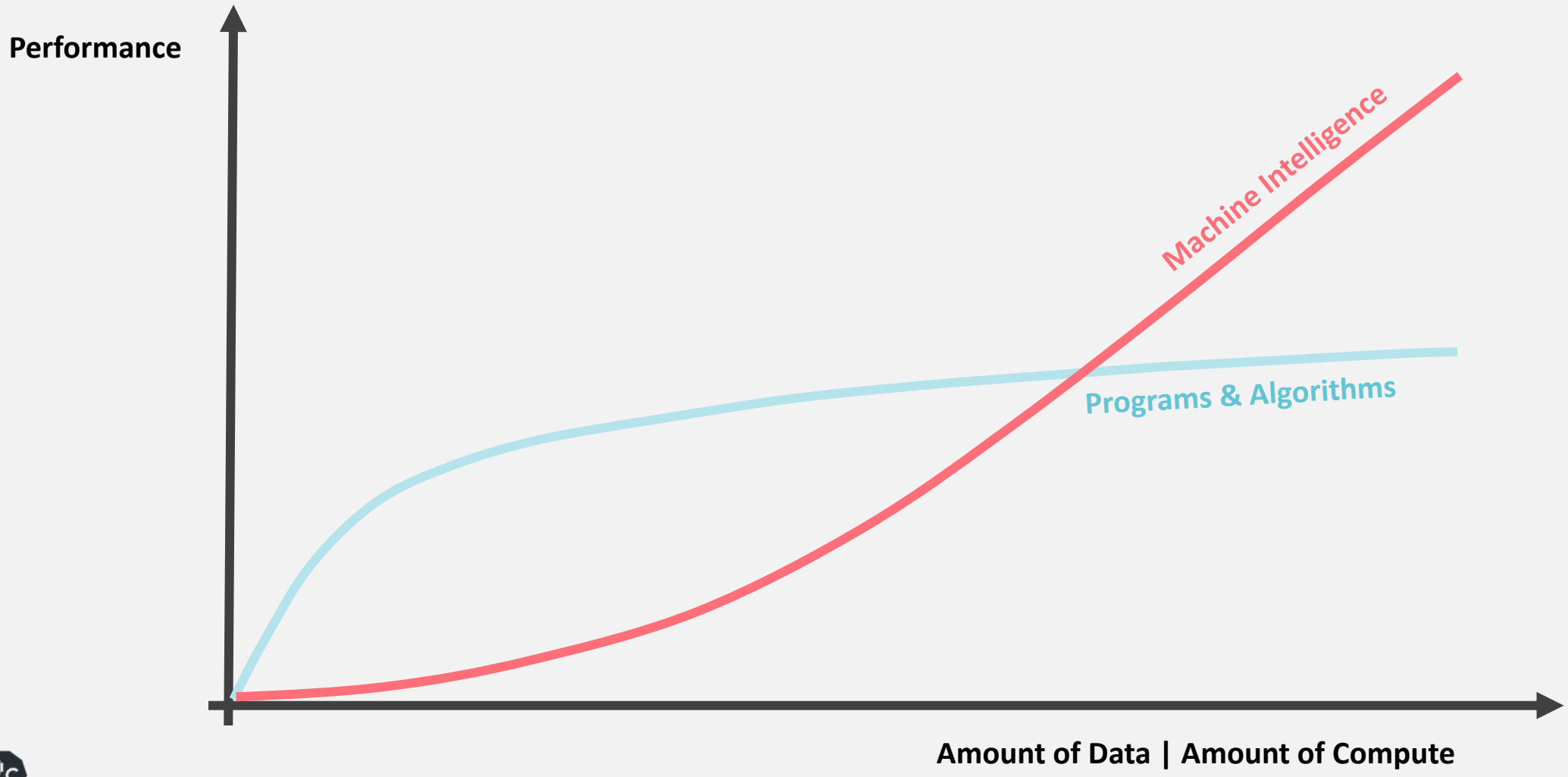
# Machine Intelligence at Supercomputer Scale

