THE BOW-2000 IPU MACHINE

The Bow-2000 IPU Machine is the building block for machine intelligence infrastructure at scale, delivering unprecedented performance and efficiency.

Based on the new Bow IPU processor, this remarkably efficient system delivers up to 1.4 petaFLOPS of AI performance in a flexible and practical modular design, a standard 1U server blade. A scalable compute and memory communication architecture enables these units to be connected in large numbers to populate data centres and expand to supercomputing scale.

Each Bow-2000 is powered by four Bow IPU processors - the first processor in the world to be manufactured using a revolutionary Wafer-on-Wafer (WoW) silicon production technology. This technology enables much more efficient power delivery, higher operating frequency and enhanced overall performance.

Bow-2000 is designed to make both training and inference of large and emerging machine learning models faster and more efficient. It is designed to support the current and future requirements of machine intelligence innovators.

The Bow-2000 IPU Machine

4x Bow IPU processors
1.4 petaFLOPS AI compute
3.6 GB In-Processor-Memory @ 260TB/s
Up to 256 GB Streaming Memory
2.8 Tbps IPU-Fabric™

Each Bow IPU Processor

World's first Wafer-On-Wafer processor
350 teraFLOPS AI Compute
0.9 GB In-Processor-Memory @ 65 TB/s
1,472 independent processor cores
8,832 independent parallel programs
10x IPU-Links™ delivering 320GB/s

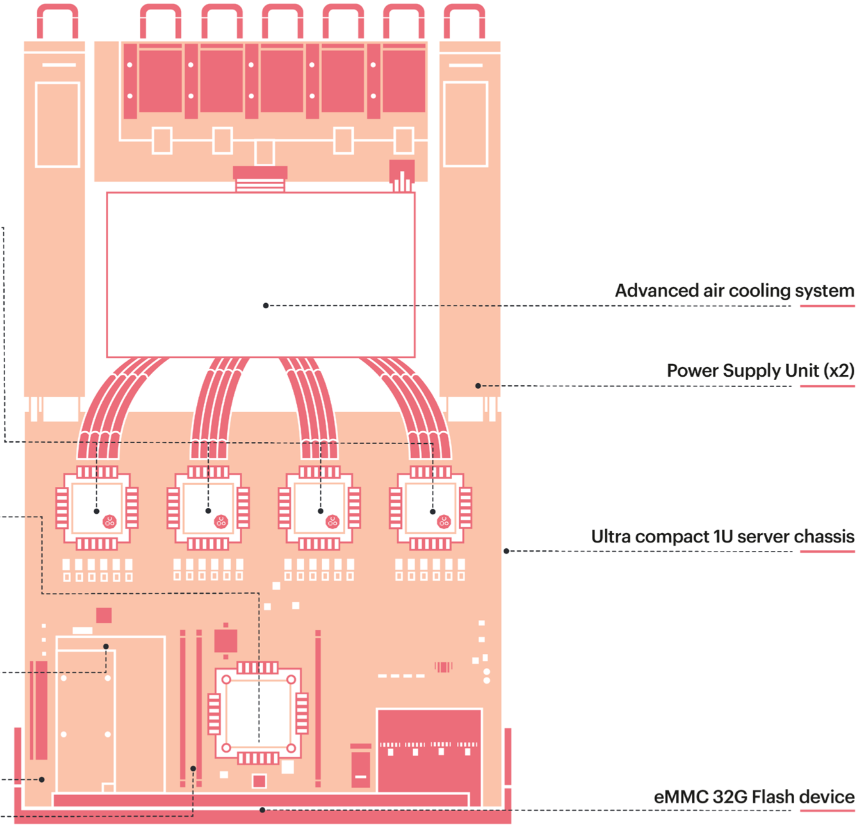
IPU-Gateway SoC

Arm Cortex-A quad-core SoC
Super low latency IPU-Fabric interconnect

Board Management Controller

PCIe FH3/4L G4x8 Slot (RNIC/SmartNIC)

DDR4 DIMM DRAM x 2



Advanced air cooling system

Power Supply Unit (x2)

Ultra compact 1U server chassis

eMMC 32G Flash device

Software first for ease of deployment

Bow-2000 supports industry standard converged infrastructure management tools such as Open BMC, Redfish, Docker containers, and orchestration with Slurm and Kubernetes, as well as many popular machine learning frameworks, including TensorFlow, PyTorch, PaddlePaddle, ONNX and much more. Additional platforms are added on a regular basis.

