GRAPHCORE

ISC HPC USER FORUM

AI MEETS HPC IN SCIENTIFIC COMPUTING USING GRAPHCORE IPUS

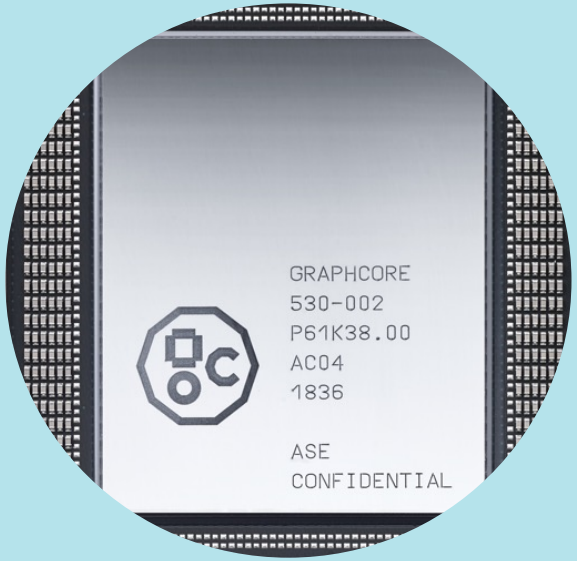
Paweł Gepner
Senior System Field Engineer

AGENDA

- WELCOME AND GRAPHCORE INTRODUCTION
- IPU ARCHITECTURE OVERVIEW
- INSIGHTS INTO POD SYSTEM
- SHORT INTRODUCTION TO THE POPLAR SDK
- ACCELERATION OF SCIENTIFIC APPLICATIONS AND ENABLING HETEROGENEITY OF AI AND HPC
- HYBRIDIZATION OF HPC AND AI SYSTEMS
- Q&A SESSION

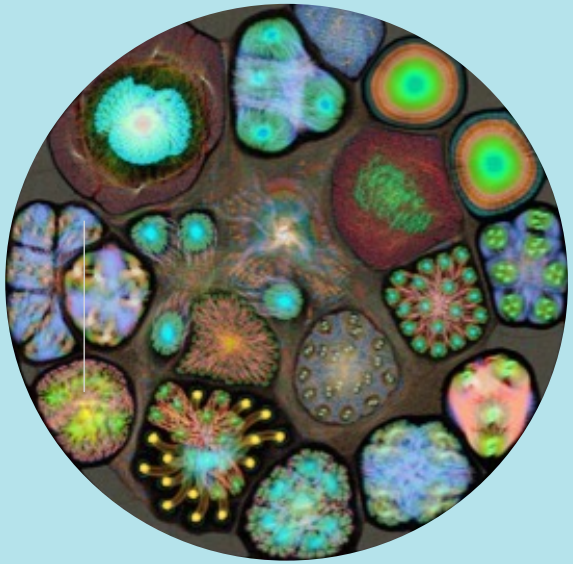
ABOUT US...

Hardware



IPU processors
designed for AI

Software



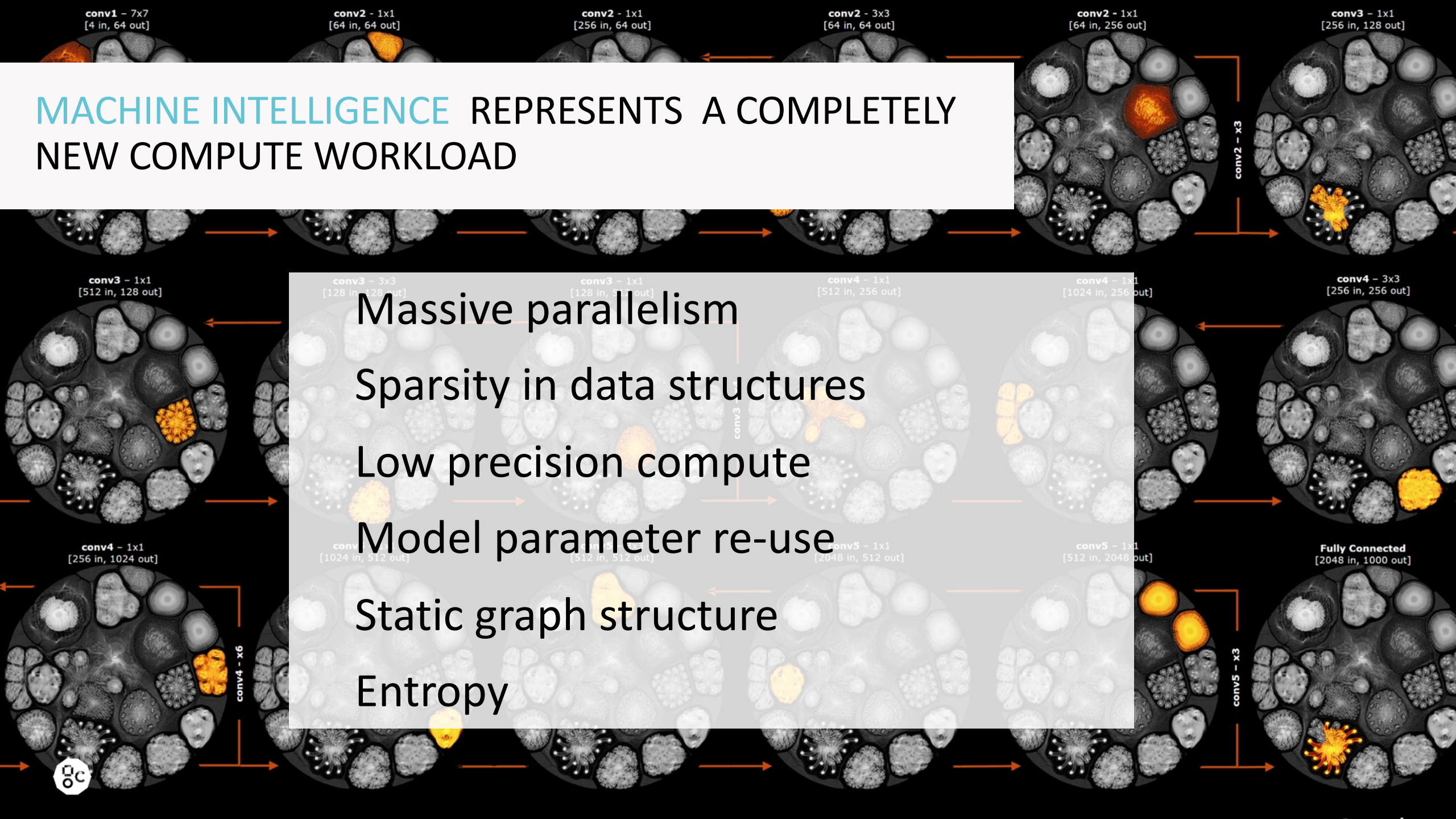
Poplar® software stack &
development tools

Platforms

M2000 and Server
IPU-POD$_{64}$ scale-out

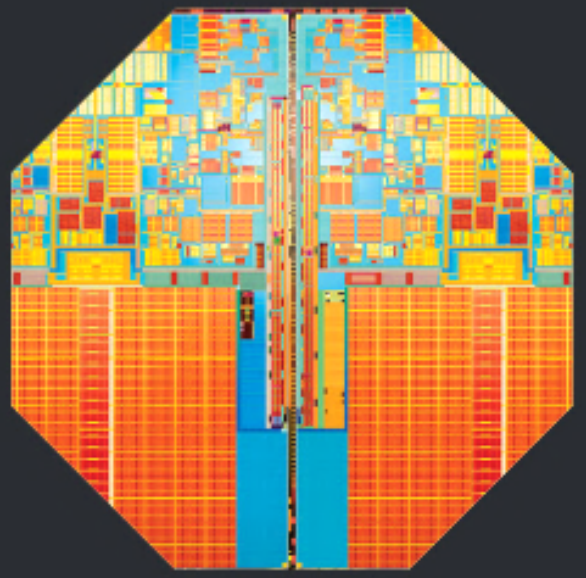
IPU ARCHITECTURE OVERVIEW

MACHINE INTELLIGENCE REPRESENTS A COMPLETELY NEW COMPUTE WORKLOAD



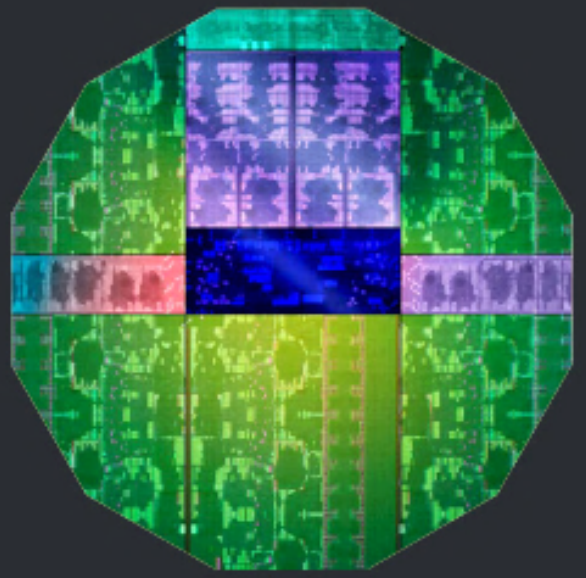
- Massive parallelism
- Sparsity in data structures
- Low precision compute
- Model parameter re-use
- Static graph structure
- Entropy

