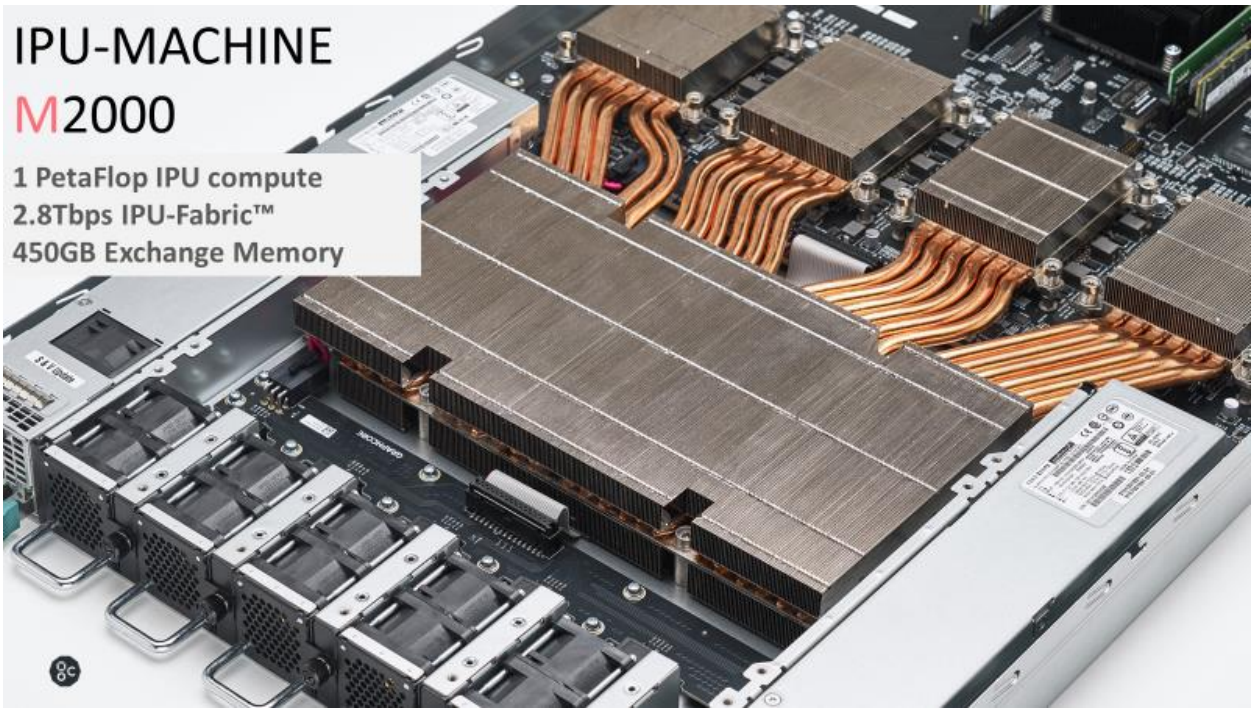
# THE GRAPHCORE SECOND-GENERATION IPU

## INTRODUCTION

Graphcore, the U.K.-based startup that launched the Intelligence Processing Unit (IPU) for AI acceleration in 2018, has introduced the IPU-Machine. This second-generation platform has greater processing power, more memory and built-in scalability for handling extremely large parallel processing workloads. The well-funded startup has a blue-ribbon pedigree of engineers, advisers and investors, and enjoys a valuation approaching $2 billion. Its first-generation hardware is now available on the Microsoft Azure cloud as well as in Dell-EMC servers. Both companies are investors. Graphcore is now betting its future on this second-generation platform, a plug-and-play building block for massive scalability that is currently unique in the industry.

## FIGURE 1: THE GRAPHCORE IPU-MACHINE




**The Graphcore IPU-Machine includes four IPU's, integrated 100GbE scale out fabric, PCIe and additional DDR memory. Each 1U appliance can deliver up to a petaflop of AI performance and 450GB of memory.**

*Source: Graphcore*

Supported by the company's Poplar development stack, the new platform highlights the 7nm Colossus MK2 in a four-way IPU-Machine appliance, and can scale up to 64,000

IPUs across 1024 racks. The fully configured AI supercomputer can deliver some 16 exaflops of AI (16-bit FP) performance. As with the first generation, the company's focus is to simplify high-performance parallel computing at significant scale.

This research paper will explore the new platform and assess its strengths and weaknesses compared to the growing cadre of potential competitors.

