GRAPHCORE

GROW: IPU-POD$_{128}$

Innovate at scale

IPU-POD™ systems are designed to accelerate large and demanding machine learning models for flexible and efficient scale out.

The IPU-POD$_{128}$ features 32 IPU-M2000™ compute blades, based on the innovative GC200 Intelligence Processing Unit (IPU). The IPU-POD$_{128}$ can deliver up to 32 petaFLOPS of AI compute.

IPU-POD$_{128}$ is the first system to utilize the new IPU-Gateway Links, the horizontal, rack-to-rack connection that extends IPU connectivity across multiple PODs.

The whole system, hardware and software, has been architected together. IPU-POD$_{128}$ supports standard frameworks and protocols to enable smooth integration into existing data center environments. Innovators can focus on deploying their AI workloads at scale, using familiar tools while benefitting from cutting-edge performance.

Disaggregation for customised compute

Machine intelligence workloads have very different compute demands. For production deployment, optimizing the ratio of AI to host compute can help to maximize performance, while improving total cost of ownership. IPU-POD systems allow flexible mapping of the number of servers and switches to the requisite number of IPU-M2000™ platforms, so deployment is better tailored to production AI workloads. IPU-POD$_{128}$ supports multiple server configurations.

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 IPUs, across IPU-M2000s, between IPU-PODs and throughout the data center. IPU-Fabric delivers high-performance low-latency communication to maximize AI application efficiency and is built to work with standard data center communication technologies.

System Specifications

IPUs	128 x GC200 IPUs
IPU-M2000s	32 x IPU-M2000
IPU Cores	188,416
Threads	1,130,496
Performance	32 petaFLOPS FP16.16 8 petaFLOPS FP32
Exchange-Memory	Up to 8.3TB (includes 115.2GB In-Processor Memory and 8.2TB Streaming Memory)
IPU-Fabric	2.8Tbps
Host-Link	100 GE RoCEv2
Software	Poplar® SDK
System Weight	900 kg + Host servers and switches
System Dimensions	32U + 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
Optional Switched Version	Contact Graphcore sales

GRAPHCORE

Built for AI developers

IPU-POD systems support industry-standard software tools. Developers can work with frameworks such as TensorFlow, PyTorch, PyTorch Lightning and Keras, as well as open standards like ONNX and model libraries like Hugging Face.

For deeper control and maximum performance, the Poplar framework enables direct IPU programming in Python and C++. Poplar allows effortless scaling of models across many IPUs without adding development complexity, so developers can focus on the accuracy and performance of their application.

At Graphcore we put power in the hands of AI developers allowing them to innovate. Our software stack supports industry open standards and much of it is open source.

Access to AI expertise

Graphcore has a global network of partners to assist users of IPU-PODs all the way from installation and application development through to production deployment. For documentation and other help, visit our website.

Straightforward deployment

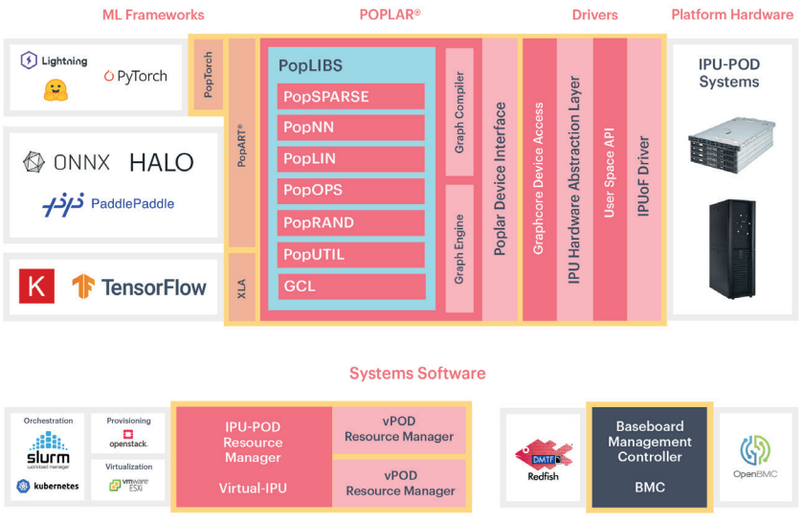
Ease of deployment has been a paramount consideration in designing the IPU-POD. The result is a solution that supports standard hardware and software interfaces and protocols, and integrates effectively with existing data center infrastructures.

IPU-PODs support a rich suite of software and tools for management and visualization based on industry-standard open source software and open APIs including OpenBMC, Redfish DTMF, IPMI over LAN, Prometheus, and Grafana.

Industry-proven management tools

Docker and Kubernetes support makes it simple to automate application deployment, scaling, and management of IPU-PODs. Virtual-IPU™ technology offers secure provisioning of IPUs to different tenants and workloads. Developers can build model replicas within and across multiple IPU-PODs and provision IPUs across many IPU-PODs for very large models.

IPU-PODs have an easy-to-use, intuitive web GUI for simplified IPU resource management. Engineers can manage status, perform system tests, and provision IPUs for workloads. IPU-PODs also integrate with a variety of cloud provisioning and management stacks, including VMWare's Radium.



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