LEGACY PROCESSOR ARCHITECTURES HAVE BEEN REPURPOSED FOR ML



CPU

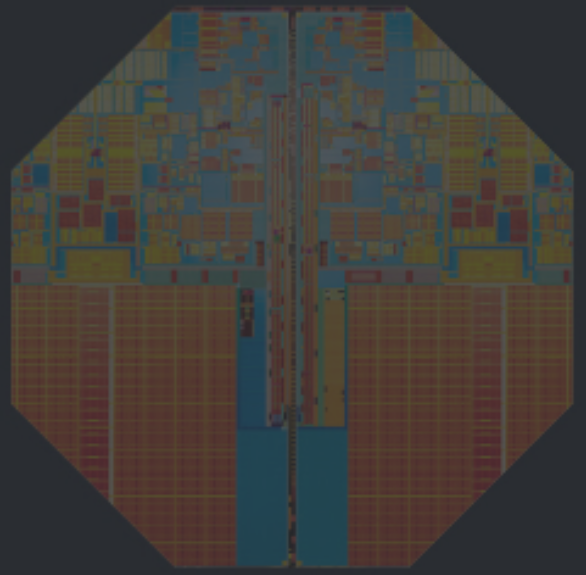
Apps and Web/
Scalar



GPU

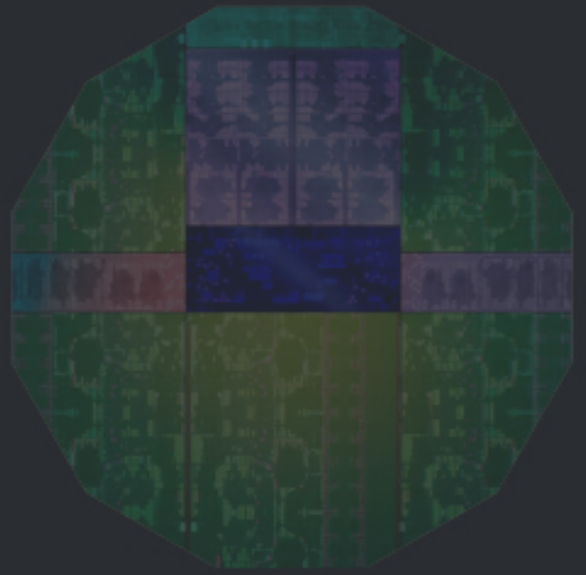
Graphics and HPC/
Vector

A NEW PROCESSOR IS REQUIRED FOR THE FUTURE



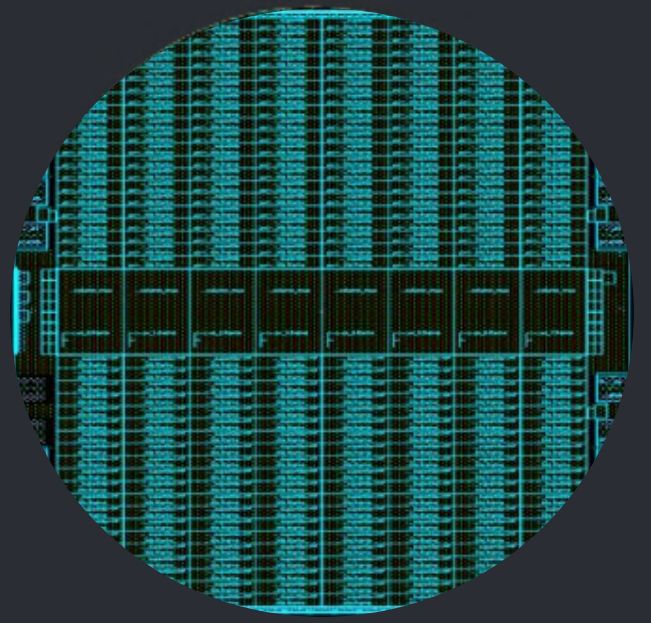
CPU

Apps and Web/
Scalar



GPU

Graphics and HPC/
Vector

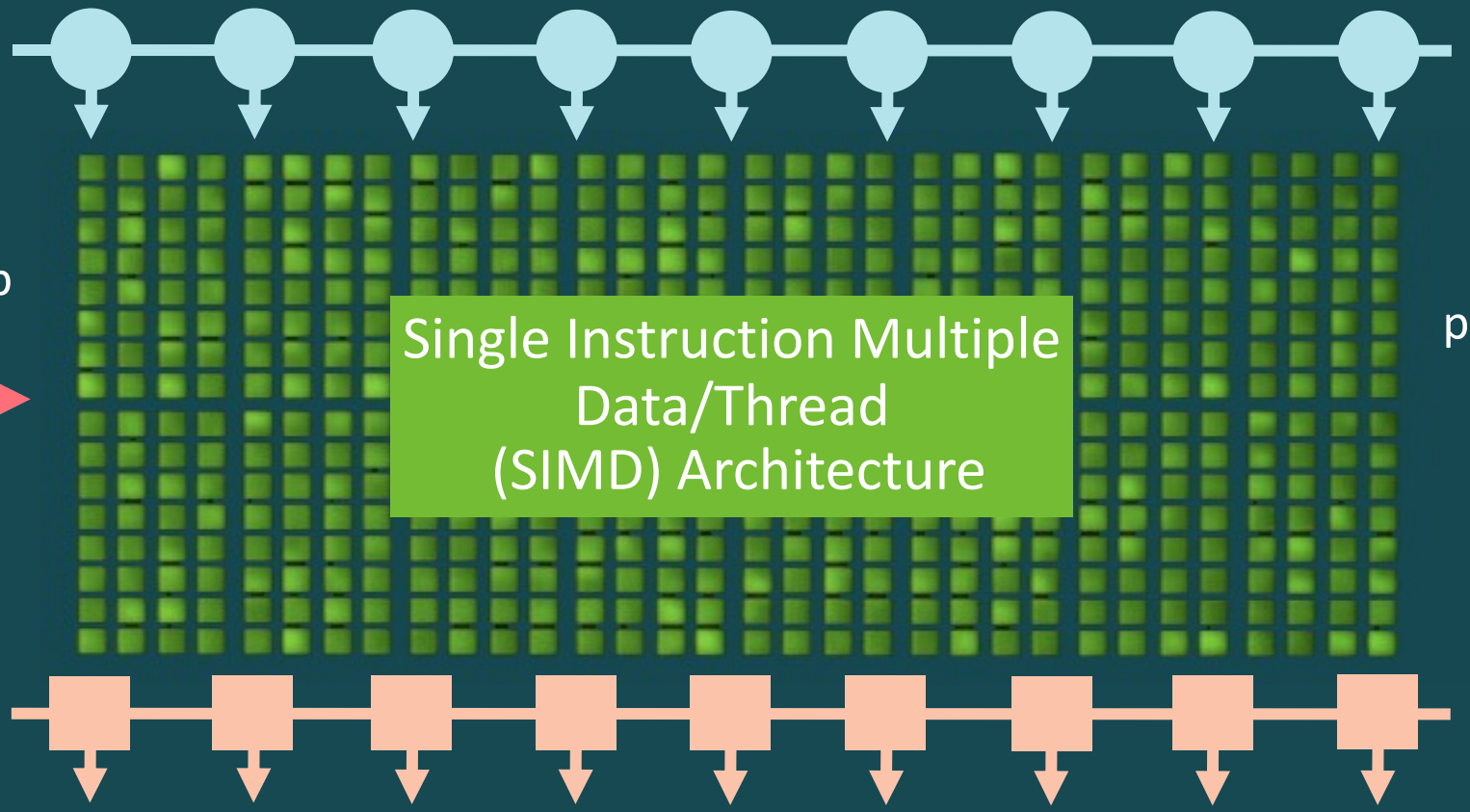


IPU

Artificial Intelligence/
Graph

TODAY'S PARALLEL MACHINES ARE INFLEXIBLE

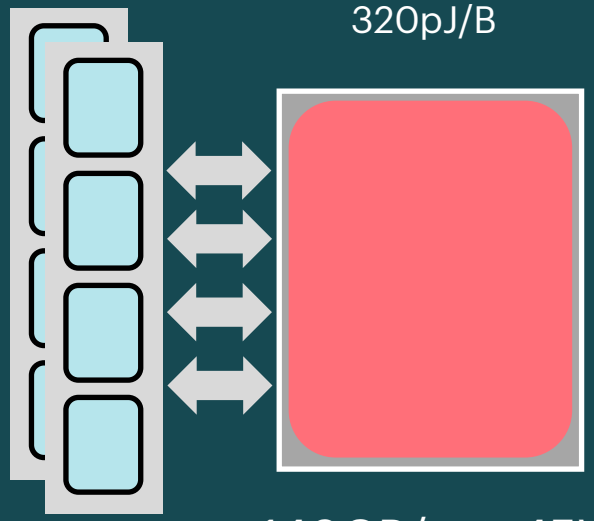
Instruction operates on long data vector
(typically 32x32 bits or 64x16 bits)