16th April 2020
WIDS Oslo

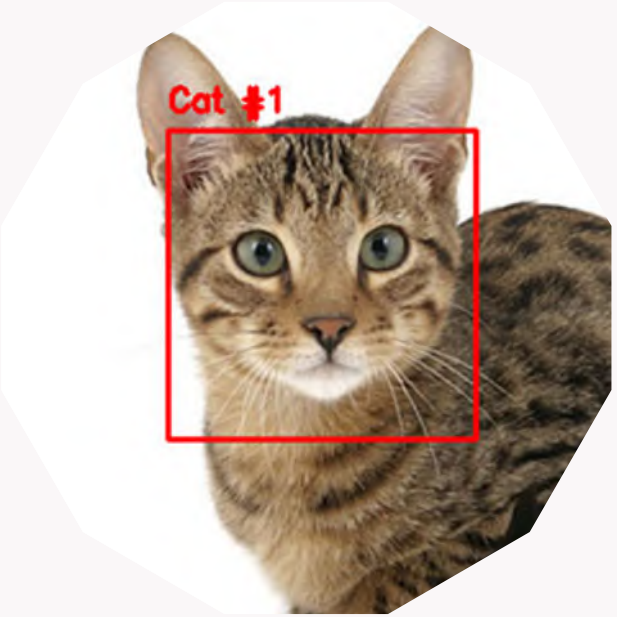
GRAPHCORE
ellied@graphcore.ai

# WE LIVE IN A NEW WORLD WHERE COMPUTERS LEARN FROM DATA



# MACHINE INTELLIGENCE IS EVOLVING

## TODAY



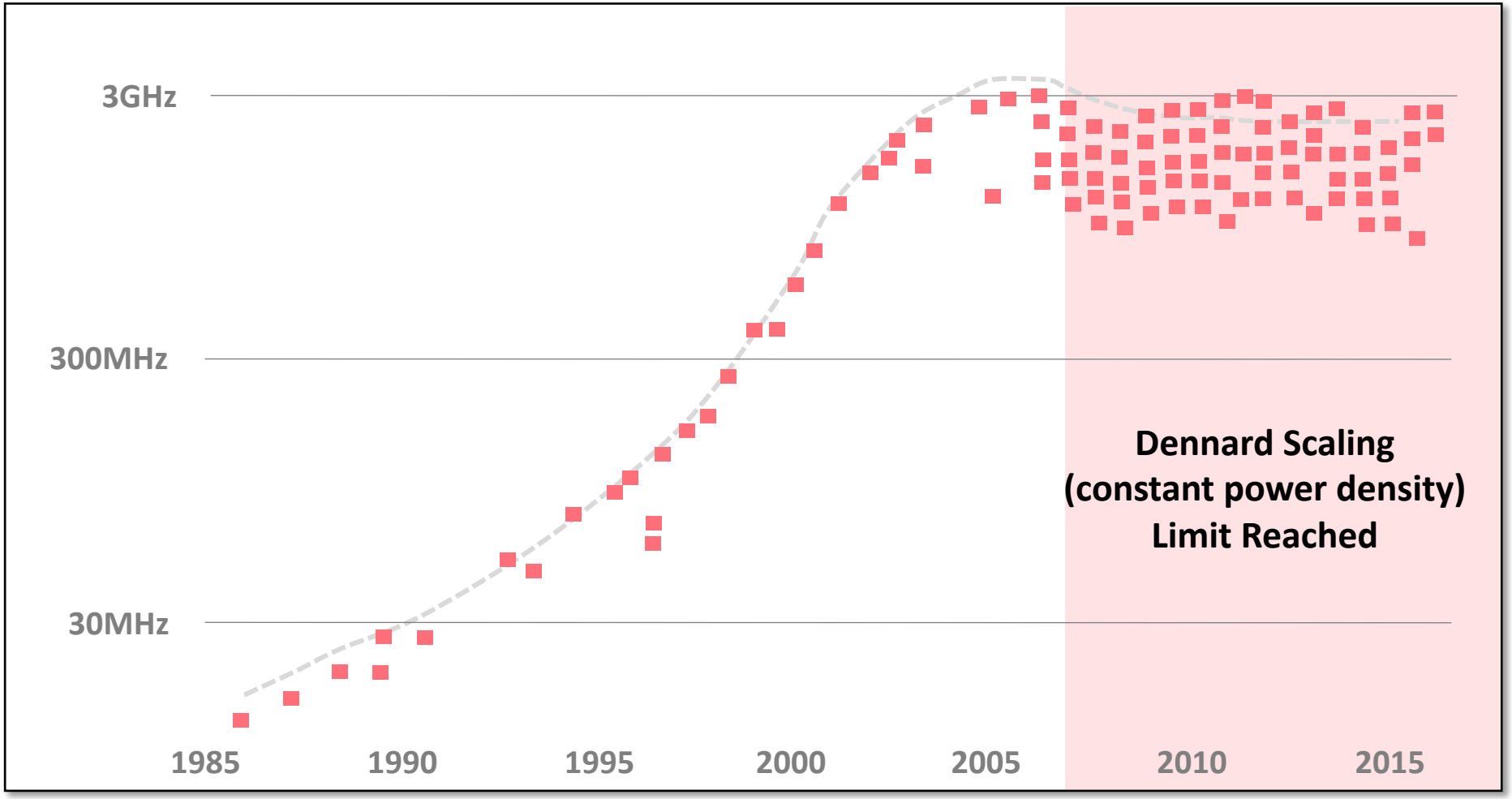

**Static Data**
**Perception**
**Train then Deploy**

## TOMORROW

**Sequence Data**
**Context and Understanding**
**Learn from Experience**


3

# PROCESSORS ARE NOT GETTING FASTER




Source: Stanford University

# THE OTHER MAJOR CHALLENGE IS MEMORY BANDWIDTH

PROCESSOR

MEMORY INTERFACE

MEMORY

If 10x faster…

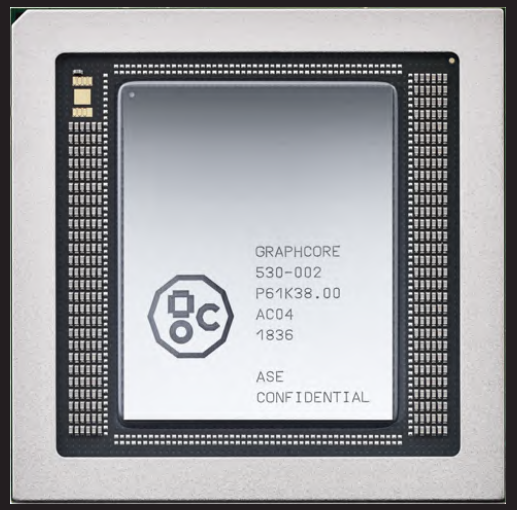
…10x faster how?


5

WIRED

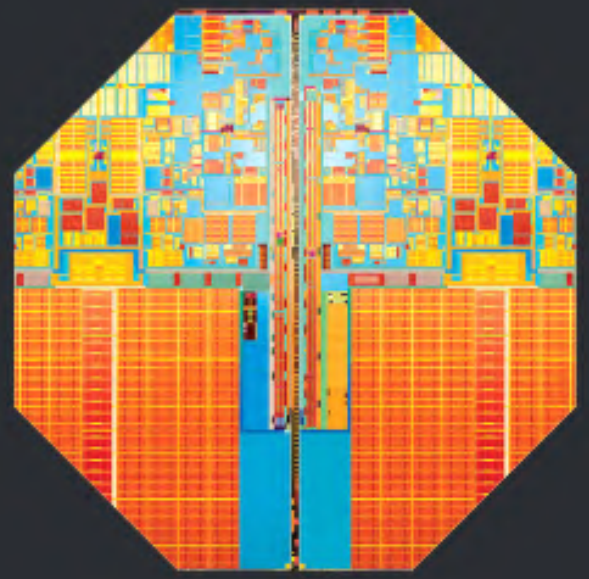
**Wired** – "How might we build machine learning systems that function more like a brain? "

**Geoff Hinton** – "I think we need to move towards a different type of computer. Fortunately I have one here…" Hinton reaches into his wallet and pulls out a large, shiny silicon chip:



