguous racks, while ensuring a low latency and a highly secure environment. Through numerous networking capabilities, Kao Data could also offer access to a wide range of private and public networks and connectivity providers, with direct on-ramps into the cloud via Megaport.

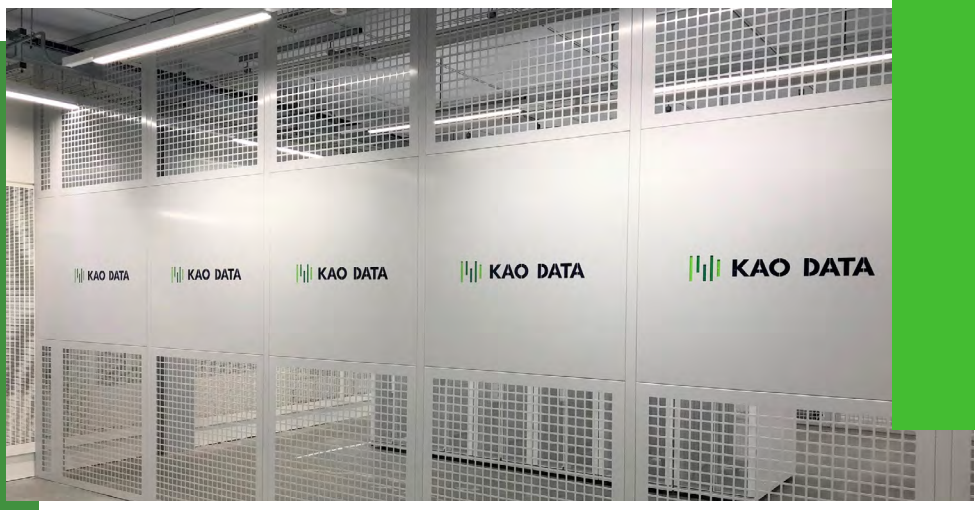
Furthermore, InstaDeep would benefit from Kao Data's ultra-efficient cooling systems, access to abundant, 100% renewable power and a market-leading low PUE; offering them competitive pricing and a cost effective solution for their specialist AI infrastructure.

"When we found Kao Data, we were excited by the design of its facility and its technical expertise in HPC and AI. During our first engagement, we quickly realised they truly understood our requirements. Their knowledge and expertise were evident and they gave us the confidence they could accommodate our specific installation requests while meeting demanding timescales."

- Nacef Labidi, Lead DevOps Engineer and Project Lead at InstaDeep.

Success: Low cost, high performance AI development

The £1M supercomputing deployment at Kao Data has provided InstaDeep with a highly scalable, low cost, and customised architecture, which creates the perfect hybrid environment to deliver exceptional performance capabilities, supporting the production, research and experimentation instances hosted in the cloud.



"We believe that our supercomputer will provide a return on investment in less than six months and will directly complement the development cycles we're hosting in the cloud. Designing our own customised environment and hosting it at Kao Data offers a cost–effective and highly accessible on-premise solution, which is optimised for the needs of our DevOps teams."

- Nacef Labidi, Lead DevOps Engineer and Project Lead at InstaDeep.

Its flexible design, with access to abundant, renewable power, delivers capacity of 28kW per rack as standard, with the ability to increase workloads to 85kW if needed. This provides an excellent development platform for prototyping and training more complex AI models, and almost limitless scalability as InstaDeep grows.

With access to multiple cloud and connectivity providers, latency is no longer an issue. Such diverse connectivity offers InstaDeep the ability to combine Kubernetes and containerisation development programmes, together with its automation pipelines, and seamlessly upload qualified models from the cluster to the cloud. This approach also ensures that the company has its own, dedicated, on-premises supercomputer that offers the flexibility to burst into cloud-based resources, should the cluster ever reach peak capacity.

Furthermore, Kao Data's location in the UK Innovation Corridor provides a secure and robust home close to the company's London HQ. This offers InstaDeep's engineers ease of access to their own infrastructure as and when it's required, while providing valuable timesaving in the development cycle. Finally, its position as the home for HPC and AI has provided InstaDeep with valuable technical expertise, alongside a facility precision engineered for industrial-scale computing.

"Our supercomputer offers massive compute capacity for our experimentation pipelines and is where all of our development workloads are hosted. Approximately 1 in 10 models go from experimental stage through to data models, which we provide to customers via the cloud. The team at Kao Data has supported us to deliver a highly efficient development platform, which is essential to meet current and future demand.

The installation at Kao Data has mitigated all of our design, engineering, physical and electrical concerns, while providing an efficient, scalable framework that delivers performance requirements. Their engineering expertise provided specific features that would alleviate many of our constraints, with competitive pricing based on our footprint and the power usage. Finally, the team's commitment to customer service is exceptional. When they commit to a specific live date, they ensure the job gets done."

- Nacef Labidi, Lead DevOps Engineer and Project Lead at InstaDeep.

Looking forward, InstaDeep's supercomputing footprint has already begun to scale. The company is considering a second phase, which would see it add new GPUs and increase the size of its supercomputer. As a global leader in AI-powered decision-making, the ability to test, develop and scale cost-effectively remains crucial to InstaDeep's business. And its location within the UK Innovation Corridor's digital ecosystem offers a high-performance, cost-effective home for its compute.