Only efficient for large blocks of dense data / large mini-batches

MEMORY TRADE-OFF

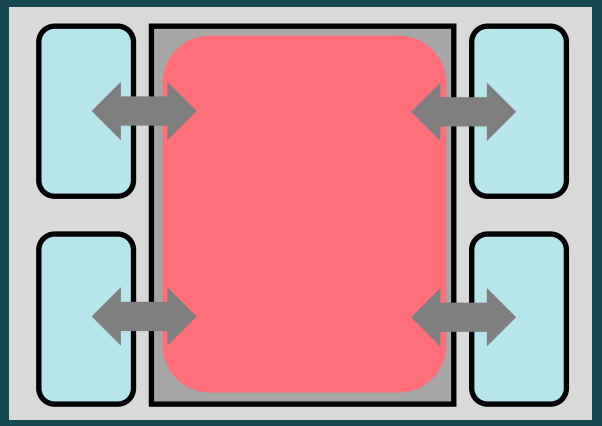
DDR4 DIMMs
320pJ/B



~140GB/s @ 45W

1x

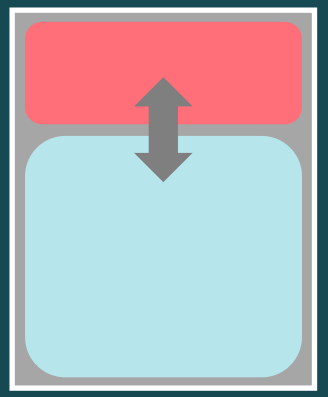
HBM2 on silicon-interposer
64pJ/B



700GB/s @ 45W

5x

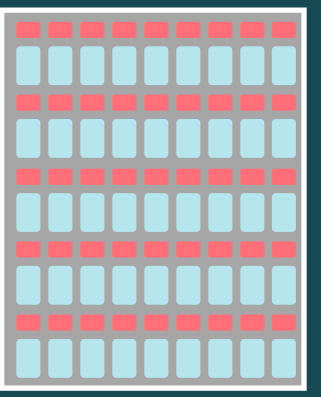
SRAM on chip
10pJ/B



4.5TB/s @ 45W

32x

distributed SRAM
on chip
1pJ/B



45TB/s @ 45W

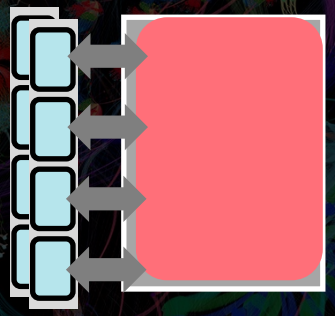
320x

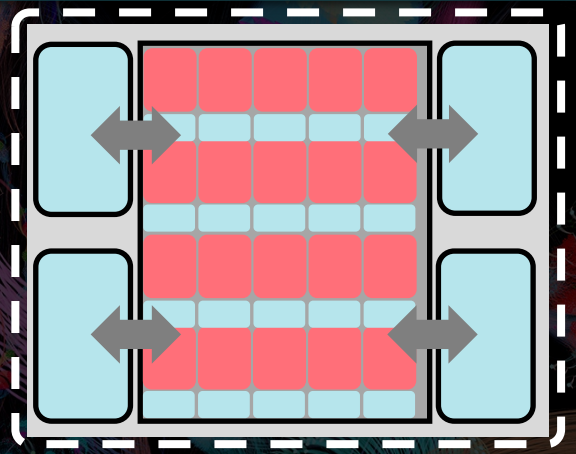
MASSIVE PARALLELISM WITH ULTRAFAST MEMORY ACCESS

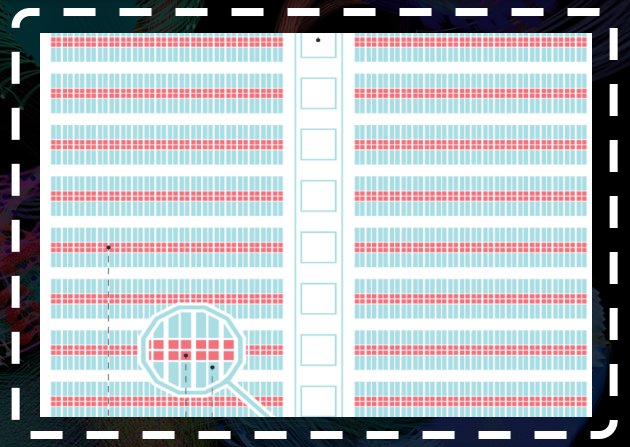
	CPU	GPU	IPU
Parallelism	Designed for scalar processes	SIMD/SIMT architecture. Designed for large blocks of dense contiguous data	**Massively parallel MIMD. High performance/efficiency for future ML trends**

Processor

Memory





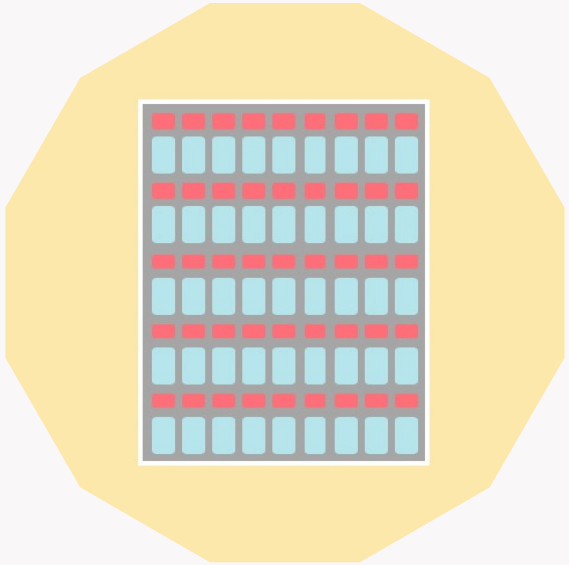


Memory Access	Off-chip memory	Model and Data spread across off-chip and small on-chip cache and shared mem.	**Model & Data in tightly coupled large locally distributed SRAM**

Generalised comparisons & illustrative diagrams

IPU PROVIDES UNIQUE SPEED AND FLEXIBILITY

Code and data are always in SRAM

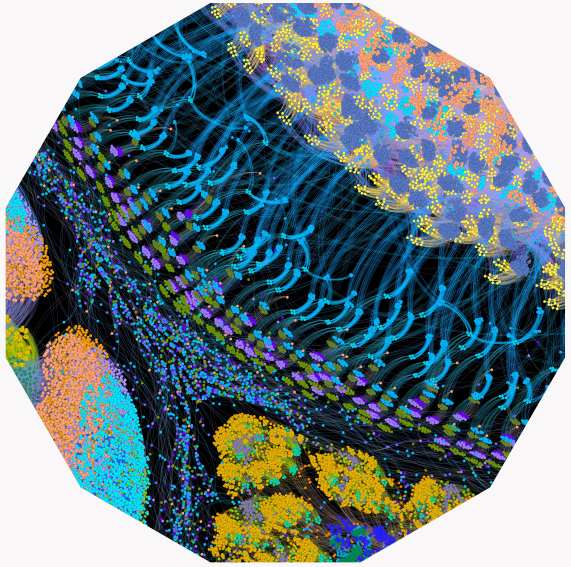


45 TB/s memory bandwidth

Parallel programs are truly independent

7,296 independent threads

Native support for computational graphs



Poplar® SDK

BULK SYNCHRONOUS PARALLEL (BSP)

