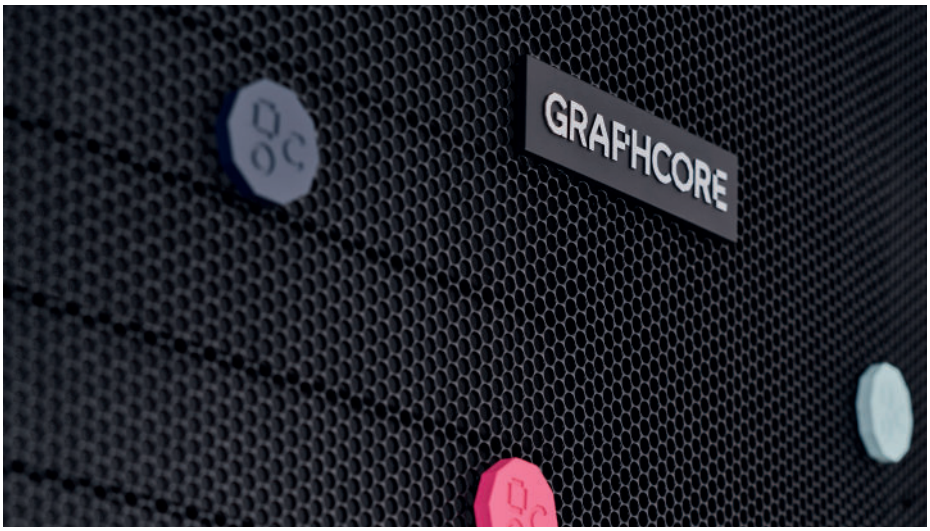
GRAPHCORE

IPU-POD$_{64}$



For Innovators making new AI Breakthroughs

Innovate at massive scale

IPU-POD$_{64}$ is Graphcore's reference architecture for scale-out for a powerful and flexible AI infrastructure design for all your AI training and inference workloads. Built around 16 IPU-M2000s delivering 16 petaFlops of AI compute, IPU-POD$_{64}$ brings together world-class IPU compute with a choice of best in class datacenter technologies and systems in flexible, pre-qualified configurations, to ensure your datacenter is operating with maximum efficiency and performance.

Disaggregated to scale with your needs

AI workloads have very different compute demands. For production deployment, optimizing the ratio of AI to host compute can maximise performance and efficiency, and improve TCO. IPU-POD$_{64}$ is a disaggregated system that separates host servers and switches from IPU-M2000 building blocks in a datacenter. With IPU-POD$_{64}$ you build the system that is ideally matched to your production AI workload.

Unmatched scale-out with IPU-Fabric

IPU-Fabric is Graphcore's innovative low-latency, all-to-all IPU interconnect. Eliminating communication bottlenecks with reliable deterministic performance it is highly efficient whatever your scale.

Data center compatibility

IPU-POD$_{64}$ brings together powerful IPU compute with a choice of best in class data center technologies and systems from leading technology providers in flexible, pre-qualified configurations, to ensure your data center is operating with maximum efficiency and performance, while making your data center AI deployments simpler and faster.

System Specifications

IPUs	64 x GC200 Mk2 IPUs
IPU-M2000s	16 x IPU-M2000
Exchange-Memory	2.1TB (includes 57.6GB In-Processor Memory and 2TB Streaming Memory)
Performance	16 petaFLOPS FP16.16 4 petaFLOPS FP32
IPU Cores	94,208
Threads	565,440
IPU-Fabric	2.8Tbps
Host-Link	100 GE RoCEv2
Software	Poplar
System Weight	450 kg + Host servers and switches
System Dimensions	16U + Host servers and switches
Host server	Selection of approved host servers from Graphcore partners.
Thermal	Air-Cooled
Optional Switched Version	Contact Graphcore sales

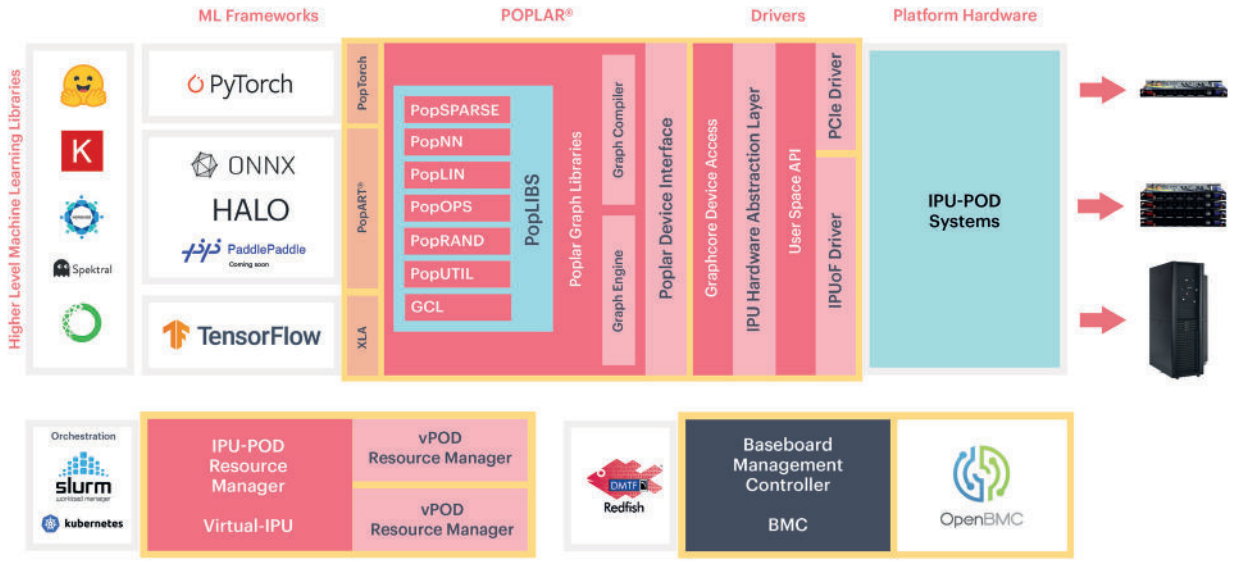
GRAPHCORE

Software First

Fully integrated and IPU-optimised, our Poplar software leverages the unique characteristics of 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 you to focus on the accuracy and performance of your application.

Virtual-IPU

Graphcore's Virtual-IPU technology offers secure provisioning of IPUs to different tenants and workloads. It lets you build model replicas within and across multiple IPU-PODs and to provision multiple IPUs across many IPU-PODs for very large models.



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-POD$_{64}$ 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 IPUs for workloads.

Industry-proven management tools

IPU-POD$_{64}$ is supported with a rich suite of software and tools for management and visualization. These tools are designed with industry-standard open source software and open APIs for straightforward datacenter IT integration including OpenBMC, Redfish DTMF, IPMI over LAN, Prometheus, and Grafana.

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

GRAPHCORE.AI