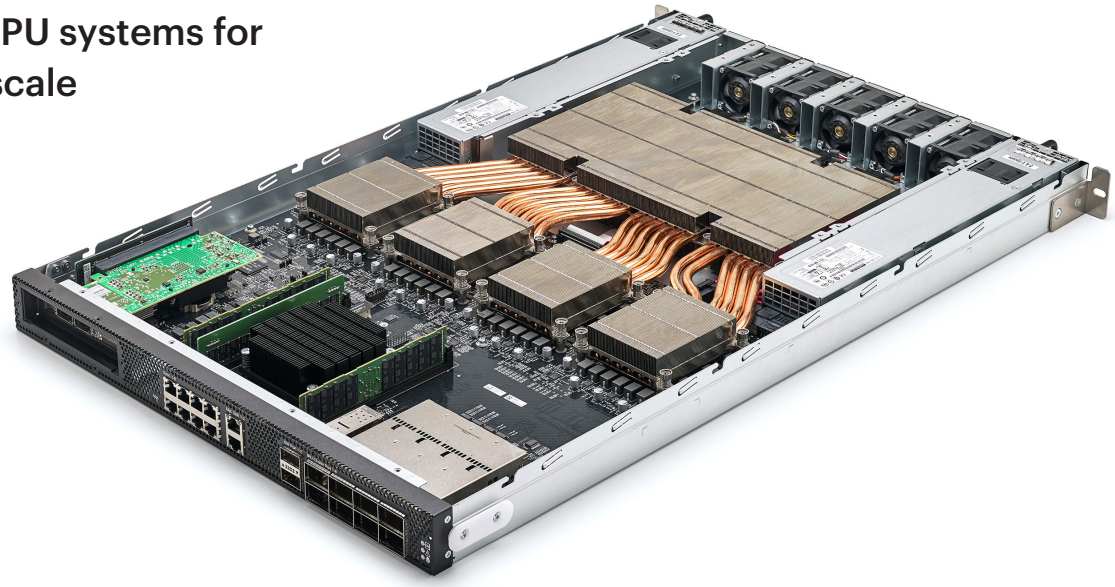
GRAPHCORE

# IPU-MACHINE: M2000

## Second generation IPU systems for AI infrastructure at scale



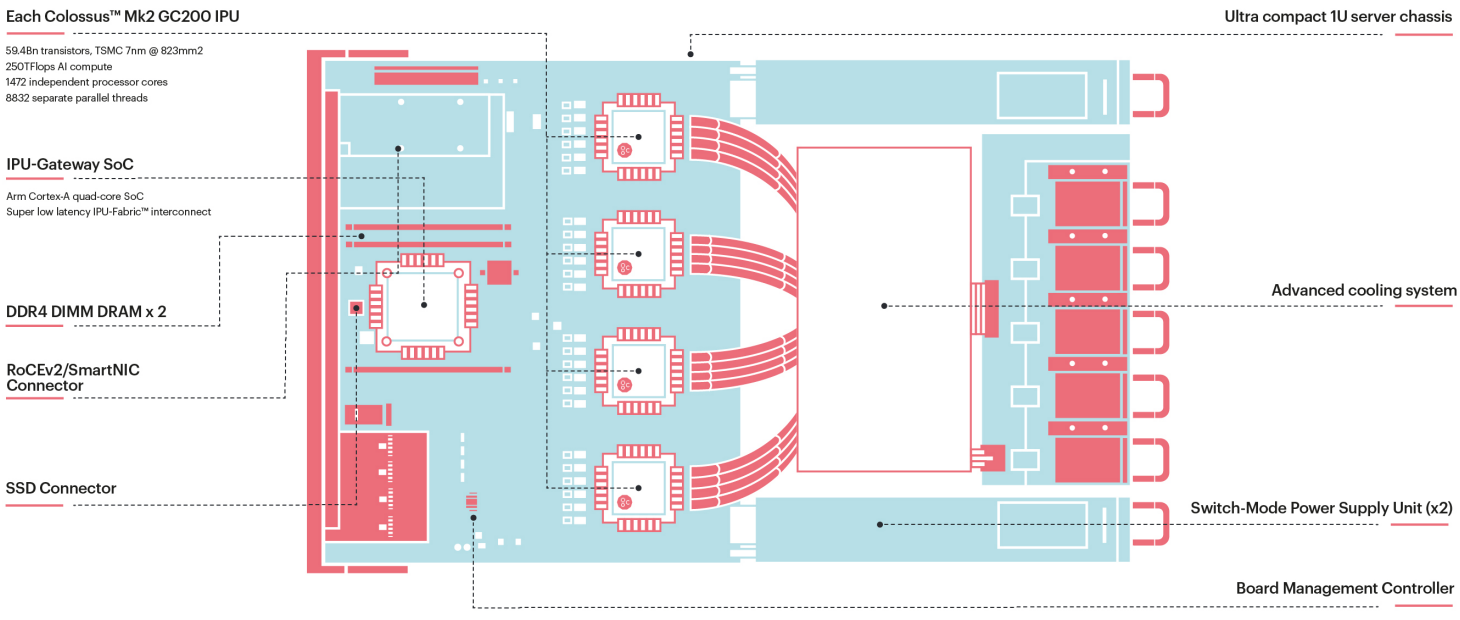
A core, new building block for AI infrastructure, the IPU-M2000 is powered by 4 x Colossus Mk2 GC200, Graphcore's second generation 7nm IPU. It packs 1 PetaFlop of AI compute, up to 450GB Exchange Memory and 2.8Tbps IPU-Fabric for super low latency communication, in a slim 1U blade to handle the most demanding of machine intelligence workloads.