Software bridging model for parallel computing

Compute

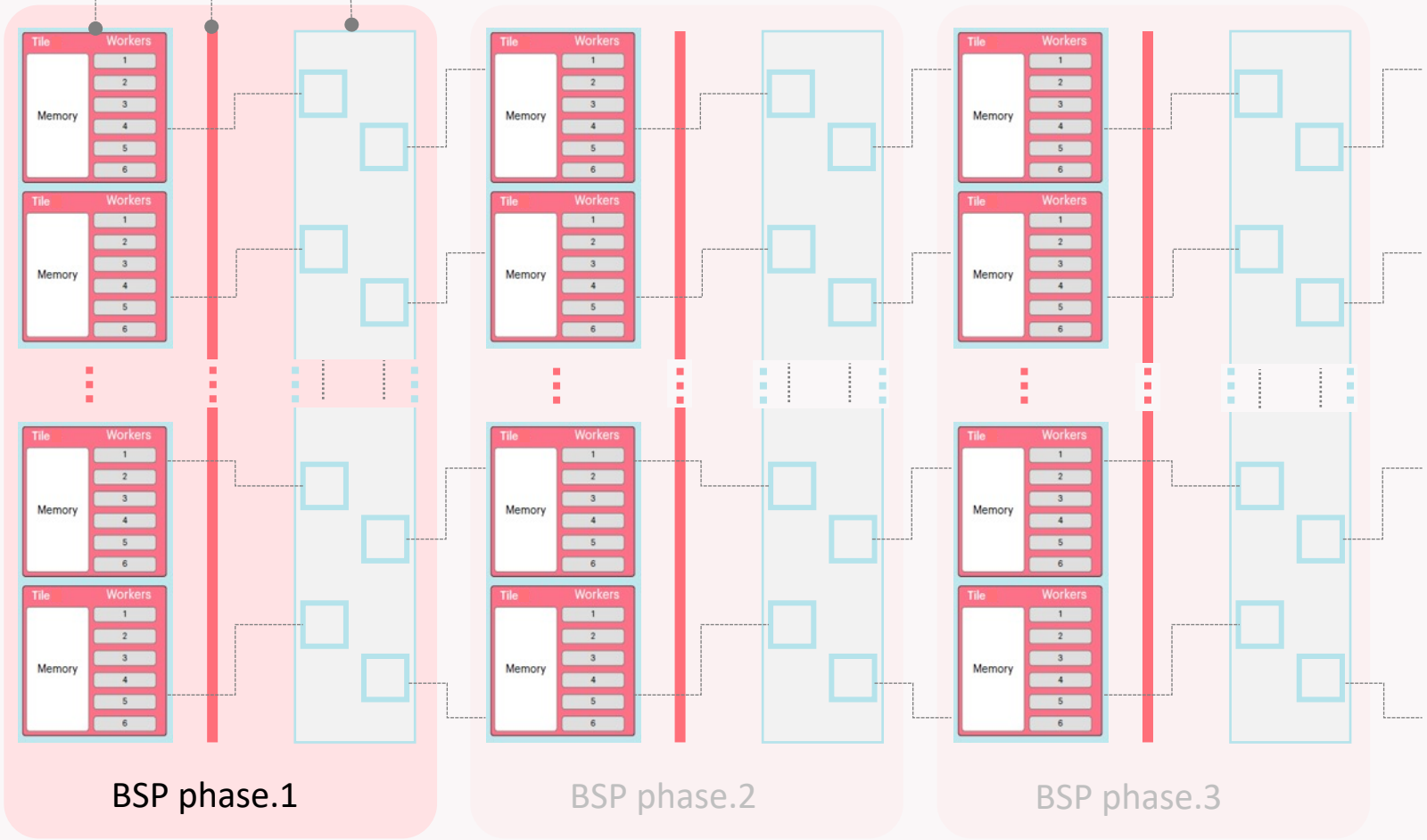
10,000s of compute threads all operating in parallel each with all the data that they need, held locally

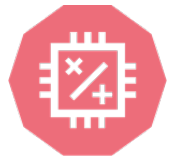
BSP Sync

All threads are synchronized

Exchange

Data is exchanged so that every thread has all the data that it needs for the next phase of Compute





