GRAPHCORE

IPU-M2000 SCALE-OUT: IPU-POD$_{64}$

Datasheet

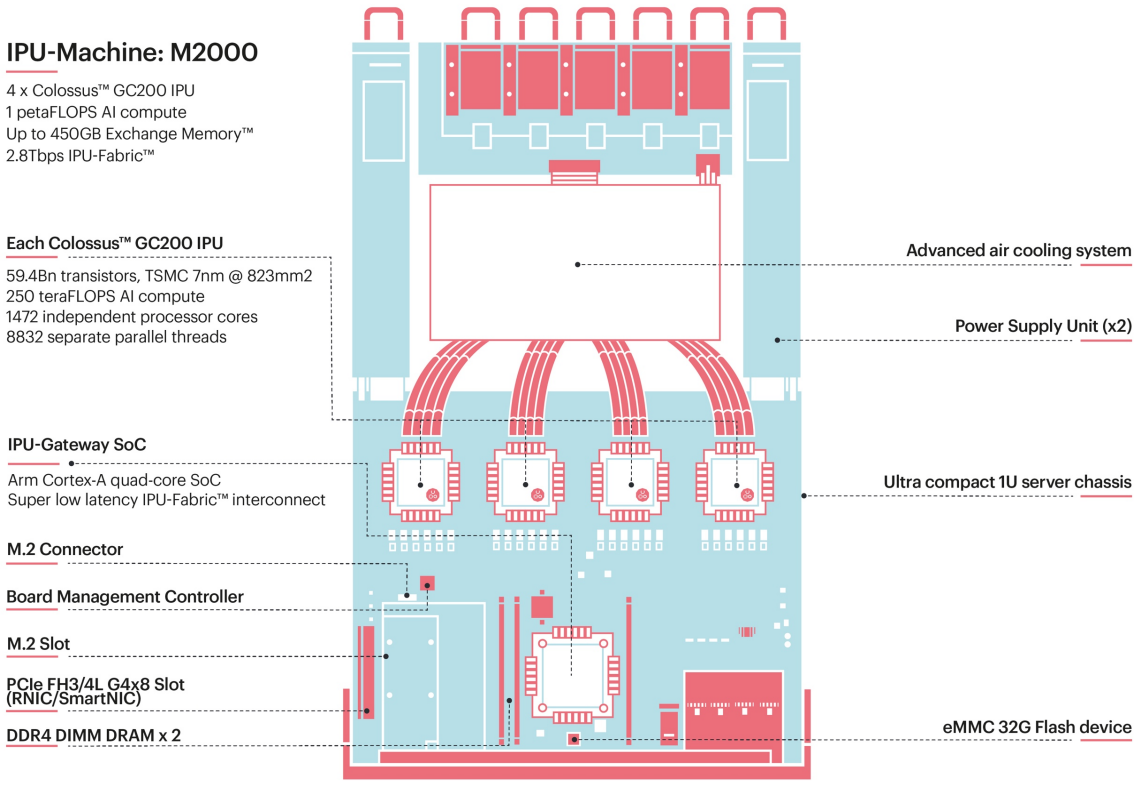
Table of contents

1 Overview

Graphcore's IPU-M2000 is designed to support scale-up and scale-out for exascale machine intelligence compute. The IPU-POD reference designs, based on the IPU-M2000, deliver scalable building blocks for the massive levels of compute in next generation machine intelligence workloads.

The IPU-POD reference design is currently available in an IPU-POD$_{16}$ configuration with 4 x IPU-M2000s, as well as an IPU-POD$_{64}$ configuration with 16 x IPU-M2000s. IPU-POD$_{64}$ racks can be scaled for systems ranging from 64 to 64K IPU processors in switched or direct 3D torus IPU-Fabric™ configurations.

Other configurations such as IPU-POD$_{32}$ and larger scale-out systems (IPU-POD$_{128}$ and IPU-POD$_{256}$) will be available in 2021 – please contact Graphcore sales for more information.



