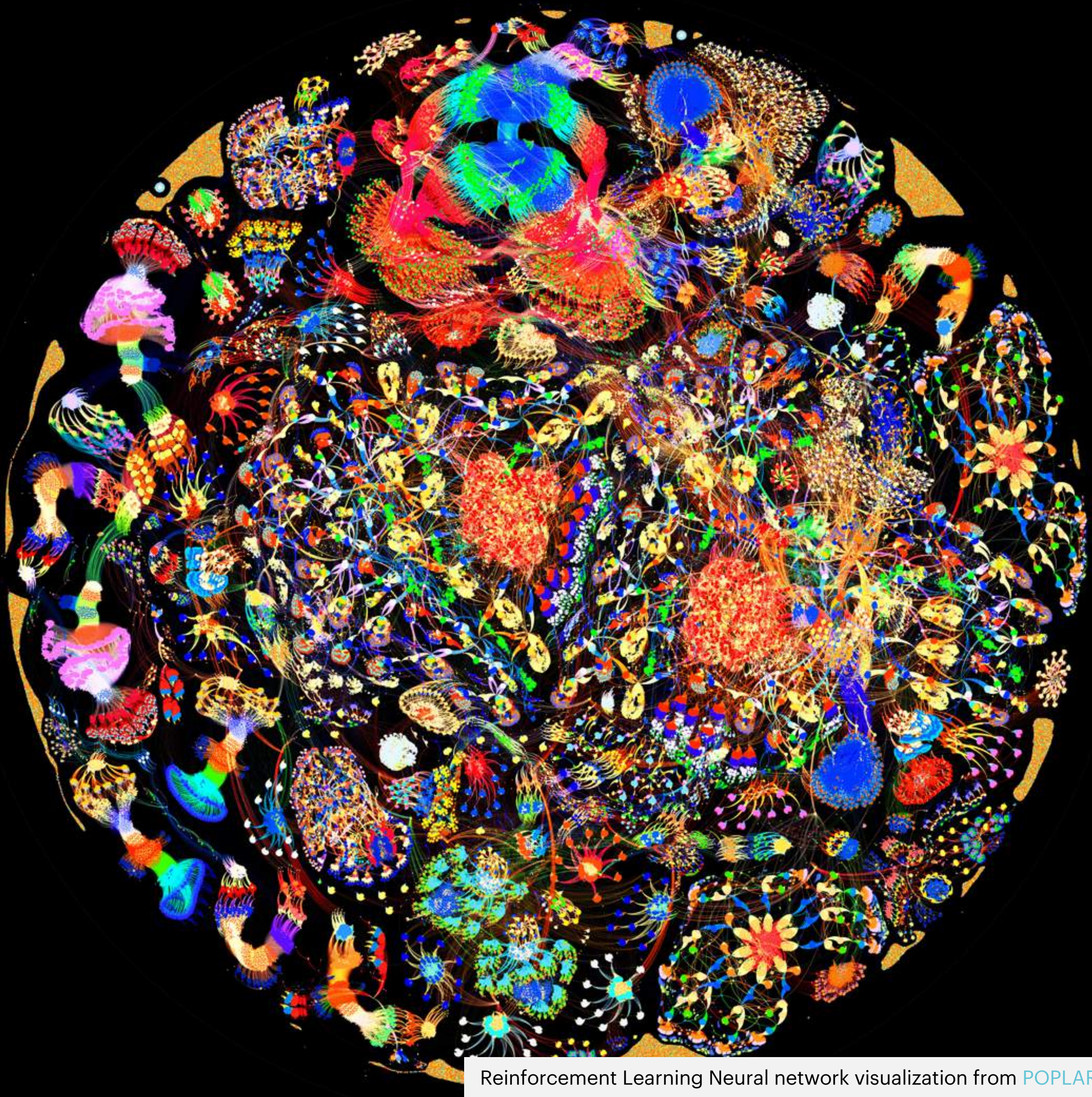
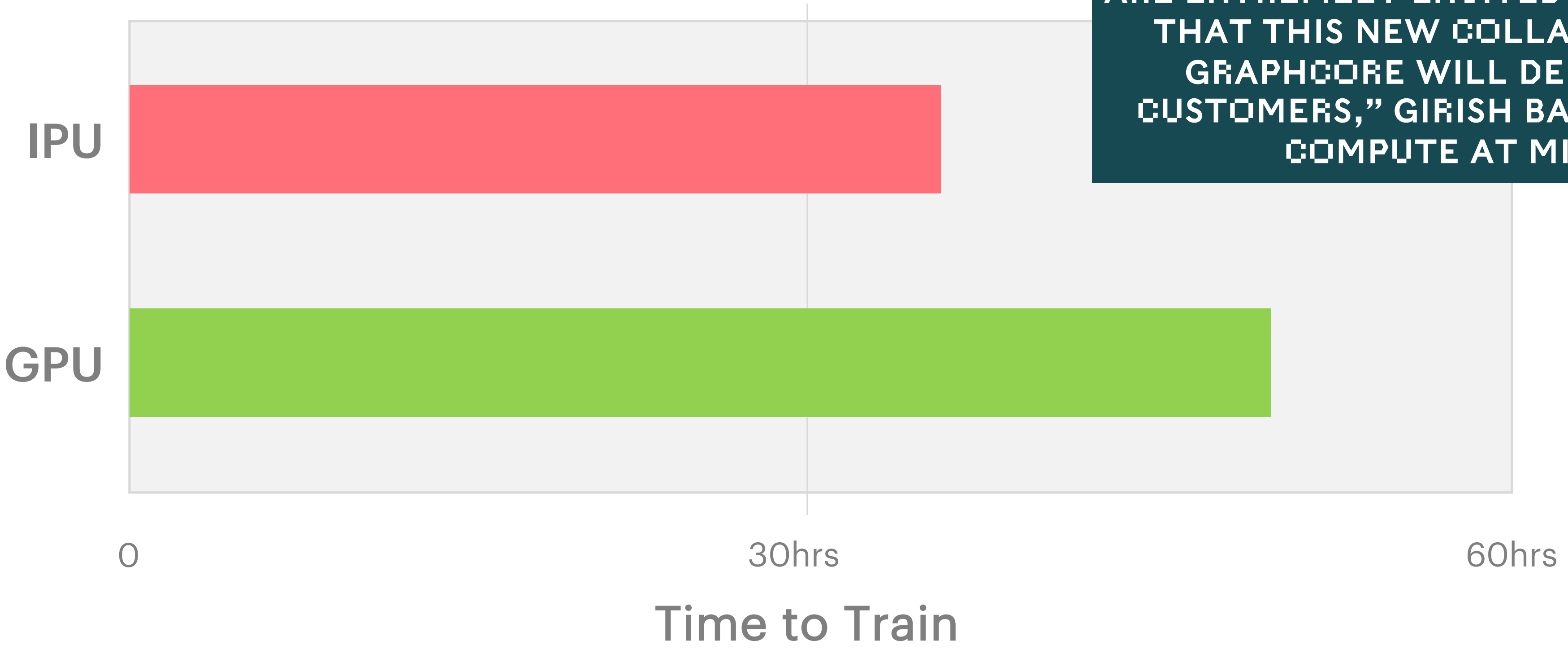
GRAPHCORE BENCHMARKS



