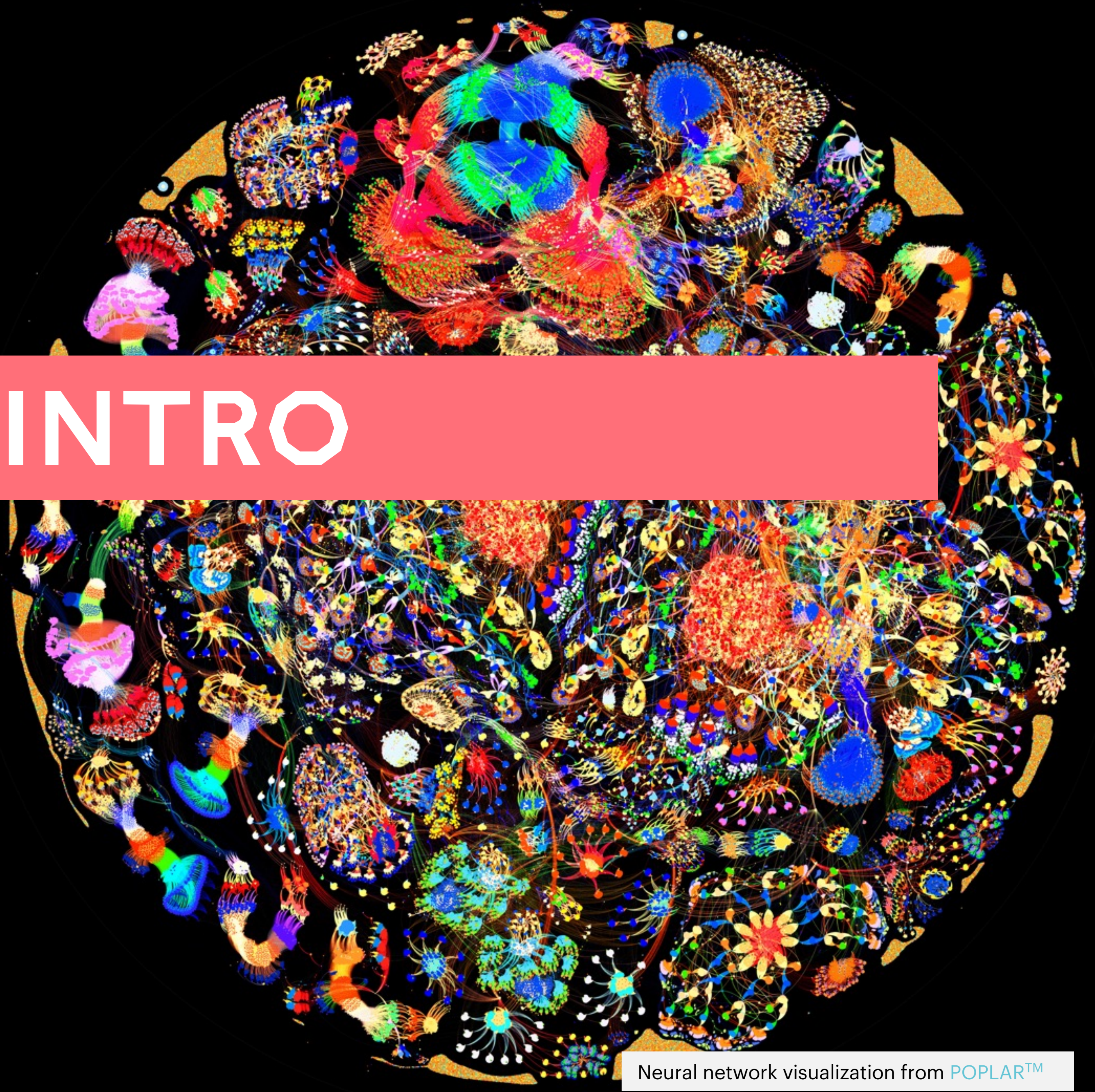
GRAPHCORE IPU INTRO

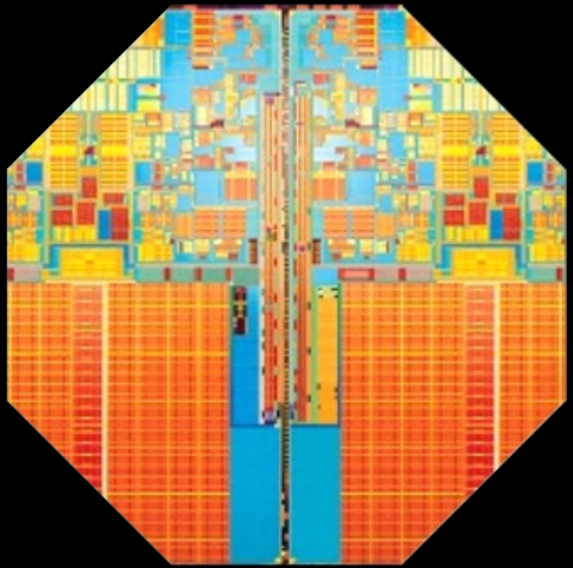
GRAPHCORE

Dr Alex Titterton
Solutions Architect



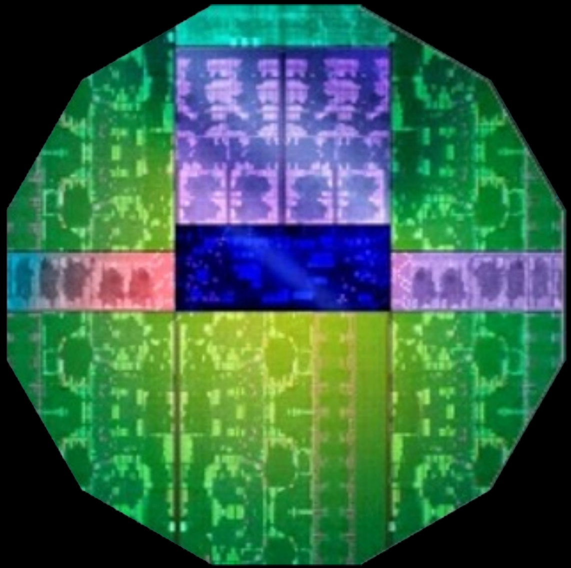
Neural network visualization from POPLAR™