## THE COLOSSUS MK2 IPU (GC200)

The new MK2 part, manufactured by TSMC, is a massively parallel 59.4 B transistor processor. It delivers some 250 Trillion Operations per Second (TOPS) across 1,472 cores and 900MB of In-Processor Memory interconnected across a 2.8Tb/s low-latency fabric. Most of the architectural design of the MK1 generation carries over to the MK2 platform, with processing tiles containing cores and on-die SRAM, interconnected over the same fabric that can extend off-die to communicate with other IPU domains.

In shrinking the original IPU design to 7nm, the Graphcore designers opted for performance and memory maximization instead of cost reduction. This is consistent with the generally held observation that AI applications remain performance-limited, not cost-sensitive in data center training applications, while edge AI inference processing is far more cost- and power-sensitive. Consequently, the MK2 provides 20% more cores, 3X more on-die SRAM and 16X more scalability than its predecessor. New system software enables enhanced scalability, deployment and management.

At a system level, the on-die IPU memory is now supplemented by up to 448GB of "streaming memory" DRAM. The MK2 IPU also gets a performance boost from a set of novel floating-point implementation techniques developed by Graphcore, called AI-Float, used to tune energy and performance for AI computation. Using the standard IEEE FP16 format, AI-Float is optimized in several ways. Its stochastic rounding enables FP16 to match FP32 performance on master weights and FP16.16 to match FP16.32 for forward and backward propagation providing 250 teraflops compute per chip. The chip also supports 62.5 teraflops single-precision FP32.

Many or even most AI models produce model parameters with high levels of sparsity. Not multiplying by a zero element can increase performance by a factor of two or more. The MK2 features new sparsity optimizations for a range of sparse patterns including block, scalar, static and dynamic sparsity. The challenge is knowing ahead of time when not to multiply. Graphcore has been able to effectively compile sparsity optimization into the graph vertex codelets.

## THE IPU-MACHINE (M2000)

Delivering AI silicon as a system instead of a chip is becoming common with accelerators since it can improve time to market by 6-12 months over customized OEM- or ODM-dependent design and testing. Graphcore productized its MK1 silicon in a two-IPU PCIe board to ease adoption and speed time to market. With the MK2 version, Graphcore took this a significant step further, delivering an appliance containing four IPU devices, called the M2000 IPU-Machine. The 1U pizza box is accessed over 100Gb Ethernet with ROCE (RDMA over Converged Ethernet) for low-latency access. Using Ethernet avoids the bottlenecks and costs of PCIe connectors and enables a flexible CPU to accelerator ratio.

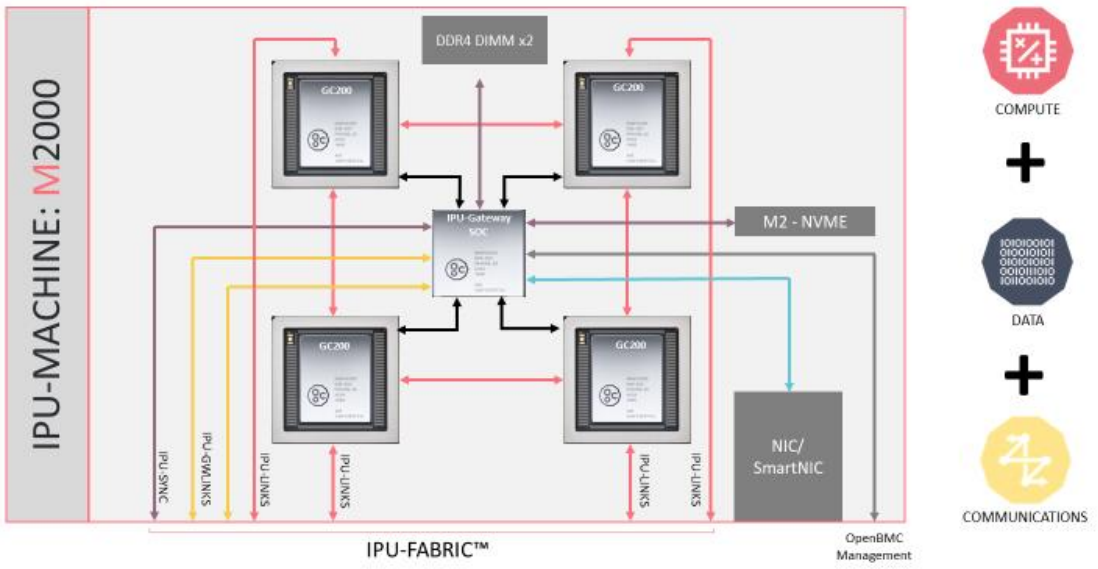
The box includes integrated scale-out networking, which enables the user to easily scale from a small system for development to massive rack deployments, all networked over standard networking at a lower cost than using InfiniBand. The IPU-Fabric connects tiles and other IPUs by tunneling over Ethernet, maintaining the same programming model and Bulk Synchronous Processing (BSP) regardless of the size of the deployment. The IPU-Machine lists for $32,450, which may sound expensive but is a good value when comparing the platform's performance to the competition.

