Bow Pod systems deliver high performance and e□iciency 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$_{256}$ system is the solution for innovators ready to grow their capacity to supercomputing scale. It delivers massive e□iciency and productivity gains by enabling large model training runs to be completed in hours or minutes instead of months or weeks. Bow Pod$_{256}$ delivers AI at scale for production deployment in enterprise data centres, as well as private and public clouds.

**Latest generation IPU**

The Bow Pod$_{256}$

## System Specifications

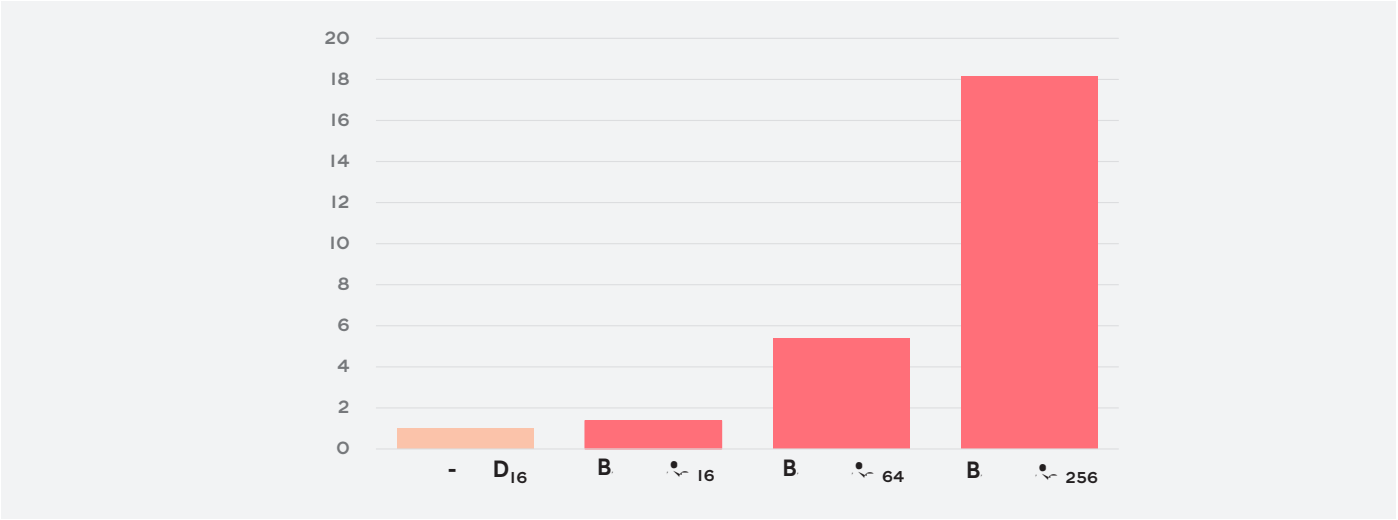
| | 256 Bow IPUs |
|---|---|
| **1U blade units** | 64 Bow-2000 machines |
| **Separate cores** | 376,832 |
| **Threads** | > 2 million |
| **Performance** | 89.6 petaFLOPS FP16.16<br>22.4 petaFLOPS FP32 |
| **Memory** | 230.4 GB In-Processor-Memory™<br>Up to 16,384 GB Streaming Memory™ |
| **Software** | Poplar® SDK |

| | |
|---|---|
| **Host-Link** | 100 GE RoCEv2 |
| **System Weight** | 1800 kg + Host servers and switches |
| **System Dimensions** | 64U + 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$_{256}$

## BOW PERFORMANCE SCALING

B D - DG A C:EFF C E E -B4




**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$_{256}$ supports multiple server configurations.

**Communication architecture built for scaling**

E□icient 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 e□iciency 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 e□ortless 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