INTELLIGENCE PROCESSING UNIT
DESIGNED FOR AI



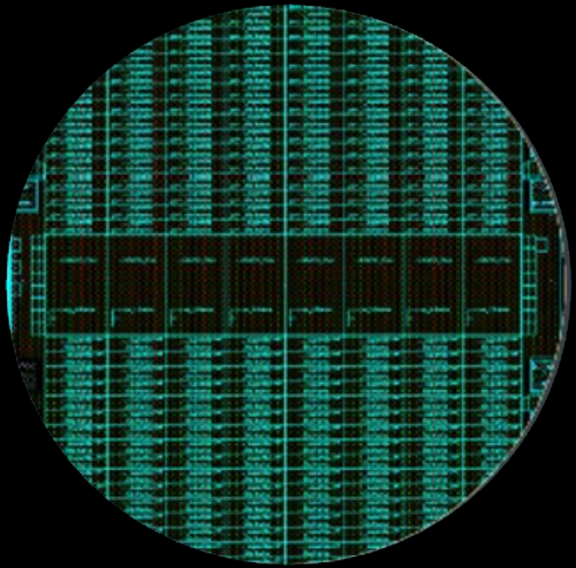
CPU

Scalar



GPU/TPU

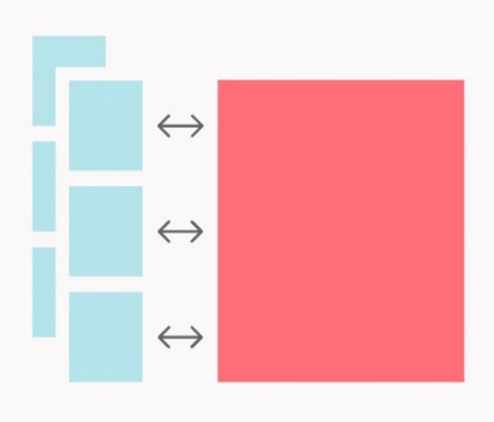
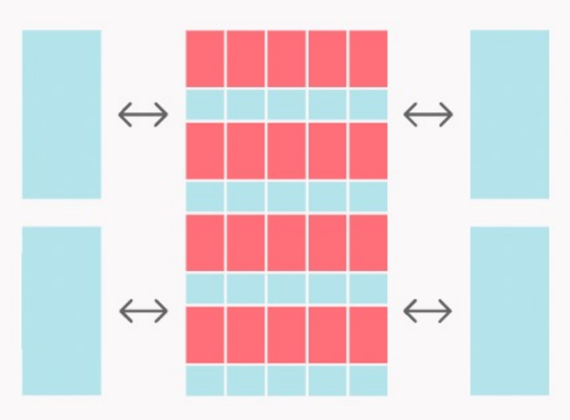
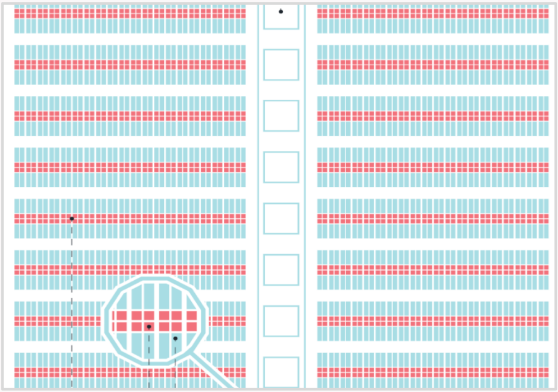
Vector



IPU

Graph

THE INTELLIGENCE PROCESSING UNIT (IPU)
WHAT MAKES IT DIFFERENT?

	CPU	GPU	IPU
Parallelism	Designed for scalar processing	SIMD/SIMT architecture. Designed for large blocks of dense contiguous data	Massively parallel MIMD architecture. High performance/efficiency for future ML trends
Processor Memory			
Memory Bandwidth	Off-chip memory	Model and Data spread across off-chip and small on-chip cache and shared memory (2TB/s for A100 HBM)	Main Model & Data in tightly coupled large locally distributed SRAM (~65 TB/s for Bow IPU)

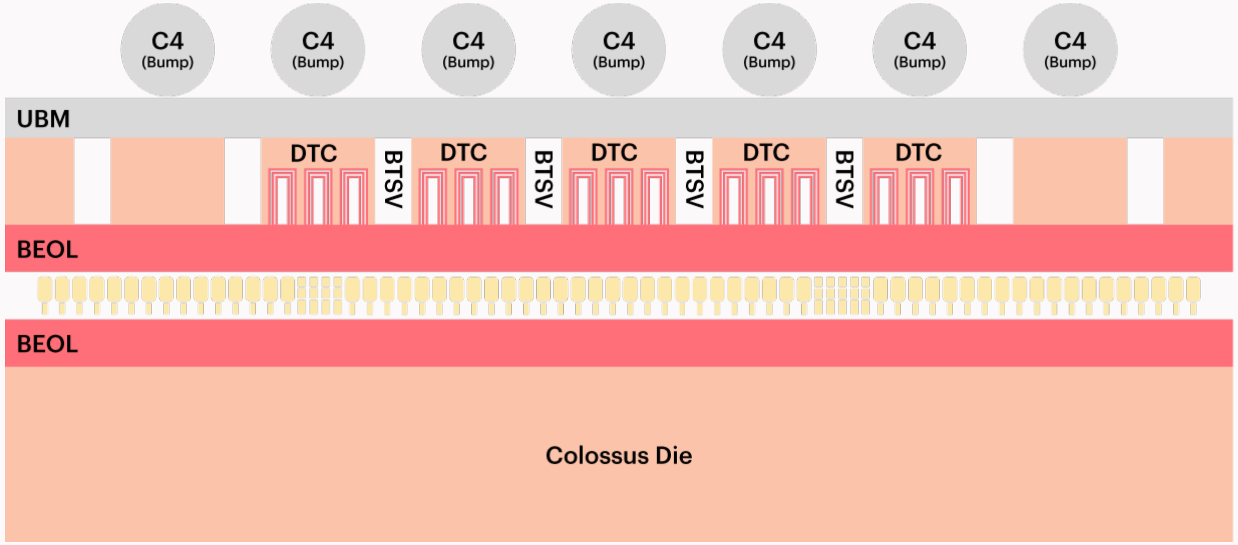
INTRODUCING THE BOW IPU

WORLD'S FIRST 3D WAFER-ON-WAFER PROCESSOR

- 3D silicon wafer stacked processor
- 350 TeraFLOPS AI compute
- Optimized silicon power delivery
- 0.9 GigaByte In-Processor-Memory @ 65TB/s
- 1,472 independent processor cores
- 8,832 independent parallel programs
- 10x IPU-Links™ delivering 320GB/s

BOW IPU: 3D WAFER-ON-WAFER PROCESSOR

GRAPHCORE + tsmc



Advanced silicon wafer stacking technology co-developed between Graphcore and TSMC

World's first commercial deployment using TSMC SoIC-WoW™ technology in Bow IPU

Enabling technology for closely coupled power delivery die to maximize application performance

