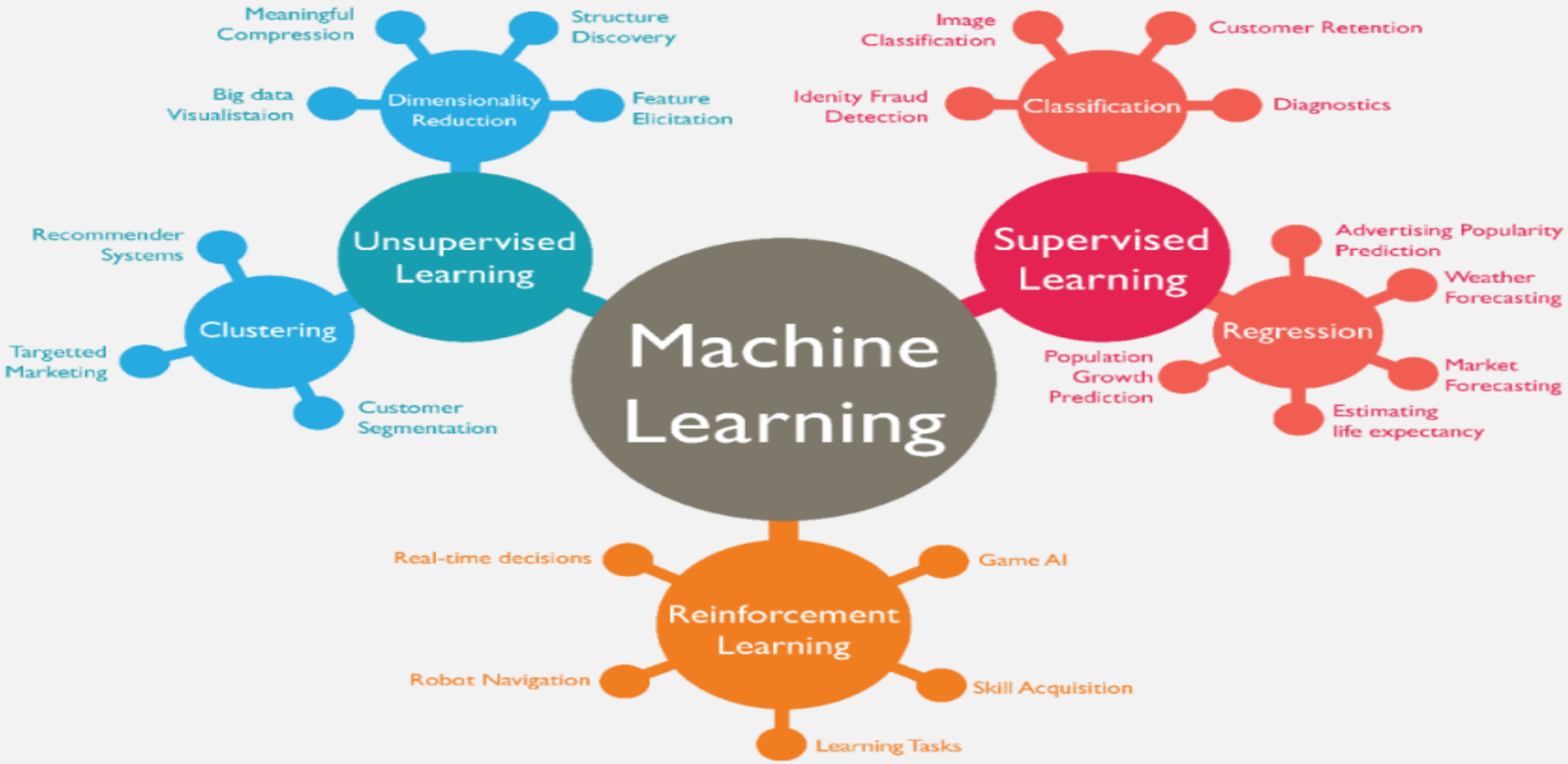
Introducing

Graph Core

Author:

AmirHossein Mohammadi

Machine Learning Bubble Chart



Learning....

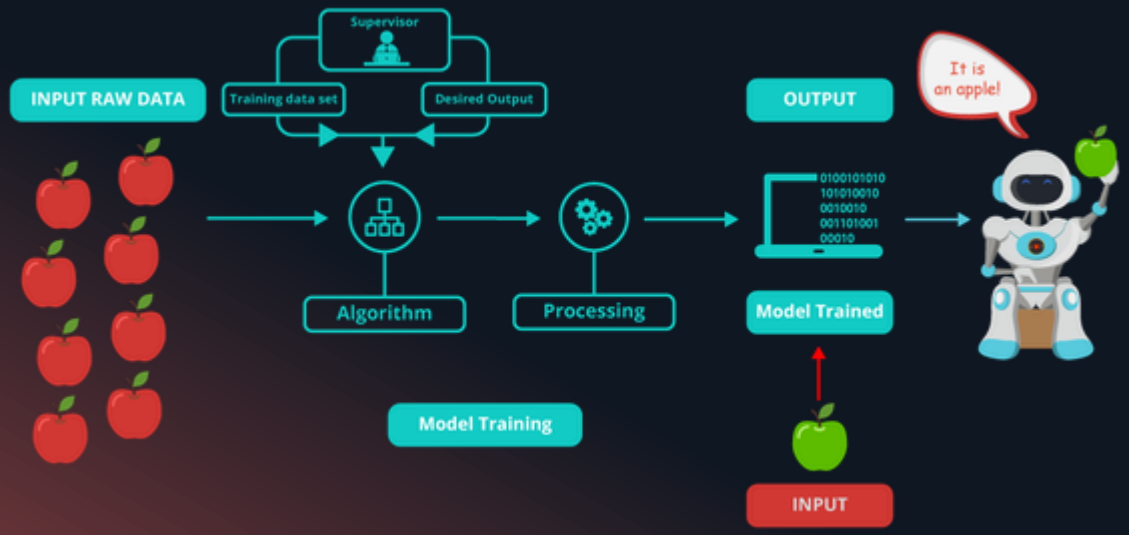
Tom Mitchell(1997)

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T ,as measured by P , improves with experience E.

- بهبود عملکرد از طریق تجربه

Supervised Learning

یادگیری تحت نظارت



مجموعه ای از نمونه های یادگیری وجود دارد که بازای هر ورودی،مقدار خروجی و یا تابع مربوطه نیز مشخص است.هدف سیستم یادگیرنده،بدست آوردن فرضیه ای است که تابع و یا رابطه بین ورودی و یا خروجی را حدس بزند

Origin	Manufacturer	Color	Decade	Type	Example Type
Japan	Honda	Blue	1980	Economy	Positive
Japan	Toyota	Green	1970	Sports	Negative
Japan	Toyota	Blue	1990	Economy	Positive
USA	Chrysler	Red	1980	Economy	Negative
Japan	Honda	White	1980	Economy	Positive
Japan	Toyota	Green	1980	Economy	Positive
Japan	Honda	Red	1990	Economy	Negative

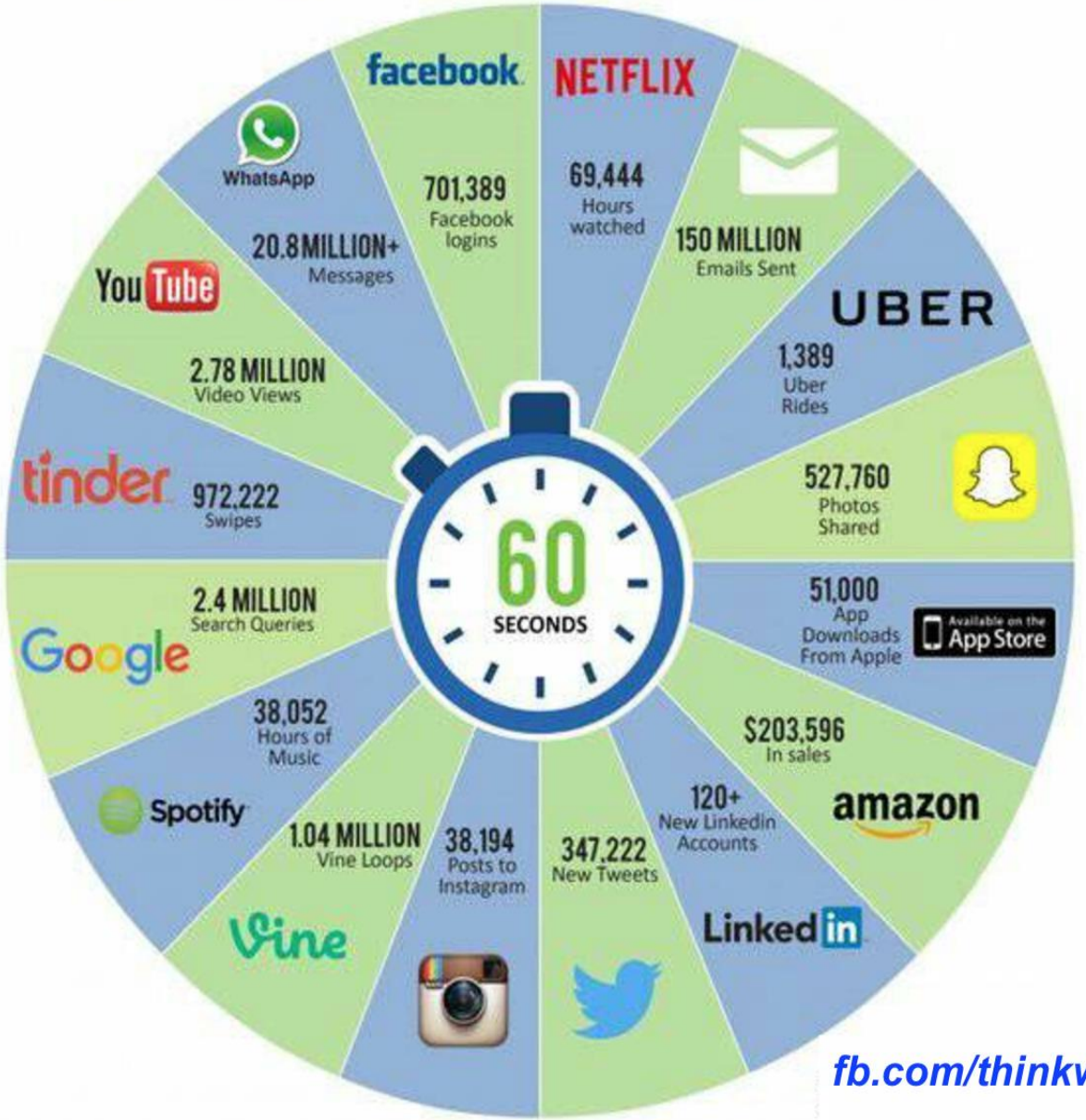
Iris setosa

Iris versicolor

Iris virginica

petal length, petal width, sepal length, sepal width

2016 What happens in an INTERNET MINUTE?

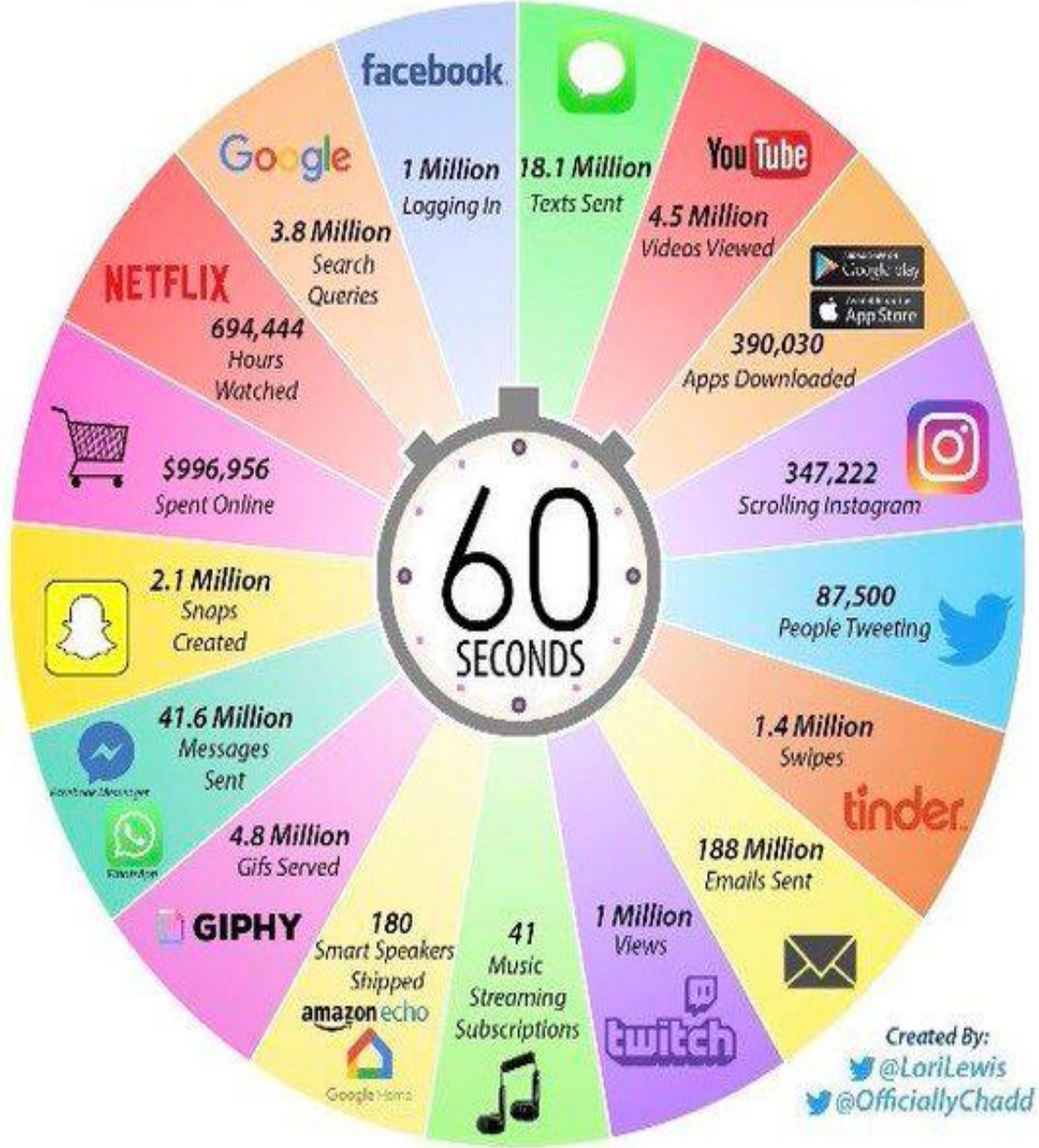


2017 This Is What Happens In An Internet Minute

fb.com/thinkwittynow

2018 This Is What Happens In An Internet Minute

2019 This Is What Happens In An Internet Minute



Large Data Means?

- 1000 **kilobytes** = 1 Megabyte
- 1000 **Megabytes** = 1 Gigabyte
- 1000 **Gigabytes** = 1 Terabyte
- 1000 **Terabytes** = 1 Petabyte
- 1000 **Petabytes** = 1 Exabyte
- 1000 **Exabytes** = 1 Zettabyte
- 1000 **Zettabytes** = 1 Yottabyte

What is Big Data?

edureka!

"Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications"

Volume

Processing increasing huge data sets

Variety

Processing different types of data

Velocity

Data is being generated at an alarming rate

Value

Finding correct meaning out of the data

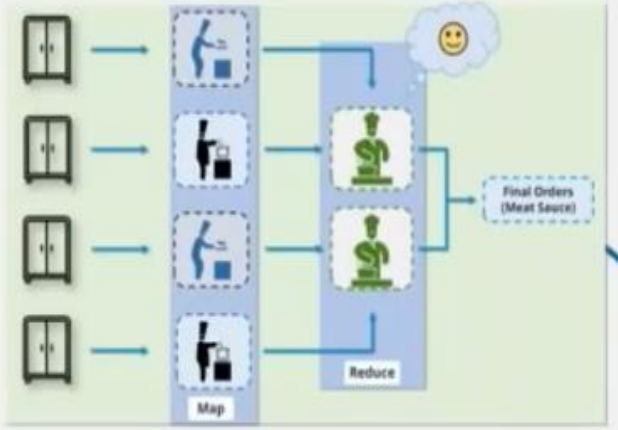
Veracity

Min	Max	Mean	SD
4.3	?	5.84	0.83
2.0	4.4	3.05	[illegible]
15000	7.9	1.20	0.43
0.1	2.5	?	0.76

Uncertainty and inconsistencies in the data

Need of a Framework

edureka!

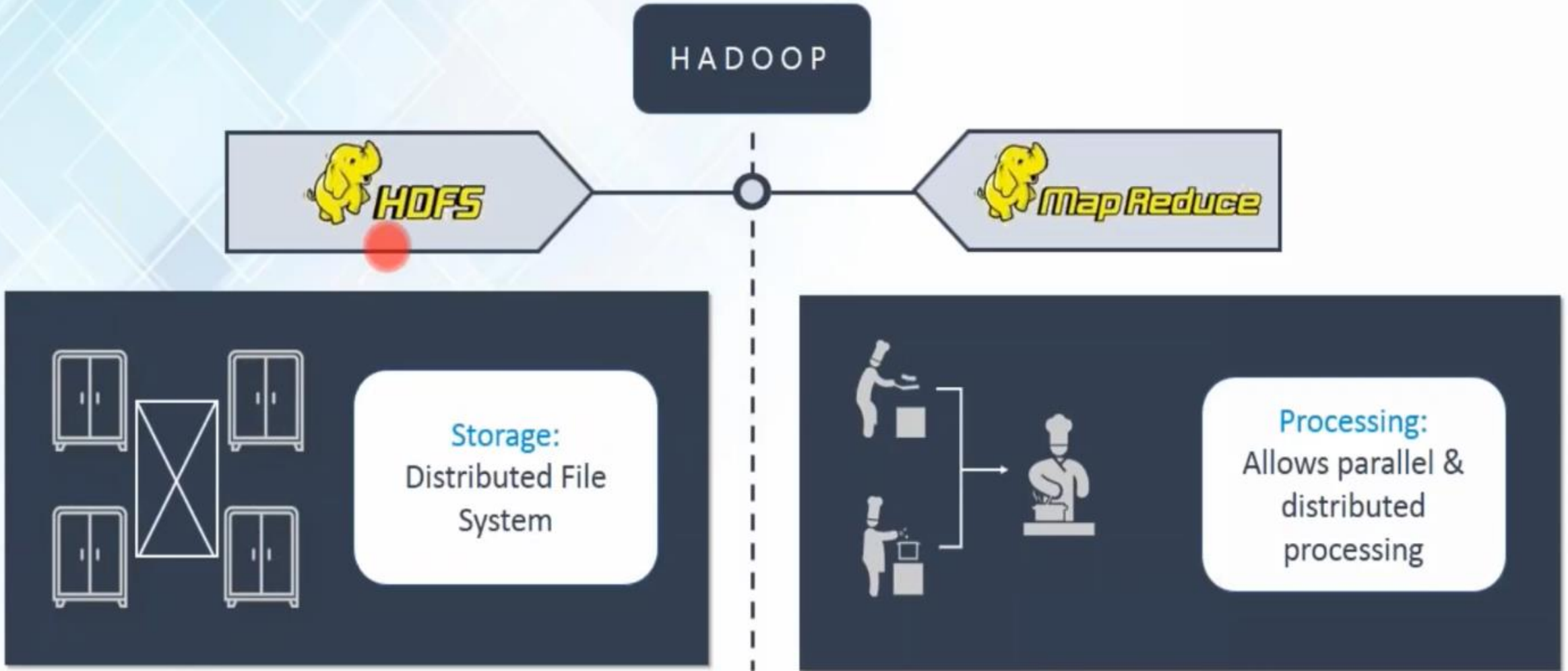


Do we have a framework that works like that ?