In addition to plug-and-play scaling, the IPU-Machine supplements In-Processor Memory with DDR memory available to the four IPUs. Customer feedback on the MK1 must have indicated that the next generation would need substantially more memory to run the extremely large models that are under development. The new IPU-Machine provides 450GB of memory to handle these larger models. Since model size is doubling every 3.5 months, according to OpenAI.org, this memory architecture could be a game-changer, providing 100X the bandwidth and 10X the capacity found in High Bandwidth Memory (HBM2) at a significantly lower cost.

The memory model for the IPU-Machine is also quite different from that found in CPUs or many AI accelerators such as GPUs. Instead of a memory hierarchy that requires swapping data and code from host memory store to the accelerator's memory, the Poplar Graph Compiler creates the deterministic code-memory relationships in both the memory on the tile and the DDR memory on the Machine. The IPU tile in the graph vertex knows where the data resides and accesses it directly. No caching, no swapping, no pre-fetch and no incremental latencies are incurred. In fact, the IPU-Machine can be used in stand-alone mode for inference processing without any attachment to a host server. And thanks to the BSP model first introduced in the MK1 compiling both computation and communication, the network communication overhead is kept to a

minimum compared to traditional messaging or shared memory constructs commonly used for parallel processing. The IPU-Machine includes a Gateway chip which provides access to the DRAM, two 100Gbps IPU-Fabric Links, a PCIe slot for standard SmartNICs, two 1GbE OpenBMC management interfaces, and access to an M.2 slot.

## FIGURE 2: IPU-MACHINE M2000 ARCHITECTURAL DIAGRAM




**Graphcore will deliver the IPU silicon via a 1U IPU-Machine appliance to enable easy scaling and deployment.**

*Source: Graphcore*

## THE SECOND-GEN IPU-FABRIC

Built-in fabrics are becoming a necessity for AI accelerators since model sizes are increasing dramatically, some containing billions of parameters. These large models must be distributed across hundreds or thousands of processors to solve problems in a reasonable time. To address this need, some companies provide proprietary fabrics on their accelerators, such as NVIDIA's NVLink, while others such as Habana Labs, acquired by Intel in 2019, depend on standard Ethernet. Graphcore's hybrid model uses a proprietary IPU-Link fabric to communicate across the tiles in an IPU and adjacent rack IPUs, while tunneling the IPU-Link protocol across standard 100GbE for rack-to-rack scale-out supporting larger configurations. The IPU-Machine design plays a huge role here, enabling plug-and-play scaling, with or without host CPUs (for inference processing), across a massive infrastructure.