COMPUTE

COLOSSUS MK2

the worlds most complex processor

59.4Bn transistors, TSMC 7nm @ 823mm^2

250TFlops AI-Float | 900MB In-Processor-Memory™

1472 independent processor cores

8832 separate parallel threads

>8x step-up in system performance vs Mk1

GC200 IPU

GC200 IPU PROCESSOR

IPU-Tiles™

1472 independent IPU-Tiles™ each with an IPU-Core™ and In-Processor-Memory™

IPU-Core™

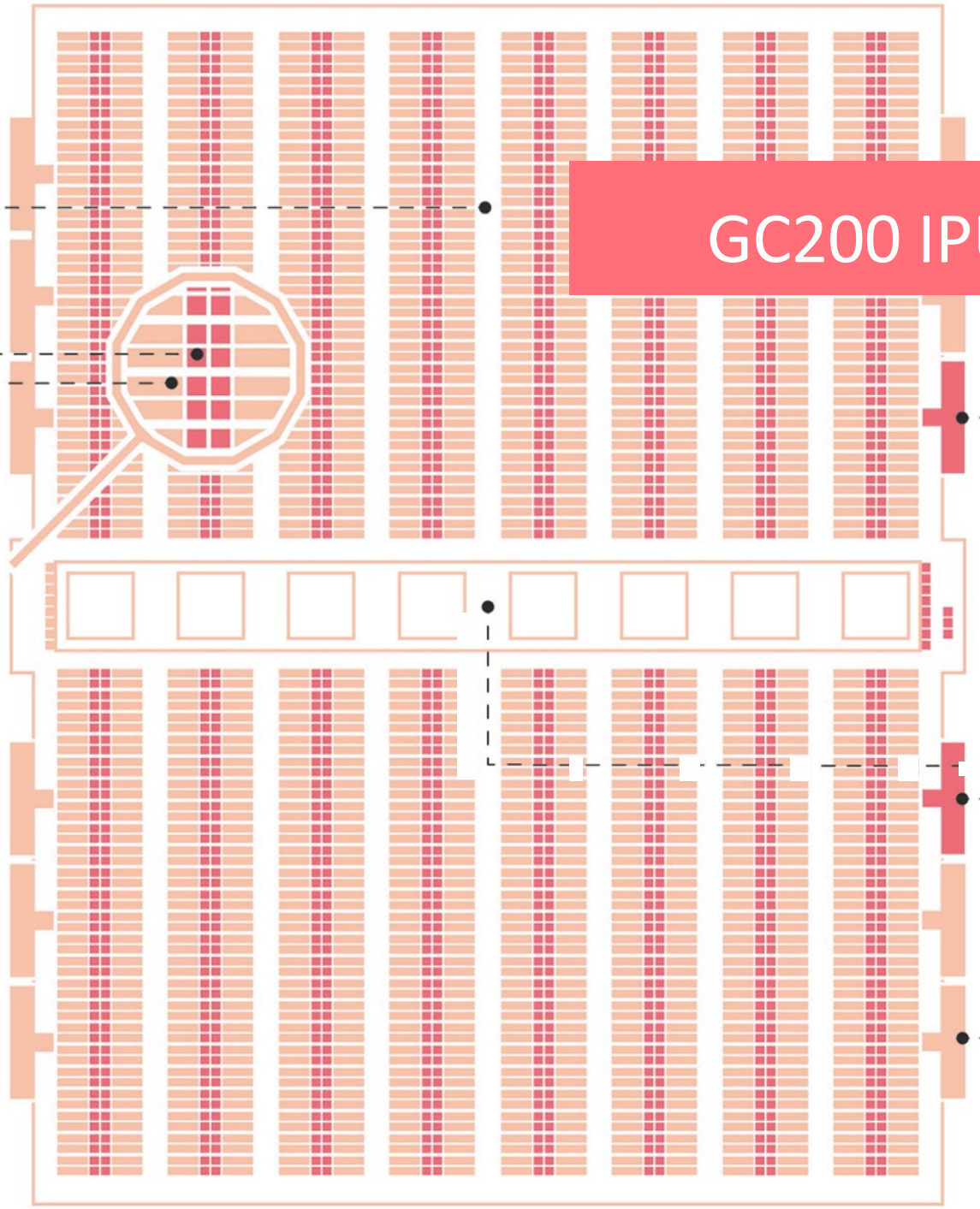
1472 independent IPU-Core™

8832 independent program threads executing in parallel

In-Processor-Memory™

900MB In-Processor-Memory™ per IPU

47.5TB/s memory bandwidth per IPU



IPU-Exchange™

8 TB/s all to all IPU-Exchange™
Non-blocking, any communication pattern

PCIe

PCI Gen4 x16
64 GB/s bidirectional bandwidth to host

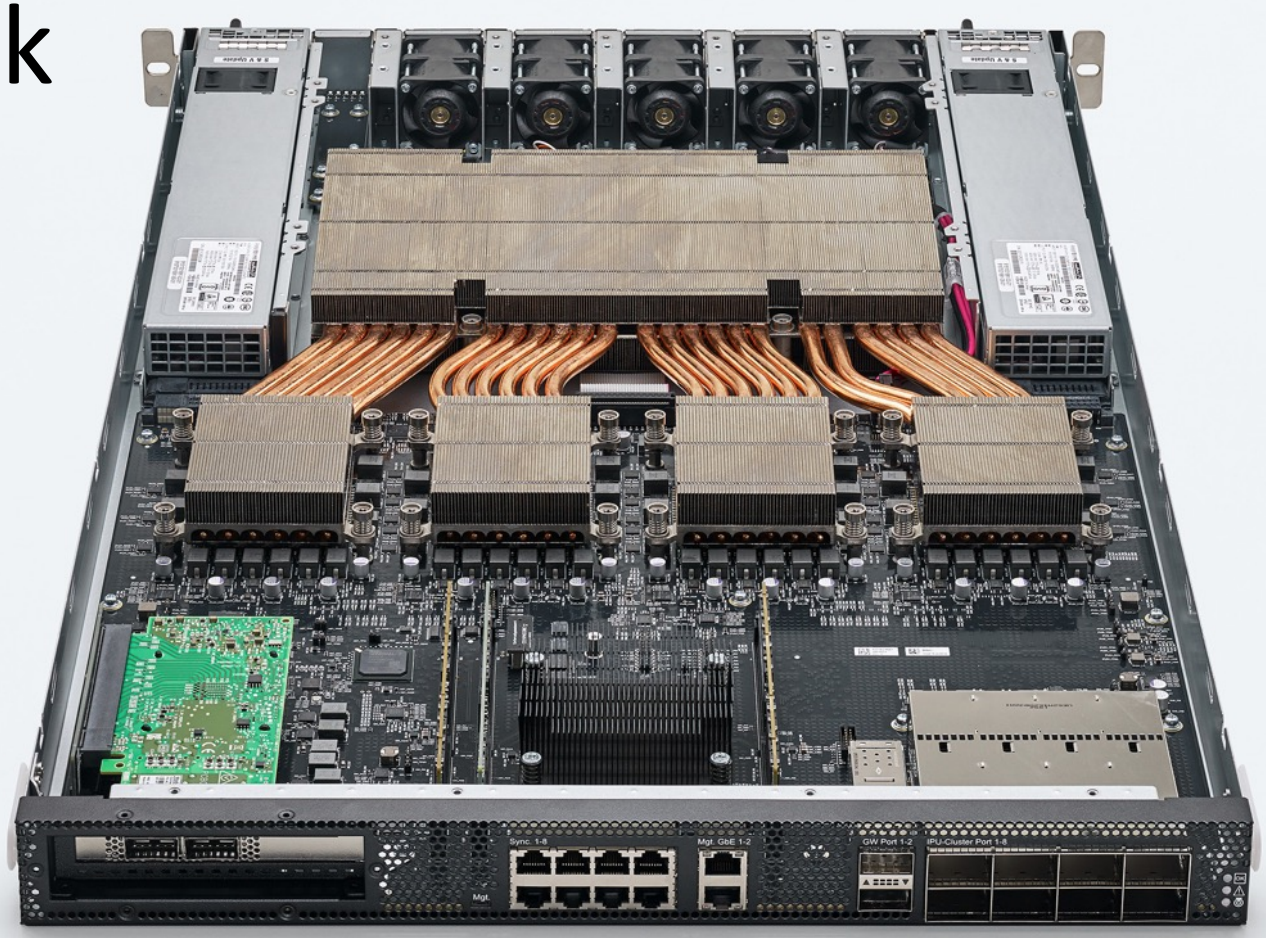
IPU-Links™

10 x IPU-Links,
320GB/s chip to chip bandwidth

14

IPU POD SYSTEMS

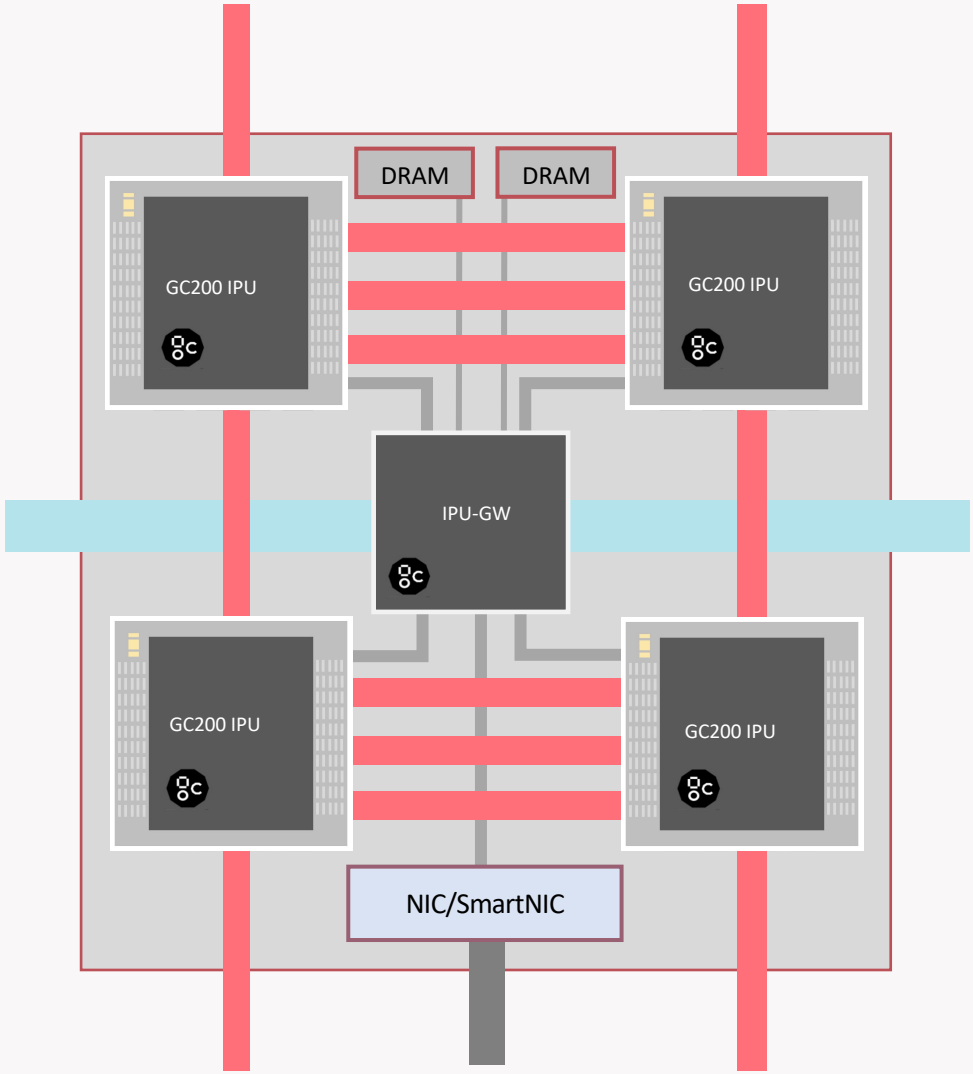
IPU-M2000
PODs Building Block



1U Machine Intelligence Compute Blade

1 PetaFlop IPU compute
2.8Tbps IPU-Fabric™

IPU-M2000 - POD BUILDING BLOCK



COMPUTE

4x GC200 IPUs

- 1 $PFLOP_{16.16}$ compute
- 5,888 processor cores
- > 35,000 independent parallel threads

DATA

Up to ~450GB Exchange Memory™

- Up to 448GB Streaming Memory™ DRAM
- 3.6GB In-Processor-Memory™

COMMUNICATIONS

IPU-Fabric for Compiled-In Networking

- Host-Link – 100GE link to Poplar Server
- IPU-Link – 4x 512Gbps for intra IPU-POD64 communication
- GW-Link - 2x 100bps Gateway-Links for inter IPU-POD64 communication (0.4 Tbps bidir.)
- Sync-Link IPU-POD hardware sync signal

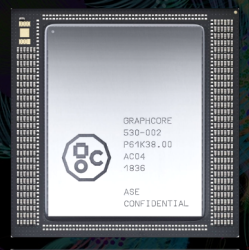
x16 IPU-Link 64GB/s

100GE Host-Link Network I/F

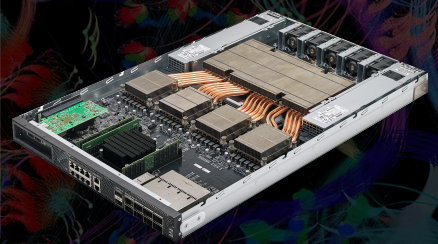
100Gbps IPU-GW Link

x8 PCIe G4 32GB/s

GRAPHCORE IPU SCALEOUT



X4



X16

X1024

GC200
(Processor)

IPU-M2000
(shelf)