Apache Hadoop: Framework to Process Big Data

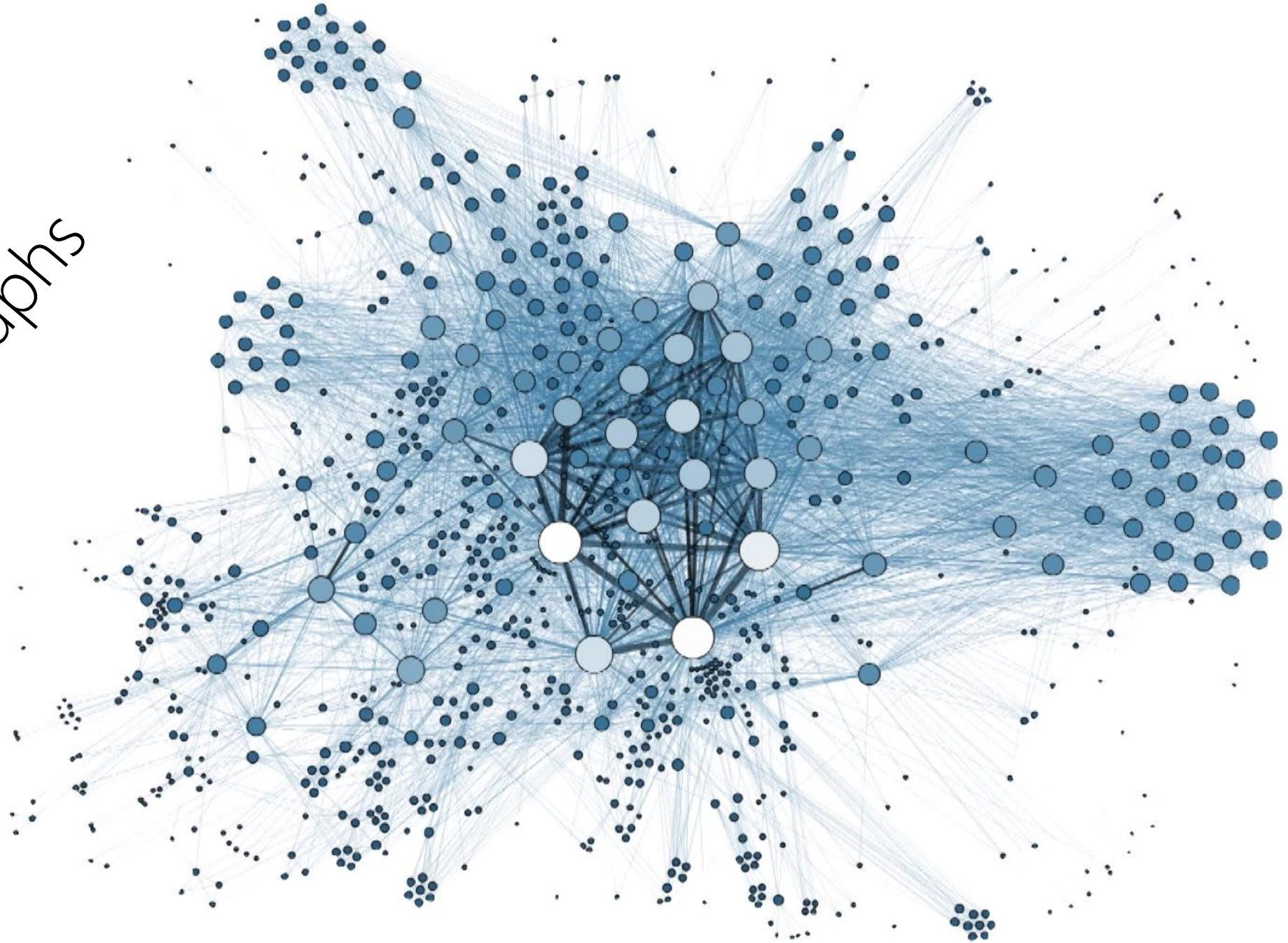
edureka!

Hadoop is a framework that allows us to store and process large data sets in parallel and distributed fashion



good idea but....

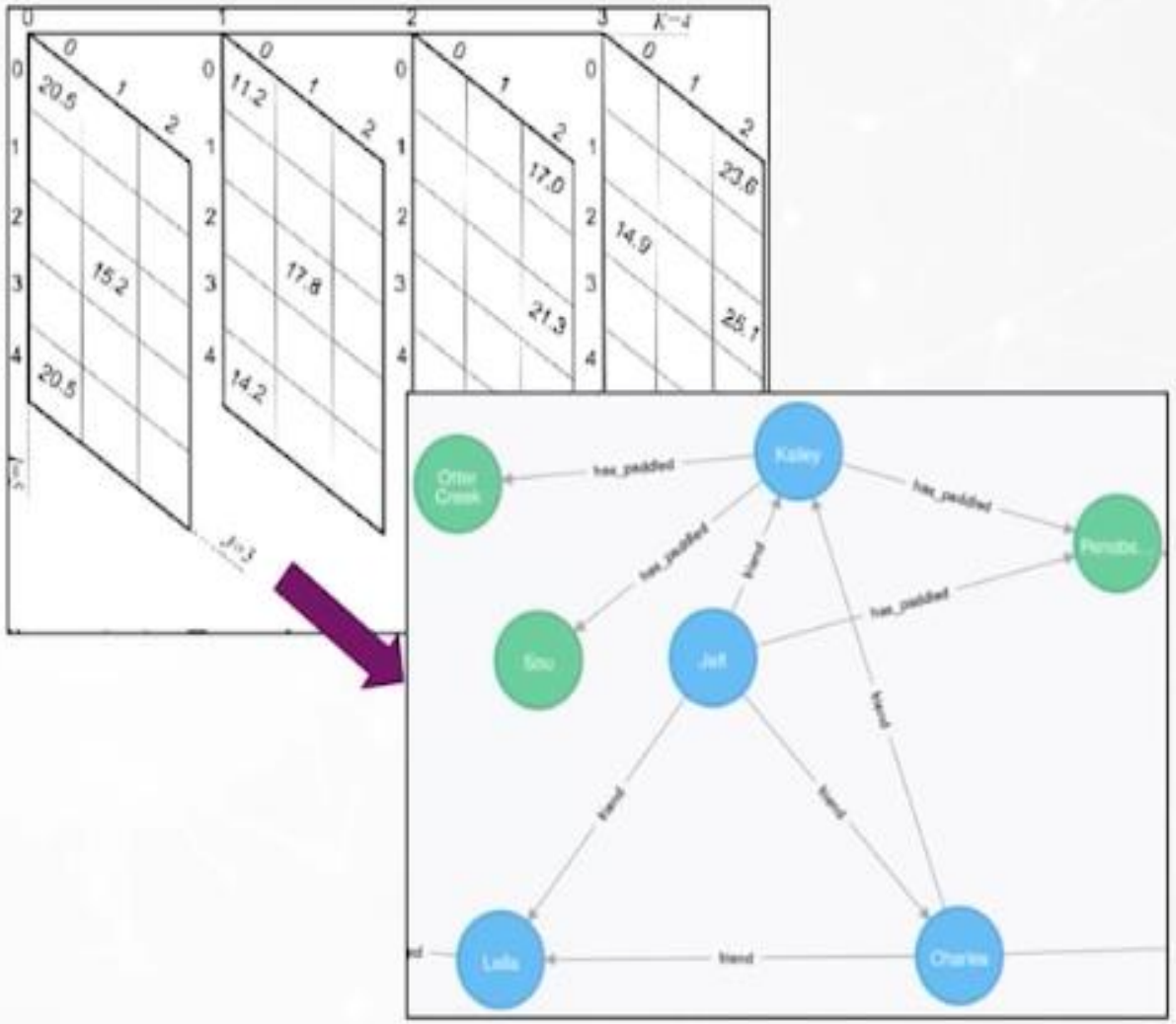
Graphs

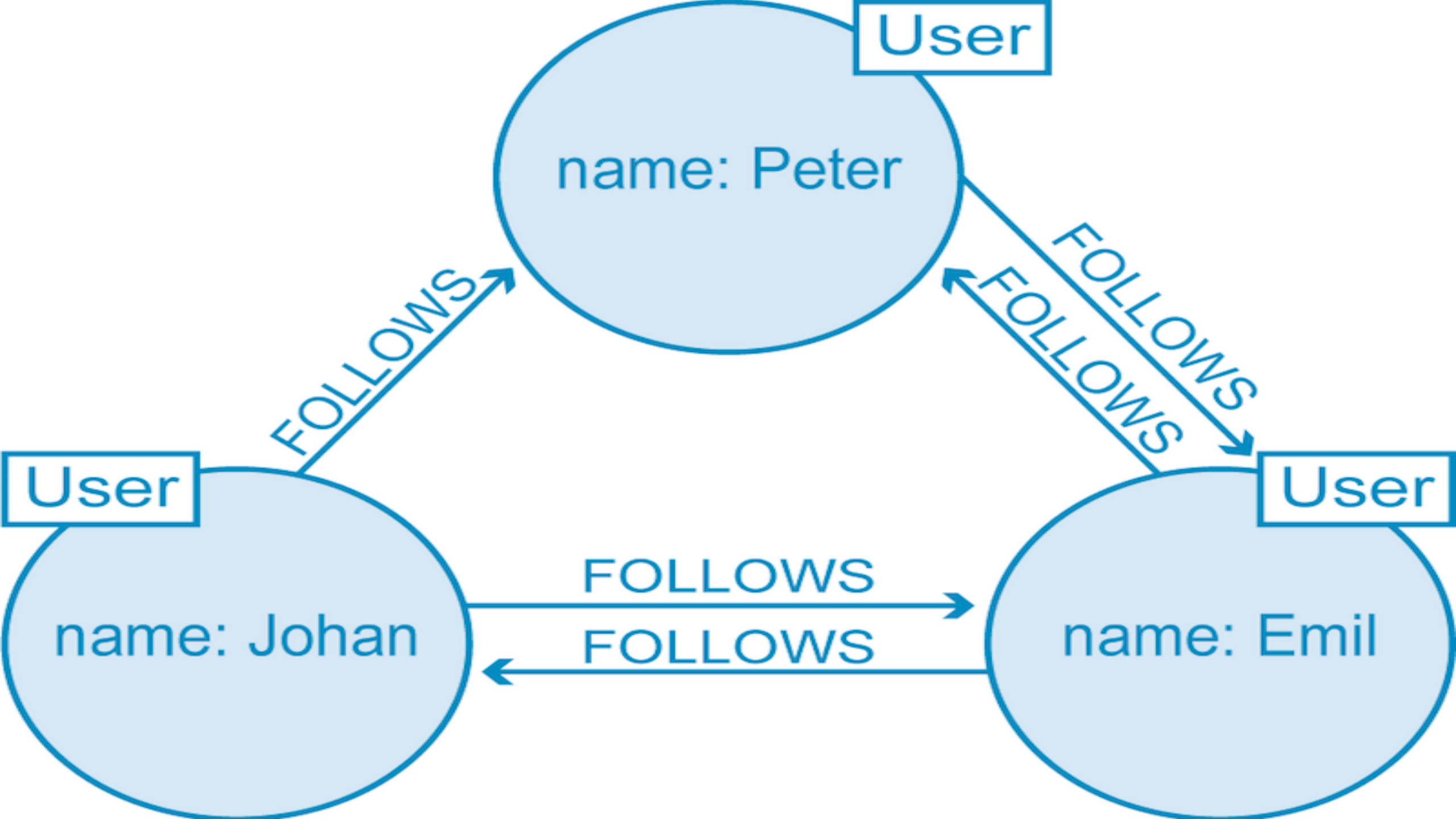


Model Optimization

Brute force is as inelegant as it sounds

- 56% of enterprise CIOs say iterative model training is the largest ML challenge[1]
- Renting more and more GPU time is not the answer – not every problem is "embarrassingly parallel"
- Table joins bog down data pipelines
- Sparse matrix compression methods are inefficient (more tables!)




User
name: Peter
FOLLOWS
FOLLOWS
FOLLOWS
FOLLOWS
User
name: Johan
User
name: Emil
FOLLOWS
FOLLOWS


THATS
ENOUGH!


Access
Permission
ID Number
B0213648
52
28
One Time Password
0218
access permission


48
mph
Autonomous
Mode
NETWORK
DRIVER

https://www.google.com/search?client=firefox-b-d&q=iran

Google

iran

All Images Maps News Videos More Settings Tools

About 1,180,000,000 results (0.57 seconds)

Top stories

L'Iran a bien arrêté l'anthropologue franco-iranienne Fariba Adelkhah

Le Monde

12 hours ago

Iran rejects suggestion its missile programme is negotiable

BBC

1 hour ago

At summit with Russia, Israel and US demanded Iran leave Lebanon, Iraq …

The Times of Israel

2 hours ago

More for iran

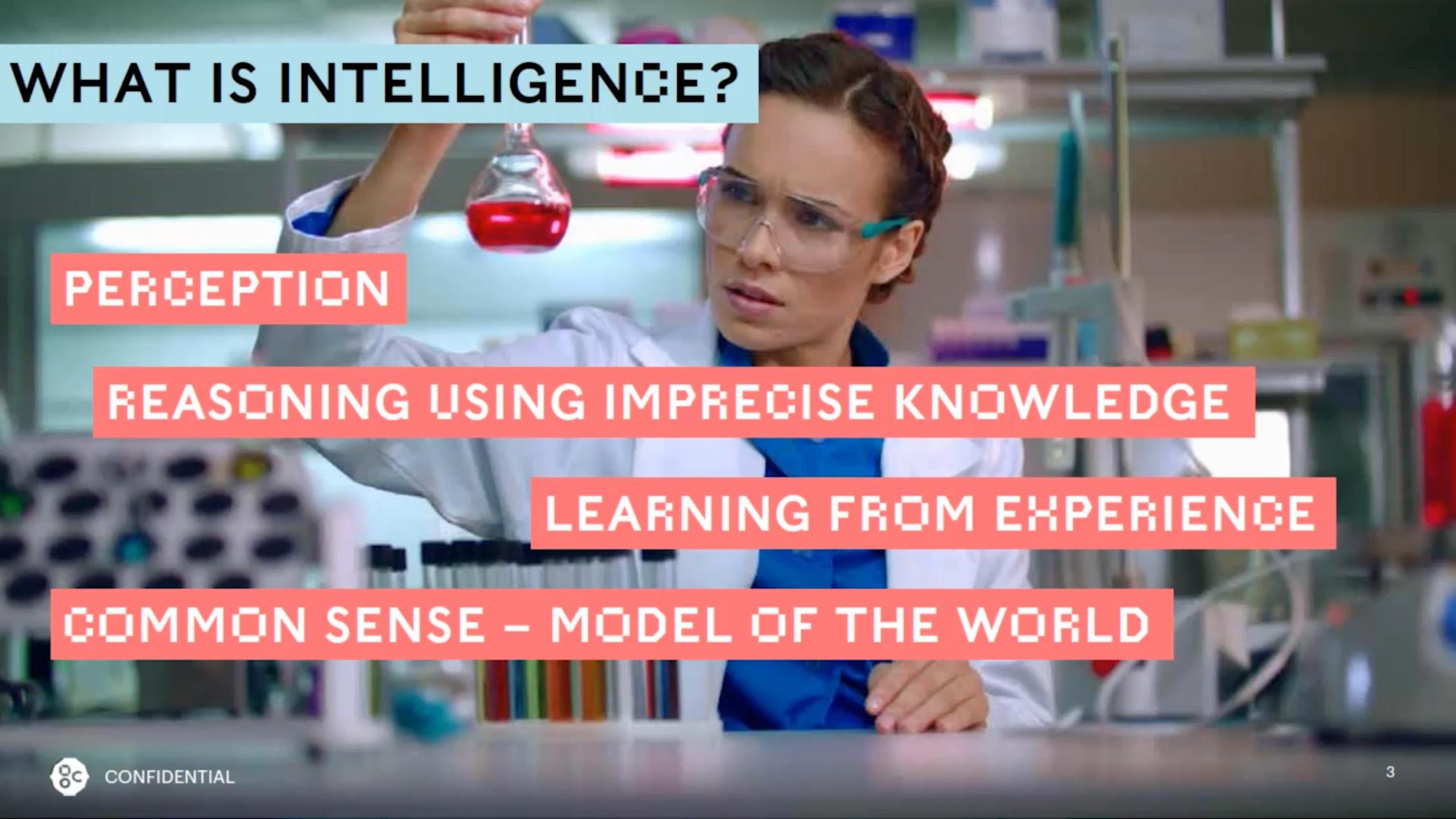
Iran

Country in the Middle East

Iran, also called Persia, and officially the Islamic Republic of Iran, is a country in Western Asia. With 82 million inhabitants, Iran is the world's 18th most populous country. Its territory spans 1,648,195 km², making it the second largest country in the Middle East and the 17th largest in the world. Wikipedia