Developers, system integrators and data center engineers are therefore able to use the tools they are already familiar with and focus on fast innovation and deployment.

Poplar® Software

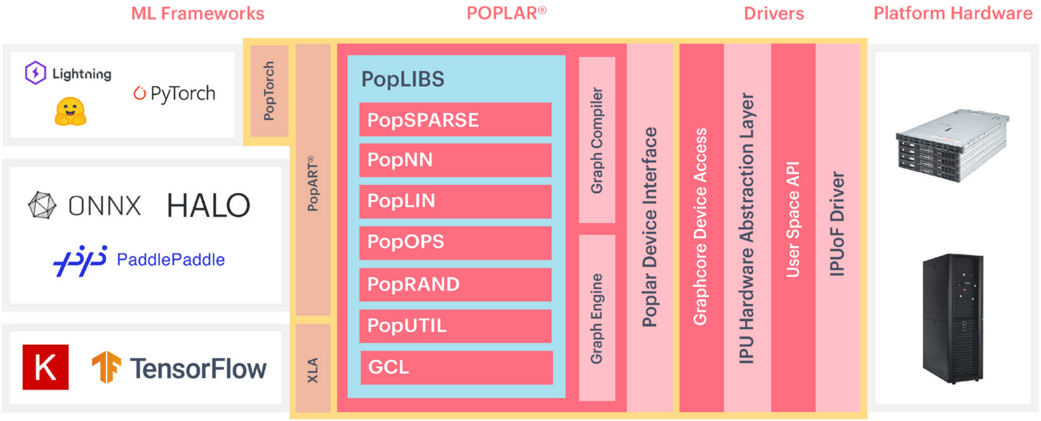
The Poplar SDK is a comprehensive software stack which was developed alongside the IPU to enable innovators to directly access and benefit from it. Poplar makes the management of IPUs at scale as simple as programming a single device, allowing the user to focus on the data and the results.

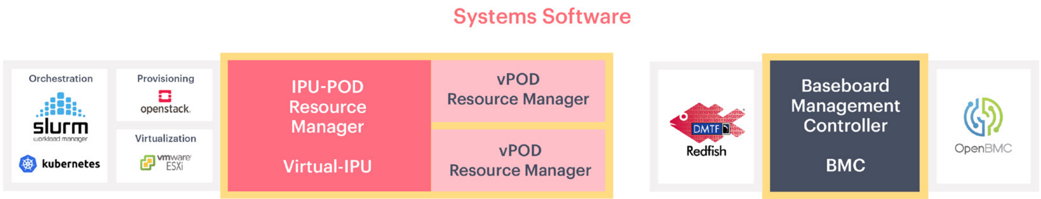
A state-of-the-art compiler simplifies IPU programming by handling all the scheduling and work partitioning of large models, including memory control. The Graph Engine builds the runtime to execute workloads efficiently across all available IPU processors, blades and Pods.

Along with running large models across sizeable IPU-based systems, it is possible to dynamically share workloads via the Virtual IPU software. While thousands of Bow-2000 machines in the system can work together on large model training, simultaneously the remaining machines can be allocated for inference and production deployment.

More help at hand

With direct access to expert support from Graphcore AI engineers, and availability of readily optimised models in our model garden, you will be up and running in no time.





Key Features™

Compute

- 4 x Bow IPUs
- 1.4 petaFLOPS of AI performance
- 5888 independent processor cores

Memory

- Up to ~260GB of memory comprised of:
  - Up to 256GB Streaming Memory™
  - 3.6GB In-Processor-Memory™

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
- Supporting all major ML frameworks

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 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