GRAPHCORE SOFTWARE

ML APPLICATIONS

- NLP/TRANSFORMERS
- IMAGE CLASSIFICATION/CNNS
- OBJECT DETECTION
- LARGE MODELS
- MLPERF
- CONDITIONAL SPARSITY
- GNNS

DEVELOPER ECOSYSTEM

- TUTORIALS
- CODE EXAMPLES
- DOCUMENTATION
- VIDEOS
- NATIVE IPU CODERS PROGRAM
- APPS PORTFOLIO

POPLAR® SDK

FRONTENDS
- JUPYTER NOTEBOOKS
- INFERENCE DEPLOYMENT TOOLKIT

FRAMEWORKS
- ONNX
- HALO
- Keras
- PaddlePaddle

FW BACKENDS
- XLA
- POPART+
- POPDIST
- PARTITIONER
- POPIR
- POPIT

POPLAR®
- POPLIBS
- GCL
- POPLAR
- GRAPH ENGINE
- GRAPH COMPILER

DRIVERS
- GC DEVICE ACCESS LAYER
- IPUOF DRIVER
- PCIe DRIVER

POPVISION TOOLS

- GRAPH ANALYZER
- SYSTEM ANALYZER
- DEBUGGER
- DEVELOPMENT ENVIRONMENT

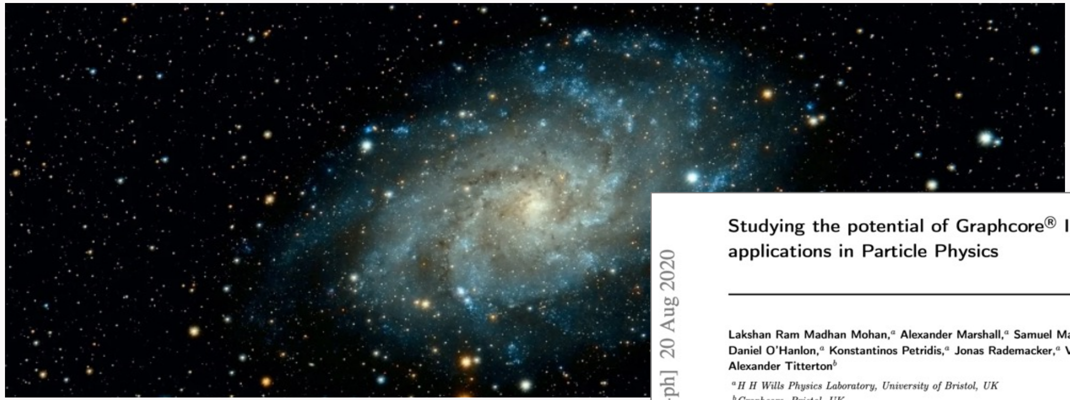
SYSTEM SOFTWARE

- V-IPU
- SYSTEM MONITORING
  - PROMETHEUS
  - GRAFANA
- JOB DEPLOYMENT
  - K8S
  - SLURM

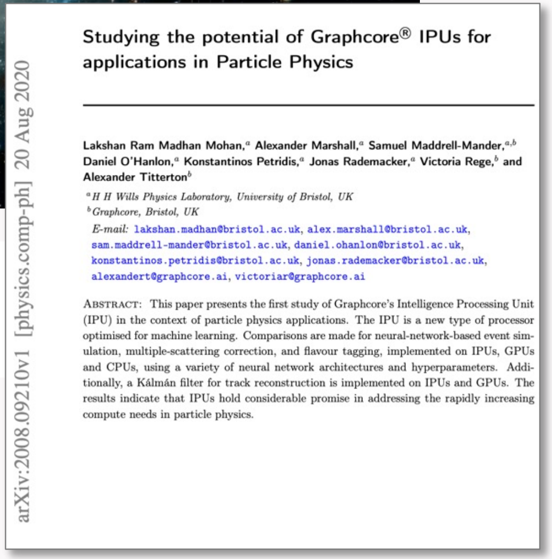
GRAPHCORE RESEARCH COLLABORATIONS (SMÖRGÅSBORD)

ResNet50 POPLAR™ compute graph

IPUs in Research



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



arXiv:2008.09210v1 [physics.comp-ph] 20 Aug 2020

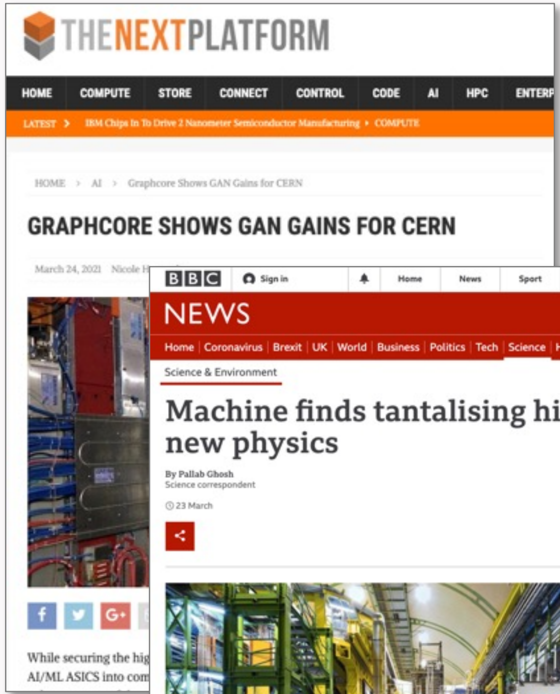
Studying the potential of Graphcore® IPUs for applications in Particle Physics

Lakshan Ram Madhan Mohan,[a] Alexander Marshall,[a] Samuel Maddrell-Mander,[a,b] Daniel O'Hanlon,[a] Konstantinos Petridis,[a] Jonas Rademacker,[a] Victoria Rege,[b] and Alexander Titterton[b]

[a] H H Wills Physics Laboratory, University of Bristol, UK
[b] Graphcore, Bristol, UK

E-mail: lakshan.madhan@bristol.ac.uk, alex.marshall@bristol.ac.uk, sam.maddrell-mander@bristol.ac.uk, daniel.ohanlon@bristol.ac.uk, konstantinos.petridis@bristol.ac.uk, jonas.rademacker@bristol.ac.uk, alexandert@graphcore.ai, victoriar@graphcore.ai

ABSTRACT: This paper presents the first study of Graphcore's Intelligence Processing Unit (IPU) in the context of particle physics applications. The IPU is a new type of processor optimised for machine learning. Comparisons are made for neural-network-based event simulation, multiple-scattering correction, and flavour tagging, implemented on IPUs, GPUs and CPUs, using a variety of neural network architectures and hyperparameters. Additionally, a Kálmán filter for track reconstruction is implemented on IPUs and GPUs. The results indicate that IPUs hold considerable promise in addressing the rapidly increasing compute needs in particle physics.



University of BRISTOL

CERN

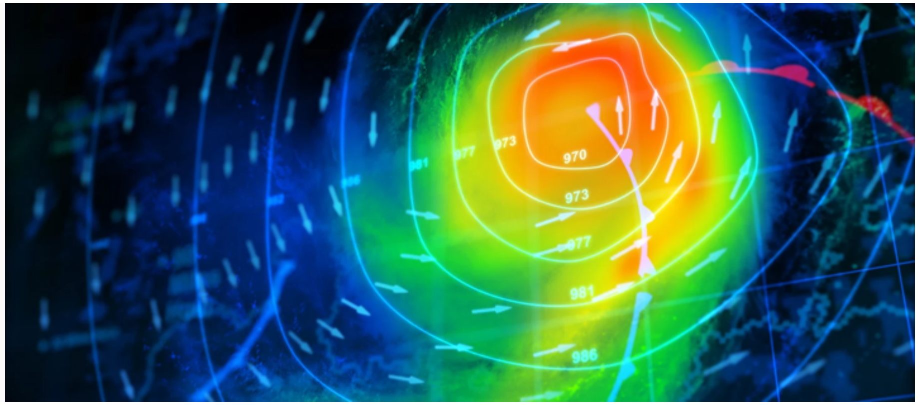
https://www.graphcore.ai/resources/research-papers

Using AI to accelerate HPC Scientific Applications

Mar 09, 2022 \ AI, HPC

AI FOR SIMULATION: HOW GRAPHCORE IS HELPING TRANSFORM TRADITIONAL HPC

Written By:
Alex Titterton



SHARE:

SUBSCRIBE

Get Updates

For many years High Performance Computing (HPC) techniques have been used to solve the world's most complex scientific problems across a wide range of applications, from modelling Higgs boson decay at the Large Hadron Collider to using Monte-Carlo simulation to predicting whether the weather will improve.

However, due to the immense complexity of the calculations involved in many of these applications, researchers are often waiting a long time for simulation results to arrive. Speeding up these workflows by simply running the same programs on more powerful hardware can be very expensive, with a large cost often giving only a modest improvement in performance.

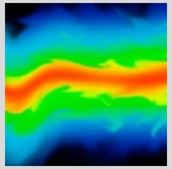
Clearly, a new approach is required to efficiently speed up these workloads, and many researchers are turning to surrogate machine learning models.

A surrogate model is a machine learning model intended to imitate part of a traditional HPC workflow,

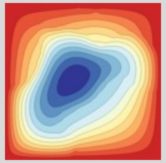
For more information, see our technical blog post:
https://www.graphcore.ai/posts/ai-for-simulation-how-graphcore-is-helping-transform-traditional-hpc