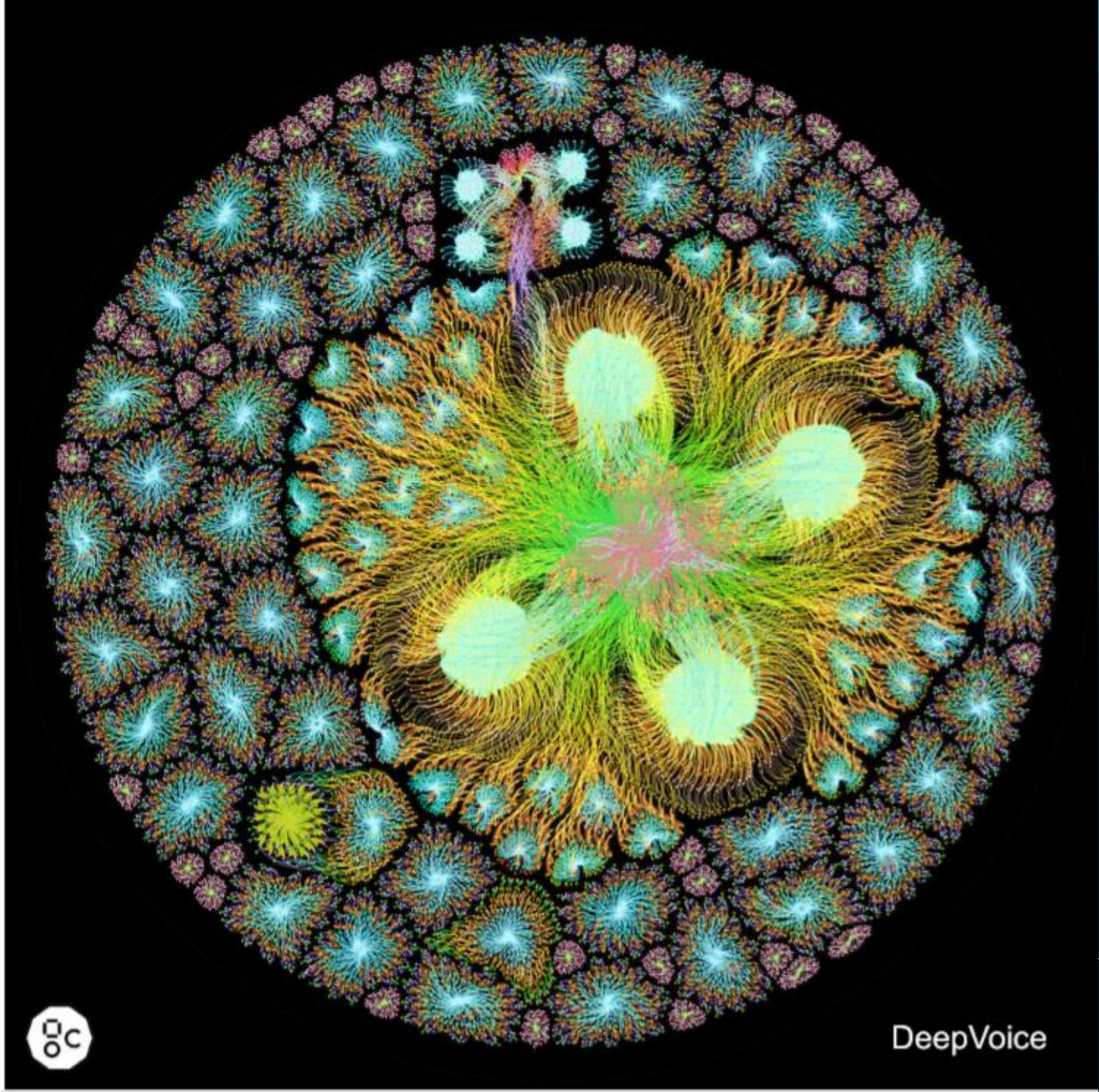

NEW
IDEA

WE HAVE DEVELOPED A NEW KIND OF HARDWARE THAT WILL LET INNOVATORS CREATE THE NEXT GENERATION OF MACHINE INTELLIGENCE

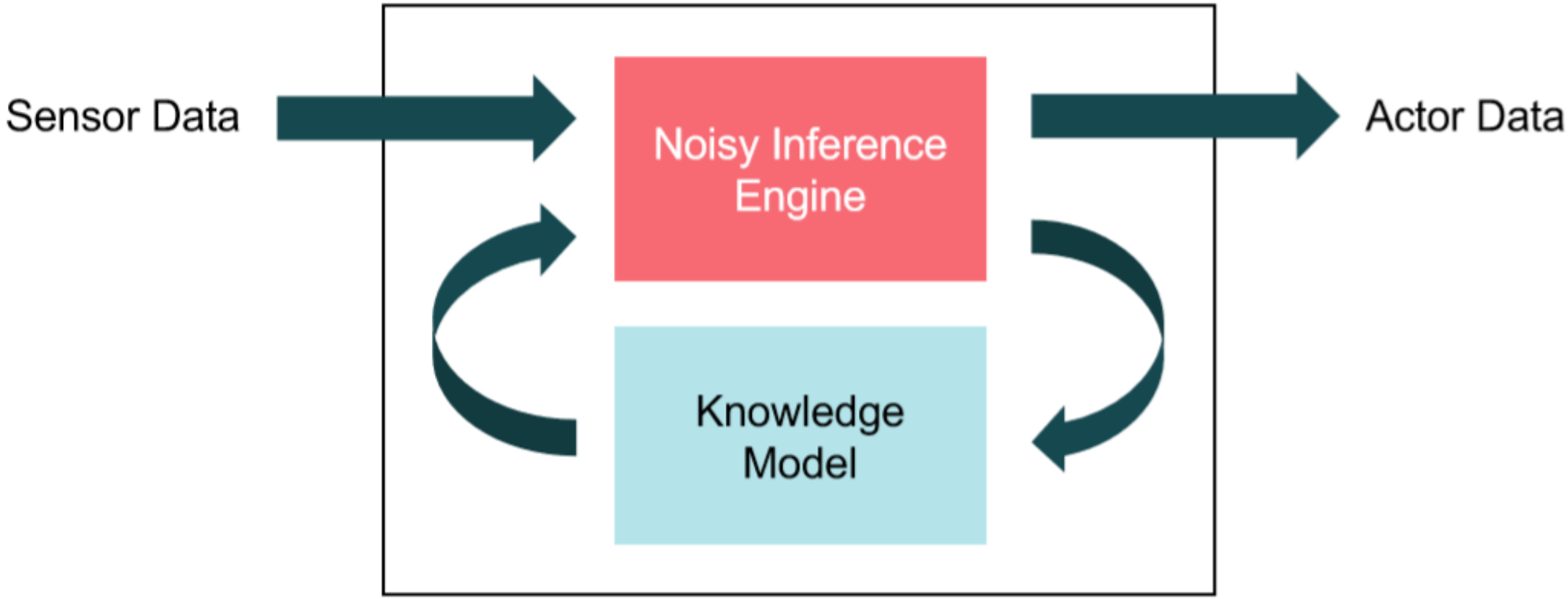

WHAT IS INTELLIGENCE?
PERCEPTION
REASONING USING IMPRECISE KNOWLEDGE
LEARNING FROM EXPERIENCE
COMMON SENSE – MODEL OF THE WORLD

GRAPHCORE

Intelligence Processing Unit



INTELLIGENCE MACHINE

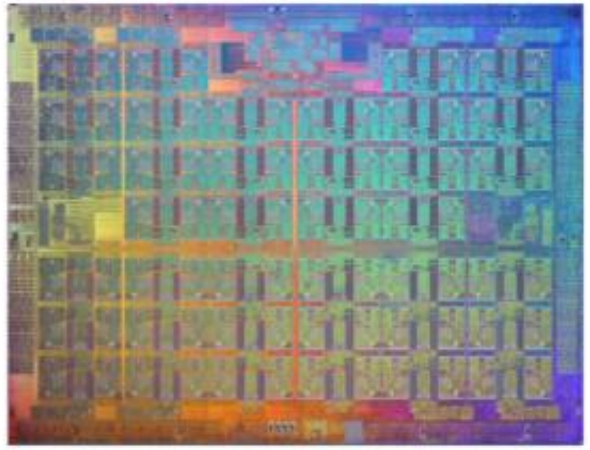


A canonical intelligent agent is a sequence-to-sequence translator.

Learning is inference (of model structure and parameters).

Inference is stochastic optimization, of some cost function.

INTELLIGENCE REQUIRES A NEW ARCHITECTURE

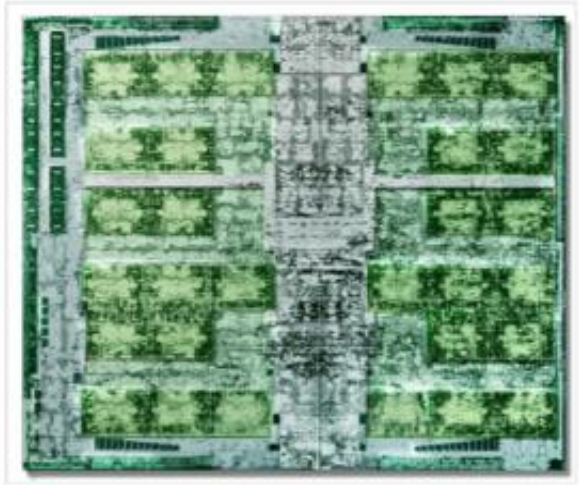


CPU

Scalar

Designed for office apps

Evolved for web servers

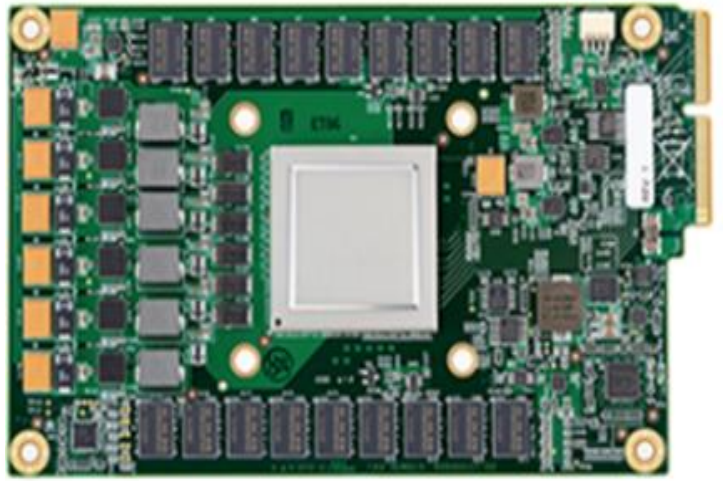


GPU

Vector

Designed for graphics

Evolved for HPC



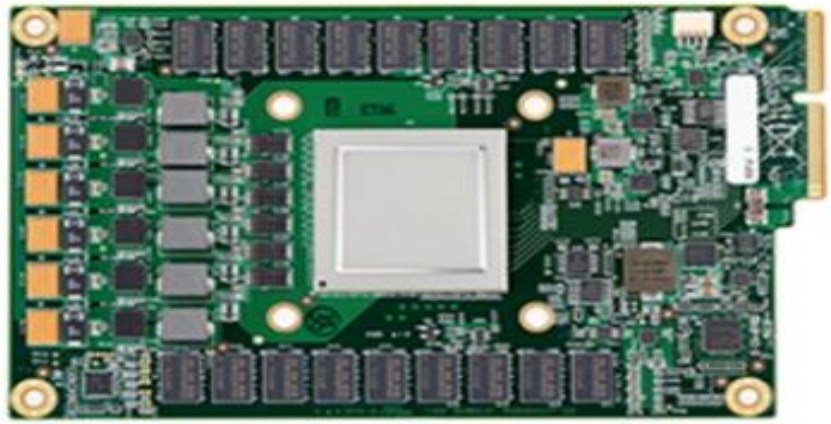
TPU

Tensor

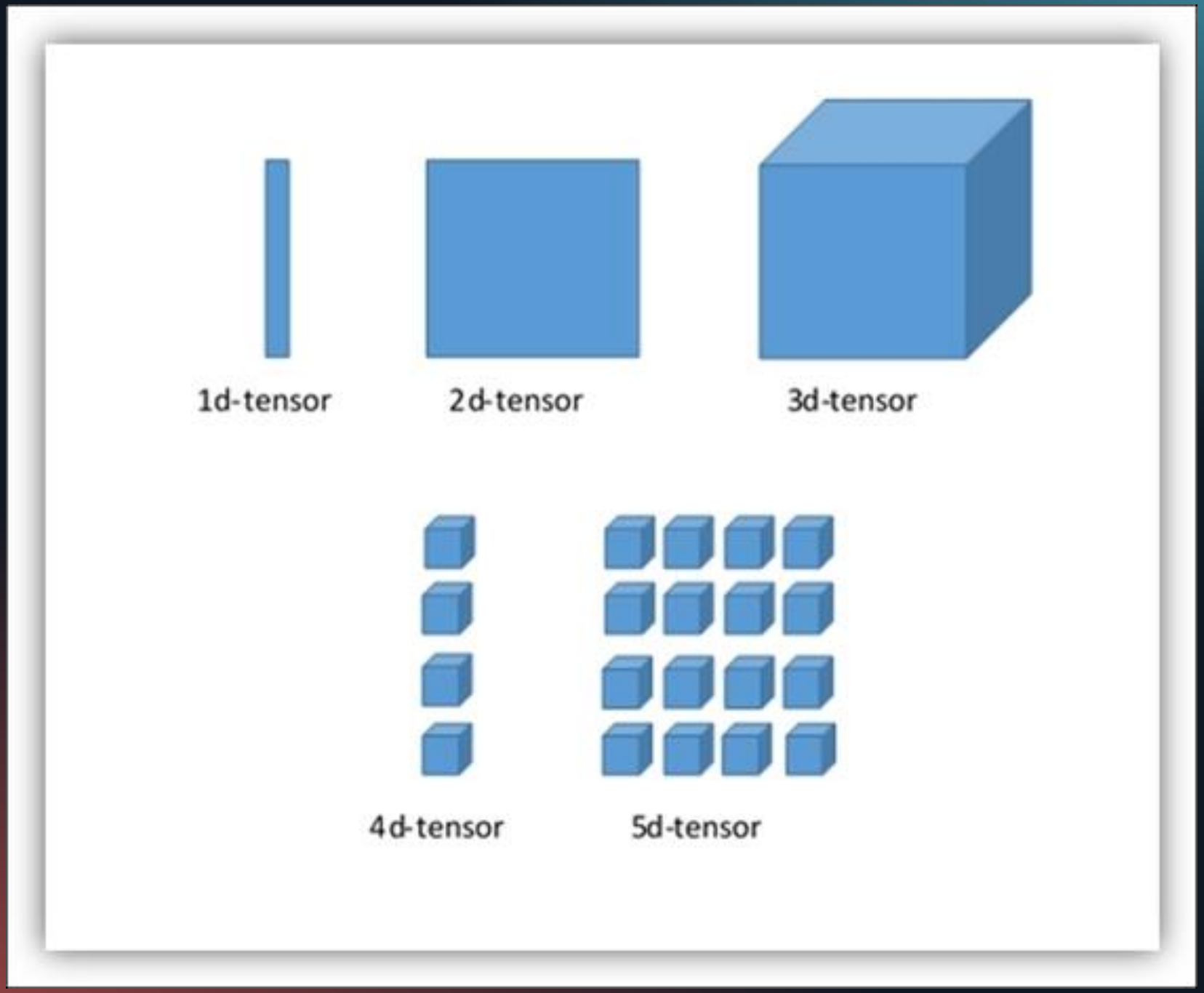
Designed for intelligence

Screenshot

TPU: Tensor Processing Unit



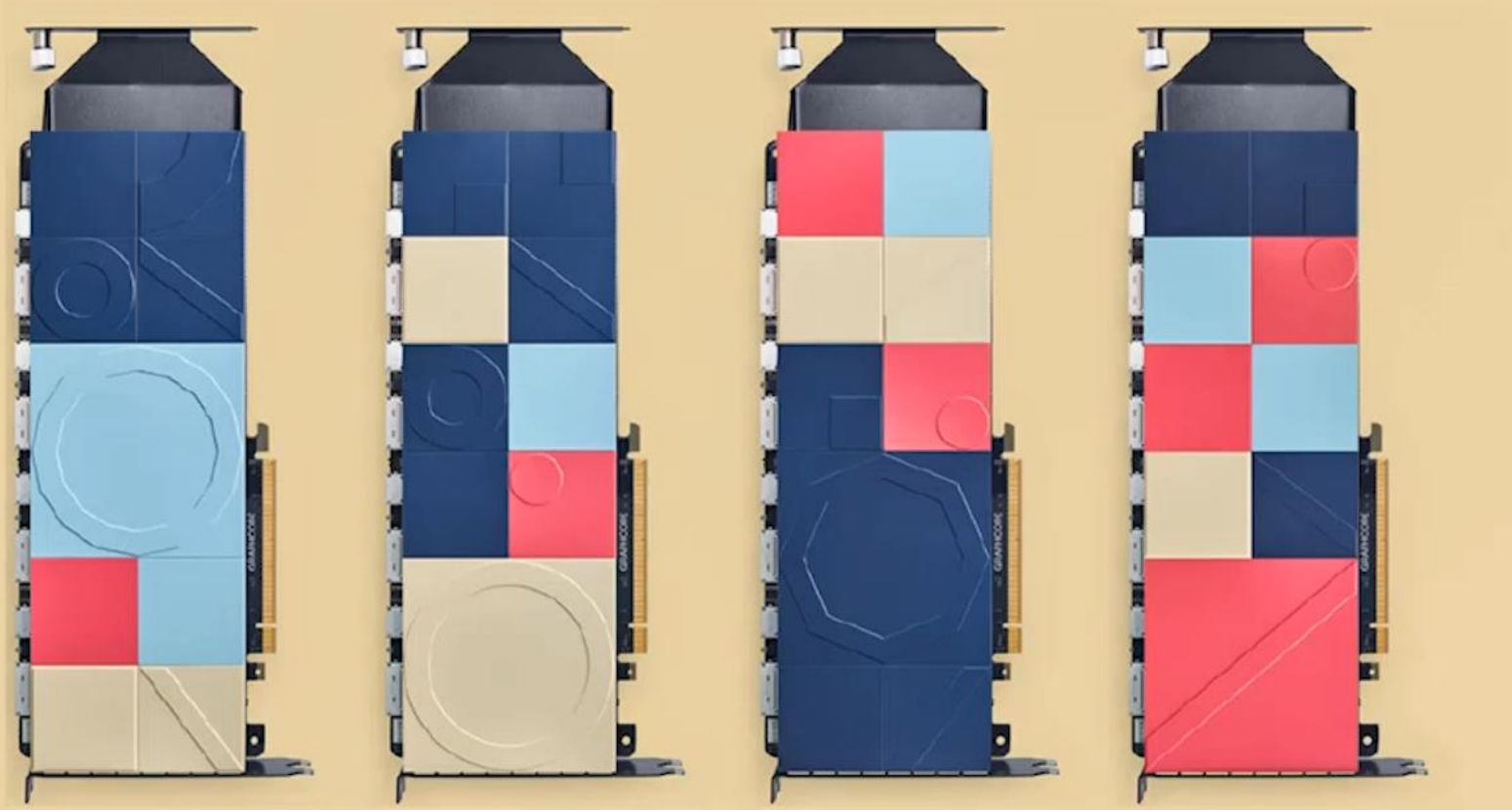
- quantization: 8-bit integers
- CISC ('Complex instruction set computing') design: implements high-level instructions that run more complex tasks
- matrix processor: processes hundreds of thousands of matrix operations in a single clock cycle
- May 2017: second generation -- 'Cloud TPU'