2 Product description

2.1 IPU-POD$_{64}$ reference design for scale-out

Graphcore's reference design assembles sixteen IPU-M2000s into an IPU-POD$_{64}$ delivering 16 petaFLOPS of AI compute. The IPU-POD$_{64}^{TM}$ can be used individually (64 IPU processors) or as a building block for larger systems with up to 64,000 IPU processors delivering 16 exaFLOPS of AI compute. Virtualization and provisioning software allow the AI compute resources to be elastically allocated to users and be grouped for both model-parallel and data-parallel AI compute. The IPU-POD$_{64}$ reference design combines the sixteen IPU-M2000s with network switches and a host server in a pre-qualified rack configuration (switches and host server not provided by Graphcore). The pre-qualified IPU-POD$_{64}$ system assumes the following default components:

- 1 – 4 Dell R6525 host server(s) with dual-socket AMD Epyc2 CPUs (more server options to be qualified for future versions). Default number of servers is 1, however up to 4 host servers are supported depending on workload - please speak to Graphcore sales
- 1 Arista 7060X ToR switch (32x100G + 2 10G)
- 1 Arista 7010T management switch (48p 1G+ 4x1/10G)

The Dell R6525 1U server is the default server which is fully qualified by Graphcore as a host for IPU-POD$_{64}$ systems. Full configuration details can be found in the "Approved servers" document (https://docs.graphcore.ai/projects/graphcore-approved-server-list/). Please contact Graphcore sales for alternative offerings from our reseller and OEM partners.

The IPU-POD$_{64}$ is characterized by the following features:

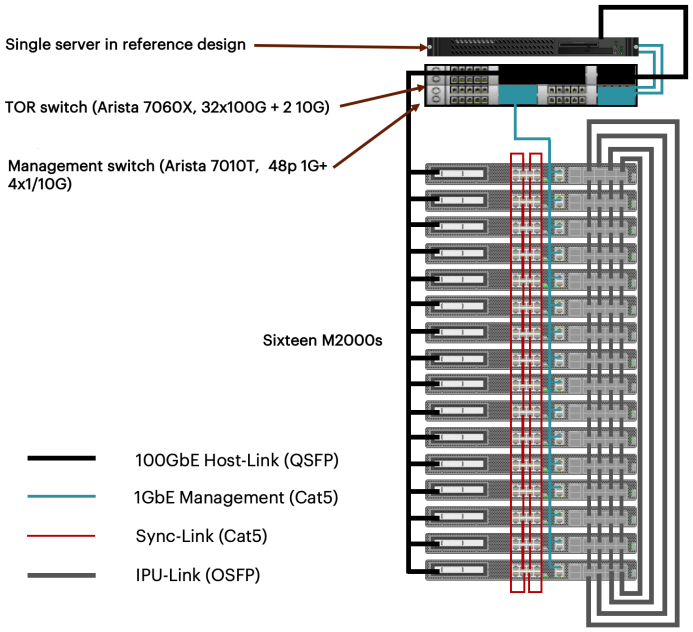
- Disaggregated host architecture allows for different server requirements based on workload
- 16 petaFLOPS (FP16.16) of AI compute, 4 petaFLOPS @ FP32, and up to 7.23 TBytes of Exchange Memory
- 2D-torus IPU-Link topology
- Scalable to 1,024 IPU-POD$_{64}$ supporting 65,536 GC200 IPU processors

A high-level view of the IPU-POD$_{64}$ cabling is shown in the figure below.

IPU-POD$_{64}$ REFERENCE DESIGN

IPU-POD$_{64}$ with default options for host server and switches

- Sixteen IPU-M2000 platforms
- Reference architecture supports different server requirements based on workload
- IPU-POD$_{64}$ configuration:
  - 64 IPUs
  - 16 PFLOPs @ FP16.16
  - ~58GB IPU In-Processor Memory
  - ~7TB Streaming Memory
- IPU-POD host disaggregation
  - Flexibly connect required host server compute over fabric
- 2D-torus topology
  - Maximizes bandwidth across IPU-Links
  - All-Reduce 2x faster than mesh topology
- Scalable to 64K GC200 IPUs