Relevant Application Areas

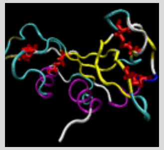
High Energy Physics



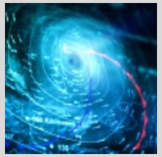
Computational Fluid Dynamics



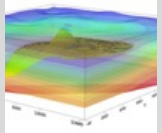
Partial Differential Equations



Protein Folding



Weather Forecasting

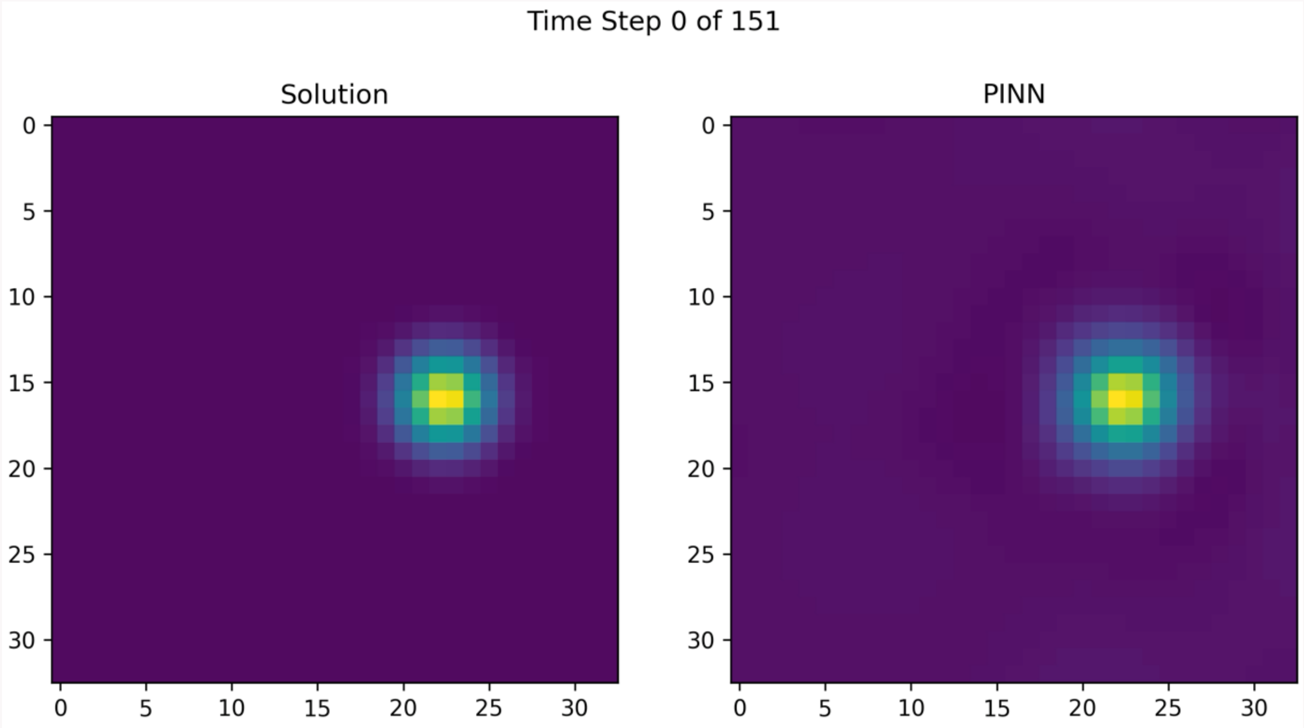


Oil & Gas Exploration Simulation

Science and Technology Facilities Council
Hartree Centre

Physics-Informed Neural Networks (PINNs)

UK Atomic Energy Authority



RESULTS ORIGINALLY SHOWN AT SC22, SOLVING A 2D WAVE EQUATION USING A PHYSICS-INFORMED NEURAL NETWORK IMPLEMENTED IN TENSORFLOW 2.

2XIPU FOUND TO BE 11X FASTER THAN 1XA100 GPU, AT SIMILAR MONETARY & ENERGY COST.

WORK DONE IN COLLABORATION WITH STFC HARTREE AND
THE UK ATOMIC ENERGY AUTHORITY

Platform	Time to Train / seconds (20k Epochs)	Speedup vs GPU
2 Bow IPUs	41	11x
A100 GPU	530	-

GRAPHCORE

HOT TOPICS IN AI RESEARCH

ResNet50 POPLAR™ compute graph

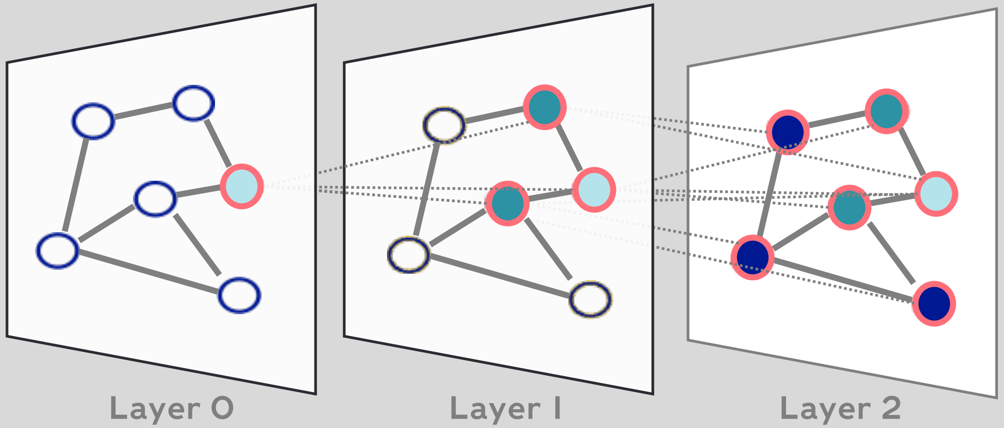
IPU for GNNs

The next frontier in AI/ML

GRAPHCORE

WHAT IS A GNN ?
GRAPH NEURAL NETWORK

GNNS ARE USED TO SOLVE GRAPH PREDICTION TASKS



Visualisation of how a node accumulates information from neighbouring nodes through the layers of the GNN

GNNs broadly follow a recursive neighbourhood aggregation (or message passing) scheme, where each node aggregates feature embeddings of its neighbours to compute its new feature embeddings[1]

[1] summary derived from 'How Powerful are Graph Neural Networks?' MIT/Stanford - https://arxiv.org/pdf/1810.00826.pdf

GNN USE CASES & APPLICATIONS

HEALTHCARE

DISEASE PREDICTION
RNA–disease association
Disease–gene association
COVID-19 spread prediction

DRUG DISCOVERY
Protein structuring
Protein function prediction
Protein / drug interaction
Drug response prediction

MEDICAL IMAGING
Image segmentation
Abnormal detection
Brain connectivity research
Surgical image analysis

PATIENT RISK PREDICTION
Mining EHRs (health records)

INTERNET

SOCIAL NETWORK ANALYSIS
Social influence prediction

RECOMMENDER SYSTEMS
User and item representations

FAKE NEWS DETECTION
Rumor detection + link classification

IDENTITY RESOLUTION
Real-time personalization & advertising

FINANCE

FRAUD DETECTION
Credit card, insurance, loan fraud

TRADE MARKET PREDICTION
Trader connection surveillance

RISK & COMPLIANCE
Risk analytics + compliance reporting

DATA MIGRATION
Data mapping and consolidation

INTEREST RATE RISK
Leveraging credit scores, employment, income, and other socio-economic factors

SCIENTIFIC RESEARCH

PARTICLE PHYSICS
Particle physics simulation

CHEMICAL PHYSICS
QSOR (structure-odor) modeling

TELECOMMS

WIRELESS COMMUNICATION
Power control
Resource allocation
Channel control
Link scheduling