1d-tensor
2d-tensor
3d-tensor
4d-tensor
5d-tensor

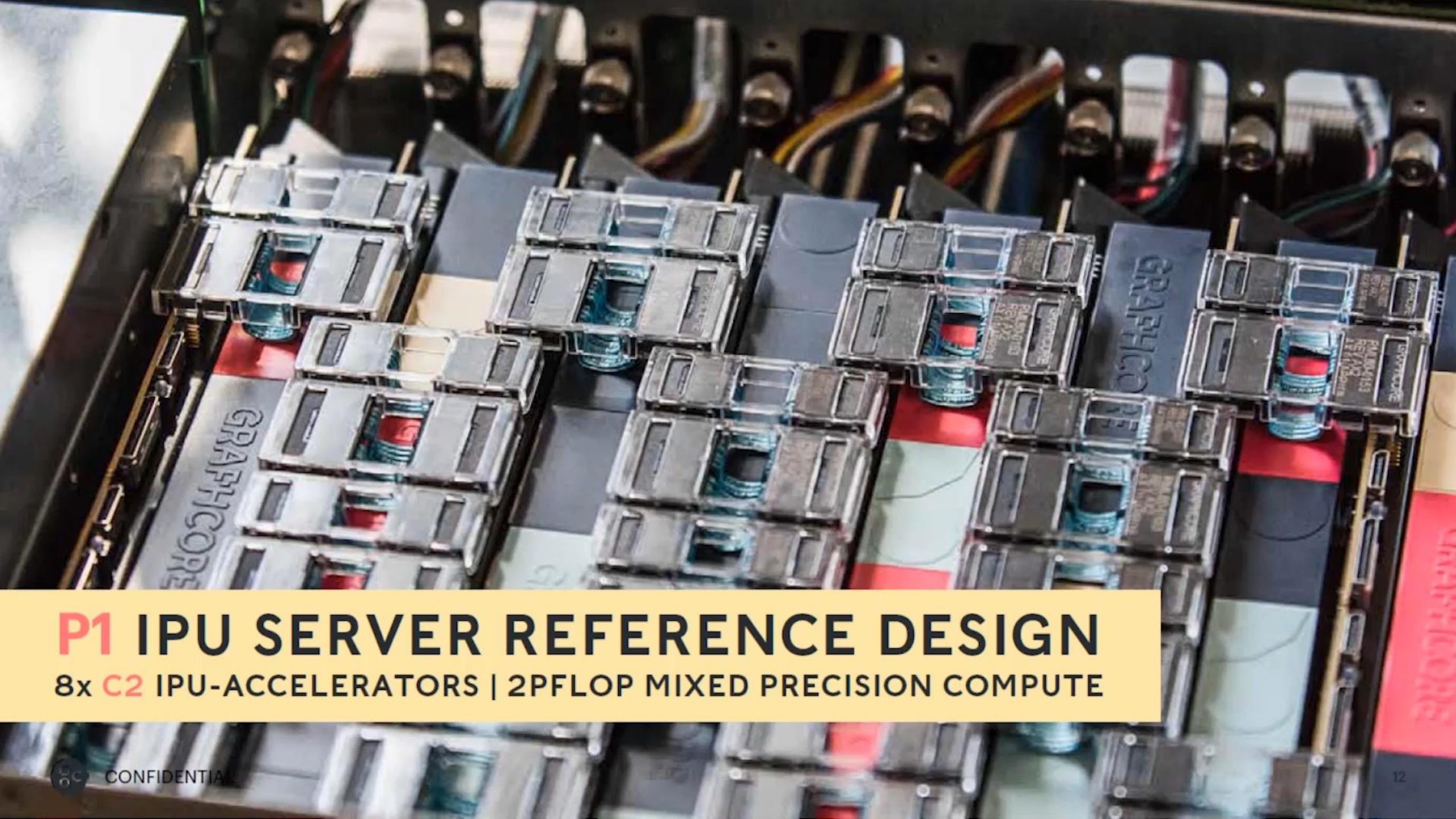

GRAPHCORE
530-0002
P60Y44-05
AA54
1807
33
SME
ES
CONFIDENTIAL
GRAPHCORE
530-0002
P60Y44-05
AA54
1807
33
SME
ES
CONFIDENTIAL

C2 IPU ACCELERATOR

100 terabyte per sec



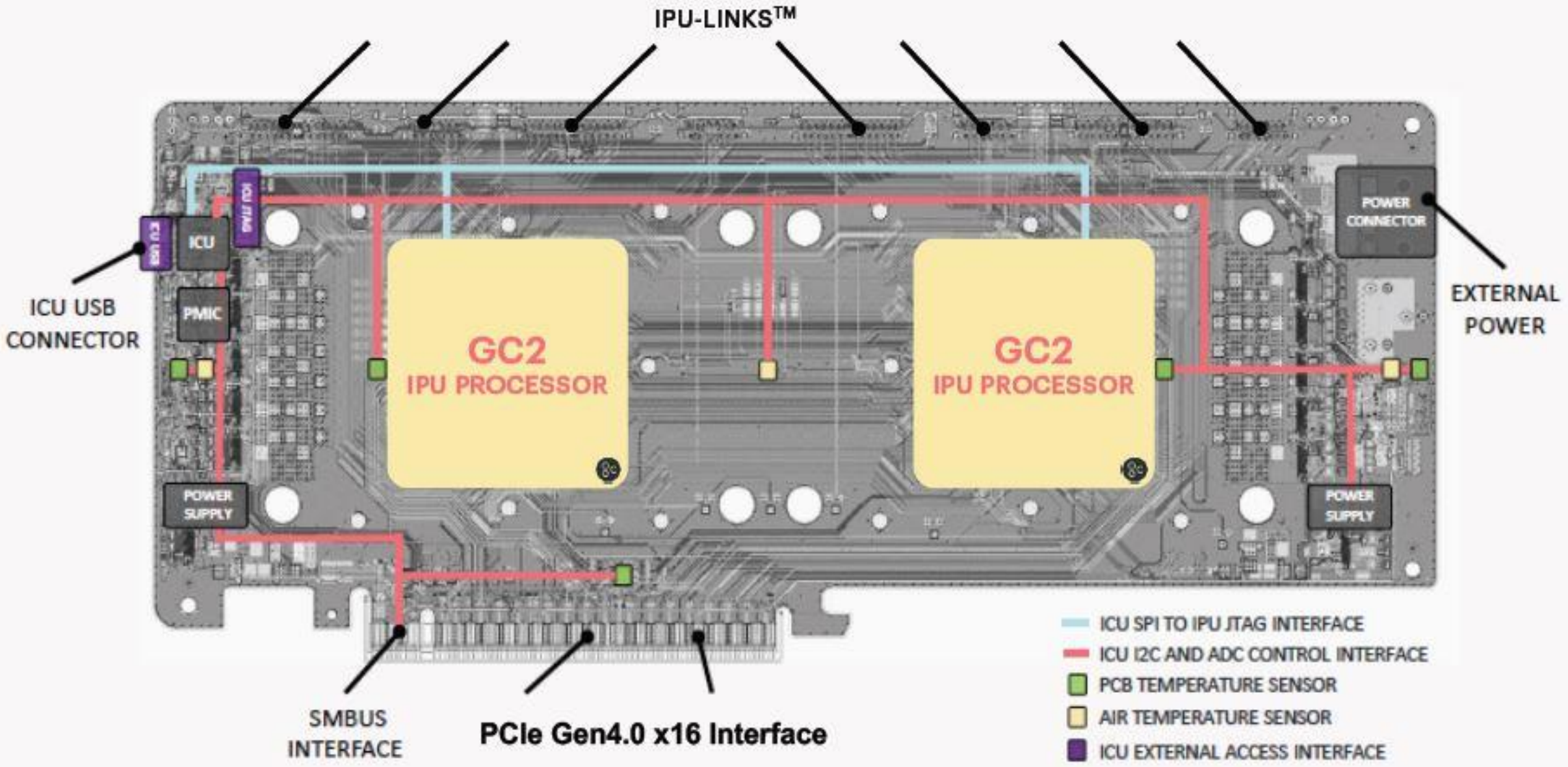
DOUBLE WIDTH PCIE CARD WITH 2 – COLOSSUS GC2 IPU PROCESSORS
CARD-TO-CARD **IPU-LINKS™** *(2.5TBps)*
250 TERA-FLOP MIXED PRECISION IPU COMPUTE @ 300W



P1 IPU SERVER REFERENCE DESIGN

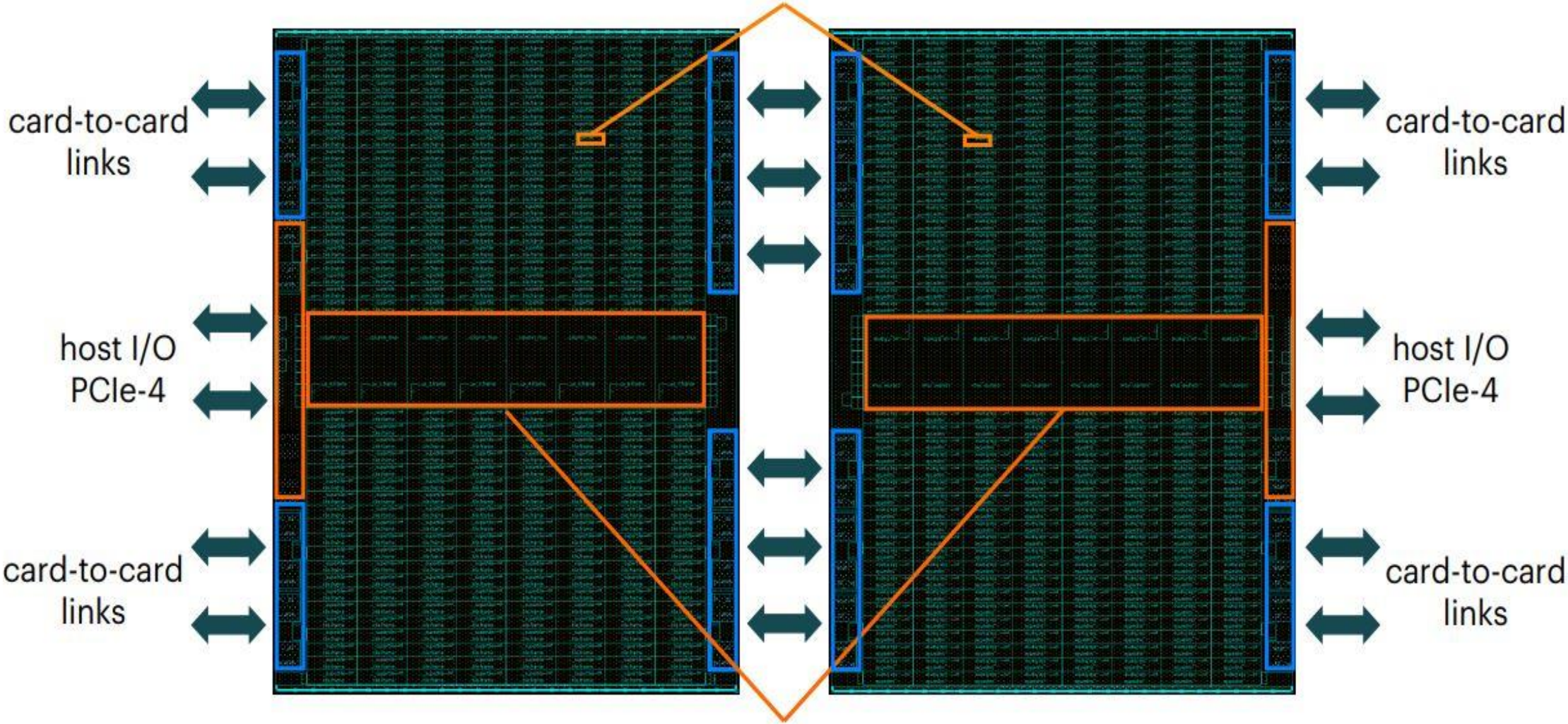
8x C2 IPU-ACCELERATORS | 2PFLOP MIXED PRECISION COMPUTE

C2 IPU-PROCESSOR PCIe CARD



GRAPHCORE IPU

2000> processor tiles >200Tflop ~600MB



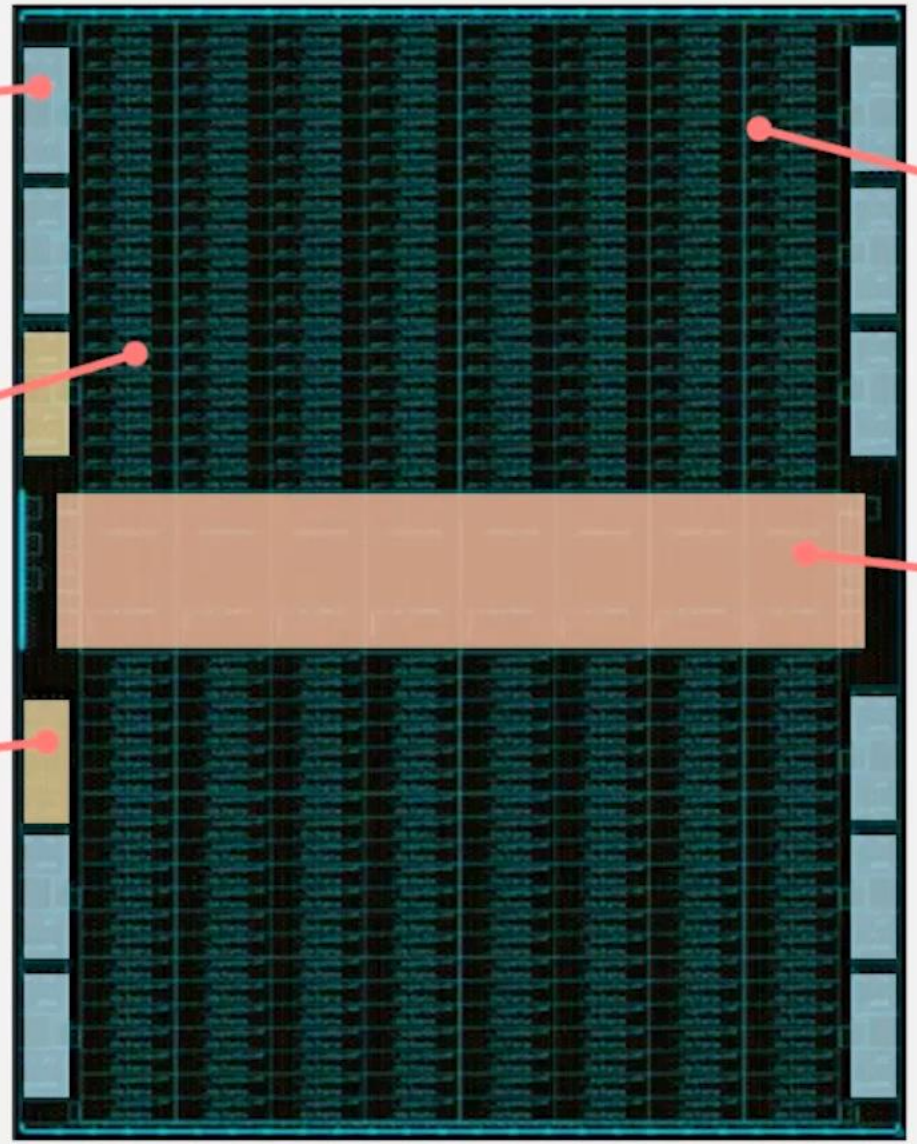
COLOSSUS GC2

the worlds most complex processor chip with 23.6Bn transistors

80 - IPU-LINKS™
2.5Tbps chip-to-chip bandwidth

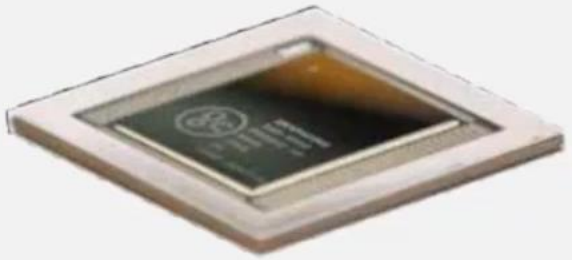
300MB IN-PROCESSOR-MEMORY™
>30TB/s memory bandwidth per chip
the whole model held inside the processor

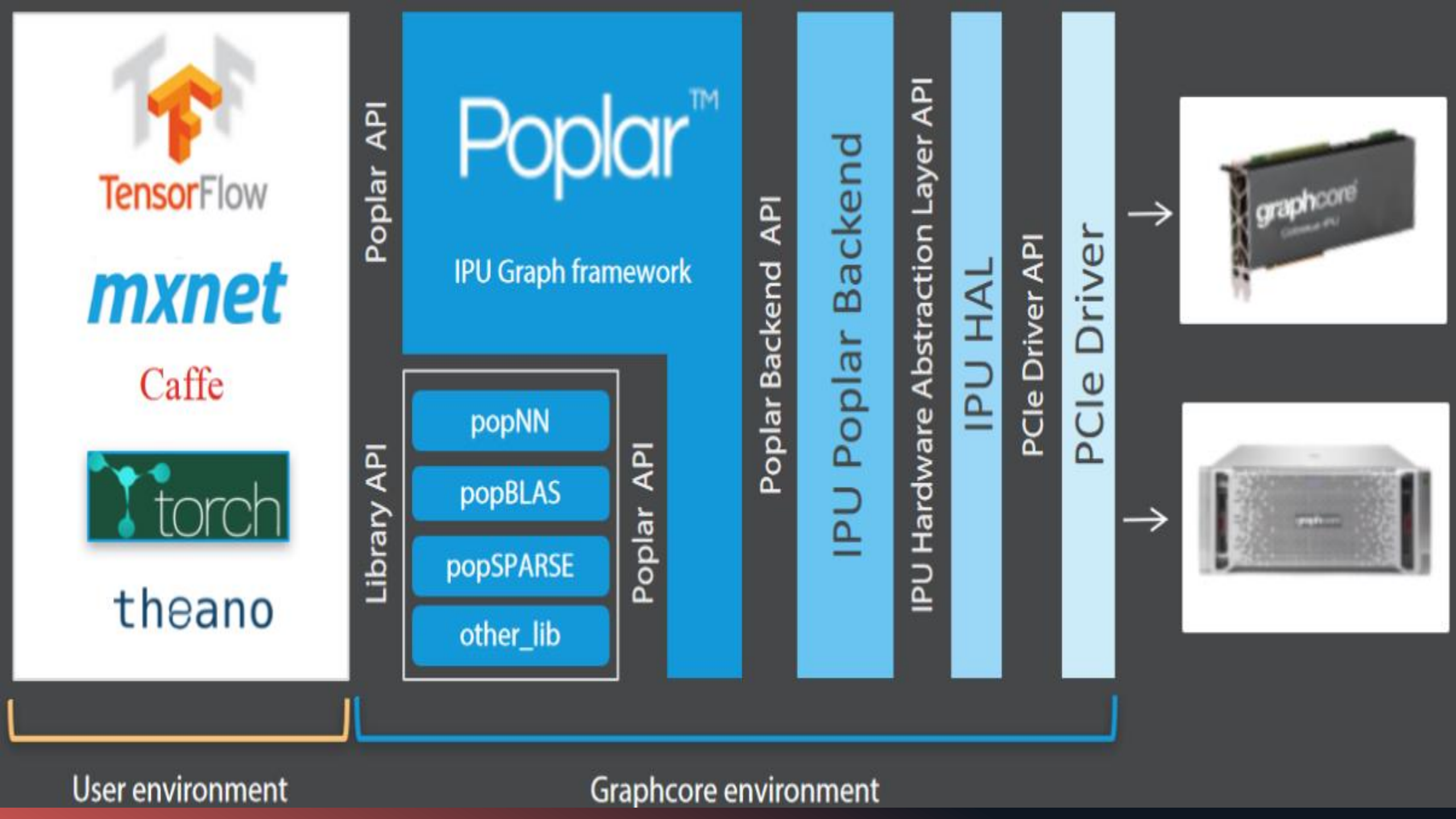
PCIe Gen4.0 x16
31.5GB/s bandwidth to host