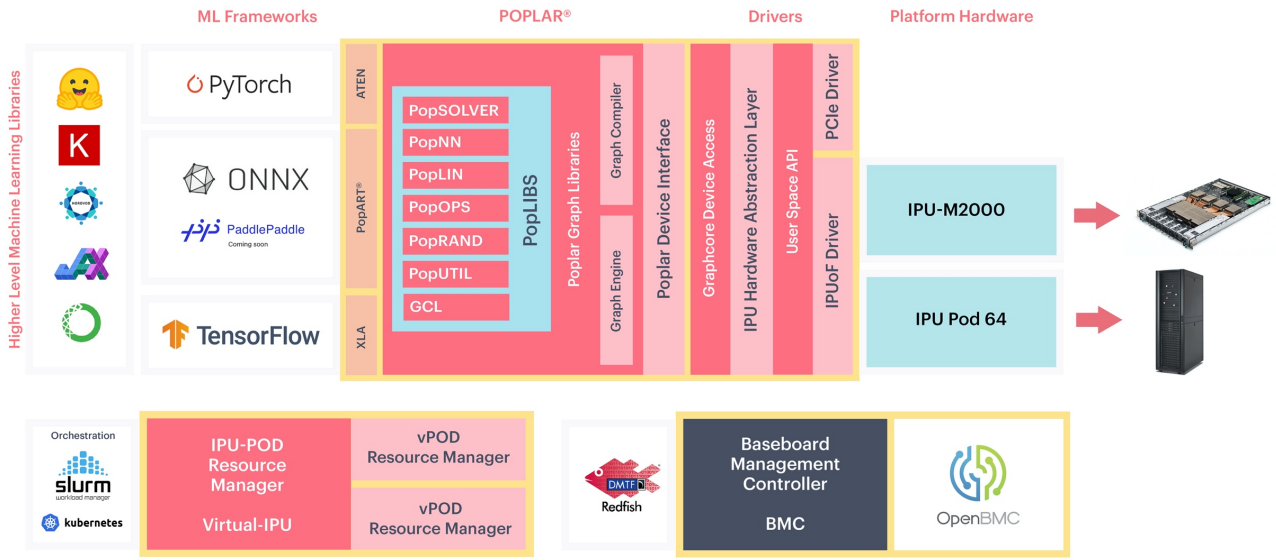
The IPU-POD$_{64}$ reference design is available as a full implementation through Graphcore's network of reseller and OEM partners.

Alternatively, customers may directly implement the IPU-POD$_{64}$ reference design with the help of the "IPU-POD$_{64}$ build and test guide" available from the Graphcore documents page. The associated "IPU-POD$_{64}$ installation and integration guide" provides more details about IPU-POD$_{64}$ power, thermal characteristics and data centre implementation requirements.

2.2 Software

IPU-M2000s are fully supported by Graphcore's Poplar® software development environment, providing a complete and mature platform for ML development and deployment. Standard ML frameworks including TensorFlow, ONNX, and PyTorch are fully supported along with access to PopLibs through our Poplar C++ API. Note that PopLibs, PopART and TensorFlow are available as open source in the Graphcore GitHub repo https://github.com/graphcore. PopTorch provides a simple wrapper around PyTorch programs to enable the programs to run seamlessly on IPUs. The Poplar SDK also includes the PopVision™ visualisation and analysis tools which provide performance monitoring for IPUs - the graphical analysis enables detailed inspection of all processing activities.