GOVERNMENT

CRIME PREVENTION
Predicting crime associations

FRAUD DETECTION
Anti-money laundering & tax fraud

CONTACT TRACING
Disease contact tracing

TRANSPORT

TRAFFIC FORECASTING
Traffic speed/time prediction

GAMING

FRAUD DETECTION
Collusion detection in gaming

RECOMMENDER SYSTEMS
Online game recommendations

MANUFACTURING

BILL OF MATERIALS
360 degree BOM analysis

TRACEABILITY
Product recall tracing e.g. cars

MASTER DATA MGMT
RDF graph data modelling

SchNet GNN

Modelling Quantum Interactions in Molecules

Pacific Northwest
NATIONAL LABORATORY

Try on Paperspace

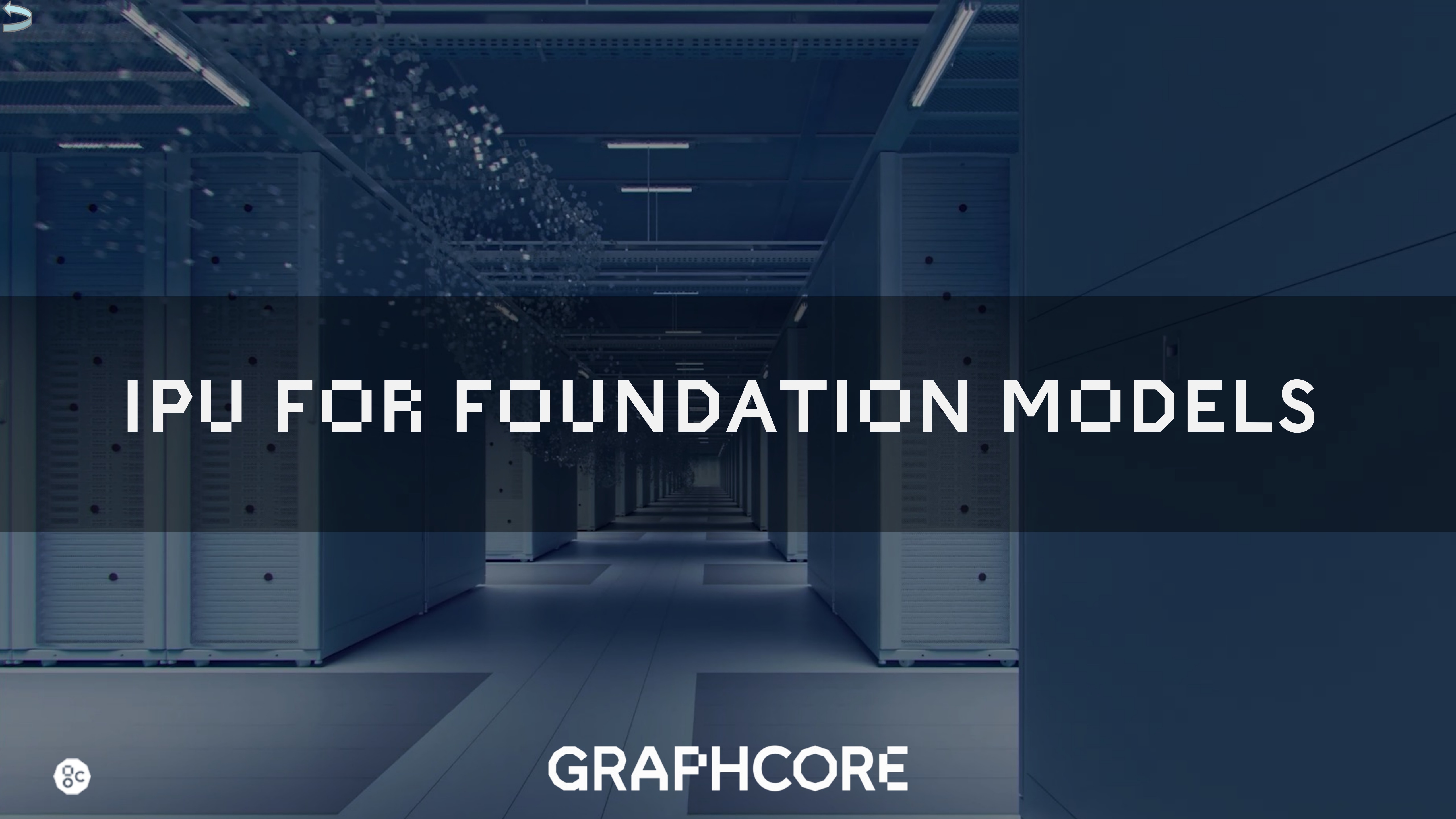
Graphcore engineers successfully trained the **SchNet**[1] model on IPUs on the **500k water clusters** dataset[2], to predict the **potential energy per cluster**.

Preliminary results show a time-to-train of **98 minutes** on 2xIPU-M2000, compared with >**60 hours** on 4xV100 GPUs in PNNL's original paper[2].

[1] JK. T. Schütt1, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller. "SchNet – A deep learning architecture for molecules and materials" *J. Chem. Phys.* **148**, 241722 (2018).

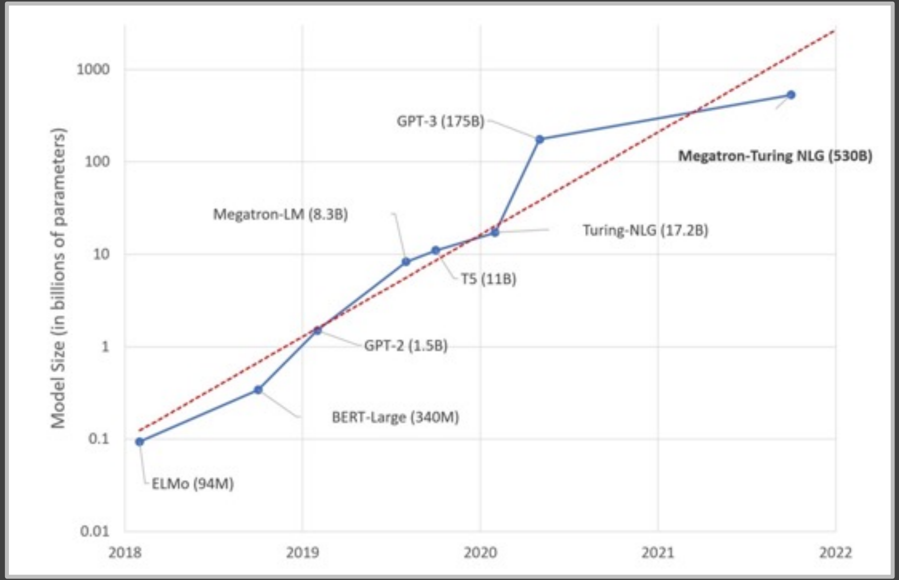
[2] Jenna A. Bilbrey, Joseph P. Heindel, Malachi Schram, Pradipta Bandyopadhyay, Sotiris S. Xantheas, and Sutanay Choudhury. "A look inside the black box: Using graph-theoretical descriptors to interpret a Continuous-Filter Convolutional Neural Network (CF-CNN) trained on the global and local minimum energy structures of neutral water clusters" *J. Chem. Phys.* **153**, 024302 (2020).

IPU FOR FOUNDATION MODELS



FOUNDATION MODEL TRENDS

- Models are getting much bigger to deliver ever higher demands on improved accuracy & performance
  - This growth is exponential for dense models
- Multimodal models broaden the learning capability by incorporating different modalities (e.g. linguistic, visual, aural)
  - => larger model demands
- Larger dense models mean more compute, more power, more cost
- Counter to this are economic and societal drivers to reduce energy consumption & cost

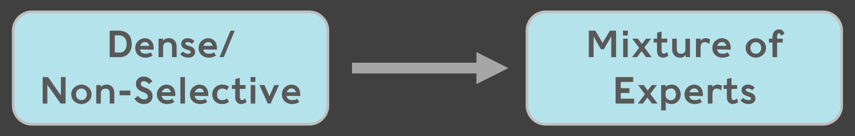


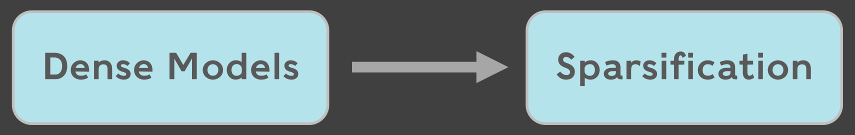
Exponential trend of SOTA NLP models:
Source: Microsoft/NVIDIA https://arxiv.org/abs/2201.11990