1216 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



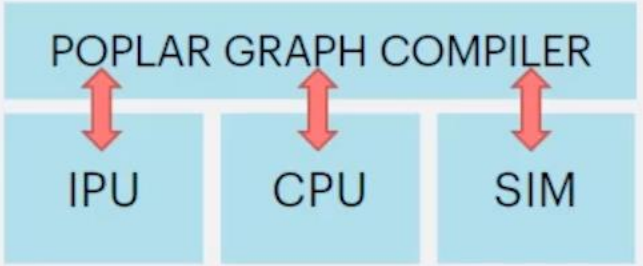

TensorFlow
mxnet
Caffe
torch
theano
Poplar API
Poplar™
IPU Graph framework
Library API
popNN
popBLAS
popSPARSE
other_lib
Poplar API
Poplar Backend API
IPU Poplar Backend
IPU Hardware Abstraction Layer API
IPU HAL
PCIe Driver API
PCIe Driver
graphcore
User environment
Graphcore environment

POPLAR®

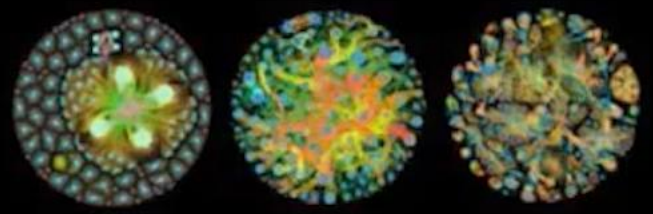
Software stack

SEAMLESS INTERFACE TO INDUSTRY STANDARD ML FRAMEWORKS

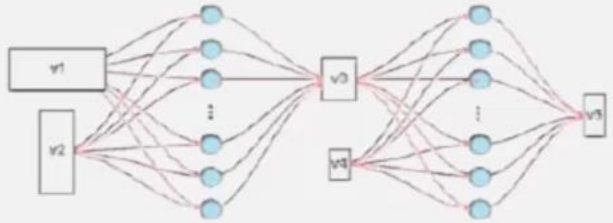
OPTIMIZED GRAPH MAPPING AND CODE COMPILER BUILT USING LLVM



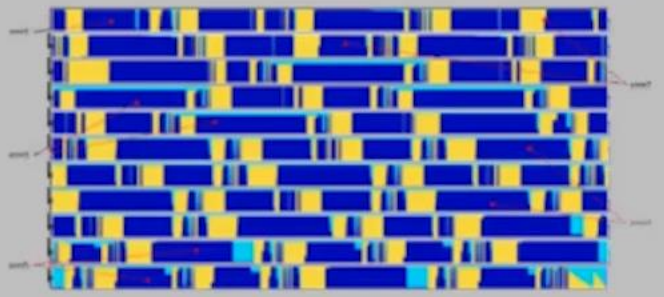
DRIVERS, UTILITIES AND GRAPH ENGINE FOR EXECUTION



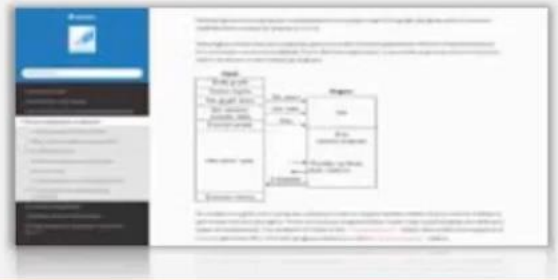
POPLAR® GRAPH PROGRAMMING FRAMEWORK (C++ & PYTHON) AND OPEN SOURCE ML LIBRARIES



ADVANCED VISUALIZATION AND DEBUG TOOLS



COMPREHENSIVE USER DOCUMENTATION, EXAMPLES, APPLICATION NOTES AND TUTORIALS



Don't invent a new processor architecture with less than 20 year utility

- In 2005 SVMs were hot
- In 2010 RFs were hot
- In 2015 NNs were hot (again)
- In 2020 ...?

At least until we understand intelligence better, we need a machine which exploits parallelism, has a simple programming abstraction, and is agnostic to model structure.

Intelligence Machine Desiderata

Computation on graphs

- massive parallelism.
- sparse data access: gather/scatter.

Low precision arithmetic

- mixed-precision floats.

Static graph structure

- compiler can partition work, allocate memory, and schedule messages.

Entropy generative

- noise in hardware.

What might limit machine Performance?

Compute

- Rate of arithmetic

Memory data bottleneck

- Bandwidth and latency for parameters and activations

Memory address bottleneck

- Rate of scatter/gather addressing for sparse data

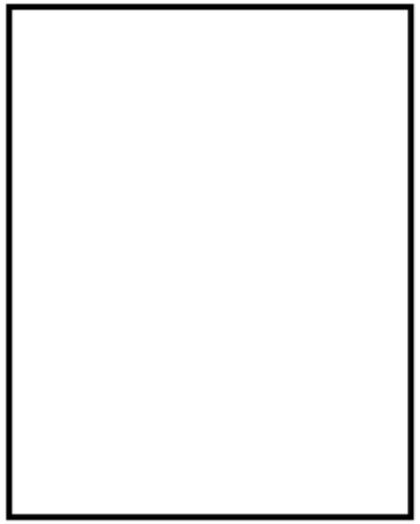
Entropy

- Rate of generation of random numbers

Power

- ...

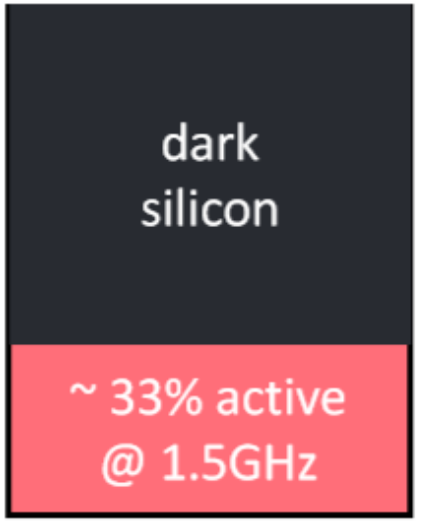
ALL LOGIC CHIPS ARE POWER LIMITED



Largest manufacturable
die ~825mm²

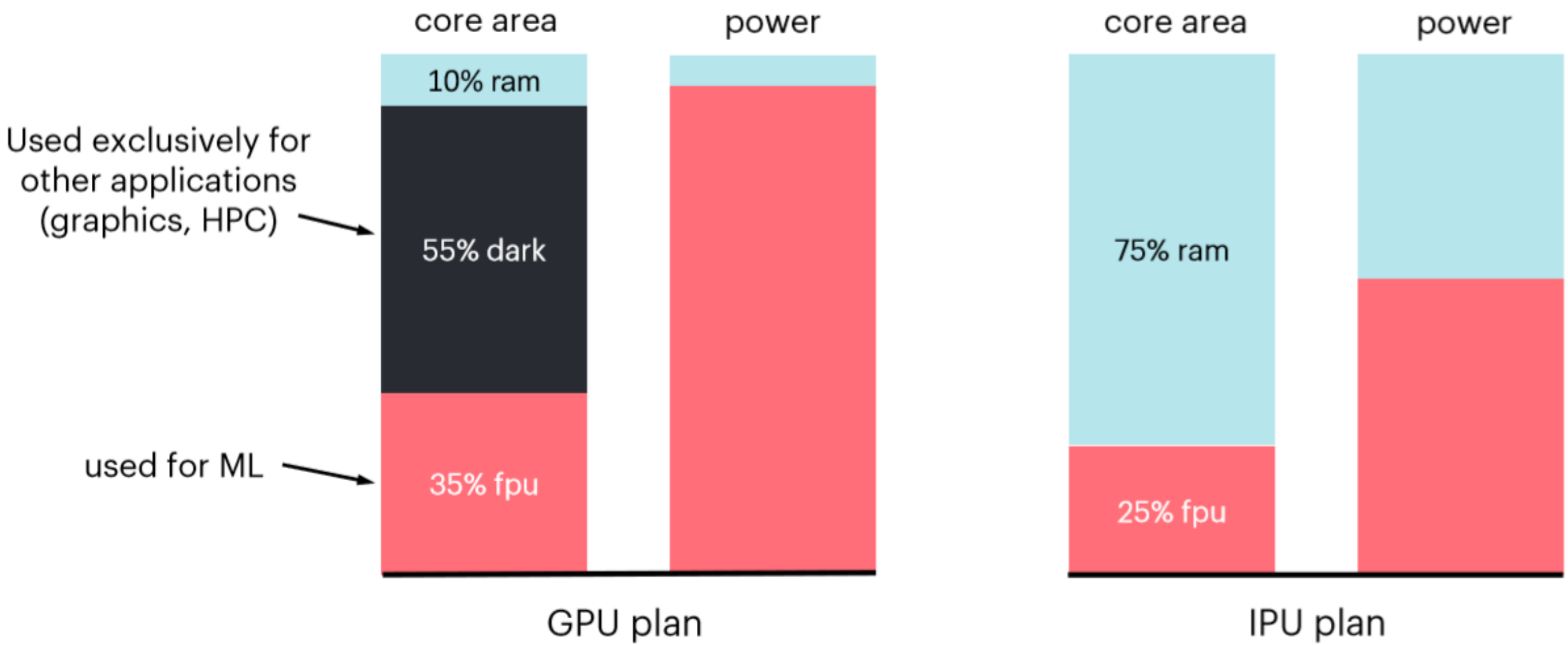
16nm 1Pflop16.32
<700mm²

1000W



8cm² die @ 200W

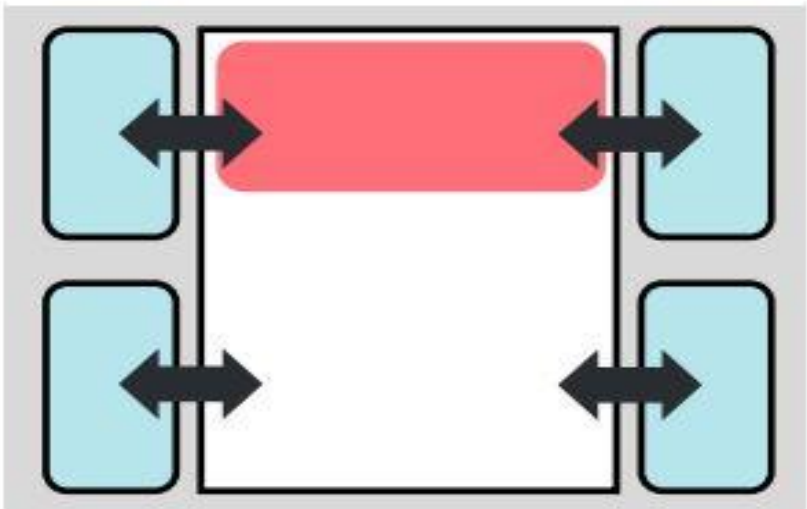
GPUS USE DARK SILICON TO SERVE MULTIPLE MARKETS, IPUS USE IT TO LOCALIZE MEMORY



MEMORY BANDWIDTH @ 240W

DRAM on interposer

180W GPU + 60W HBM2

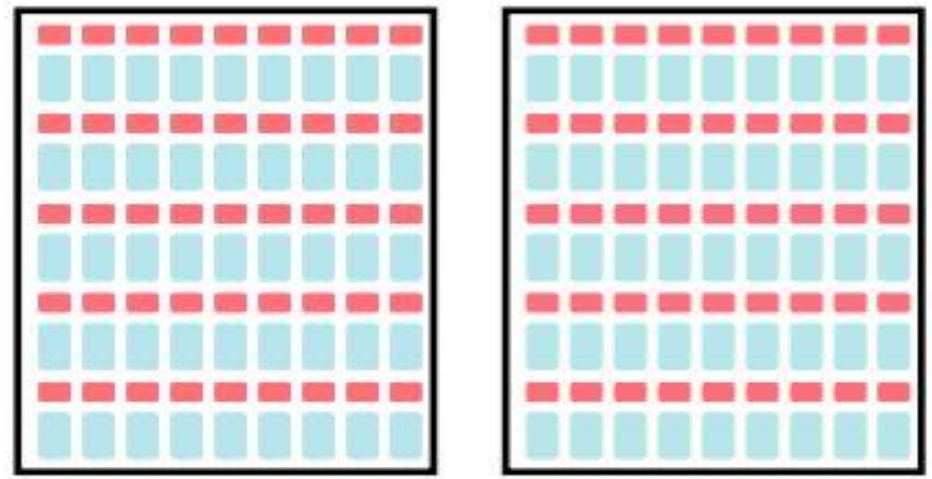


16GB @ 64pJ/B

900GB/s

Distributed SRAM on chip

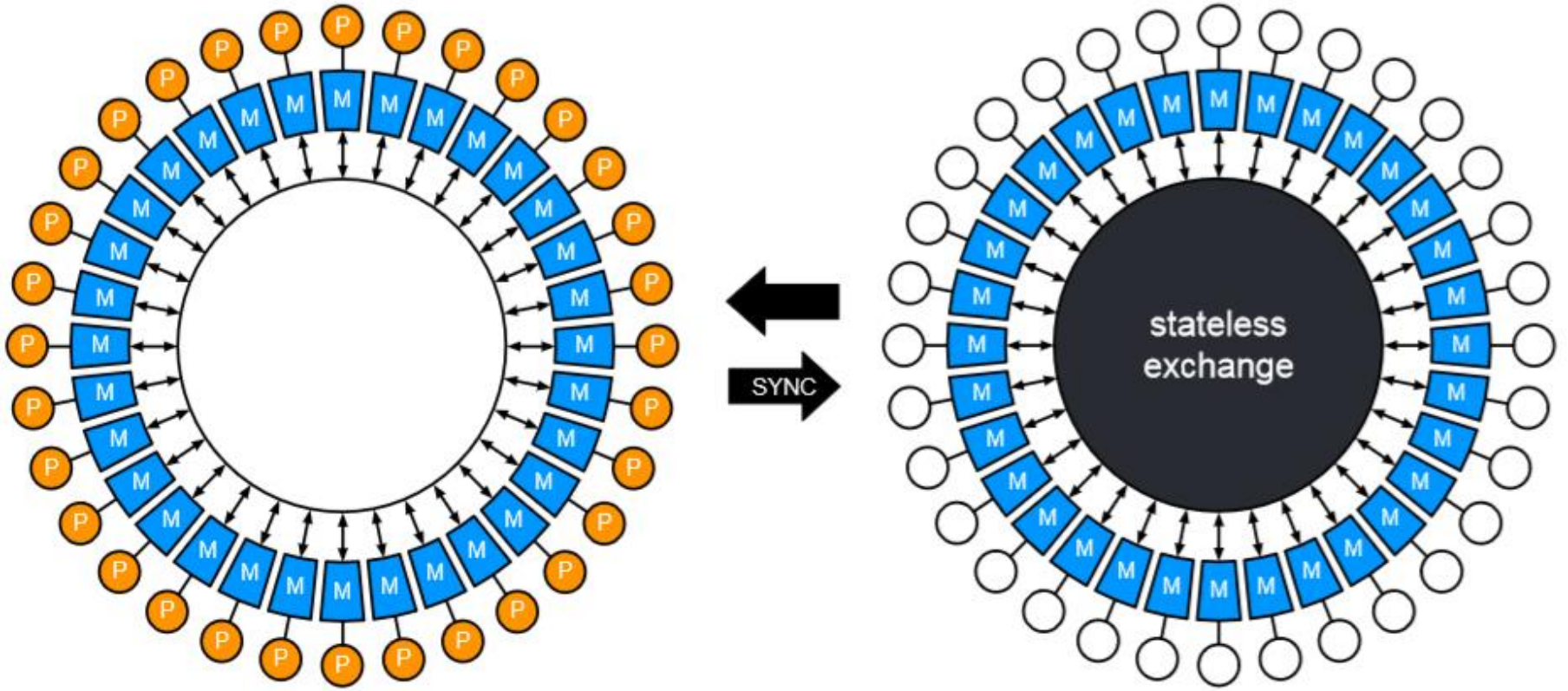
2x IPU (75W logic + 45W ram)