## FIGURE 3: IPU-MACHINE STACKING OPTIONS

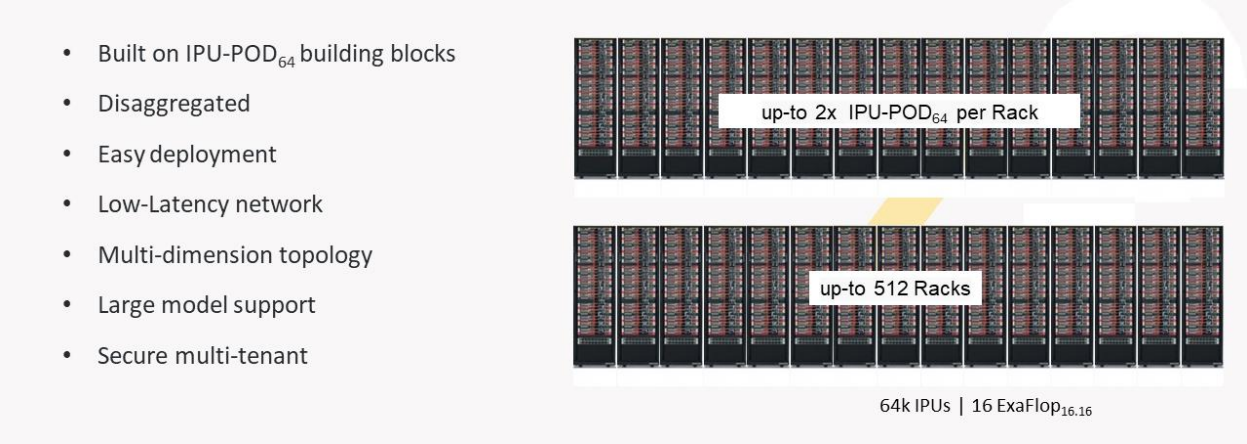

**The IPU-Machine enables plug-and-play scaling, and Graphcore is providing reference architectures for users to ensure tested configurations.**

*Source: Graphcore*

The Graphcore fabric enables a flexible disaggregation model, allowing the user to configure an array of accelerators on the fly without being constrained by a fixed ratio of CPUs to accelerators as found in most other systems. For example, natural language models require very little CPU interaction and utilization, where a 1-to-100 ratio of CPUs to accelerators is adequate. On the other hand, convolutional neural networks can require a 1-to-4 or 1-to-8 ratio since the functions − such as averaging − take place on scalar CPU cores. By leveraging 100Gb Ethernet, elastic configurations are easy to deploy and simple to use, enabling scaling up to 64,000 IPUs. In our view, this disaggregated scaling model is perhaps the most significant feature of the second-generation Graphcore IPU platform.

Machine learning exposes parallelism in three dimensions: data (or batch) parallelism, layer parallelism for transform, pooling, etc., and tensor parallelism across multiple instances. Ideally this is implemented across a multi-dimensional fabric, and the 3D-Ring topology supported by the IPU-Fabric efficiently enables this with 1-to-1 direct communications at extremely low latency. The fabric also exposes collectives such as all-reduce, all-gather and broadcast to facilitate application development.

## FIGURE 4: IPU-POD FOR SUPER COMPUTING SCALE

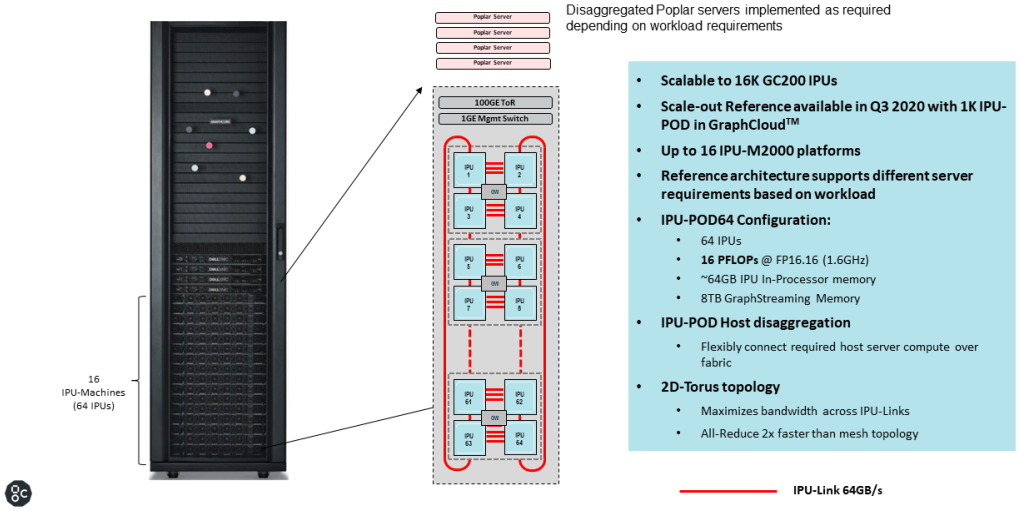



**Graphcore envisions entire data centers filled with interconnected IPU-Machines in interconnected IPU-PODs. Multi-tenancy support could make this attractive to cloud-scale service providers such as Microsoft Azure.**

*Source: Graphcore*

The disaggregated scalability model enables a wide range of deployment options, several of which have been standardized with reference implementations. While Graphcore has not disclosed customers that are building out a full-scale 16-exaflop implementation, we suspect that some are exploring this level of deployment.