4 IPUs
1 PetaFlop

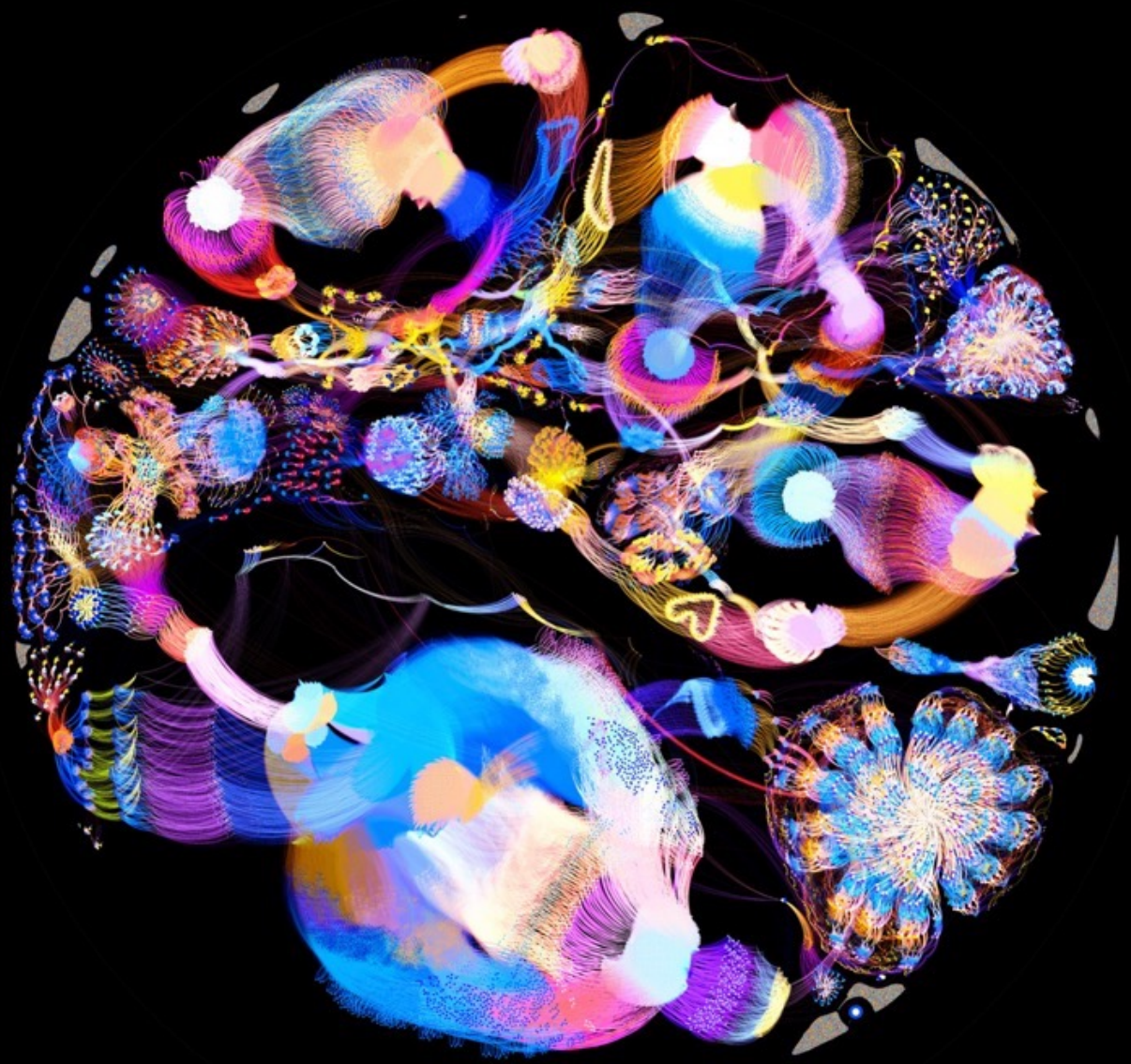
IPU-POD$_{64}$
(Rack)

64 IPUs
16 PetaFlop

IPU-POD$_{64k}$
(1024 Racks)

64k IPUs
16 ExaFlop

SOFTWARE

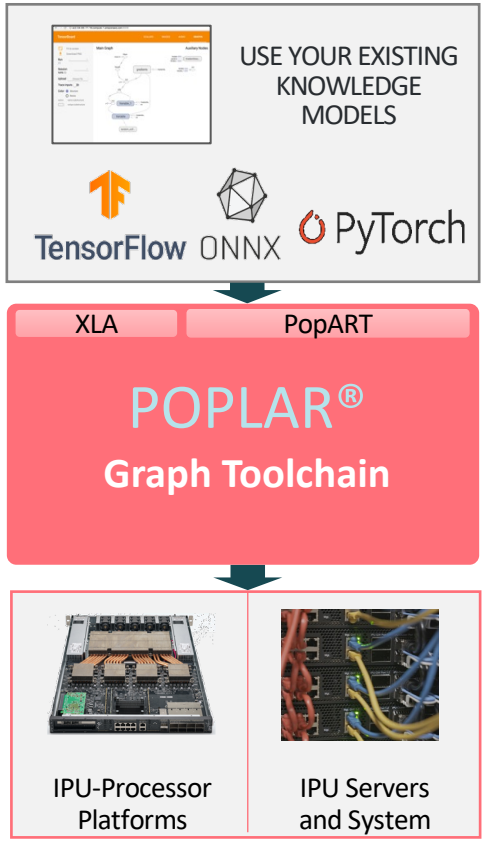


POPLAR™ COMPUTE GRAPH VISUALISATION

POPLAR EASE OF USE

TensorFlow PYTORCH Azure docker

Keras ONNX kubernetes



Open & Extensible Poplar Libraries

Get access to 50+ optimised functions for common ML models and 750 high performance compute elements. Modify and write custom libraries.

Straightforward Deployment

Pre-built Docker containers with Poplar SDK, Tools and frameworks images to get up and running fast.

ML Frameworks Support

Support for standard ML frameworks: TensorFlow 1 and 2, ONNX and PyTorch with PaddlePaddle coming soon.

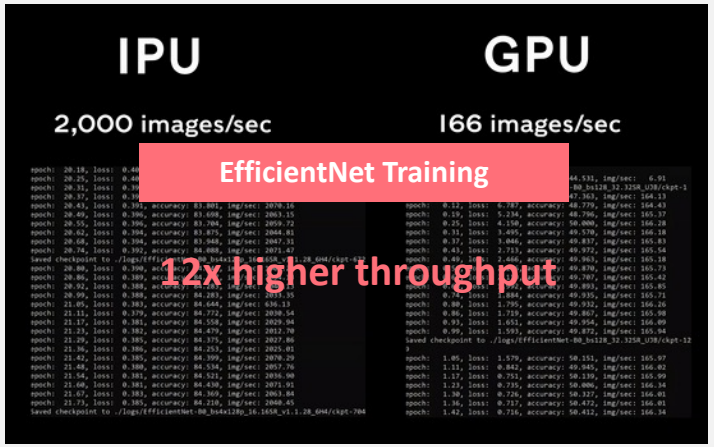
Standard Ecosystem Support

Ready for production with Microsoft Azure deployment, Kubernetes orchestration, Docker containers and Hyper-V virtualization & security.

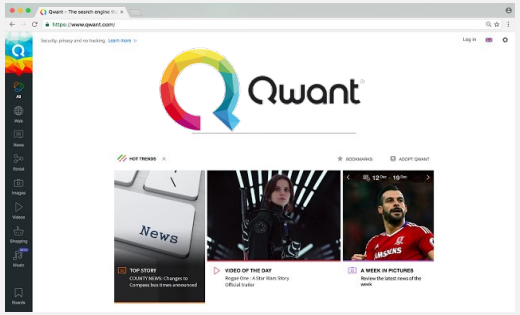
IPU ADVANTAGE

Microsoft

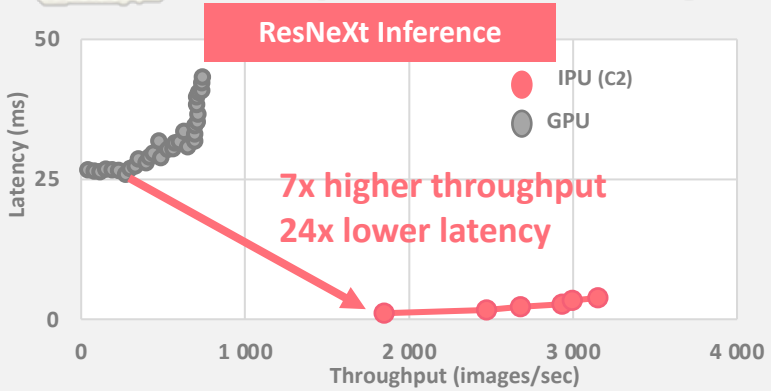
Analysis of Chest X-Ray images using variant of EfficientNet-B0
Training for COVID-19/pneumonia classification model development



Images & Work presented by Microsoft AI & Advanced Architectures Research Lead, Sujeeth Bharadwaj, PhD at Intelligent Health Inspired conference May 2020