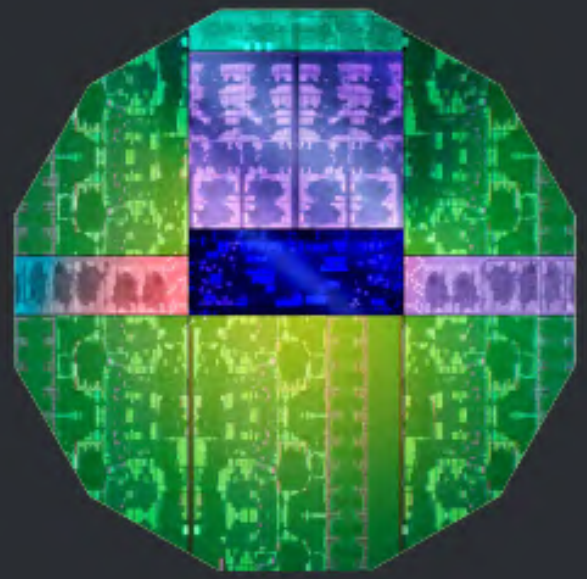

an IPU processor from Graphcore

# A NEW PROCESSOR IS REQUIRED



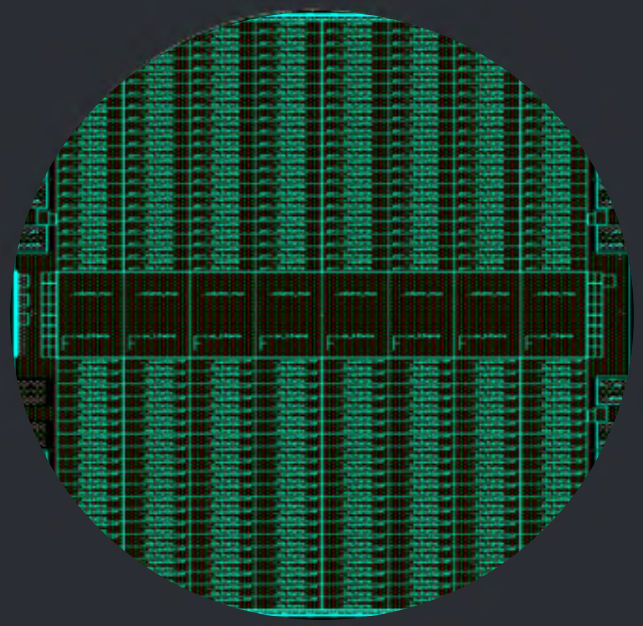
**CPU**

Apps and Web
Scalar



**GPU**

Graphics and HPC
Vector



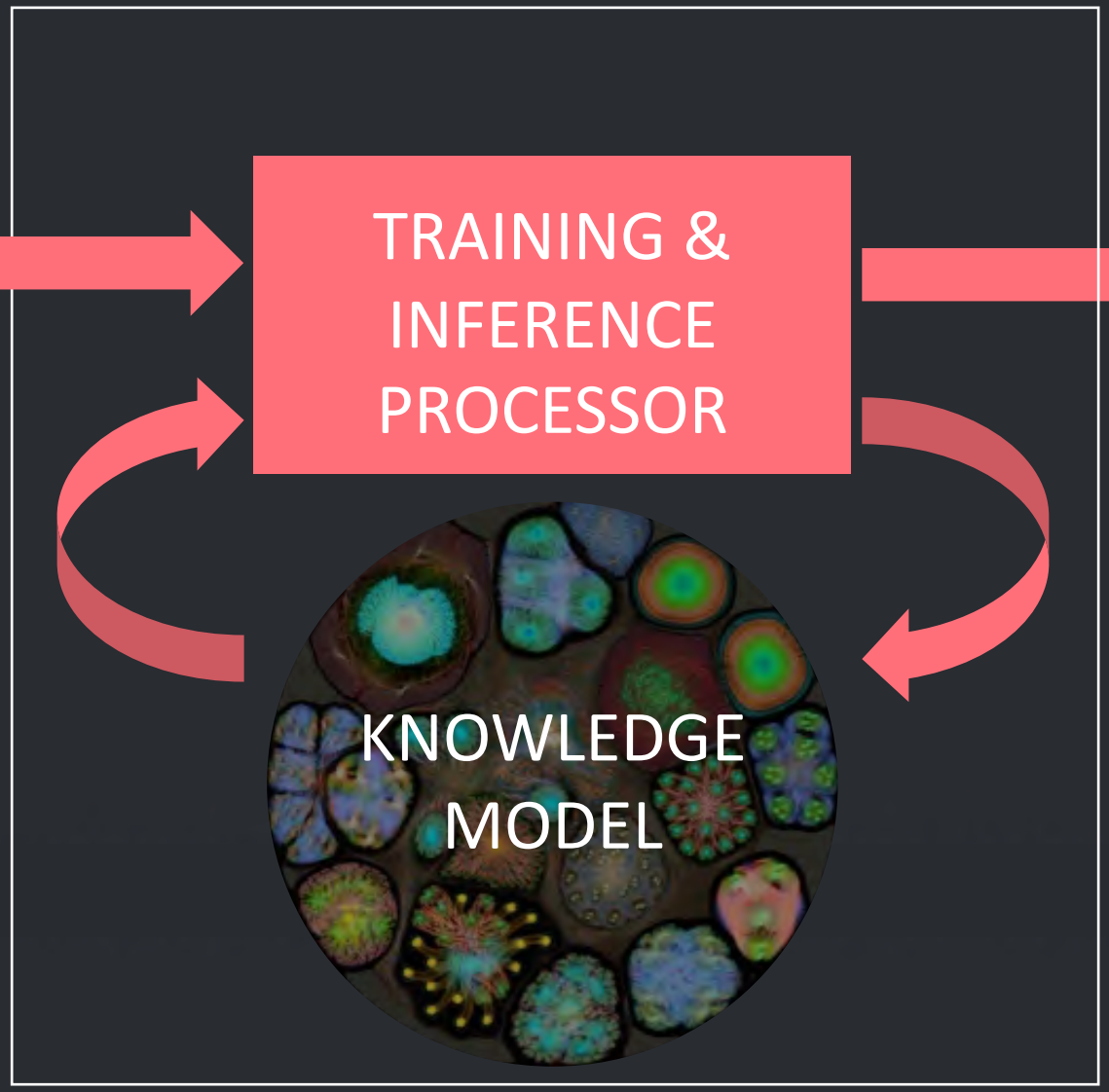
**IPU**

Machine Intelligence
Graph

# WHAT IS REQUIRED IN MACHINE INTELLIGENCE COMPUTE?

TRAINING DATA / SENSOR DATA