## FIGURE 5: IPU-POD64 REFERENCE ARCHITECTURE




**The fabric links support efficient communications within a rack and across the datacenter, using existing Ethernet equipment.**

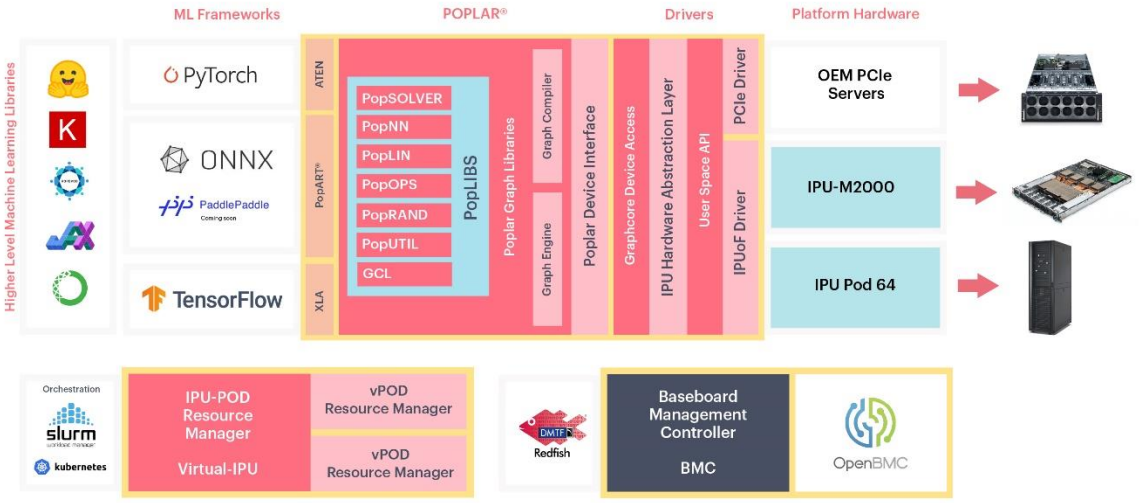
*Source: Graphcore*

The IPU-Fabric and BSP programming model enable scalability. As we explored in our research on the Poplar software stack, the ability to compile to a “virtual IPU” instance and then deploy on an arbitrary configuration should be very attractive to designers of massively parallel applications, including but not limited to deep neural network models. The 2D torus interconnect in a rack of IPU-Machines enables efficient all-reduce and other collective operations.

## UPDATES TO THE GRAPHCORE SOFTWARE STACK

Graphcore has made numerous enhancements to the Poplar software stack. Naturally, when attempting to scale to the level afforded by the MK2, advanced management software becomes critical. Graphcore has extended its software to include the Graphcore Communications Library (GCL) and plugins to industry-standard tools as SLURM, Kubernetes, Grafana and Prometheus for scheduling, virtualization, security and monitoring. Control-plane SW provides interface to virtual IPUs, carving up the physical pod into multiple virtual pods. Cloud service providers will be pleased to see support for multi-tenancy, isolating users by disabling links between virtual pods while overlay networks provide instance isolation.