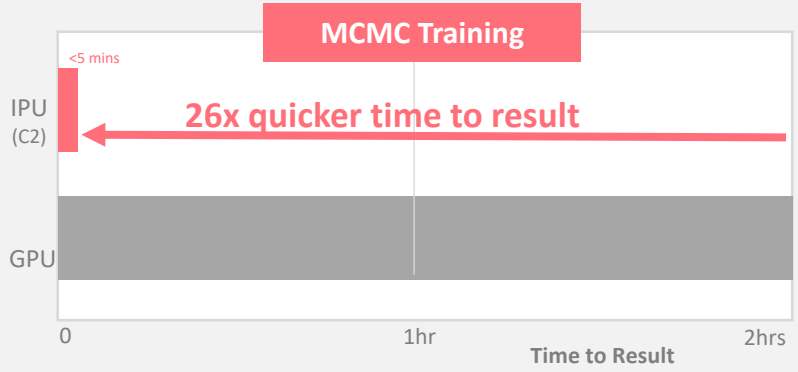


"we are seeing a significant improvement – with 6x higher performance - in our image search capability using ResNext on IPUs, out of the box"
CEO, Qwant



carmot CAPITAL

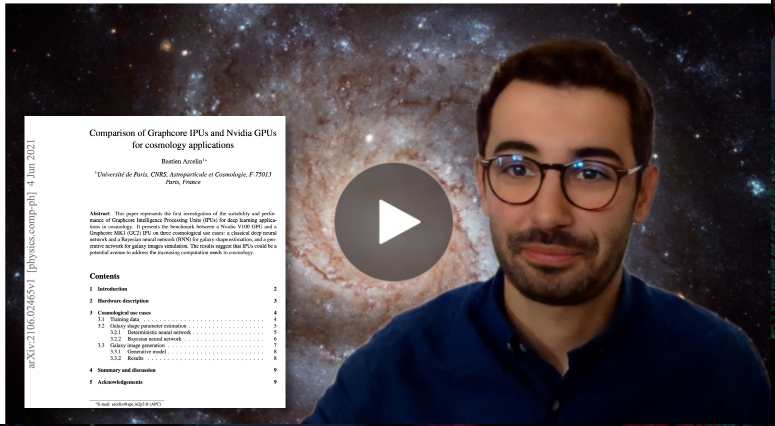
"We were able to train one of our proprietary models in 4.5minutes instead of 2hours. That's 26x faster time to train than other leading platforms"
George Sokoloff, Founder & CIO Carmot Capital

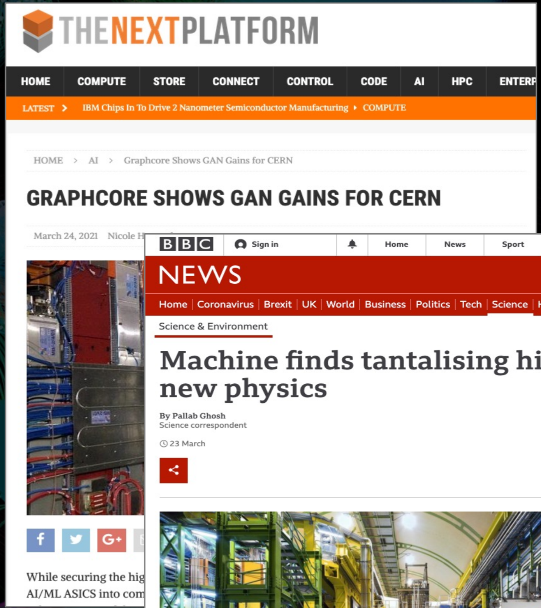


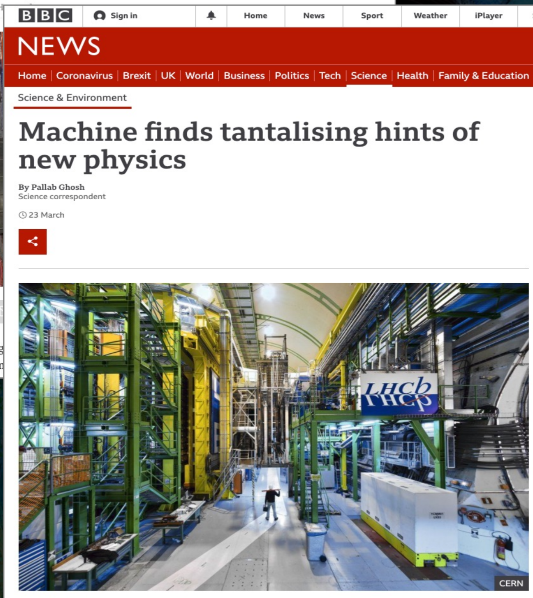
IPU FOR RESEARCH

IMPERIAL COLLEGE LONDON ACCELERATE CLASSICAL COMPUTER VISION PROBLEM ON IPU

UNIVERSITÉ DE PARIS ACCELERATES COSMOLOGY WITH IPUS

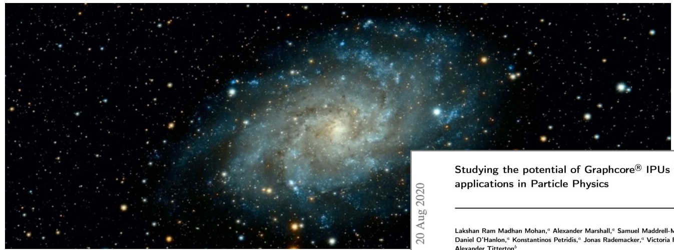


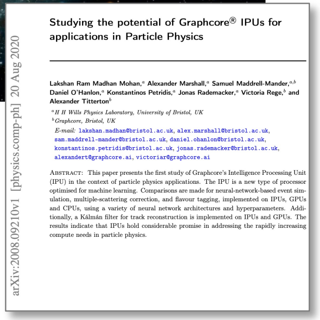




University of BRISTOL

UNIVERSITY OF BRISTOL SOLVES SCIENTIFIC PROBLEMS WITH NEW IPU-BASED AI SYSTEMS

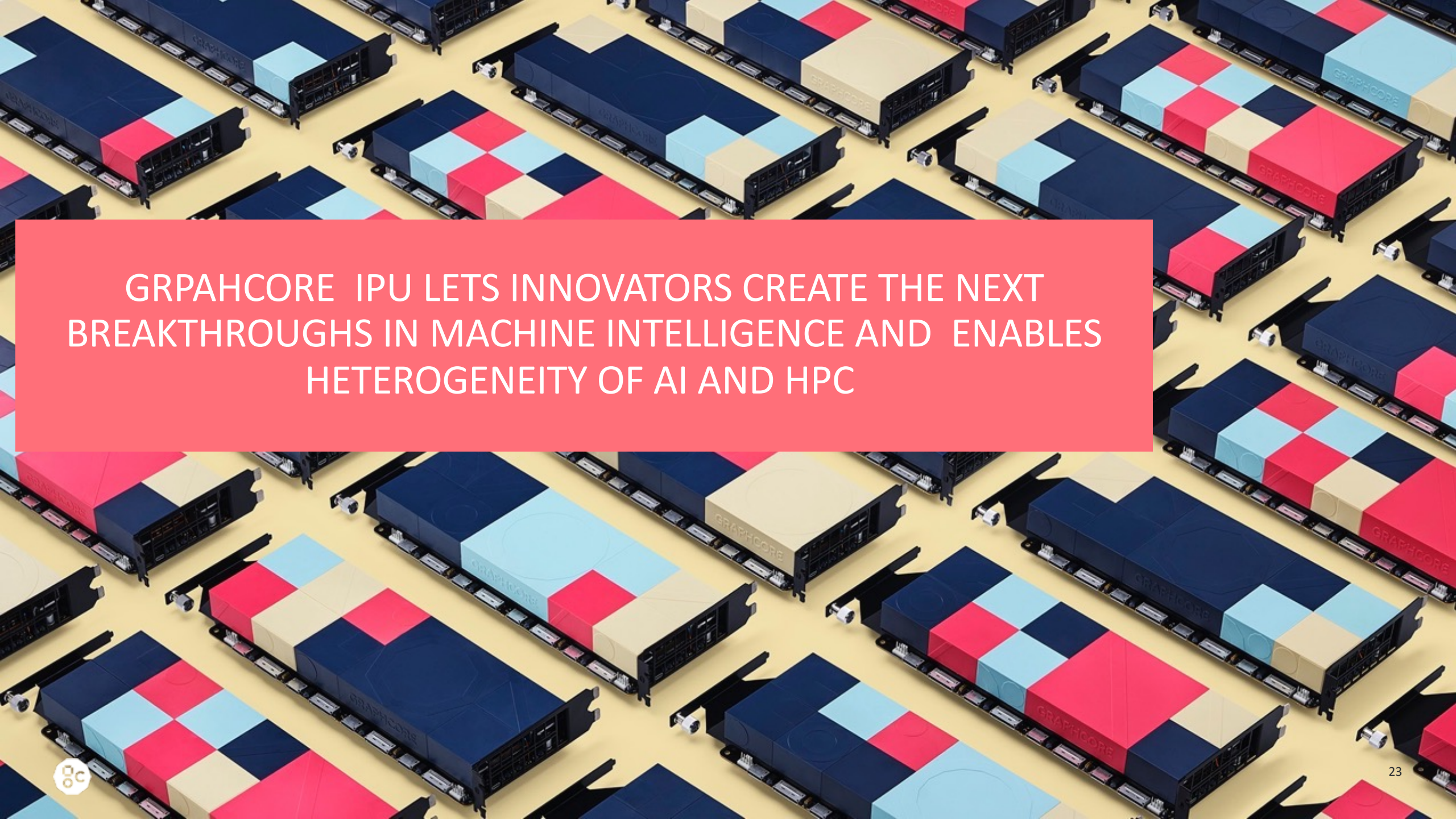




UNIVERSITY OF BRISTOL TACKLES CHALLENGES IN PARTICLE PHYSICS WITH GRAPHCORE'S IPU

CERN

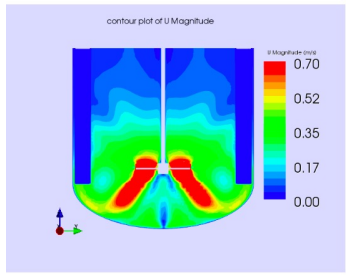
https://www.graphcore.ai/resources/research-papers