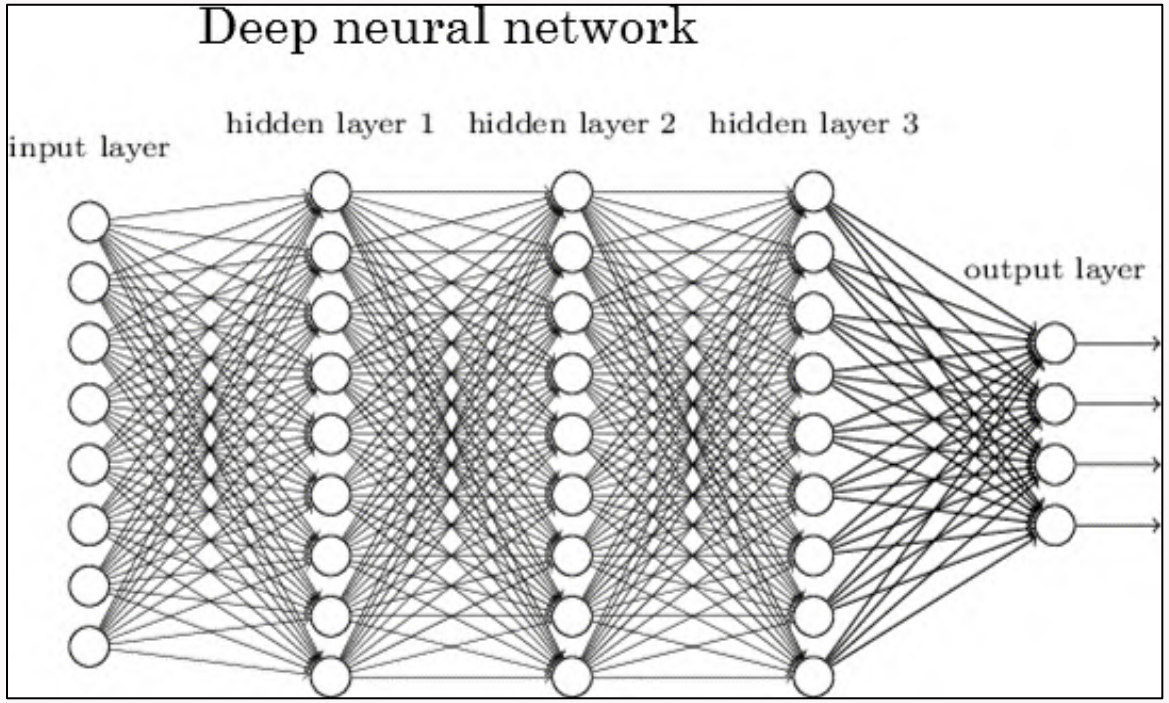
Learn parameters (and model) from data



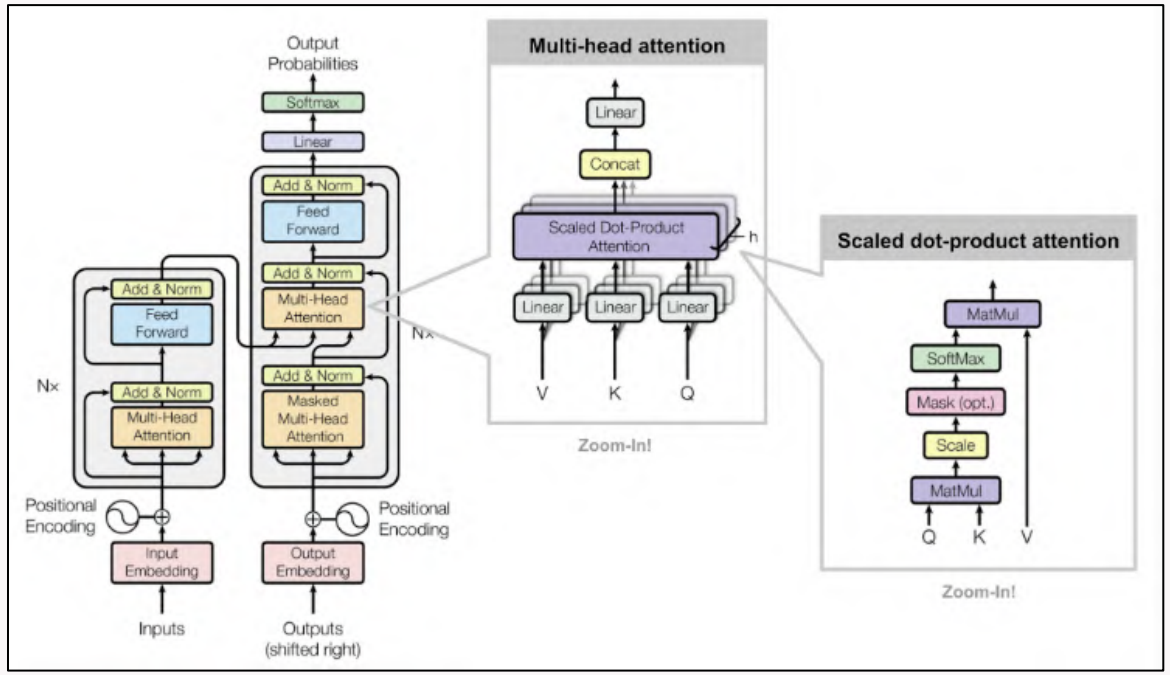

RESULT

Infer outputs using knowledge model

# A MACHINE INTELLIGENCE MODEL UNDER THE HOOD



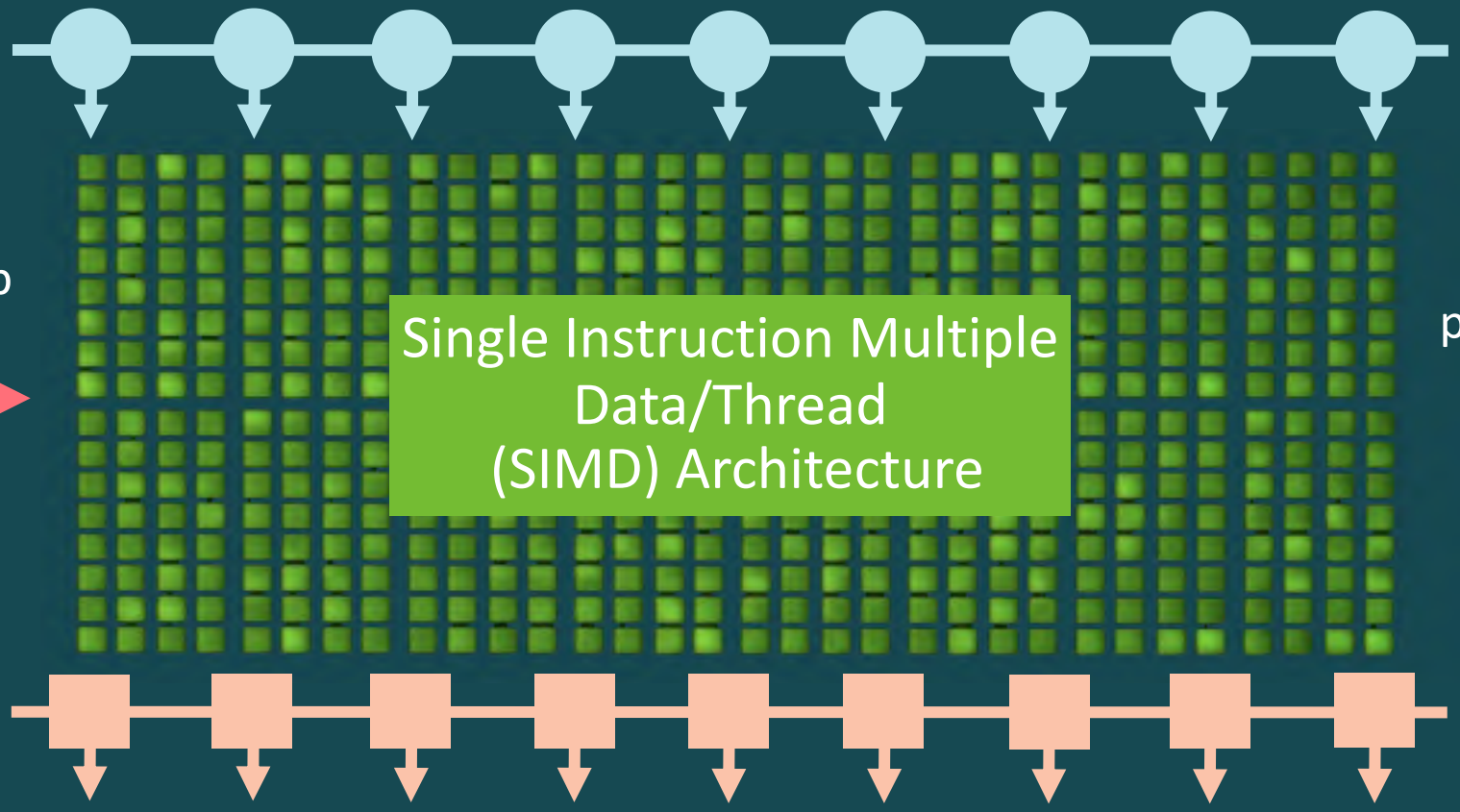

Sparse connection of input features




For SOA models such as BERT there can be very millions of parameters and hundreds of thousands of connections

