GRAPHCORE

BOW POD$_{16}$

Explore | Build | Grow

Bow Pod systems deliver high performance and efficiency for machine intelligence deployment at scale. They are designed to accelerate the large and complex models of today while also providing a platform for innovators to explore and invent the solutions of tomorrow.

The Bow Pod$_{16}$ system is ideal for exploration. It provides all the power, performance and flexibility required to fast track the prototyping stage, rapidly moving to production. Bow Pod$_{16}$ is the ideal starting point for building better, more innovative AI solutions with IPUs, in any machine learning field, including language and vision, exploring GNNs and LSTMs or creating something entirely new.

Latest generation IPU

The Bow Pod$_{16}$ system features 4 Bow-2000 machines, each containing 4 of our new pioneering Bow IPU processors. This innovative IPU is the world's first processor to be manufactured using Wafer-on-Wafer (WoW) technology, taking the benefits of the proven IPU technology to the next level.

Performance and efficiency

Bow Pod$_{16}$ delivers up to 5.6 petaFLOPS of AI compute as well as industry leading efficiency, all thanks to the use of innovative silicon technologies, a compute and memory architecture focused on efficiency and scale-out, and a software- and application-first approach in solution deployment.

Smooth deployment and short time to market

The whole system, hardware and software, has been architected together. Bow Pod$_{16}$ supports all standard frameworks and protocols to enable straightforward integration into existing data centre environments, as well as private and public clouds. A wide selection of market leading server platforms and high-performance storage appliances designed for AI have been tested and validated to offer choice, in addition to short configuration and deployment times for system aggregators. Innovators can focus on deploying their AI workloads at scale, using familiar tools and frameworks while unlocking cutting-edge performance and efficiency.

System Specifications

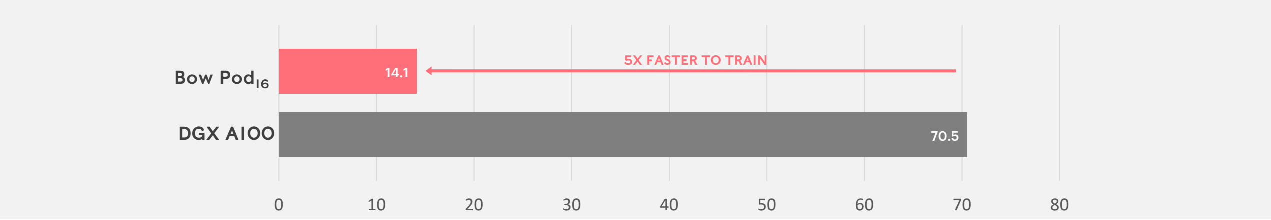
Processors	16 Bow IPUs
1U blade units	4 Bow-2000 machines
Separate cores	23,552
Threads	141,312
Performance	5.6 petaFLOPS FP16.16 1.4 petaFLOPS FP32
Memory	14.4 GB In-Processor-Memory™ Up to 1 TB Streaming Memory™
Software	Poplar® SDK

Host-Link	100 GE RoCEv2
System Weight	66 kg + Host servers and switches
System Dimensions	4U + Host servers and switches
Host server	Selection of approved host servers from Graphcore® partners.
Storage	Selection of approved solutions from Graphcore partners.
Thermal	Air-Cooled

BOW POD$_{16}$

Disaggregation for customised compute

Machine intelligence workloads have very diverse compute demands. For production deployment, optimising the ratio of AI to host compute can help maximise performance, while improving total cost of ownership. Bow Pod systems allow flexible mapping of the number of servers and switches to the requisite number of Bow-2000 machines, so deployment is better tailored to production AI workloads. Bow Pod$_{16}$ supports multiple server configurations.



EfficientNet-B4 Training Time To Train Performance
IPU-POD Platforms | Preliminary Results (Pre-SDK2.5) | G16-EfficientNet-B4 Training
DGX A100 (A100-SXM4-80GB) | TensorFlow | Mixed Precision | https://developer.nvidia.com/deep-learning-performance-training-inference

Communication architecture built for scaling

Efficient data access and transfer can unlock greater AI performance. IPU-Fabric is an innovative communication architecture for system-wide data transfer, extending high-speed interconnect within individual Bow IPUs, across Bow-2000s, between Bow Pods and throughout the data centre. IPU-Fabric delivers high-performance low-latency communication to maximise AI application efficiency and is built to work with standard data centre communication technologies.

Platform for AI developers

TensorFlow, PyTorch, PaddlePaddle, and many other popular ML frameworks are supported and available as open source, along with the comprehensive PopLibs™ library, for community driven collaboration and innovation. For developers who want full control to exploit maximum performance, the Graphcore Poplar SDK enables direct IPU programming in C++.

Designed for deployment at scale

Pre-built Docker containers with Poplar SDK tools and frameworks images let innovators get up and running fast. Various common frameworks for container orchestration, platform visualisation and provisioning are also supported, including Slurm, Kubernetes and OpenStack.

Software First

Fully integrated and IPU-optimised, Poplar software leverages the unique characteristics of the IPU architecture to build AI applications of unrivalled performance and flexibility. Poplar allows effortless scaling of models from one to thousands of IPUs without adding development complexity, allowing innovators to focus on the accuracy and performance of the application.

Access to AI expertise

A wealth of experience and support for installation, production and application development is available globally from Graphcore AI experts and from our elite partner network.

Ready to experience the next level in Machine Intelligence?

Connect with our partners below to assess your AI infrastructure requirements and solution fit.
Still have questions? Contact Graphcore directly at info@graphcore.ai