GRPAHCORE IPU LETS INNOVATORS CREATE THE NEXT BREAKTHROUGHS IN MACHINE INTELLIGENCE AND ENABLES HETEROGENEITY OF AI AND HPC

Acceleration of Scientific Applications and Enabling Heterogeneity of AI and HPC to Improve Simulation Time

Data Generation

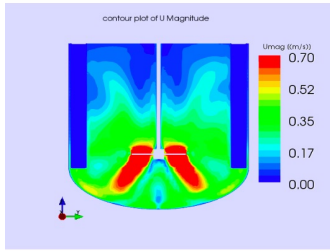
1



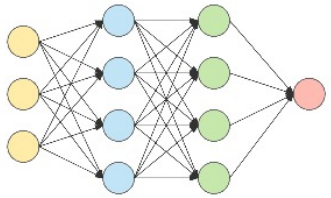
(a) CFD

Model design & ML training

2

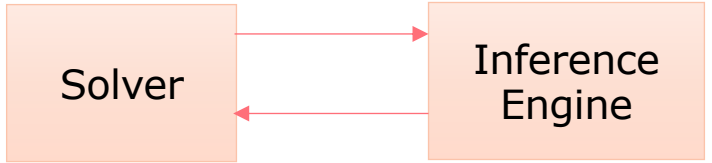


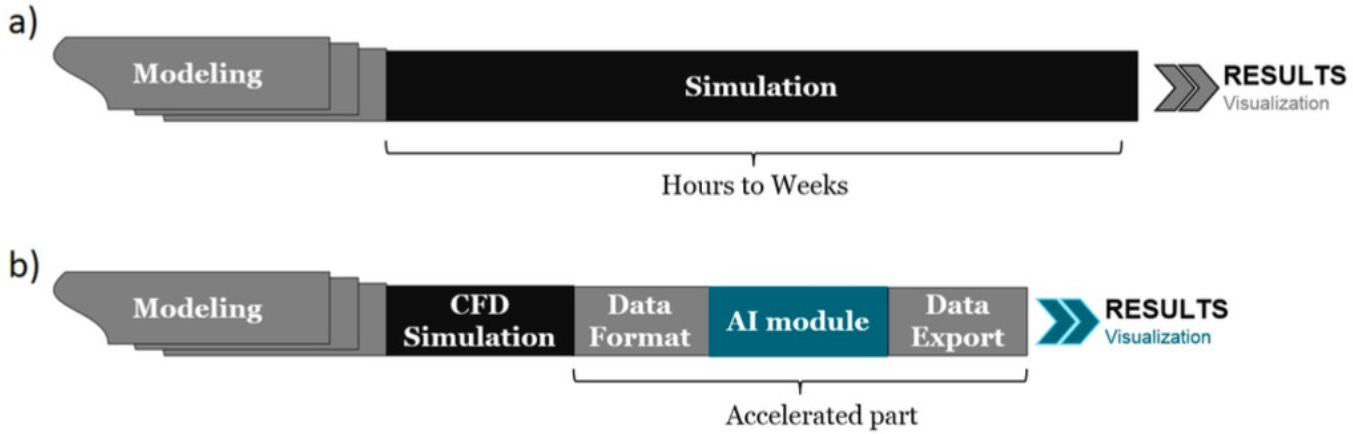
(b) AI



Model Inference & Integration

3





<u>Areas where this approach makes sense and where it has already been proven</u>

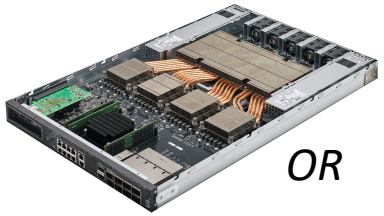
- HEP -High Energy Physics
- CFD -Computational Fluid Dynamics
- PDE -Partial Differential Equation
- PF - Protein folding
- MC - Monte Carlo simulation
- WRF - Weather Research and Forecasting
- O&G – Oil and Gas Exploration Simulation

Hybridization of HPC and AI systems

For typical HPC workloads

CPUs and GPUs type
of configurations

For AI workloads



OR

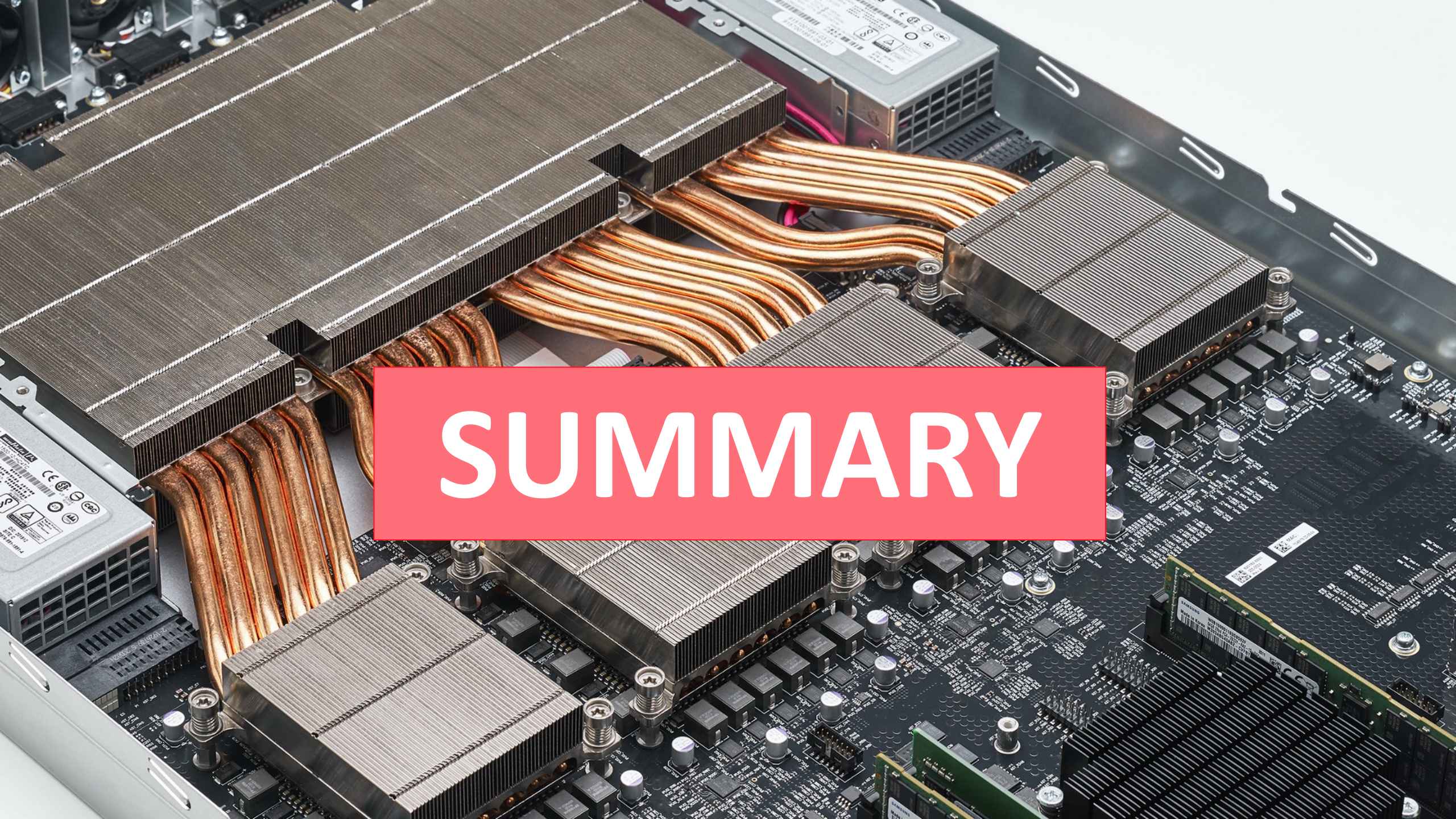
IPU-POD16
IPU-POD64
IPU-POD512

For new heterogeneous types of workloads

CPUs and GPUs type
of configurations

IPU-POD16
IPU-POD64
IPU-POD512

The New Way of Performance Acceleration for Exascale Systems


SUMMARY