600MB @ 1pJ/B

90,000GB/s

BULK SYNCHRONOUS PARALLEL



Compute Phase

stateful codelets execute on local memory state

Exchange Phase

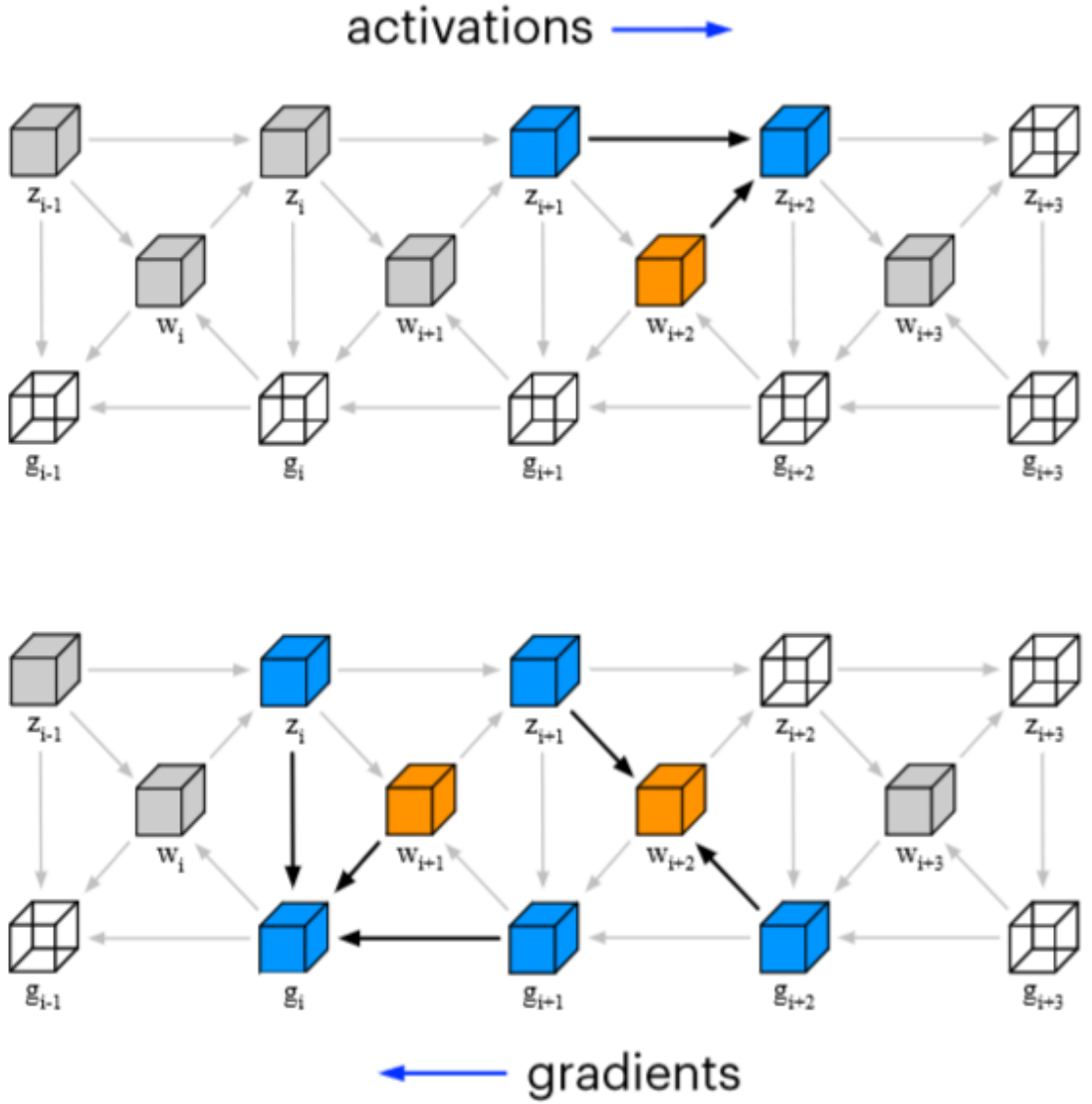
memory-to-memory data movement, no compute, no concurrency hazards

RE-COMPUTE WHAT YOU CAN'T REMEMBER

In back-propagation, most memory is consumed by storing all activations in forward pass, for gradient and weight update calculation in backward pass.

Alternatively, re-compute the activations from sparse snapshots.

Trades most storage for one repeat of forward pass compute.

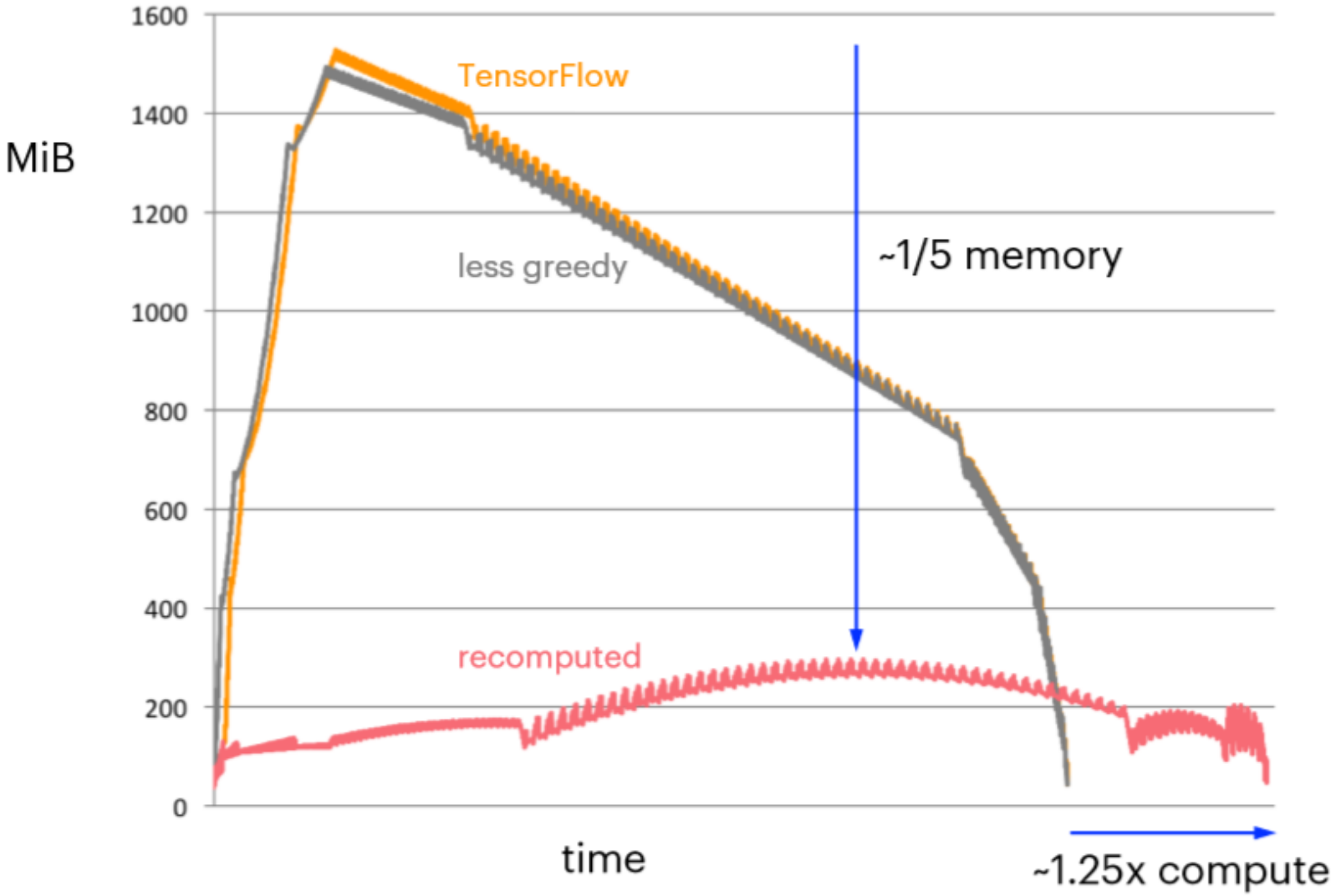


COMPUTE/MEMORY TRADE-OFF IN DENSENET-201 TRAINING

Naive strategy: memorize activations only at input of each residue block

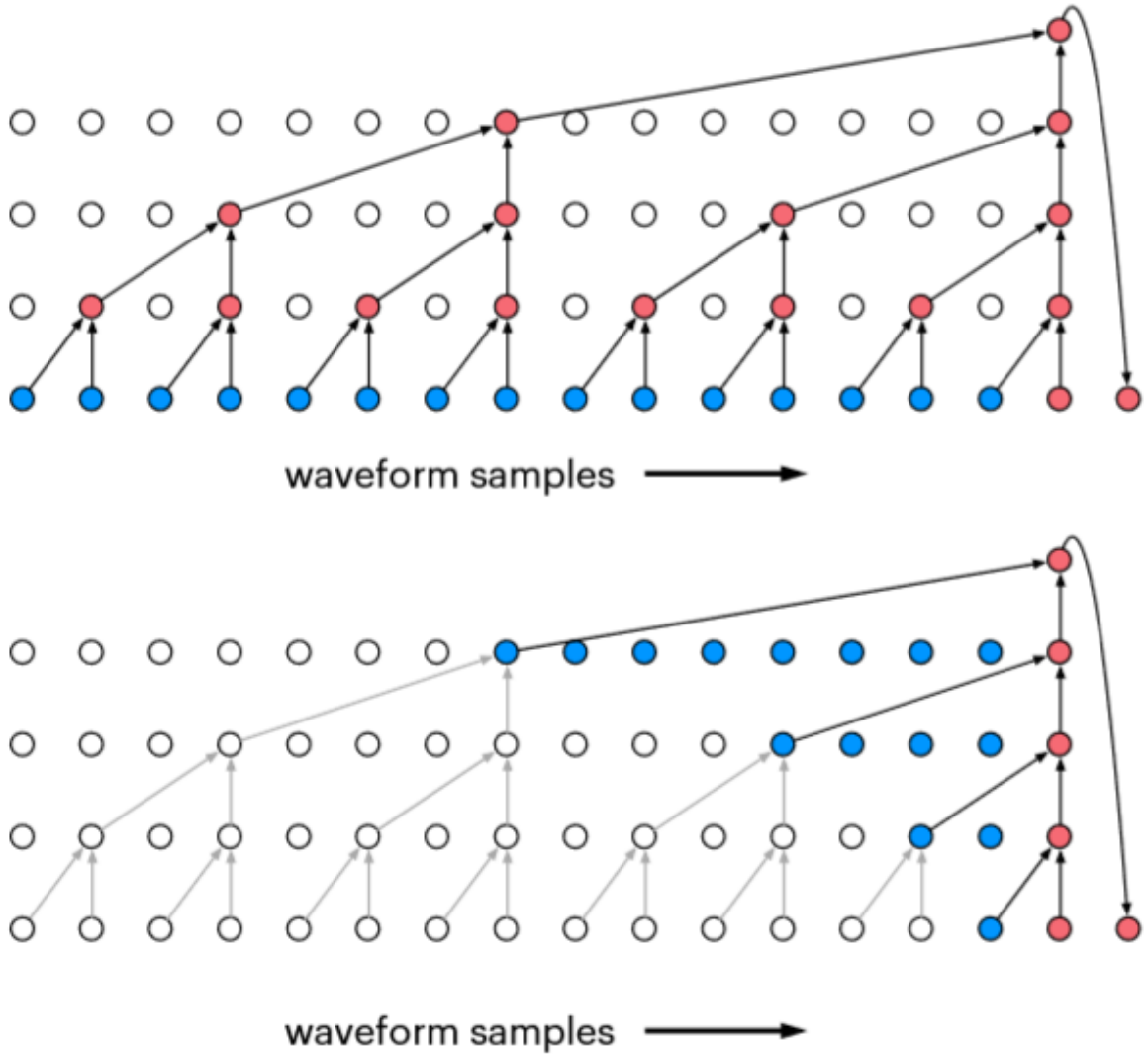
Batch=16 executing on CPU, recording total memory allocated for weights + activations.

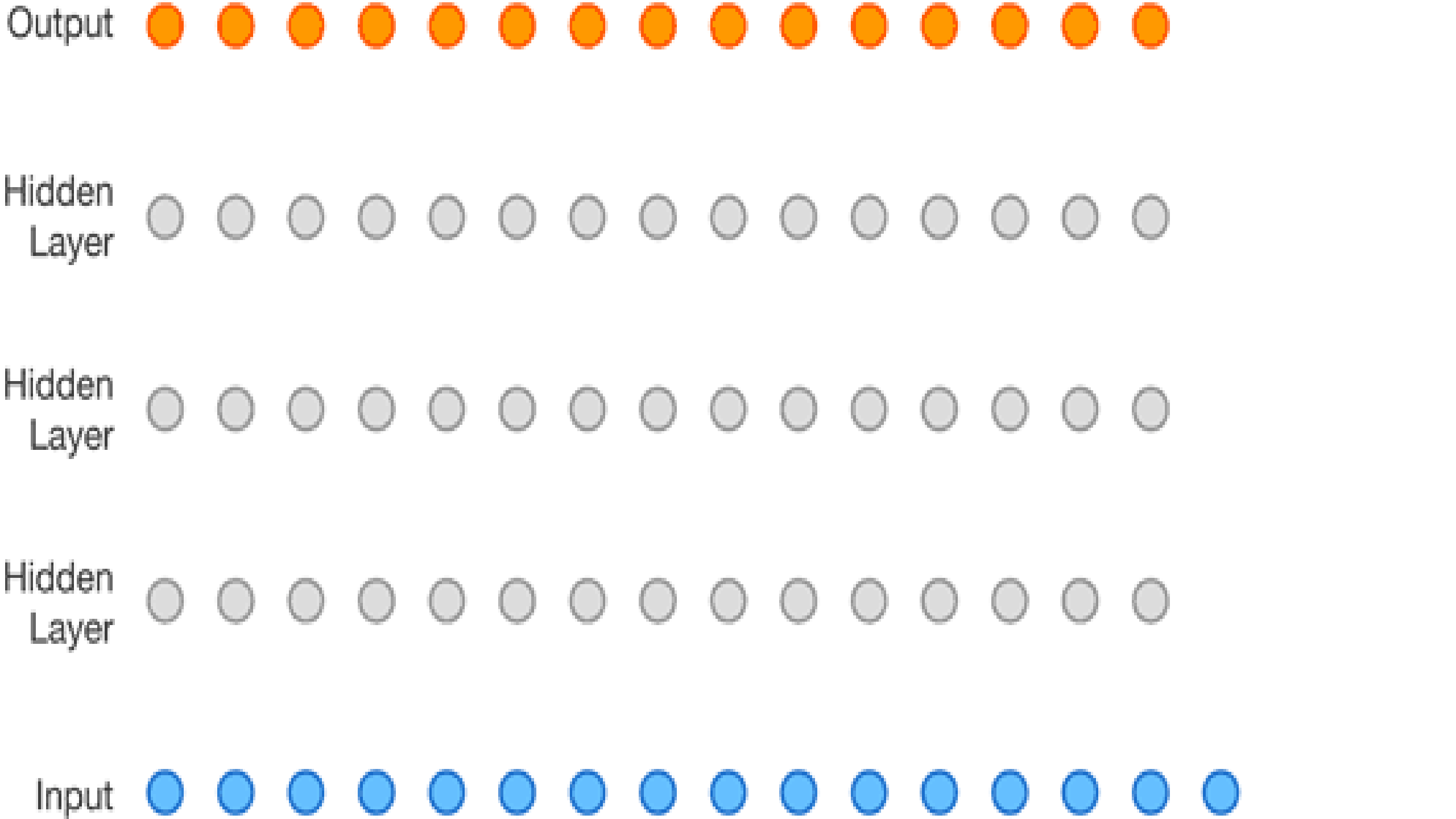
f16 weights and activations, single weight copy.