# TODAY'S MACHINES ARE NOT ALWAYS SUITED TO THIS TASK

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



Data vector in

All units perform the same Instruction on each program step

Single Instruction Multiple Data/Thread (SIMD) Architecture

Small number of processors *(~80)* each with a wide vector unit

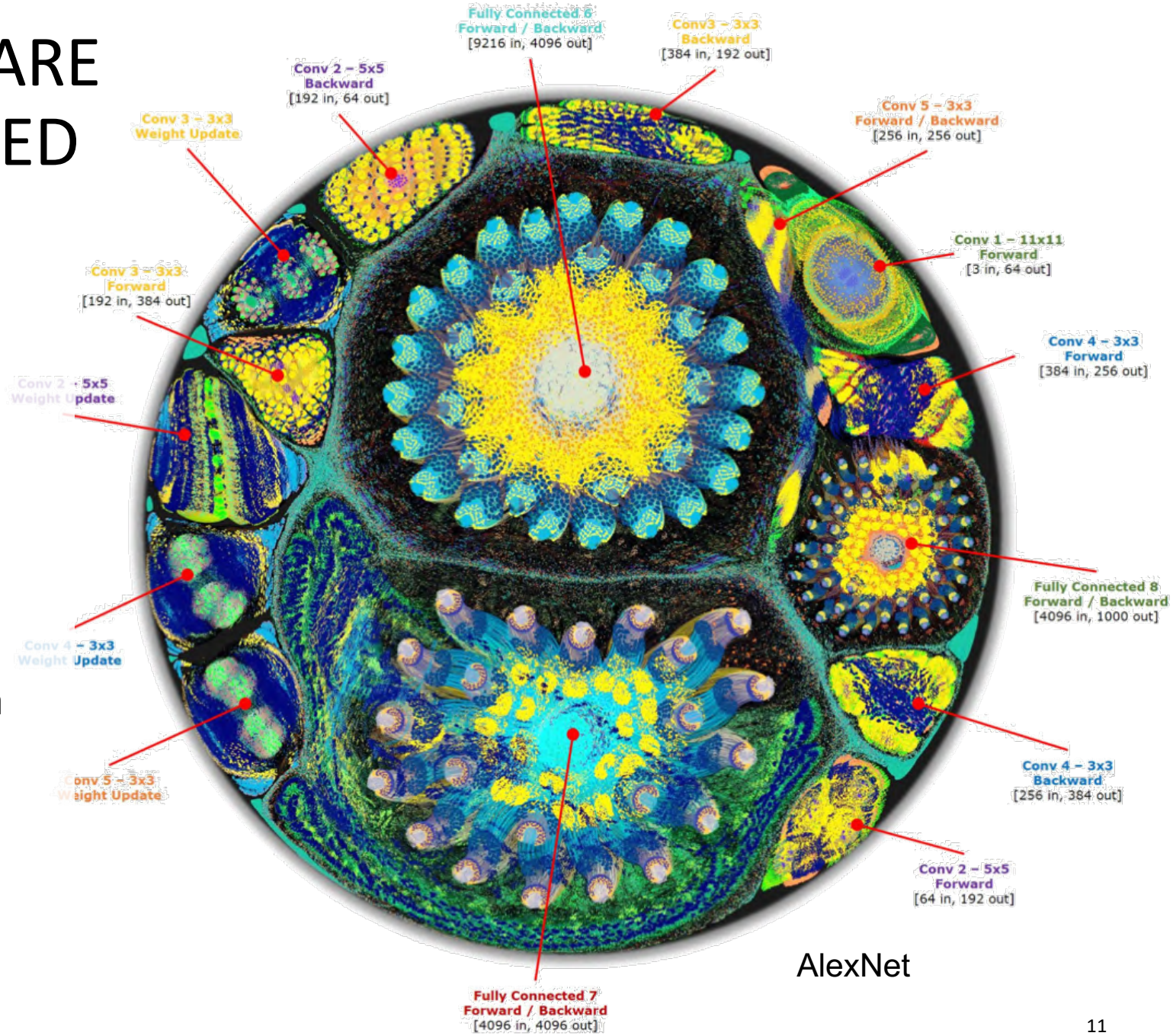
Data vector out

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

# KNOWLEDGE MODELS ARE NATURALLY REPRESENTED AS HUGE GRAPHS OF TENSORS

We need a processor that

- is massively parallel
- has ultra fast communication in between processors
- can deal efficiently with sparse data
- has a huge amount of memory bandwidth
- can deal with stochasticity (random number generation)