IMPROVING MODEL EFFICIENCY

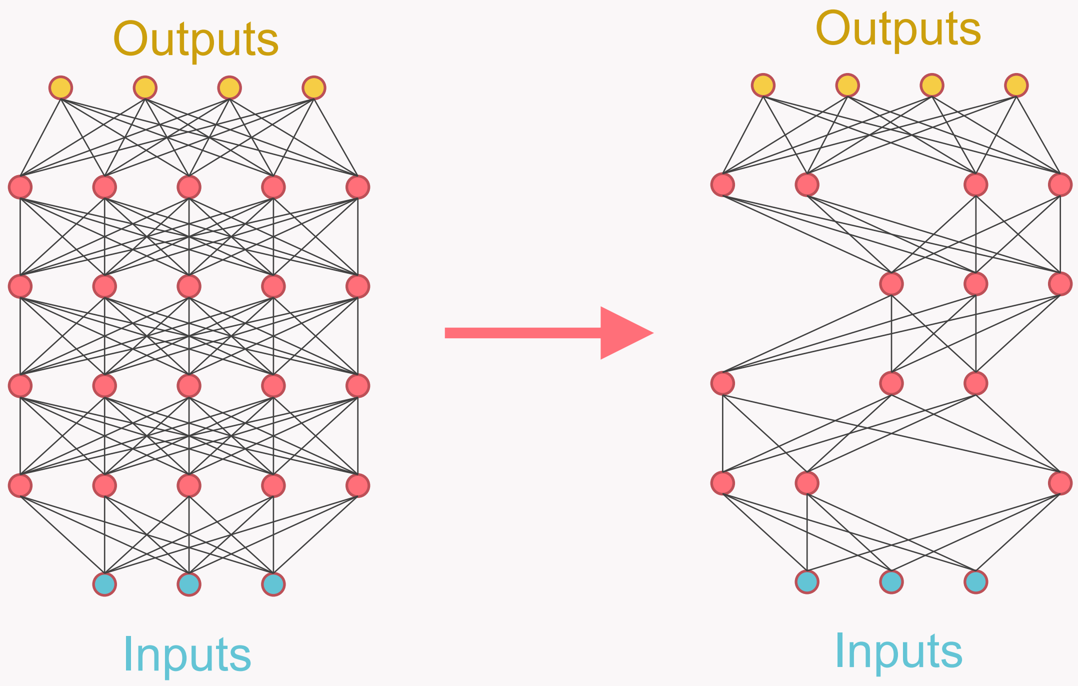
- Selectivity / Conditional Models
  - Models need to become selective (or conditional), such as Mixture of Experts (MoE) based models
  - Different parts of models are only used when needed
  - This can help reduce compute growth to linear instead of exponential

- Sparsification of models
  - Only incur cost of compute when required
    - ➢ Lower memory requirement
    - ➢ Fewer multiplications
    - ➢ Lower power





WHAT IS SPARSITY?



Only a small proportion of connections are key to model behaviour

So we can prune and re-train to create a "sparse" model

This is beneficial on a processor like the IPU that can do the sparse computation efficiently


THE 'GOOD' COMPUTER
GRAPHCORE

ROADMAP TO ULTRA-INTELLIGENCE AI

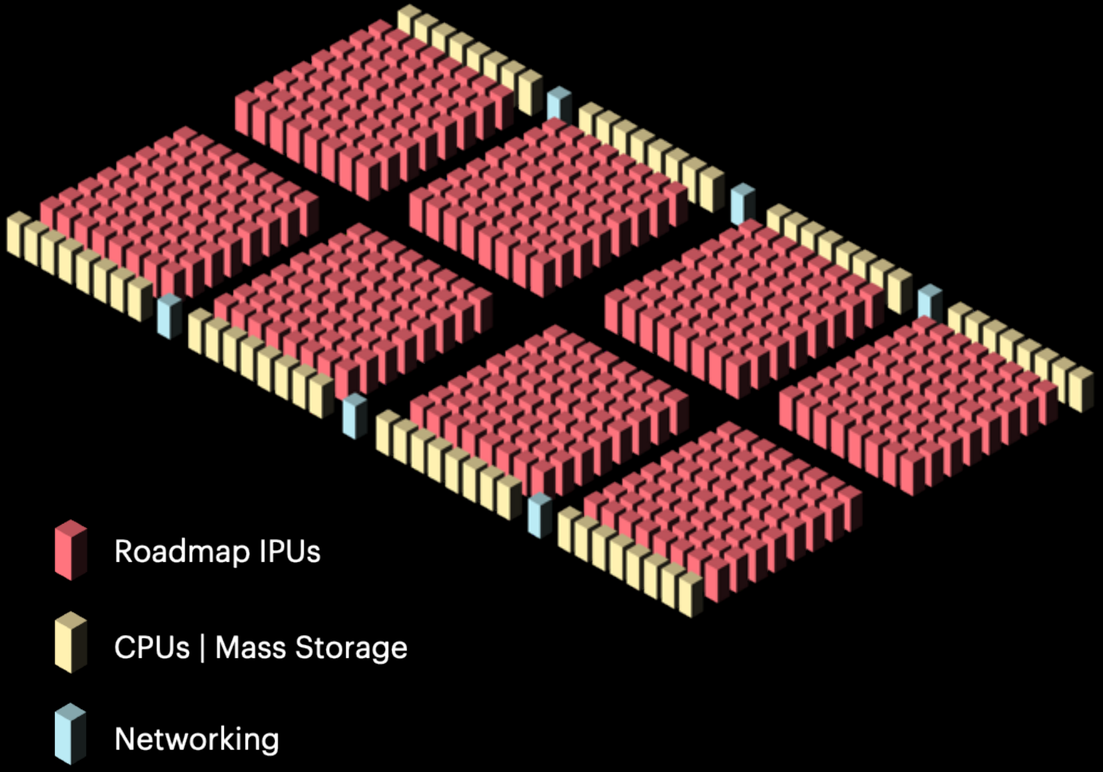
Human brain has around 100 billion neurons

With 100Tn+ synapses, equivalent to parameters in an AI model

Current largest AI models are around 1Tn parameters

Graphcore is developing an Ultra-Intelligence Machine
that will surpass the parametric capacity of the brain

THE 'GOOD' COMPUTER



Over 10 Exa-Flops of AI floating point compute from 8,192 roadmap IPUs

3D Wafer-on-Wafer logic stack

Up to 4 PB of memory with bandwidth of over 10 PB/s

Enabling AI models to be developed with 500 Tn parameters

Fully supported by Poplar® SDK

GRAPHCORE

APPLICATIONS

Very AI Large-Models:

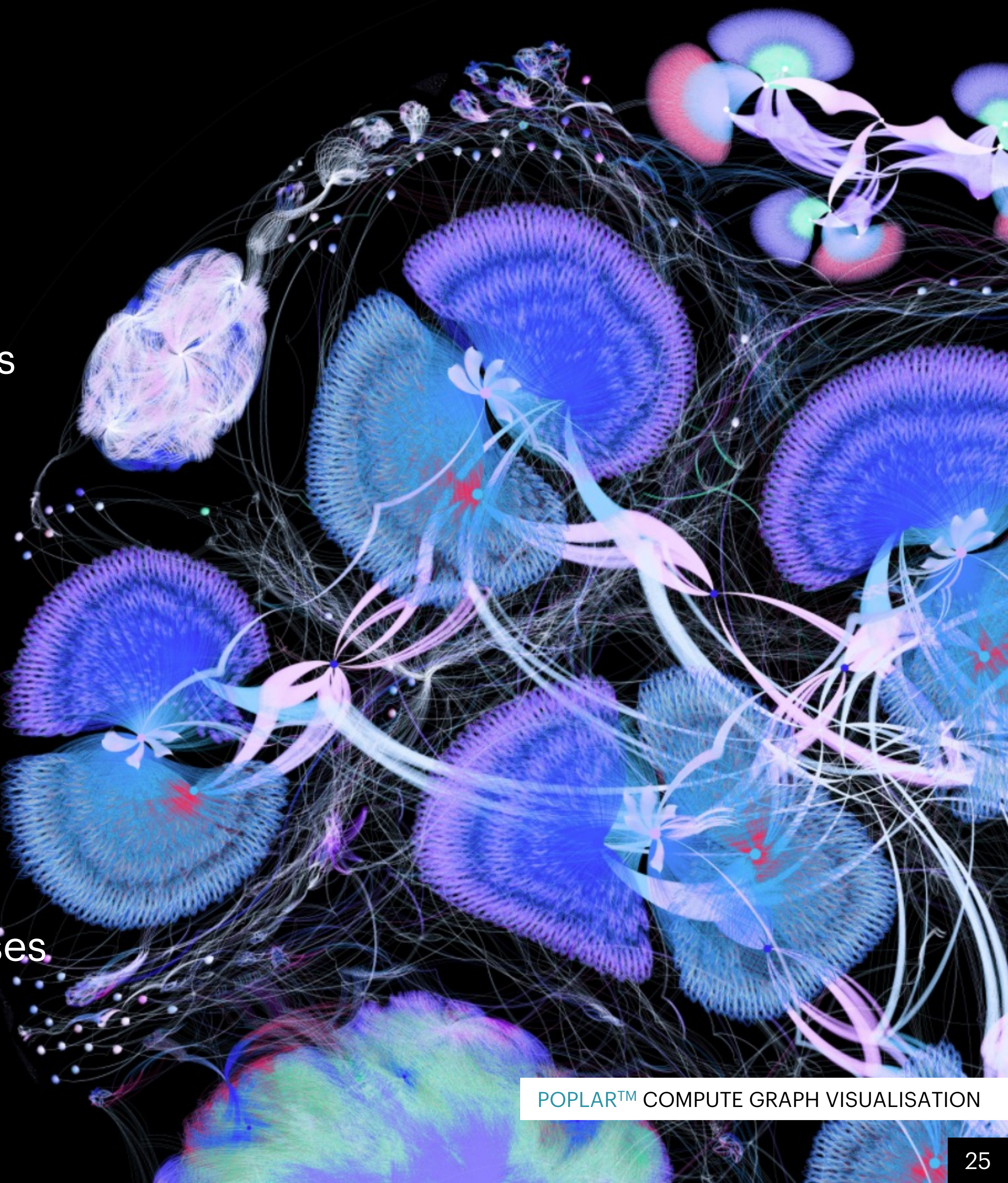
- Muti-trillion parameter model training and inference
- Next-generation, large, conditional and sparse models