COMPUTE/MEMORY TRADE-OFF IN WAVENET INFERENCE

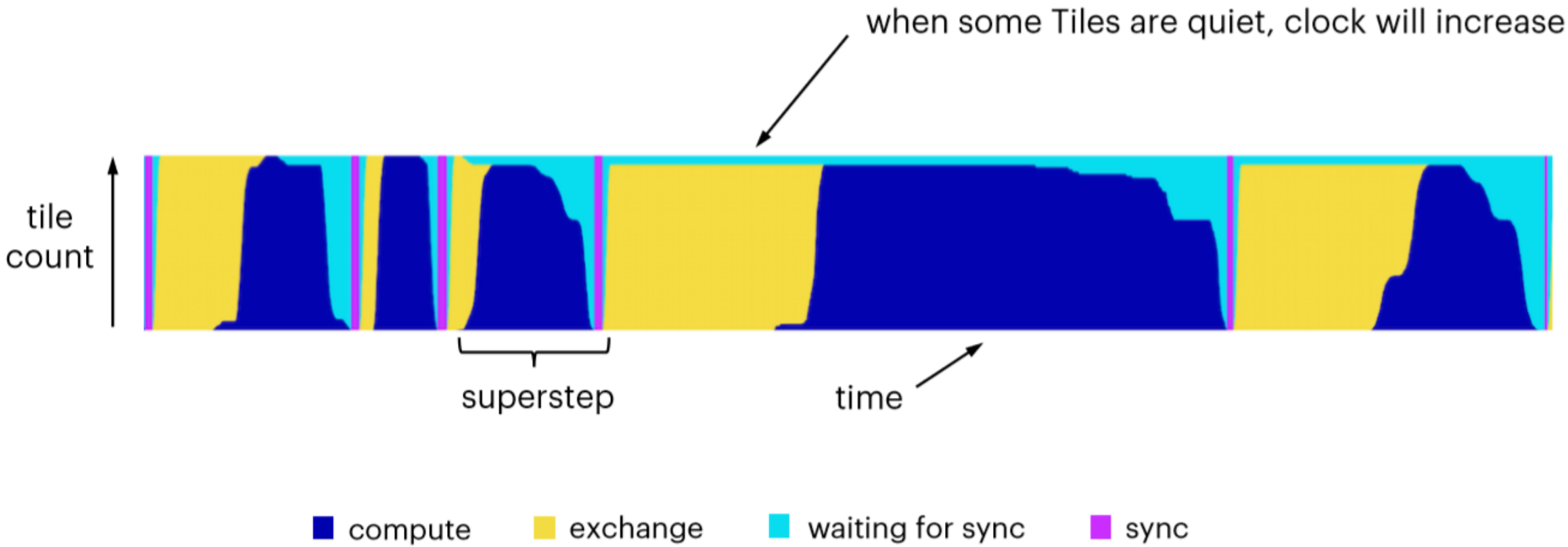
- memorized
- (re)computed



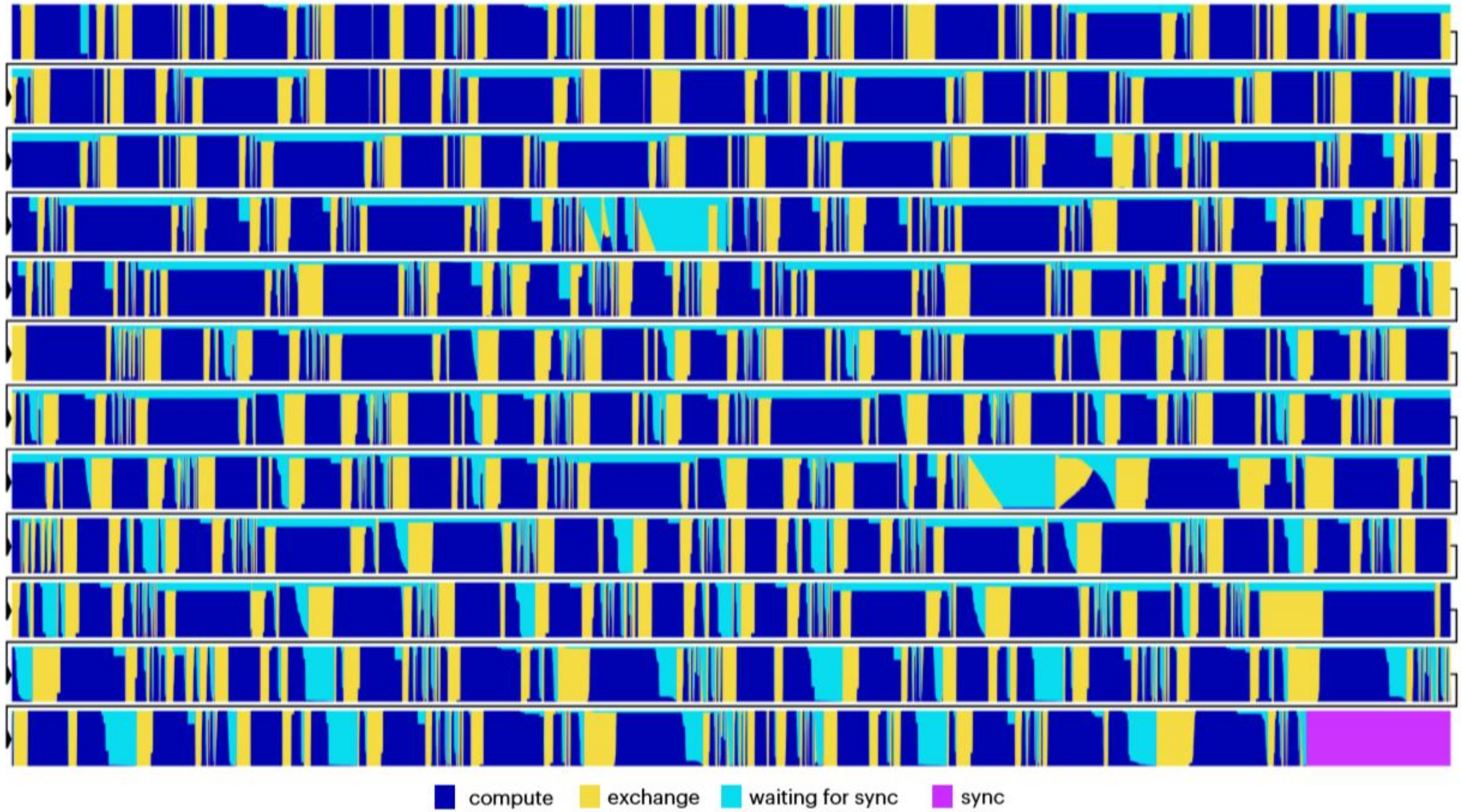

Output
Hidden Layer
Hidden Layer
Hidden Layer
Input

BSP EXECUTION TRACES

Poplar® summary view of workload balance



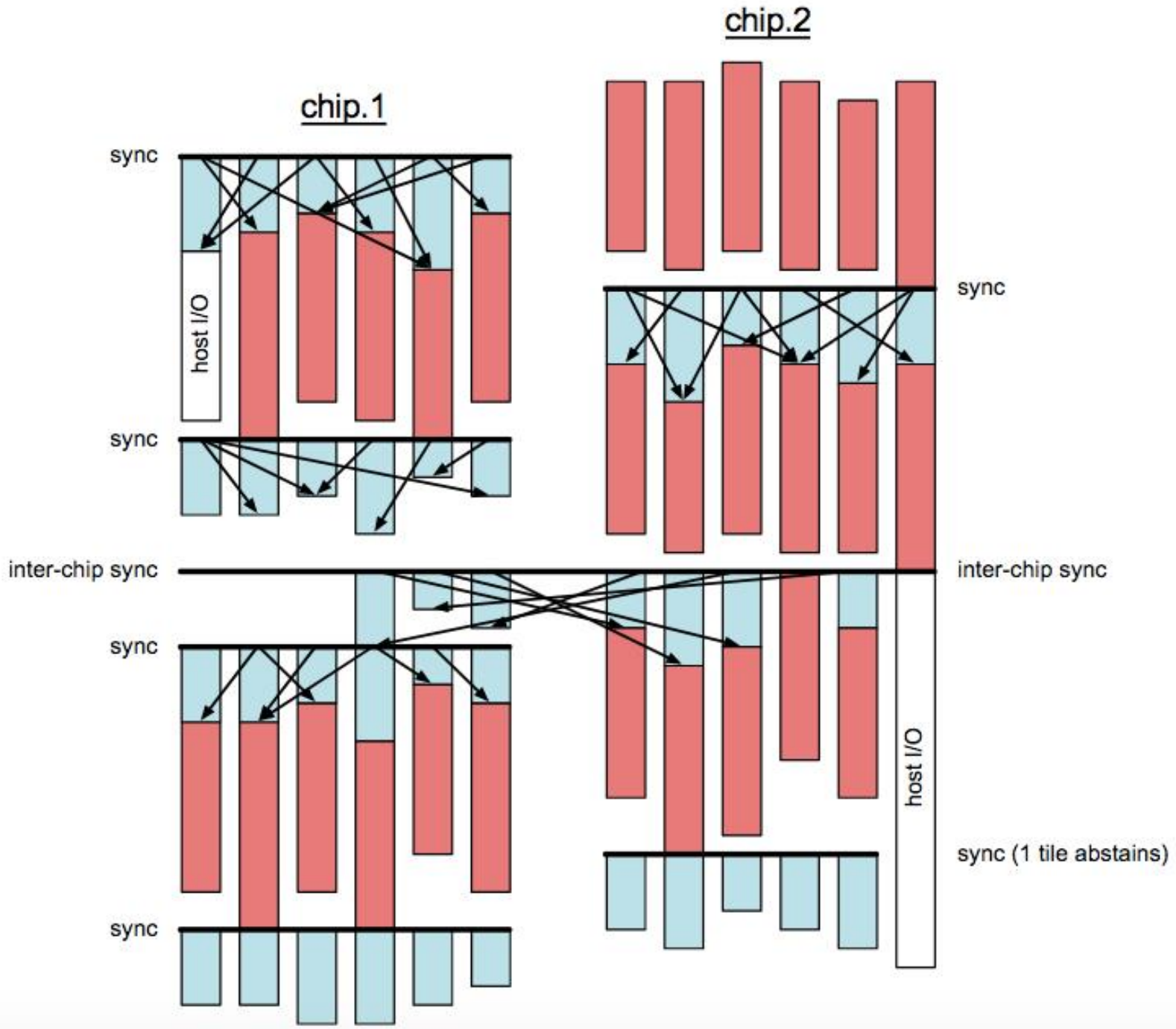
BSP TRACE: RESNET-50 TRAINING, BATCH=4



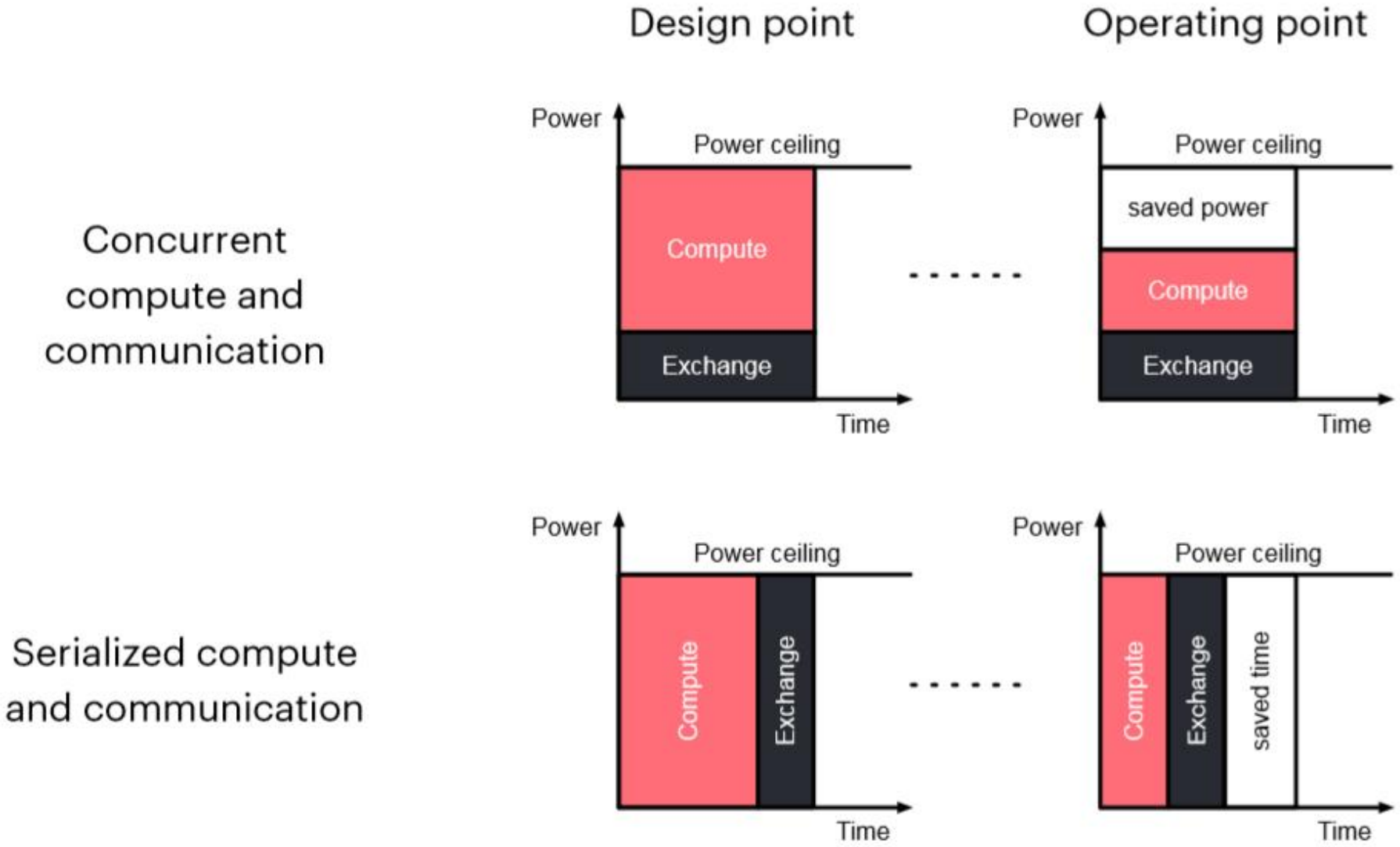
Massively parallel computing with no concurrency hazards