## FIGURE 6: POPLAR SDK

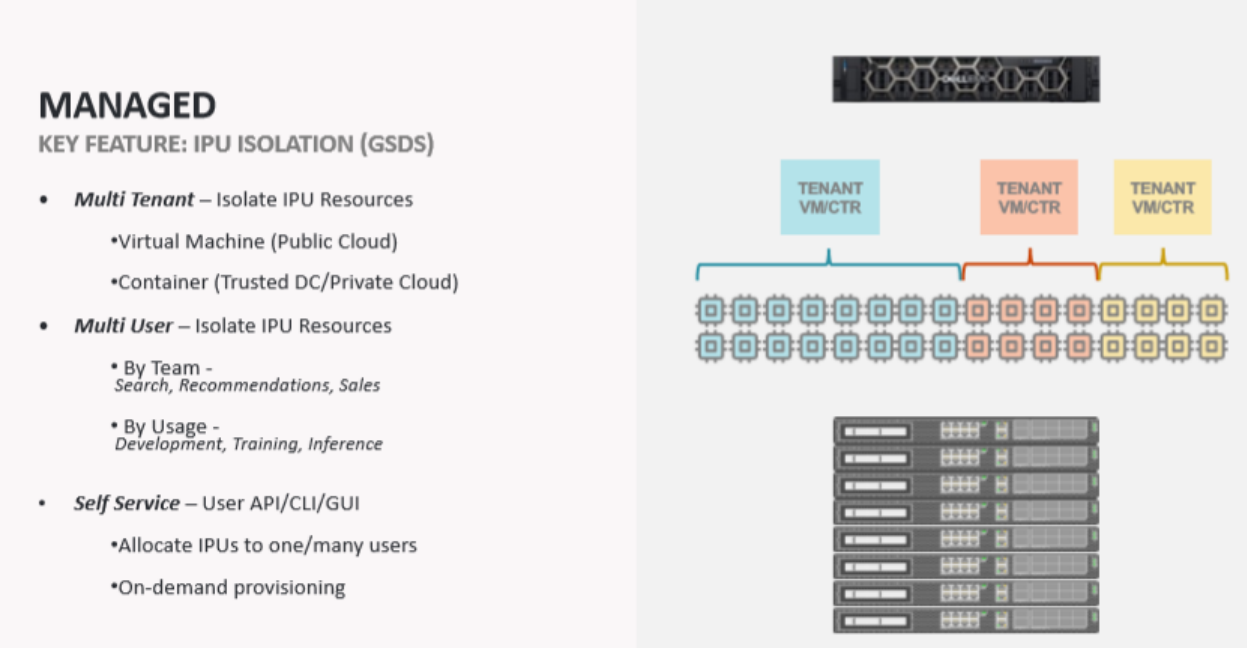



**Poplar has been enhanced to support the IPU-Machine concept and a communications library (GCL) to support scale-out communications. Changes to Poplar are highlighted in blue.**

*Source: Graphcore*

Graphcore has also added auto self-discovery and management to support the IPU-Machine pod deployment and execution. The new self-service interface and APIs for multi-tenancy and multi-user usage models should help Graphcore penetrate cloud service providers such as Microsoft Azure, an early adopter and investor in Graphcore.

## FIGURE 7: IPU-POD SUPPORT FOR MULTI-TENANCY




**API, CLI and GUI interfaces are now available to support multi-tenancy in the IPU-Pods, critical for providing a secure and private shared infrastructure.**

*Source: Graphcore*

## PERFORMANCE

Graphcore says the new four-chip IPU-Machine delivers 7-9X the performance of the two-chip predecessor PCIe card in training neural networks and more than 8X the performance in inference processing. So a chip-to-chip comparison would likely put the MK2 at a 3-4X improvement. While impressive, this performance increase is overshadowed, in our opinion, by the massive increase in memory for larger models, and the plug-and-play scalability, fabric and management of the IPU-Machine and Pods.

## FIGURE 8: COLOSSUS MK2 PERFORMANCE


**Users can expect roughly 7-9X more performance per card over that of the first-generation MK1. However, the scalability and memory capacity should open possibilities of new model development and deployment at a scale rare in production silicon.**

*Source: Graphcore*

## CONCLUSIONS

Delivering the new Colossus MK2 IPU in a plug-and-play hardware and software platform provided by Poplar and the IPU-Machine seems like an excellent strategy for wider adoption of Graphcore technology. The company is well-funded and staffed with software and system engineering teams in addition to an innovative processor design team. We believe that their stable of advisors has provided insights born from real-world experience that will minimize dead-ends and increase the attraction of Graphcore technology. The Graphcore engineering teams have thought through the development, deployment and optimization tasks that users will face. We believe that Graphcore is one of the few startups that can challenge the status quo in AI acceleration and advance parallel processing in a meaningful way. It remains to be seen if Graphcore can also extend this architecture to inference processing and eventually to edge AI, where cost-effectiveness and software ecosystems represent potential hurdles for any startup that sees performance and scale as foundational principles.

## IMPORTANT INFORMATION ABOUT THIS PAPER

### *CONTRIBUTOR*
Karl Freund, Senior Analyst at Moor Insights & Strategy

### *PUBLISHER*
Patrick Moorhead, Founder, President, & Principal Analyst at Moor Insights & Strategy

### *INQUIRIES*
Contact us if you would like to discuss this report and Moor Insights & Strategy will respond promptly.

### *CITATIONS*
This paper can be cited by accredited press and analysts but must be cited in-context, displaying author's name, author's title, and "Moor Insights & Strategy". Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

### *LICENSING*


### *DISCLOSURES*
This paper was commissioned by Graphcore. Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

### *DISCLAIMER*
The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.