AI in Science and Industry

- Healthcare: Genomics | Proteomics | Analysis
- AI-HPC: Simulation | Modeling
- Autonomous systems
- Materials Science | Manufacturing
- Environment: Weather prediction | Smart city

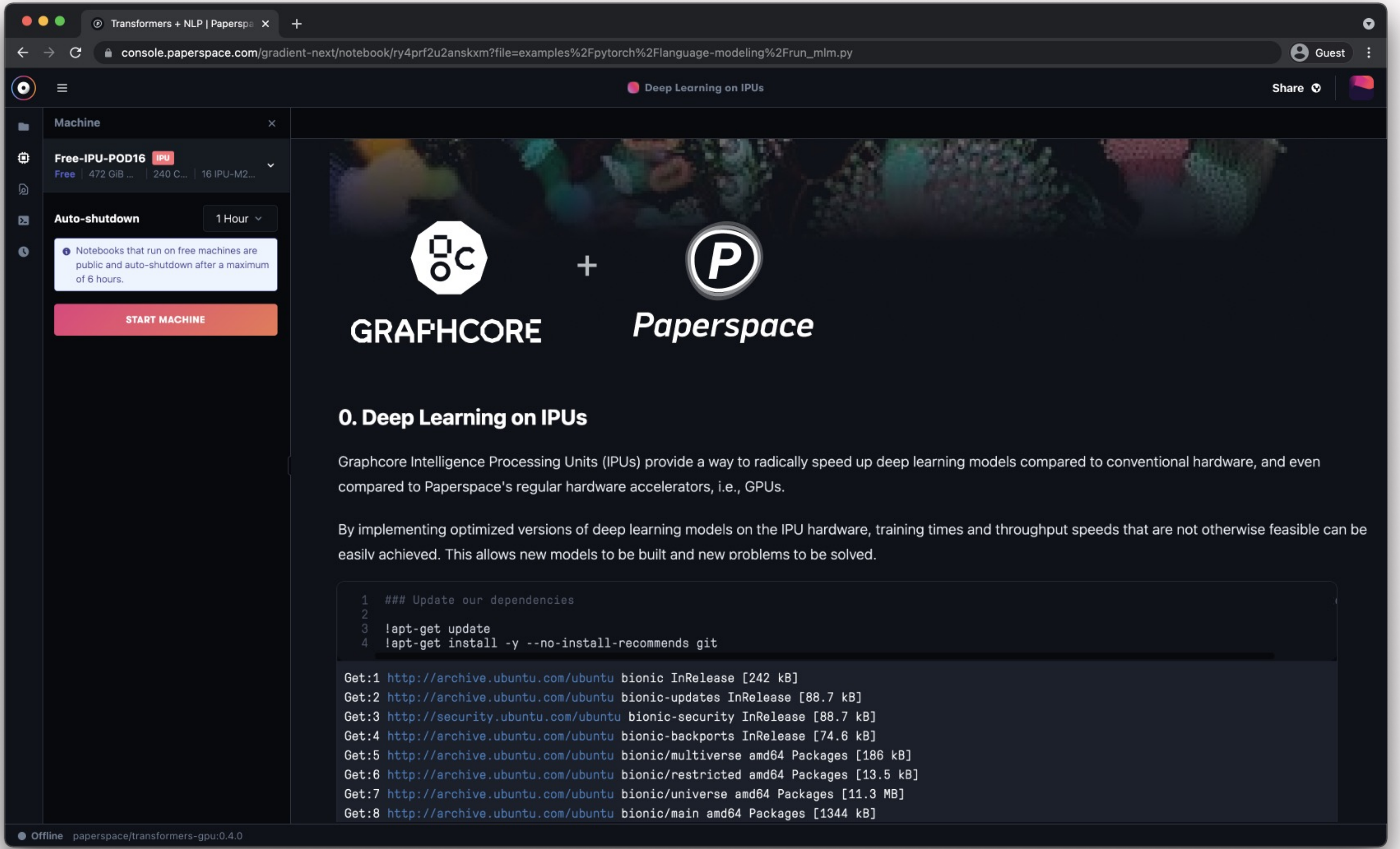
AI in Business

- Language understanding | Process automation | Bots
- Advanced big-data graph analytics and graph databases
- Next generation Recommenders