exchange phase

compute phase

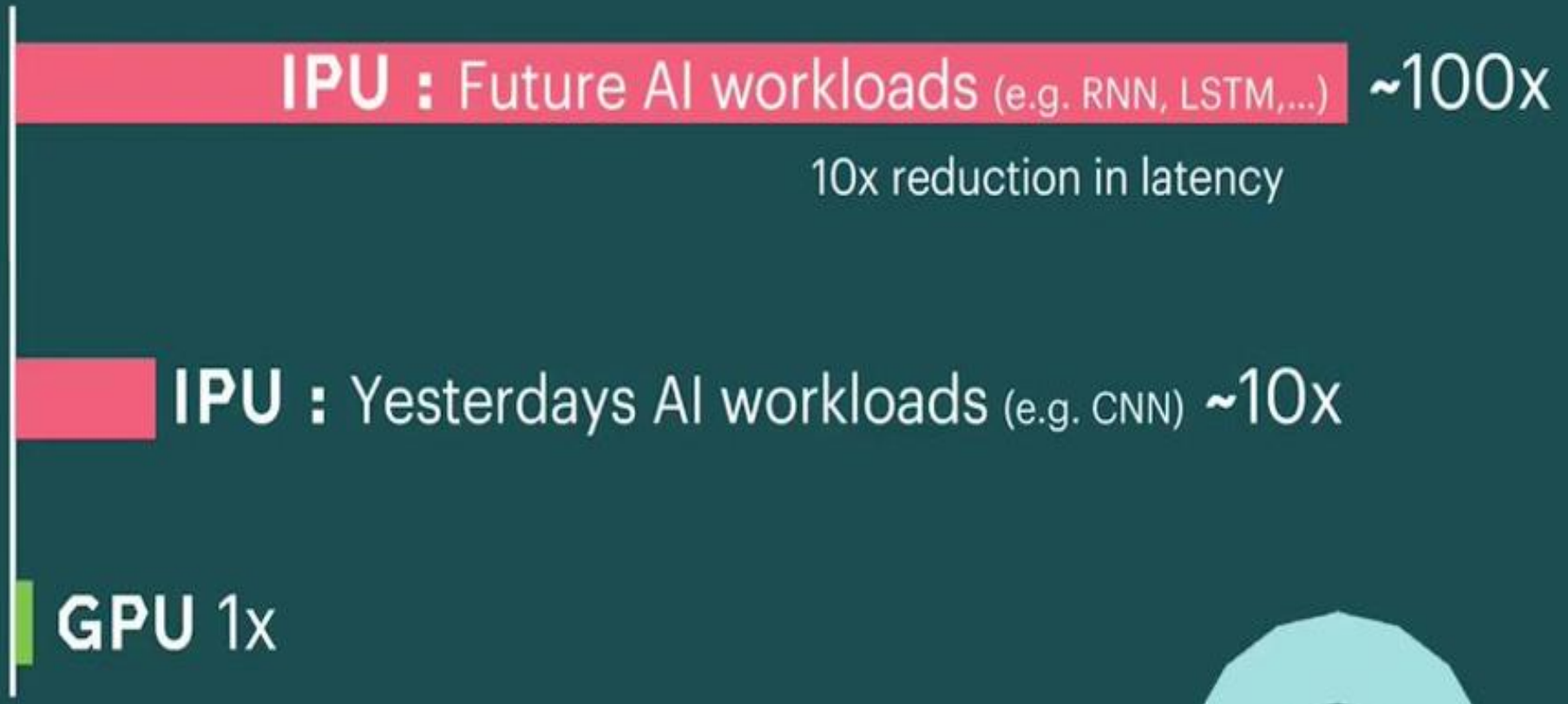


SERIALIZE COMPUTE AND COMMUNICATION

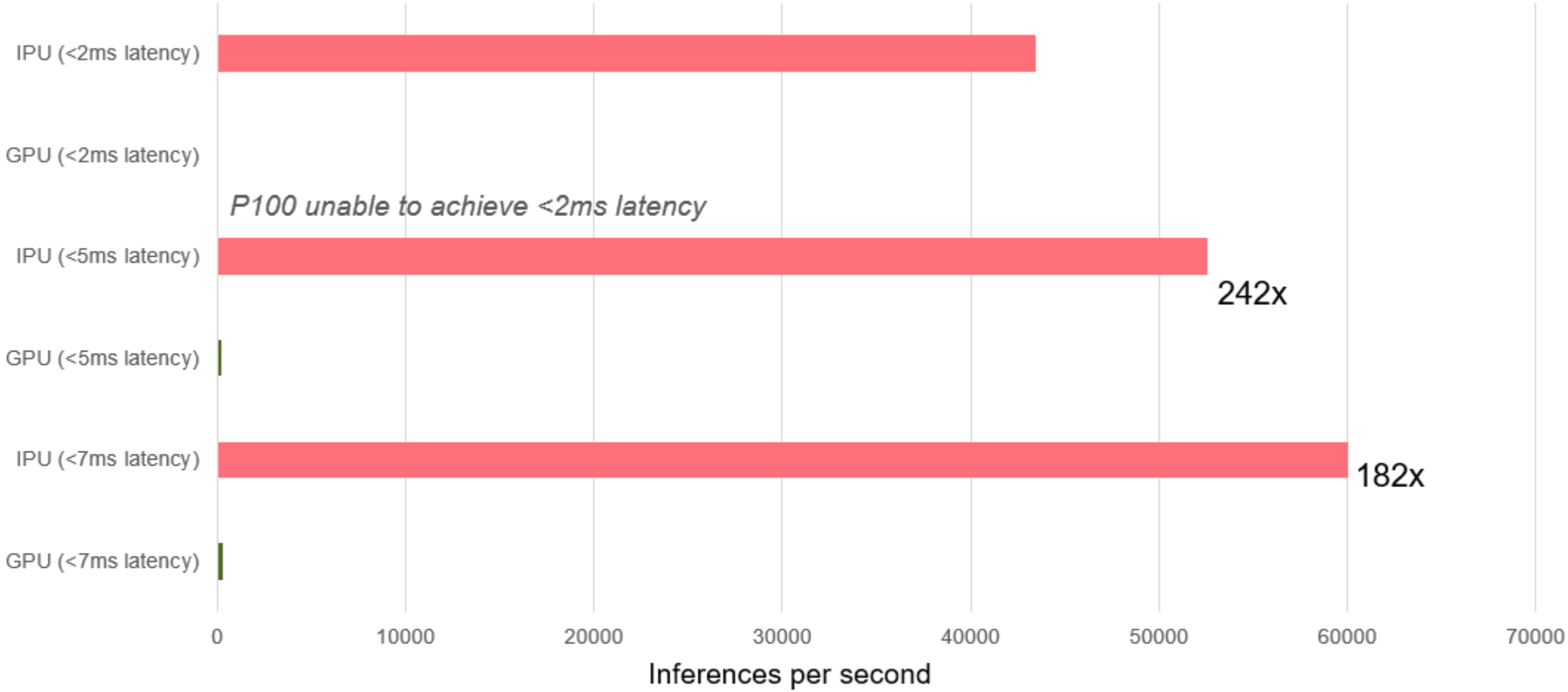


IPU ADVANTAGE

at equivalent power and form factor with latest technology

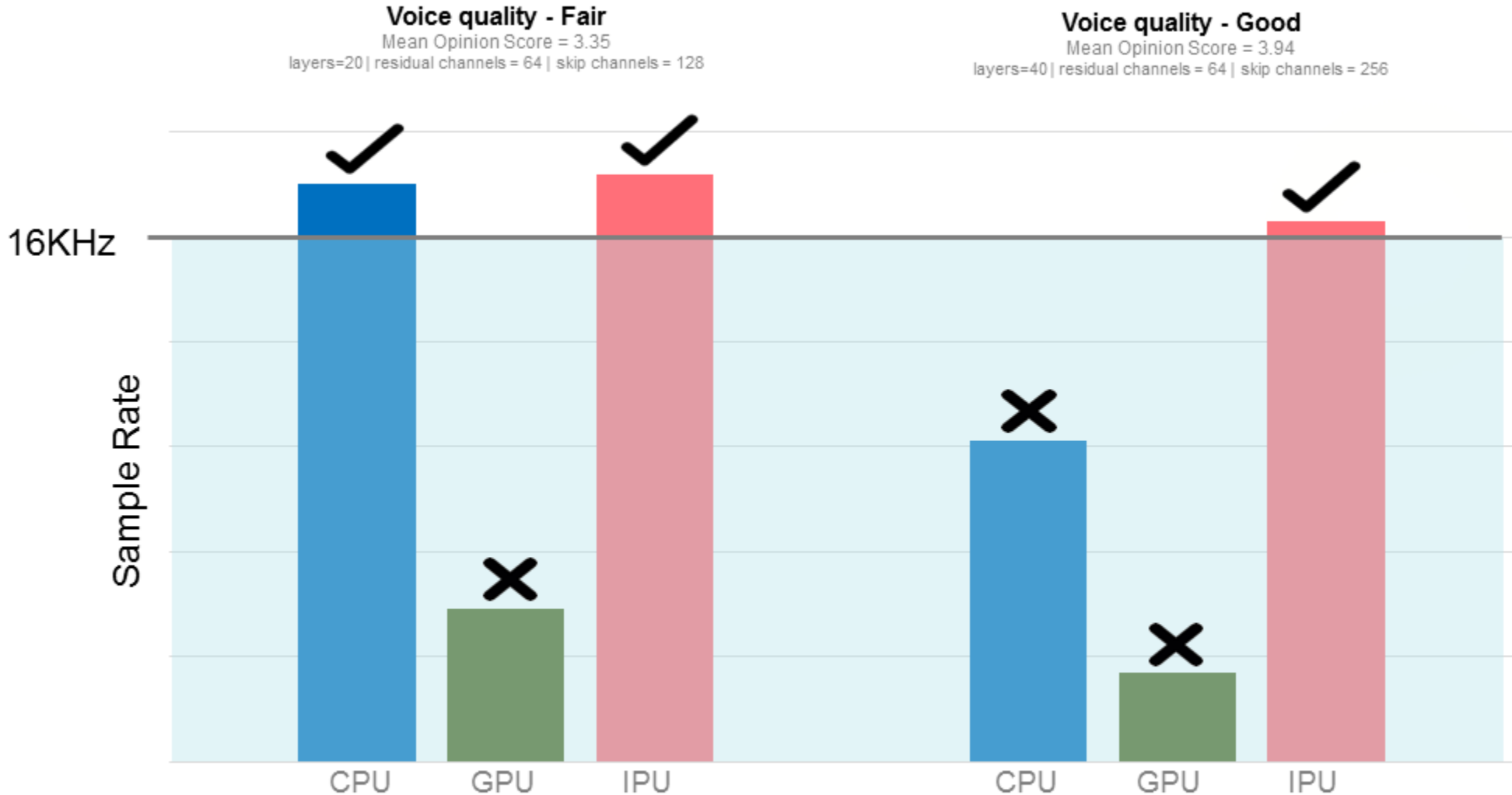


LSTM SINGLE LAYER INFERENCE



- Preliminary results
- LSTM parameters: hidden state = 1536, number of steps = 50
- IPU and GPU results using FP16 data and parameters
- IPU results on Graphcore C2 accelerator card
- GPU results on Nvidia P100-PCIE-12GB, cuDNN 7.0, cuda 8.0.61.2, half precision

DEEP VOICE: ACHIEVABLE SAMPLE RATE



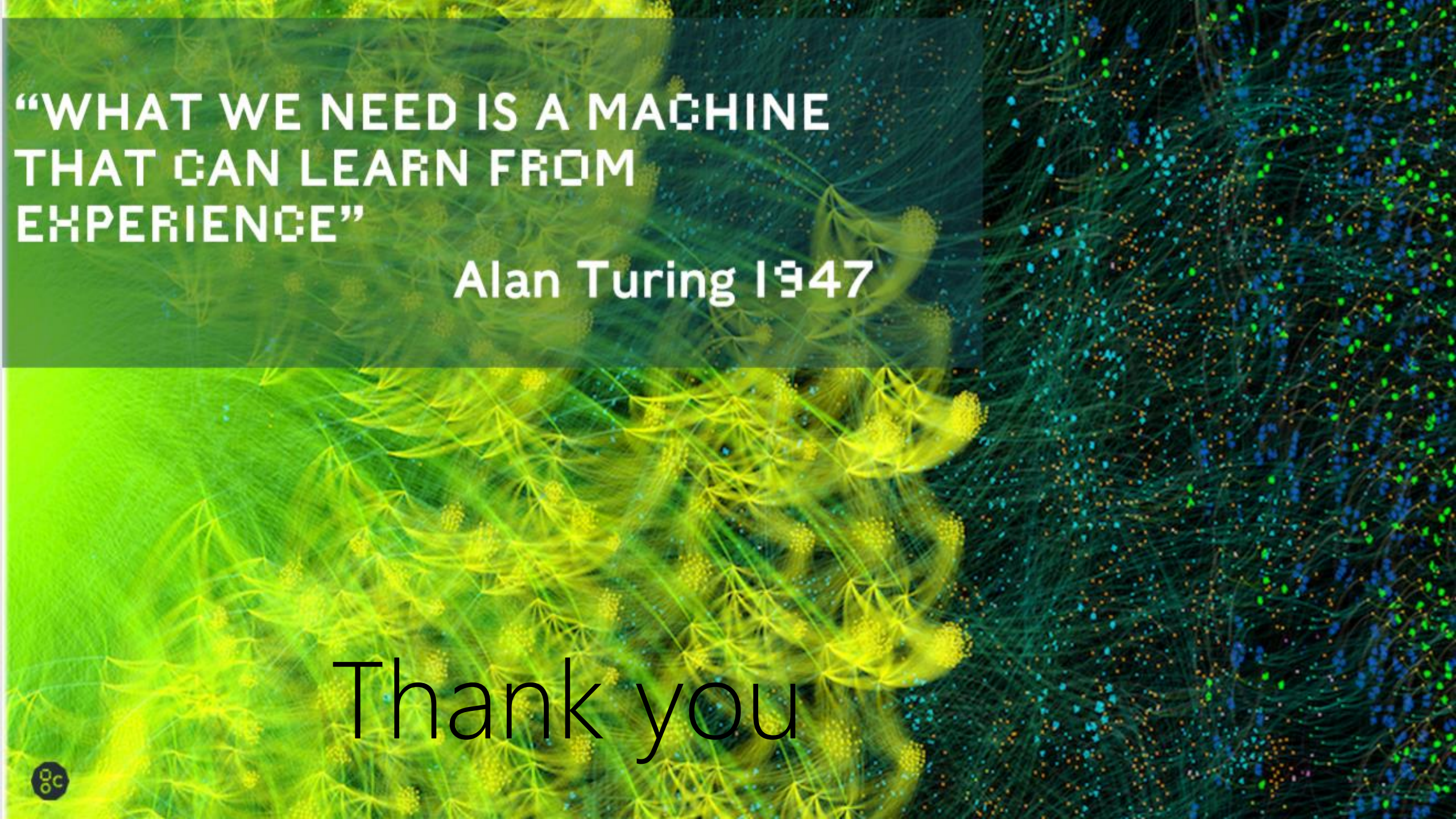
- Preliminary results
- IPU using fp16 parameters and activations
- CPU system: Intel Xeon E5-2660 v3 Haswell (2.6GHz)
- GPU system: Nvidia GeForce GTX TitanX Maxwell
- IPU system: Graphcore C2 IPU accelerator PCIe card

Benchmark: ResNet-50 ImageNet Training at 16,000 Images/Sec

8x C2 IPU Accelerator PCIe cards

54x NVIDIA Volta V100
Scaled 1/2.4x from 128x Pascal P100

Source for P100: Facebook https://arxiv.org/abs/1706.02677
Source for P100 to V100 scaling: Nvidia


“WHAT WE NEED IS A MACHINE
THAT CAN LEARN FROM
EXPERIENCE”
Alan Turing 1947
Thank you

Nigel Toon, Graphcore: A Processor for Machine Intelligence, Samsung Forum

train it so you could reduce the time it takes to train a model by

HOW IPU IS DIFFERENT...

Knowledge model held inside the processor

Thousands of processors work independently on model

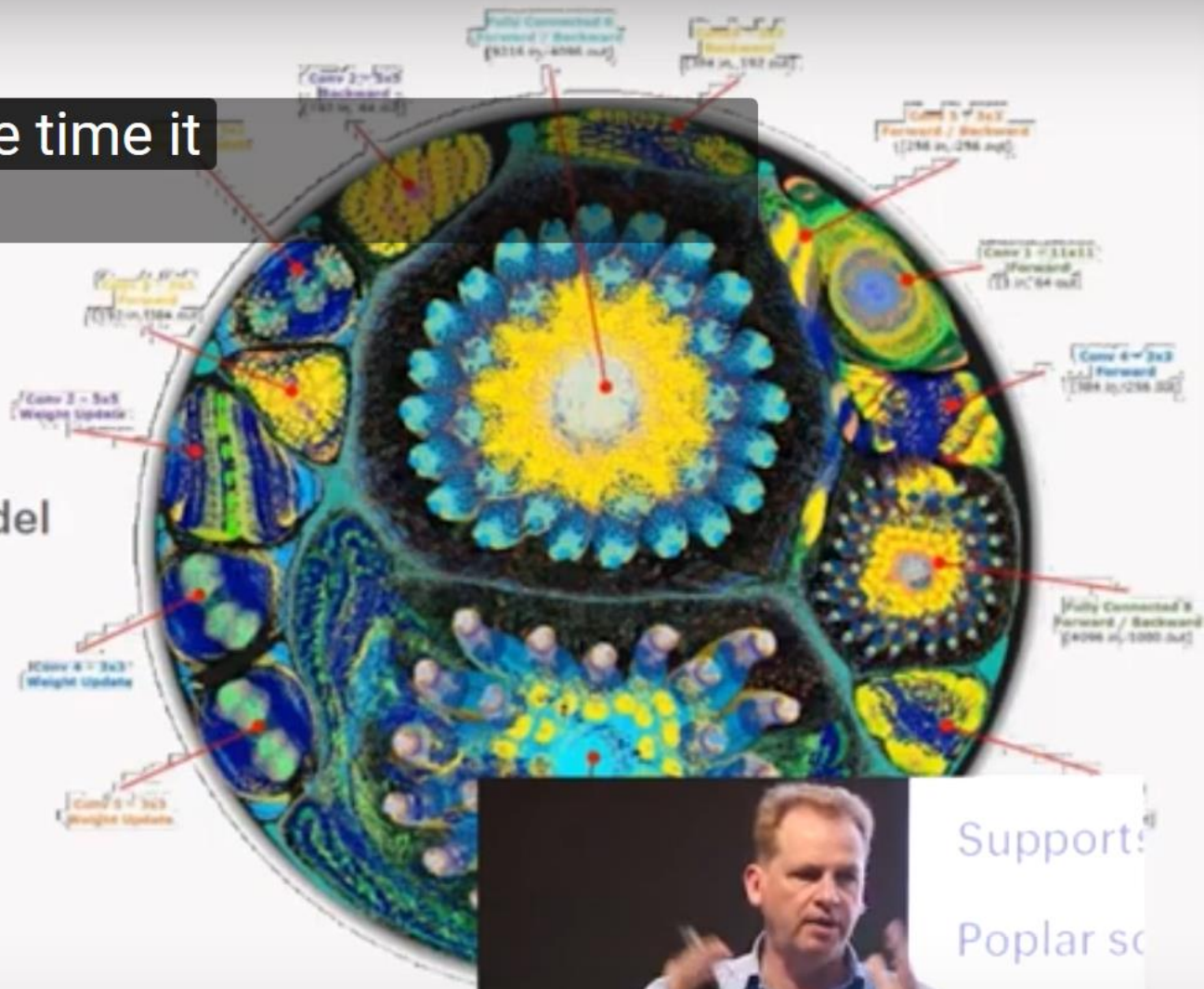
Novel arithmetic modes and entropy support

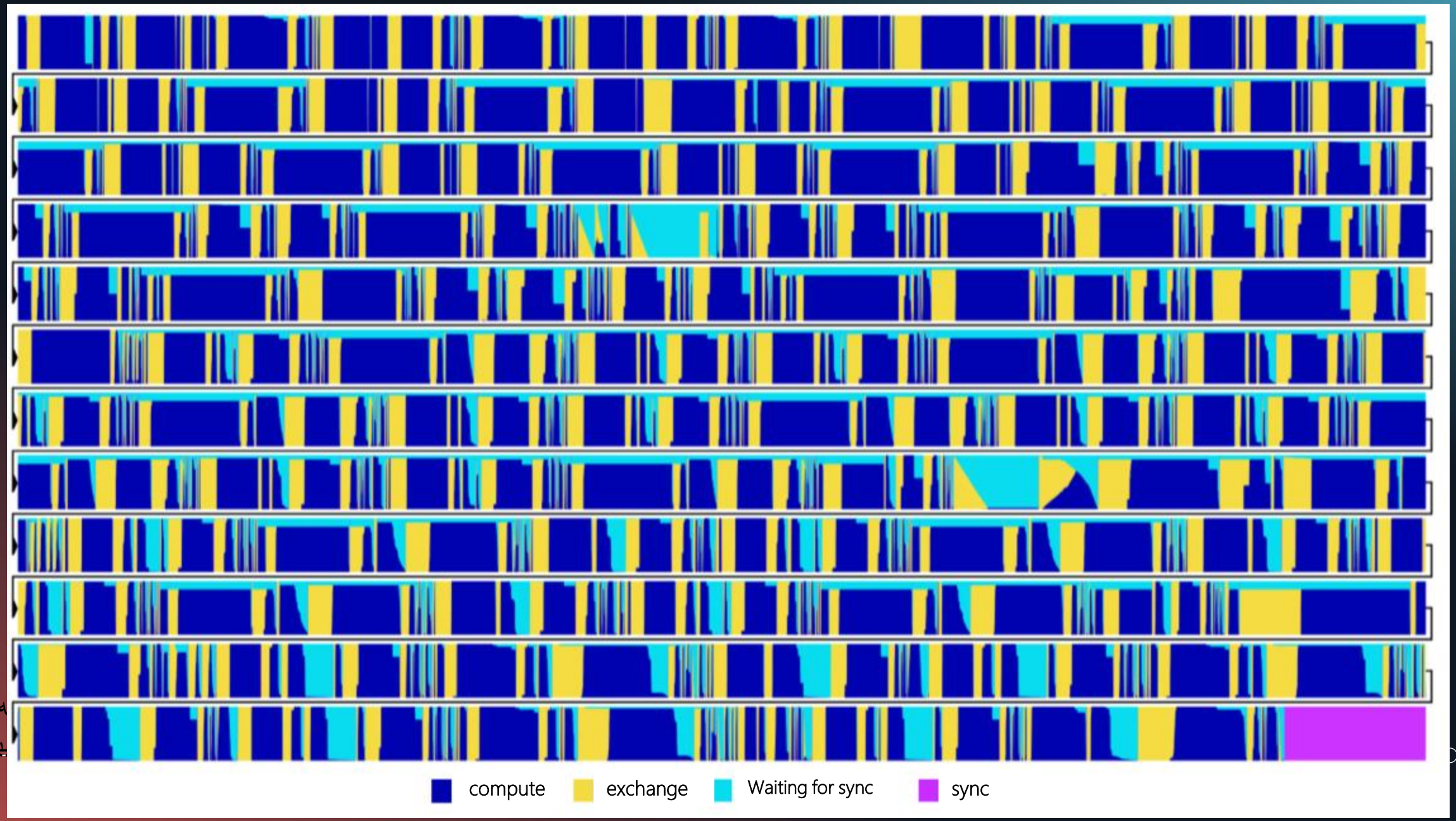
Supports existing and new model structures

Efficient for both training and inference

Supports systems that learn from experience

Poplar software stack is easy to use





ه ۱۶
یشتر

آیا آی پی یو ها قراره جایگزین بشن چه طمینانی هست که این پردازنده ها هم منسوخ نشوند؟آی پی یو یک مغز هست و بقیه پردازنده ها مثل سایر اعضای بدن

ما نه تنها به دنبال پیدا کردن خروجی به ازای داده های جدید بلکه به دنبال آپدیت کردن مدل داشنی هست که به دست آوردیم

اشتباه نکنید گراف کر یک سیستم محاسباتی بهتر از بقیه نیست بلکه یه ایده ی کاملا جدید بای پردازش است

کارهای ماشین لرنینگ همونطورکه گفته محاسبات پیچیده گرافی داره که مارو وادار به ساخت پروسسور های با این قابلیت بود.

دارن رو نسخه ۷ میلی متریش کار میکنن و اینکه این نسخه حدودا سه برابر بیشتر درون خودش حافظه داره و حدودا سرعت دو برابر و هم چنین ادعایی که دارن اینه که دارن کاری که میکن اینه که یه سری تکنیک های محاسباتی استفاده میکنند که خودش سرعت رو دو برابر میکنه
یکی از مشکلات پرفورمنسد بالا تر در چیپ های نیمه هادی مسئله پاور هست که اونها رو به ایده ی موازی سازی سیستمها رو آورده

حتی رو ۵ نانومتریش هم دارن کار میکنن

اما چرا سخت افزار مسئله ی دقت رو داریم ما نیاز به سخت افزاری دارشتیم که علاوه بر بهبود زمان دقت رو به اندازه کافی بیشتر کنه ببنید ما مسائلی داریم که بخشیش با الوریتم هایی که داریم حل میشه اما بخشیش نه

اما سوالی که پیش میاد اینه که آیا واقعا
انقدر خت افزار تونسه پیش رفت که با
گراف کر

۶ در صدر ریت ارورو

نکته ای که باز میخوام یادآوری کنم اینه که همونطور که گفتم گراف ها و الگوریتم های گرافی بسار کمک کننده بودنند اما دید ما خیلی وسیع تر بود

اینجا ماشین های خودران رو بگو

TPU
۶۰نانومتر اندازه برد

هرdie

اما سوالی که پیش میاد اینه که با این حساب ما دنبال یه پردازنده جدید هستیم به جای سی پی یو ؟
نه اینوطری نیست داستا ن بدن و مغض


و تو، از دل هر
دیواری که به رویت
کشیدند، پنجره ای
ساختی.....شاید
این بود دلیل
آفرینش تو
به وقت
۲مردا
د
تو
لد
ار
ک

Tensor flow

هر جایی که به نوعی صحبت ازماشین لرنینگ هست پیوسته
با کلمه inference مواجه هستیم

Mi

Step by step with programming

پروژه گراف کور از سال ۲۰۱۲ شروع شد در داشنگاه تورنتو