

IPU-POD₄ DA



For Innovators making
new AI Breakthroughs



Start exploring new AI horizons

IPU-POD₄ DA (Direct Attach) is the ideal platform for exploration, innovation and development. This lets AI teams make new breakthroughs in machine intelligence. An IPU-M2000, supported by a host server, deliver a powerful 1 petaFLOPS of AI compute for both training and inference workloads in an affordable, compact 2U system.

Plug-and Play with Direct Attach

IPU-POD₄ DA is designed to get you up and running in no time. A turnkey system, it represents a complete and qualified AI computing system ready for installation in your datacenter. Extensive documentation and support is provided both by AI experts at Graphcore and our elite partner network

Start small, scale big

IPU-POD₄ DA is a powerful, standalone AI compute resource. However, it also offers the opportunity for growth, on your terms. Your IPU-POD₄ DA system investment can be expanded later into a larger IPU-POD system.

AI infrastructure built to scale

Designed specifically for the communication requirements of AI workloads at scale, IPU-Fabric is Graphcore's innovative low-latency, jitter-free interconnect using industry standard IT equipment. It supports highly efficient, deterministic, all-to-all IPU interconnect across your system regardless of size.

System Specifications

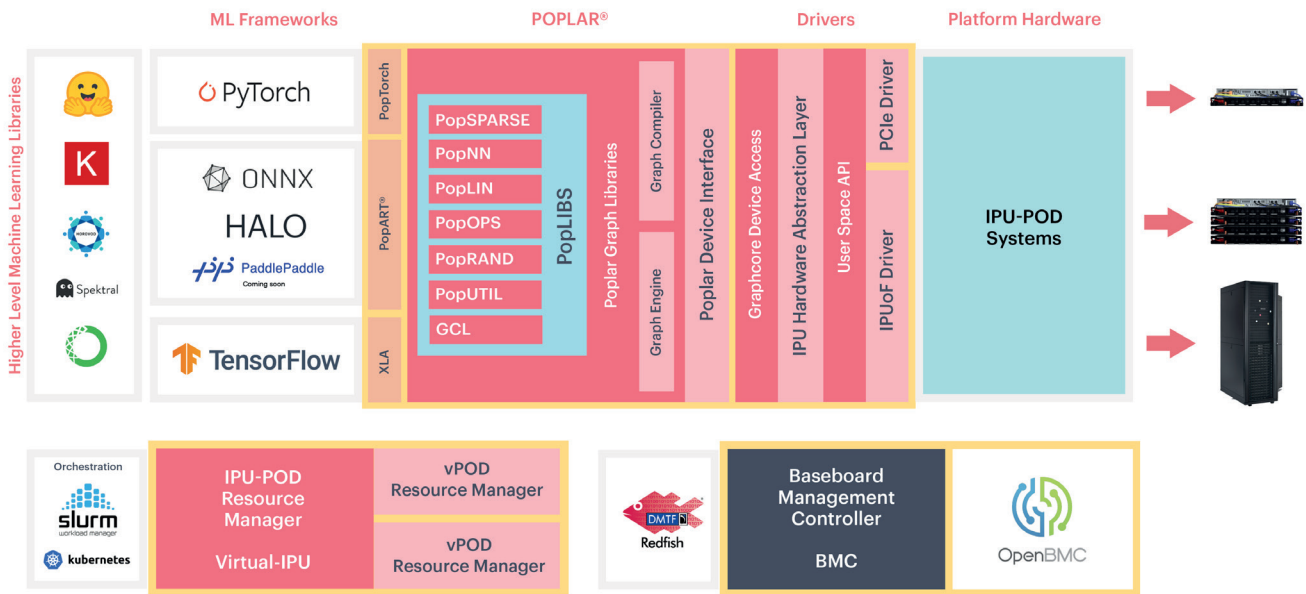
IPUs	4 x GC200 Mk2 IPU
IPU-M2000s	1 x IPU-M2000
Exchange-Memory	131.6GB (includes 3.6GB In-Processor Memory and 128GB Streaming Memory)
Performance	1 petaFLOPS FP16.16 250 teraFLOPS FP32
IPU Cores	5,888
Threads	35,328
IPU-Fabric	2.8Tbps
Host-Link	100 GE RoCEv2
Software	Poplar
System Weight	16.5 kg + Host server
System Dimensions	2U
Host server	Selection of approved host servers from Graphcore partners.
Thermal	Air-Cooled
Optional Switched Version	Contact Graphcore sales

Software First

Fully integrated and IPU-optimised, our 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 hundreds of IPU without adding development complexity, allowing you to focus on the accuracy and performance of your application.

Built for AI developers

TensorFlow, PyTorch and 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, Poplar enables direct IPU programming in C++.



Built for deployment

Pre-built Docker containers with Poplar SDK tools and frameworks images let you get up and running fast. IPU-POD₄ DA has an easy-to-use, intuitive web GUI for simplified management of IPU resources. From here you can manage status, perform system tests, and provision IPU for workloads.

Access to AI expertise

A wealth of experience and support for installation, production deployment 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