POPLAR™ COMPUTE GRAPH VISUALISATION

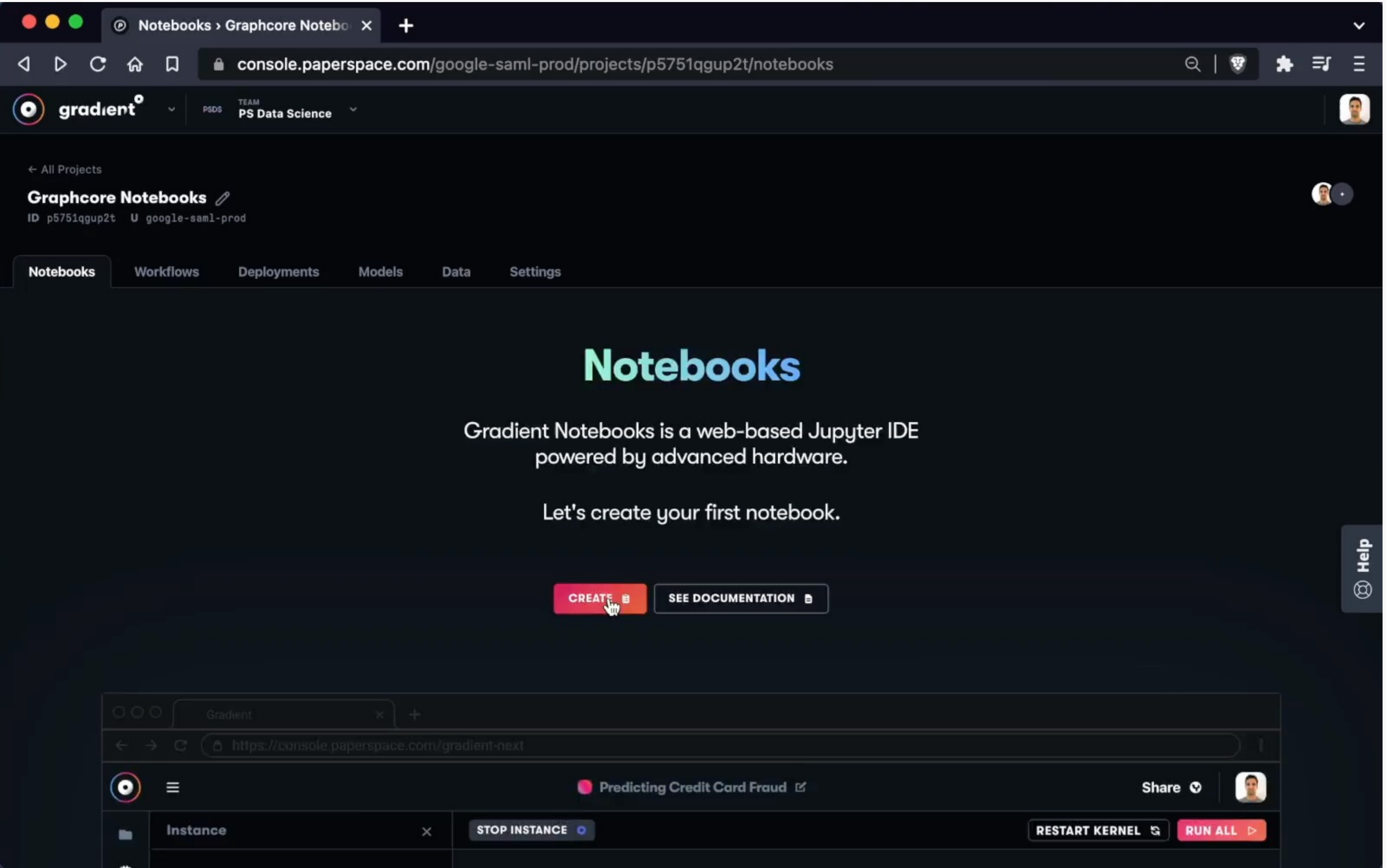
IPUS IN THE CLOUD



Paperspace

Free IPU Access:

IPUS IN THE CLOUD



Paperspace

Free IPU Access:

GRAPHCORE ACADEMIC PROGRAMME

Apply at:
graphcore.ai/academic

Test IPU hardware in the cloud at no-cost

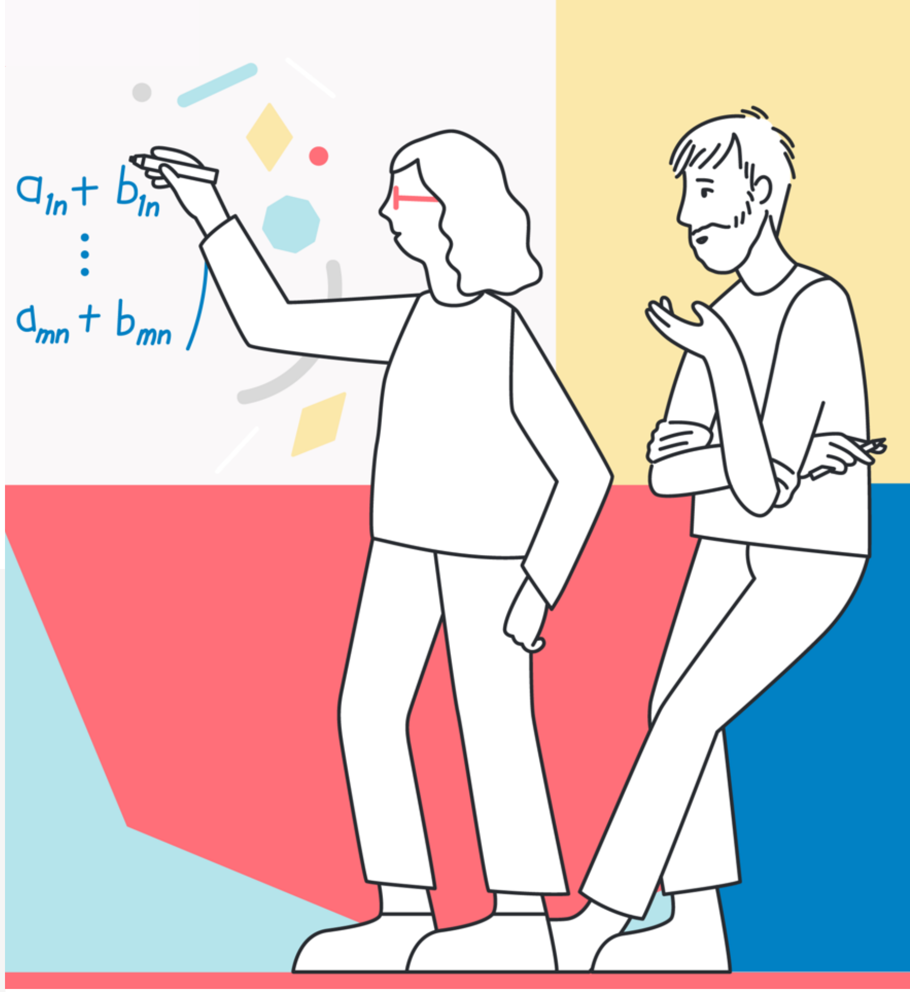
Support letters for grants & funding

Access to Poplar® & PopART® software

Support from Graphcore Researchers

RESEARCH PRIORITIES:

- Optimisation of Stochastic Learning
- New Efficient Models for Deep Learning & Graph Networks
- Sparse Training
- New Directions for Parallel Training
- Local Parallelism
- Multi-Model Training
- Conditional Sparse Computation



Apply at:
graphcore.ai/academics

GRAPHCORE ACADEMIC PROGRAMMES

Graphcore's Academic Programme provides tools and technology to educators, researchers and students to enable ground-breaking research and innovation in higher education.

MACHINE INTELLIGENCE ACADEMY

The Machine Intelligence Academy is designed to support professors, researchers, and students working in advanced AI and machine learning fields to accelerate their research and develop teaching on IPUs.

Learn more →

ACCELERATED COMPUTING ACADEMY

The Accelerated Computing Academy is aimed at C++ computer scientists looking to build new applications that transcend today's machine learning, ranging from HPC to simulation modelling, and beyond.

Learn more →

MACHINE INTELLIGENCE ACADEMY

A new age of IPU-accelerated discovery in computing

WHAT IS IT?

A collaboration initiative designed by Graphcore to enable researchers to develop novel AI techniques and accelerate their research using IPUs

WHO IS IT FOR?

Professors, researchers and students working in advanced AI, machine learning and related fields

WHY JOIN?

Members benefit from free IPU cloud credits, support letters, training workshops, engineering support, spotlight promotion and exclusive swag

Access to free IPU hardware in the cloud

Support letters for grant and funding proposals

Bespoke training workshops and educational materials

Support from Graphcore Engineers and SMEs

Project showcase and developer spotlight promotion

ACCELERATED COMPUTING ACADEMY

A new age of IPU-accelerated discovery in computing

WHAT IS IT?

A new computing academy designed by Graphcore to enable new applications that require highly parallel, high-performance compute

WHO IS IT FOR?

C++ computer scientists in academia looking to solve new problems through computationally intensive research that transcend AI and machine learning

WHY JOIN?

Members benefit from free IPU cloud credits, support letters, training workshops, internships, engineering support, spotlight promotion and exclusive swag

Access to free IPU hardware in the cloud

Support letters for grant and funding proposals

Bespoke training workshops and educational materials

Support from Graphcore Engineers and SMEs

Project showcase and developer spotlight promotion

THANK YOU

Dr Alex Titterton
alexandert@graphcore.ai

Academic programme:

Free 6-hour IPU Access:

WHY IPUS FOR GNNS

GNN REQUIREMENTS	IPU ADVANTAGE
Low arithmetic intensity due to large memory bandwidth requirements	Ultra-fast, large In-Processor Memory removes memory bandwidth contraints
GNNs often utilise multiple small graphs (or sub-graphs/clusters) & present unique gather/scatter requirements. These require memory intensive operations in parallel	Truly parallel implementation enabled by IPUs unique MIMD architecture
Graphs data structure is highly sparse, hardware capable of handling sparsity efficiently will have an advantage	Fast gather/scatter operation combined with distributed nature of IPU make sparsity its natural domain
Developers want to use high-level standard ML frameworks optimised for GNN	Standard ML Framework support including GNN focussed PyTorch Geometric
Dynamic graphs changing over time require small batch sizes	Optimised small batch size performance