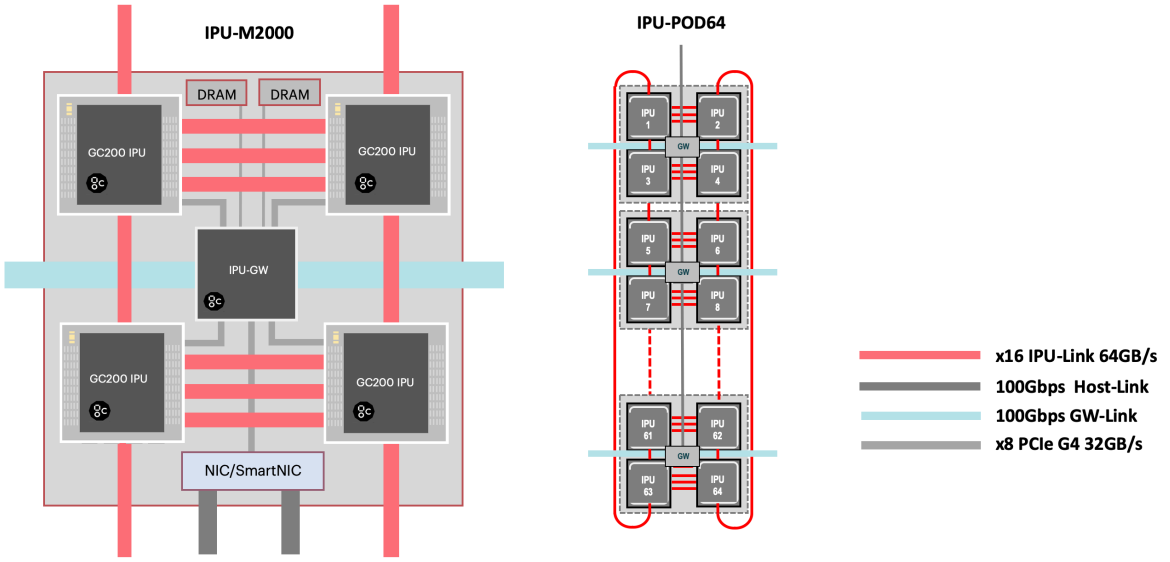
In addition to these Poplar development tools, the IPU-POD$_{64}$ is enabled with software support for industry standard converged infrastructure management tools including OpenBMC, Redfish, Docker containers, and orchestration with Slurm and Kubernetes.



Poplar SDK	Complete end-to-end software stack for developing, deploying and monitoring AI model training jobs as well as inference applications on the Graphcore IPU
ML frameworks	ONNX, TensorFlow, and PyTorch with support for Exchange Memory
Deployment options	Bare metal (Linux), VM (HyperV), containers (Docker) Supported operating systems are Ubuntu 18.04 and CentOS 7.6 For other OS options please contact sales
Host-Links	RDMA based disaggregation between a host and IPU over 100Gbps RoCEv2 NIC, using the IPU over Fabric (IPUoF) protocol Host-to-IPU ratios supported: 1:1 up to 1:64
Graphcore Communication Library (GCL)	IPU-optimized communication and collective library integrated with the Poplar SDK stack. Support all-reduce (sum,max), all-gather, reduce, broadcast Scale at near linear performance to 64k IPUs
PopVision	Visualization and analysis tools
Graphcore Virtual IPU SW	IPU-M2000 and IPU-POD$_{64}$ resource manager IPU-Fabric topology discovery and validation
Provisioning	REST API and SSH/CLI for IPU allocation / de-allocation into isolated domains (vPODs) Plug-ins for SLURM and Kubernetes (K8)
Resource monitoring	REST API and SSH/CLI for accessing the IPU-M2000 monitoring service Prometheus node exporter and Grafana (visualization) support
Lights out management	Baseboard Management Controller (OpenBMC) Dual-image firmware with local rollback support Console support, CLI/SSH based Serial-over-Lan and Redfish REST API

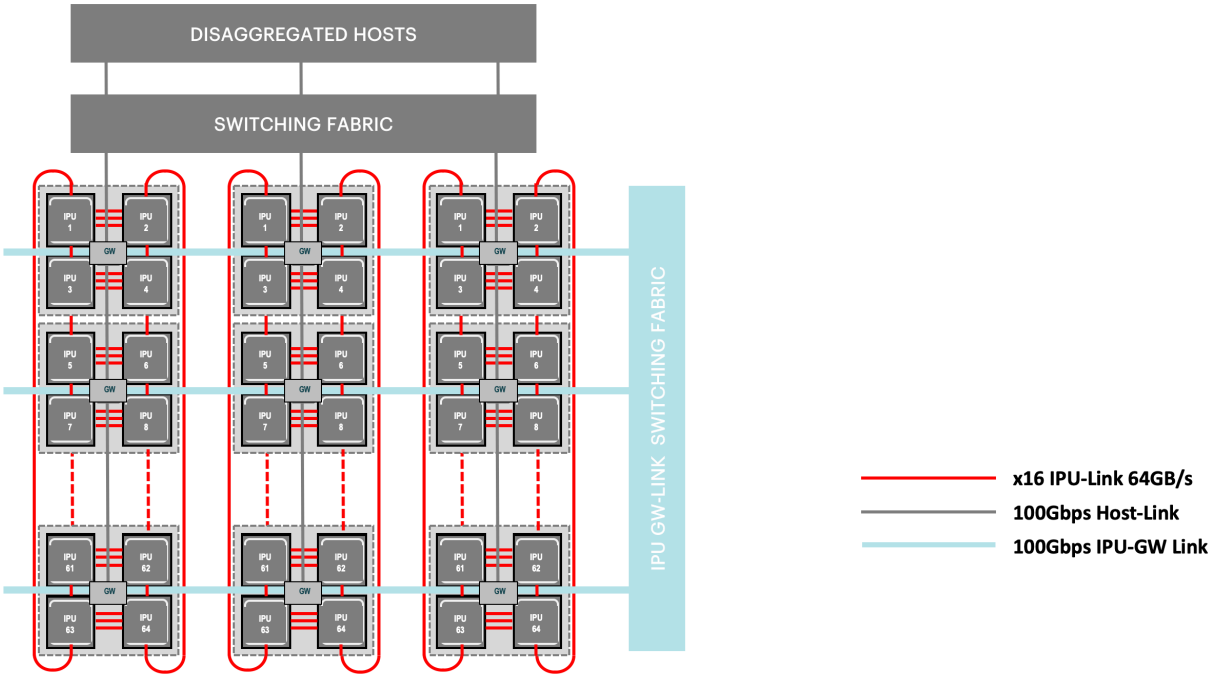
2.3 Communication for scale-out: 3D IPU-Fabric with GCL