AlexNet

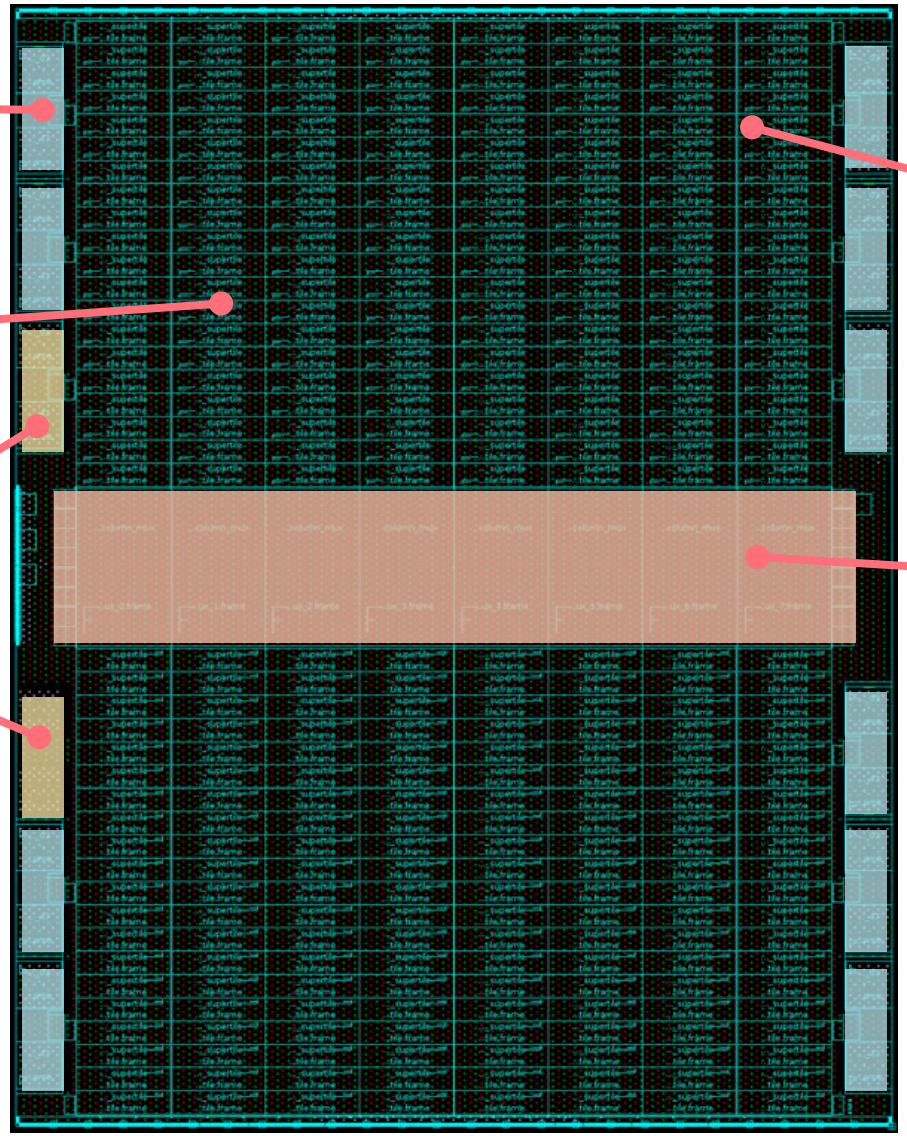
# COLOSSUS GC2

The world's most complex processor chip with 23.6 billion transistors

10x IPU-LINKS™
320GB/s chip-to-chip bandwidth

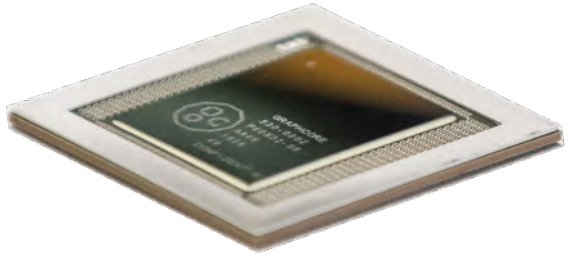
300MB IN-PROCESSOR-MEMORY™
45TB/s memory bandwidth per chip
the whole model held inside the processor

PCIe Gen4 x16
64GB/s bi-directional host communication bandwidth



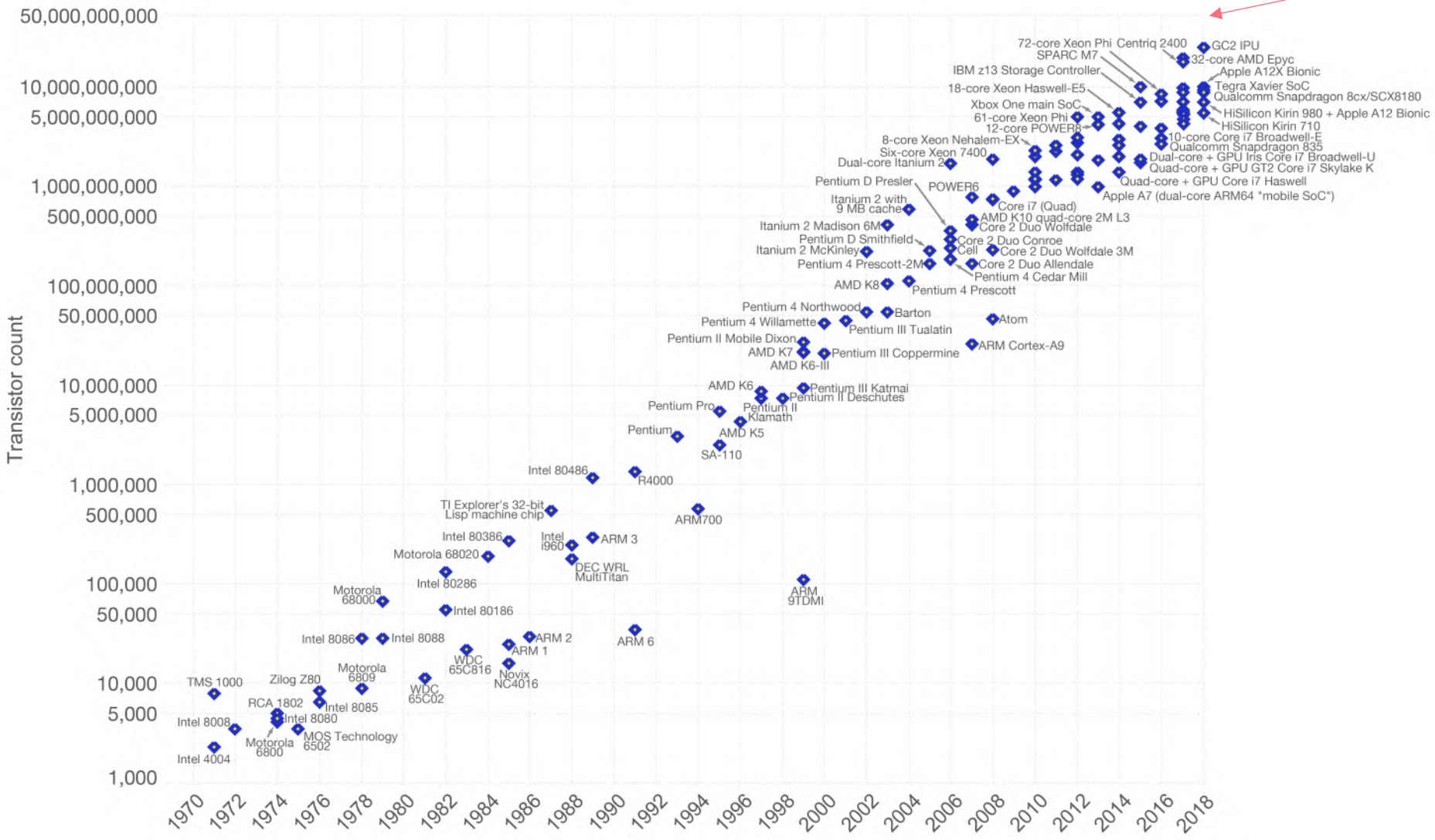
1,216 independent IPU-CORES™
each with IN-PROCESSOR-MEMORY™ tile
> 100GFLOPS per IPU-CORE™
> 7,000 programs executing in parallel

8TB/s all to all IPU-EXCHANGE™
non-blocking, any communication pattern



# Moore's Law – The number of transistors on integrated circuit chips (1971-2018)

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important as other aspects of technological progress – such as processing speed or the price of electronic products – are linked to Moore's law.



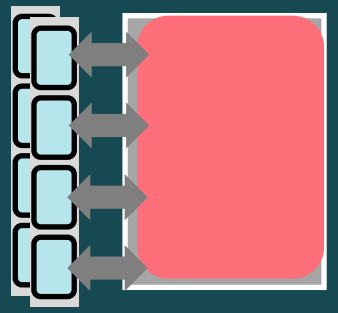

Data source: Wikipedia (https://en.wikipedia.org/wiki/Transistor_count)
The data visualization is available at OurWorldinData.org. There you find more visualizations and research on this topic.