The IPU-M2000 has a flexible, modular design, so you can start with one and scale to thousands. It works as a standalone system, eight can be stacked together or racks of 16 tightly interconnected IPU-M2000's in IPU-POD$_{64}$ systems can grow to supercomputing scale thanks to 2.8Tbps high-bandwidth, near-zero latency IPU-Fabric™ interconnect architecture, built into the box.

Designed from the ground up for high performance training and inference workloads, the IPU-M2000 unifies your AI infrastructure for maximum datacentre utilization. Get started with development and experimentation then ramp to full scale production. Available to pre-order today.

### IPU-Machine: M2000

4 x Colossus™ Mk2 GC200 IPU
1 PetaFlops AI compute
Up to 450GB Exchange Memory™
2.8Tbps IPU-Fabric™



GRAPHCORE

## Poplar Software

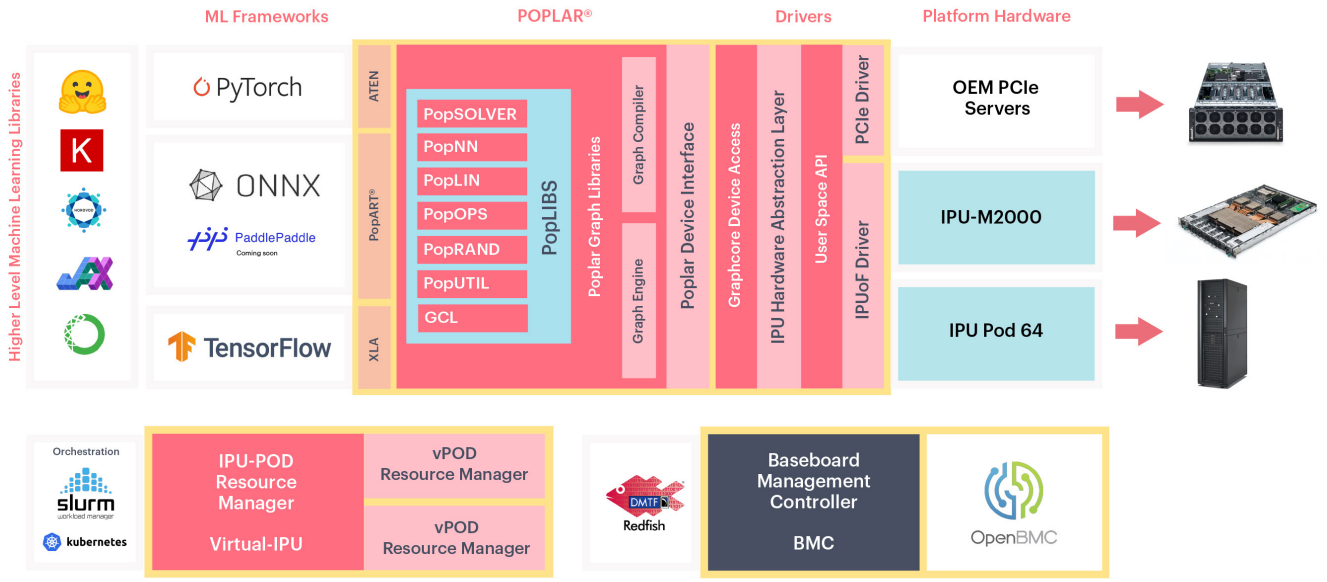
With Poplar, managing IPUs at scale is as simple as programming a single device, allowing you to focus on the data and the results.

Our state of the art compiler simplifies IPU programming by handling all the scheduling and work partitioning of large models, including memory control, while the Graph Engine builds the runtime to execute your workload efficiently across as many IPUs, IPU-Machines or IPU-PODs as you have available.

As well as running big models across large IPU configurations, we've made it possible to dynamically share your AI compute, with Graphcore's Virtual IPU software. You can have tens, hundreds, even thousands of IPUs working together on model training. At the same time, you can allocate your remaining IPU-M2000 machines for inference and production deployment.

Poplar supports standard ML frameworks including TensorFlow, PyTorch, ONNX and PaddlePaddle as well as industry standard converged infrastructure management tools for so it's easy to deploy, including Open BMC, Redfish, Docker containers, and orchestration with Slurm and Kubernetes. And, we're adding support for more platforms all the time.

With direct access to expert support from Graphcore AI engineers, you will be up and running fast.




## IPU-M2000 Key Features™

### Compute

- 4 x Colossus™ Mk2 GC200 IPU
- 1 PetaFlop AI compute
- 5888 independent processor cores

### Memory

- Up to 450GB Exchange Memory™
- 180TB/s Exchange Memory™ bandwidth

### Communications

- 2.8Tbps ultra-low latency IPU-Fabric™
- Direct connect or via Ethernet switches
- Collectives and all-reduce operations support

### IPU Gateway SoC

- Arm Cortex quad-core A-series SoC

### Form Factor

- Industry standard 1U

### Software

- Poplar SDK
- PopVision visualization and analysis tools

### Converged Infrastructure Support

- Virtual-IPU comprehensive virtualization and workload manager support
- Support for SLURM workload manager
- Support for Kubernetes orchestration
- OpenBMC management built-in
- Grafana system monitoring tool interface

## Ready to get started?

Connect with our experts to assess your AI infrastructure requirements and solution fit by contacting info@graphcore.ai

GRAPHCORE.AI