The IPU-POD$_{64}$ reference design builds on the innovative IPU-M2000 IPU-Fabric, designed to support massive scale out. The figure below shows, on the left, an abstracted view of the IPU-M2000 with the IPU-Fabric interconnects comprising IPU-Links™, GW-Links (for jitter-free IPU-to-IPU connectivity), and the Host-Link dual 100Gbps RDMA connection between the host server and each IPU-M2000. The small insert on the right shows how these interconnects are used as part of the scale-out of IPU-M2000 and the IPU-POD$_{64}$: IPU-Links join IPU processors together both within IPU-M2000s as well as between IPU-M2000s. The IPU-Link connections in the IPU-POD$_{64}$ form a 2D torus since the loops are closed top and bottom.



2.4 Host-Links and GW-Links

Host servers are disaggregated from the IPU-M2000 with Host-Links – Graphcore´s low-latency high throughout host-IPU RDMA transport using RoCEv2 . The 100Gbps Ethernet links from the RoCEv2 NICs in the IPU-M2000 connect through switches to the disaggregated host servers. The disaggregated host architecture for the IPU-POD$_{64}$ enables user-defined host:IPU ratios and allows for scalable host server utilisation depending on the machine intelligence workload. The GW-Links are used to connect multiple IPU-POD$_{64}$ either directly or through an optional ethernet switching fabric as shown below.



2.5 Technical specifications

IPU processors IPU-Cores™ Worker Threads In-Processor-Memory	64 Colossus GC200 IPU processors 94,208 565,248 57.6GB
AI compute	16 petaFLOPS AI (FP16.16) compute 4 petaFLOPS FP32 compute
Exchange Memory Default Optional (Contact sales)	 Up to 2,105.6GB (includes 57.6GB In-Processor Memory (16x 3.6GB per IPU-M2000) and 2048GB Streaming Memory (16 x 64GB DIMM x2 per IPU-M2000) Up to 4153.6GB (Includes 57.6GB In-Processor Memory (16x 3.6GB per IPU-M2000) & 4096GB Streaming Memory (16 x 128GB DIMM x2 per IPU-M2000) Up to 7225.6GB (Includes 57.6GB In-Processor Memory (16x 3.6GB per IPU-M2000) & up to 7168GB Streaming Memory (16 x 256GB DIMM x2 per IPU-M2000)
IPU-Fabric	
IPU-Links	32Tbps (64 x16 Gen4) aggregated bi-directional bandwidth for direct, 2D torus intra-rack IPU-POD$_{64}$ IPU connectivity 16 IPU-M2000s directly connected 128 standard OSFP ports with DAC cabling
GW-Links	6.4TBps (32 x 100Gbps) for inter-rack IPU-POD$_{64}$ connectivity Up-to 1024 IPU-POD$_{64}$ (direct) or 256 IPU-POD$_{64}$ (switched) can be connected Standard 100Gbps QSFP28 ports supporting industry standard transceivers (100G-DR) and DAC cabling
Internal SSDs	16 x 1TB M.2 SSD for program and data store
IPU-POD$_{64}$ internal host server(s) Default Options	 1 x Dell PowerEdge R6525 server 1 – 4 Graphcore approved server/OS options. Contact Graphcore sales for details
IPU-POD$_{64}$ internal switches Default	 1 x Arista DCS-7060CX-32S-F (100GbE ToR switch) 1 x Arista DCS-7010T-48-F (1GbE Management switch)
IPU-POD$_{64}$ internal server(s) to IPU-M2000 connectivity	1 (default) to 4 host servers have connectivity to the 16 IPU-M2000s via the IPU-POD$_{64}$ 100GbE ToR switch (Arista DCS-7060CX-32S-F) 1x 100GbE port per IPU-M2000 to connect to the ToR switch Dual (2x) 100GbE ports per host server to connect to the ToR switch
IPU-POD$_{64}$ internal management network connectivity	Aggregated in the Arista DCS-7010T-48-F 1GbE management switch are: 2x 1GbE RJ45 management ports from each of the 16 IPU-M2000s Server(s) management port(s) PDU monitoring port