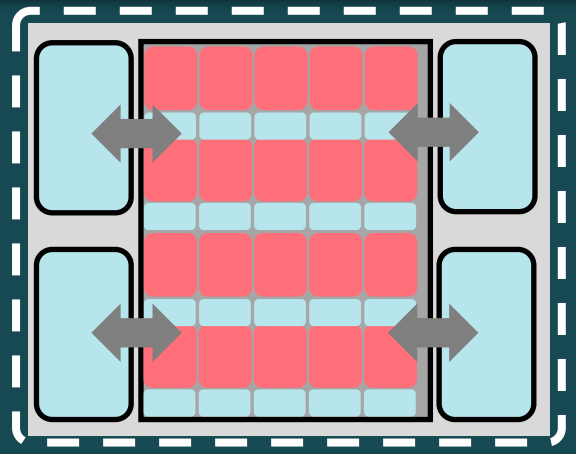
Licensed under CC-BY-SA by the author Max Roser.

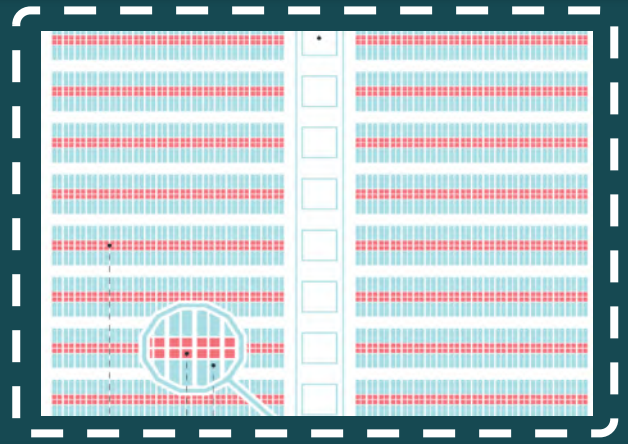
(1) Image from Wikipedia

# MASSIVE PARALLELISM WITH UNSHACKLED MEMORY ACCESS

| | CPU | GPU | IPU |
|---|---|---|---|
| **Parallelism** | Suitable for scalar processes | SIMD/SIMT architecture. Suitable for large blocks of dense contiguous data | **Massively parallel MIMD. High performance/efficiency as ML trends to sparsity & small kernels** |
| **Memory Access** | Off-chip memory<br>**1x** | Model and Data spread across off-chip and small on-chip cache and shared mem.<br>**5x – 32x** | **Model & Data in tightly coupled large locally distributed SRAM**<br>**320x** |