IPU-POD$_{64}$ thermal	Air cooled with built-in N+1 hot-plug fan cooling system in each of the individual components (IPU-M2000s, servers and switches)
Rack airflow	All IPU-POD$_{64}$ components (IPU-M2000s, server(s) and switches) are mounted for airflow direction front of rack (single door, cold aisle side) to back of rack (split door, hot aisle side)
Airflow rate	1750 CFM

IPU-POD$_{64}$ rack	
Rack	42U - 600mm (W) X 1200mm (D) x 1991mm (H)
Weight	450 kg (943 lbs)
PDU	Redundant 22kW PDUs (APC Metered Rack PDU AP8886) PDU input: IEC 60309 32 A 3P + N + E PDU whip (cord) length: 1.8 meter (exit top of IPU-POD$_{64}$ rack)
Input power (V_{ac})	100V/110V/120V/200 - 240 V_{ac} (115 - 230 V_{ac} nominal)
Power (nominal)	19kW

For information on IPU-POD$_{64}$ integration with datacentre infrastructure, please contact Graphcore sales.

2.6 Environmental characteristics

Operating temperature and humidity (inlet air)	10-32^{0}C (50 to 90^{0}F) at 20%-80% RH (*)
Operating altitude	0 to 3,048m (0-10,000ft) (**)

(*) Altitude less than 900m/3000ft and non-condensing environment

(**) Max. ambient temperature is de-rated by 1°C per 300m above 900m

2.7 Standards compliance for IPU-M2000s

EMC standards	Emissions: FCC CFR 47, ICES-003, EN55032, EN61000-3-2, EN61000-3-3, VCCI 32-1 Immunity: EN55024, EN61000-4-2, EN61000-4-3, EN61000-4-4, EN61000-4-5, EN61000-4-6, EN61000-4-8, EN61000-4-11
Safety standards	IEC62368, IEC60950
Certifications	North America (FCC), Europe (CE), UK (UKCA), Australia (RCM), Taiwan (BSMI), Japan (VCCI), South Korea (KC), China (CQC) CB-62368, CB-60950
Environmental standards	EU 2011/65/EU RoHS Directive, XVII REACH 1907/2006, 2012/19/EU WEEE Directive

3 Document details

3.1 Revision history

This document's revision history is as follows:

Version	Date	Notes
1.0	2nd of December 2020	First release

The European Directive 2012/19/EU on Waste Electrical and Electronic Equipment (WEEE) states that these appliances should not be disposed of as part of the routine solid urban waste cycle, but collected separately in order to optimise the recovery and recycling flow of the materials they contain, while also preventing potential damage to human health and the environment arising from the presence of potentially hazardous substances.

The crossed-out bin symbol is printed on all products as a reminder.

Waste may be taken to special collection site or can be delivered free of charge to the dealer when purchasing a new equivalent or without obligation to make a new purchase for equipment smaller than 25cm.

For more information on proper disposal of these devices, kindly refer to the public utility service.