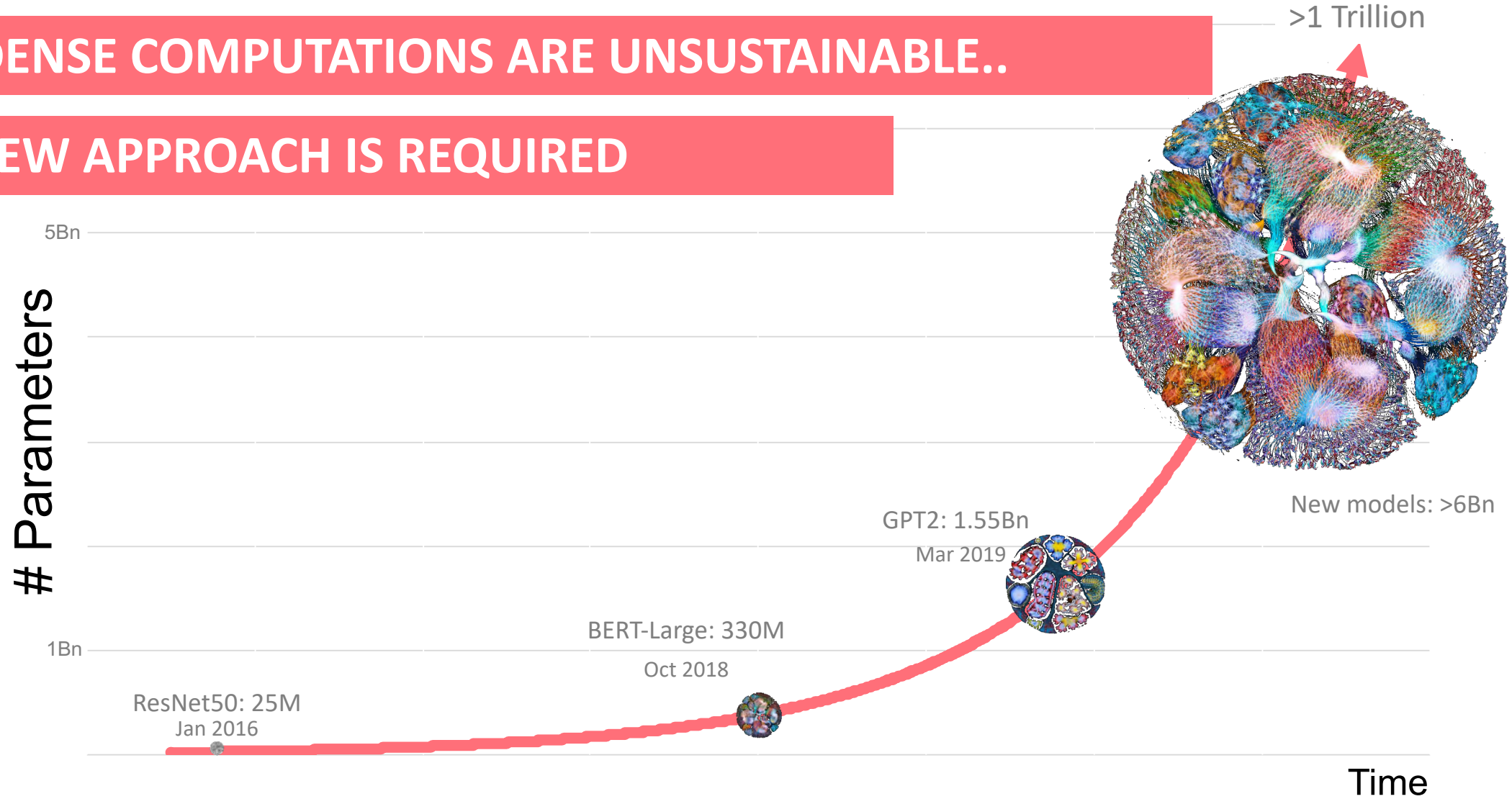
Processor Memory

# 350+ TEAM & GROWING FAST

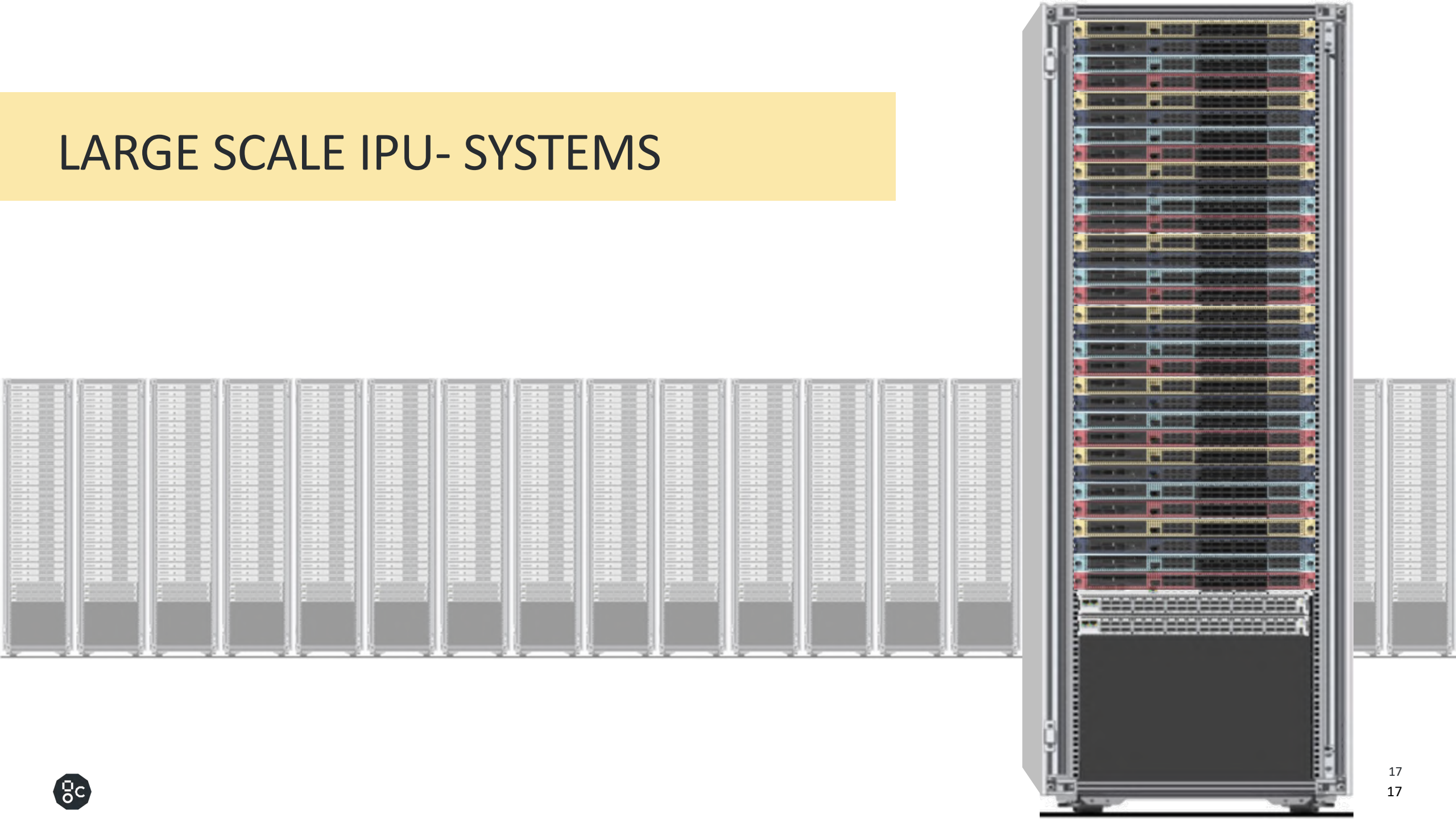
# MODEL SIZE GROWING MASSIVELY

**DENSE COMPUTATIONS ARE UNSUSTAINABLE..**

**A NEW APPROACH IS REQUIRED**



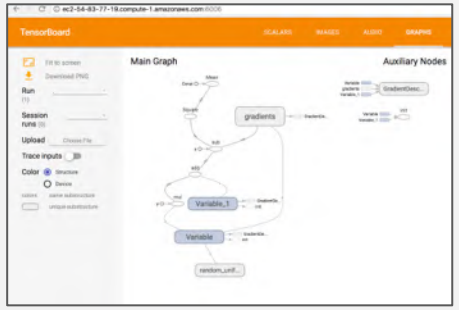
# LARGE SCALE IPU- SYSTEMS



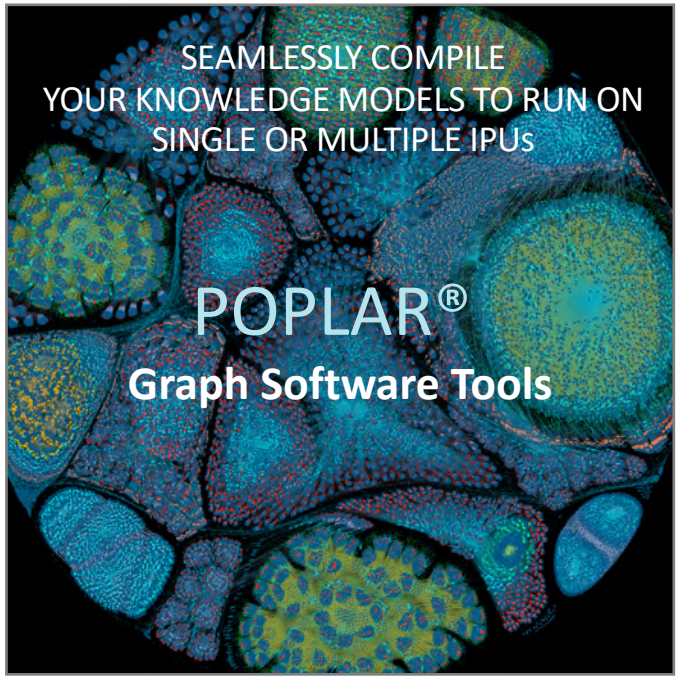

17

# SOFTWARE, SOFTWARE AND SOFTWARE!”

USE YOUR EXISTING
KNOWLEDGE MODELS

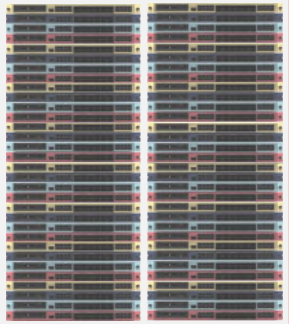



TensorFlow ONNX PYTORCH




**C2 IPU-Processor**
PCIe Card

**IPU - APPLIANCE**
Servers and
Pods/
Supercomputers



“The challenge in building any processor is software. Far too often people build interesting processors and then afterwards try and work out how they’re going to program it. The reality is you need to work out how you’re going to program it and then build a processor that will support that efficiently. “ Nigel Toon, Graphcore CEO

| | NATURAL LANGUAGE PROCESSING | IMAGE/VIDEO PROCESSING | TIME SERIES ANALYSIS | RECOMMENDERS/ RANKING | PROBABILISTIC MODELS |
|---|---|---|---|---|---|
| USE CASES | Question answering<br>Text generation<br>Sentiment analysis<br>Semantic role labeling | Action recognition,<br>Object detection,<br>Image classification | Anomaly detection,<br>Forecasting,<br>Finance use-cases | Collaborative filtering<br>Ad ranking<br>Search result recommendation | Risk management,<br>Algorithmic trading |
| EXAMPLE | BERT | ResNet<br>ResNext | LSTM<br>GRU<br>RNN<br>MLP + Feature Embedding | Deep Autoencoder | Markov Chain<br>Monte Carlo (MCMC) |

GitHub

github.com/graphcore/examples

# Thank you

We are hiring
AI Engineers
in **Oslo**

https://www.graphcore.ai/careers

GRAPHCORE

ellied@graphcore.ai