Reinforcement Learning Neural network visualization from POPLAR™

BERT-BASE : TRAINING

>25% Faster Time To Train : 36.3 hours on IPU @ 20% lower power

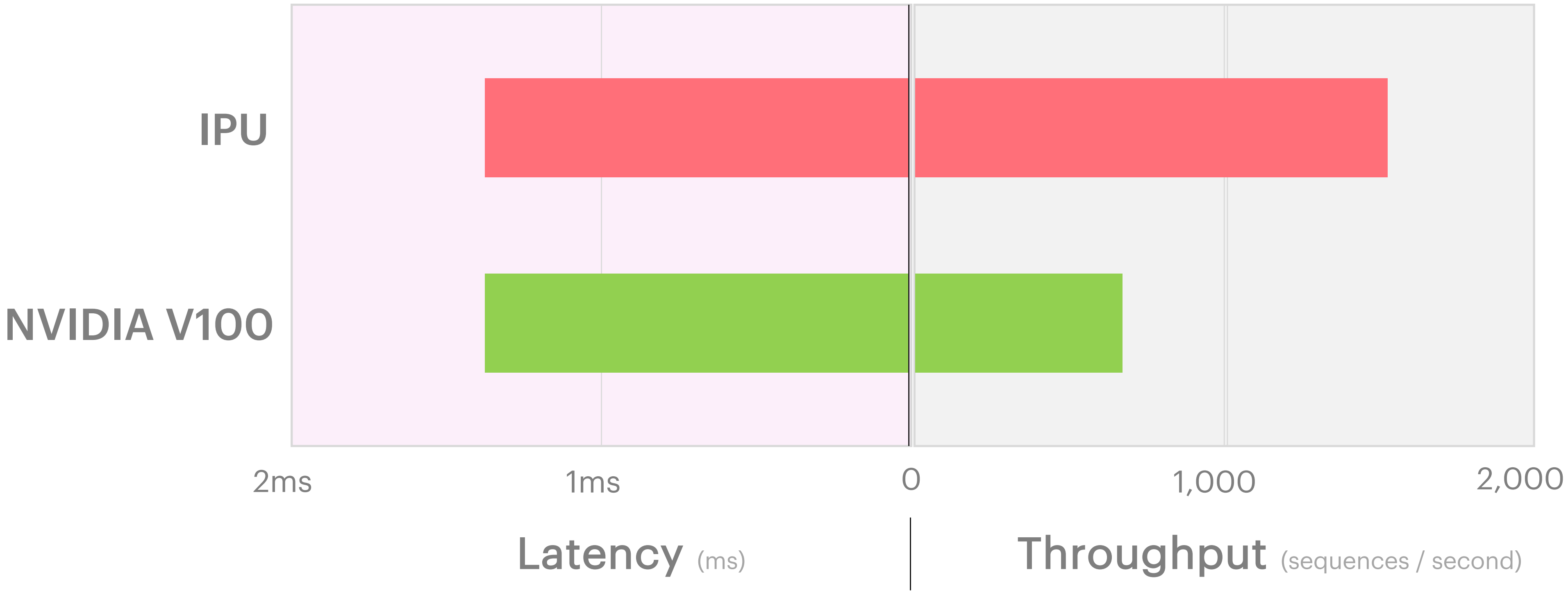


“NATURAL LANGUAGE PROCESSING MODELS ARE HUGELY IMPORTANT TO MICROSOFT. WE ARE EXTREMELY EXCITED BY THE POTENTIAL THAT THIS NEW COLLABORATION WITH GRAPHCORE WILL DELIVER FOR OUR CUSTOMERS,” GIRISH BABLANI, VP AZURE COMPUTE AT MICROSOFT

NOTES:
BERT-Base | Wikipedia dataset + SQuAD 1.1 (EM).
IPU: DSS8440, 7x Graphcore C2 – using PopART (SDK1.1.45) Ph1 SL=128, Ph2 SL=384
GPU: 8x Leading GPU system using PyTorch

BERT-BASE : INFERENCE

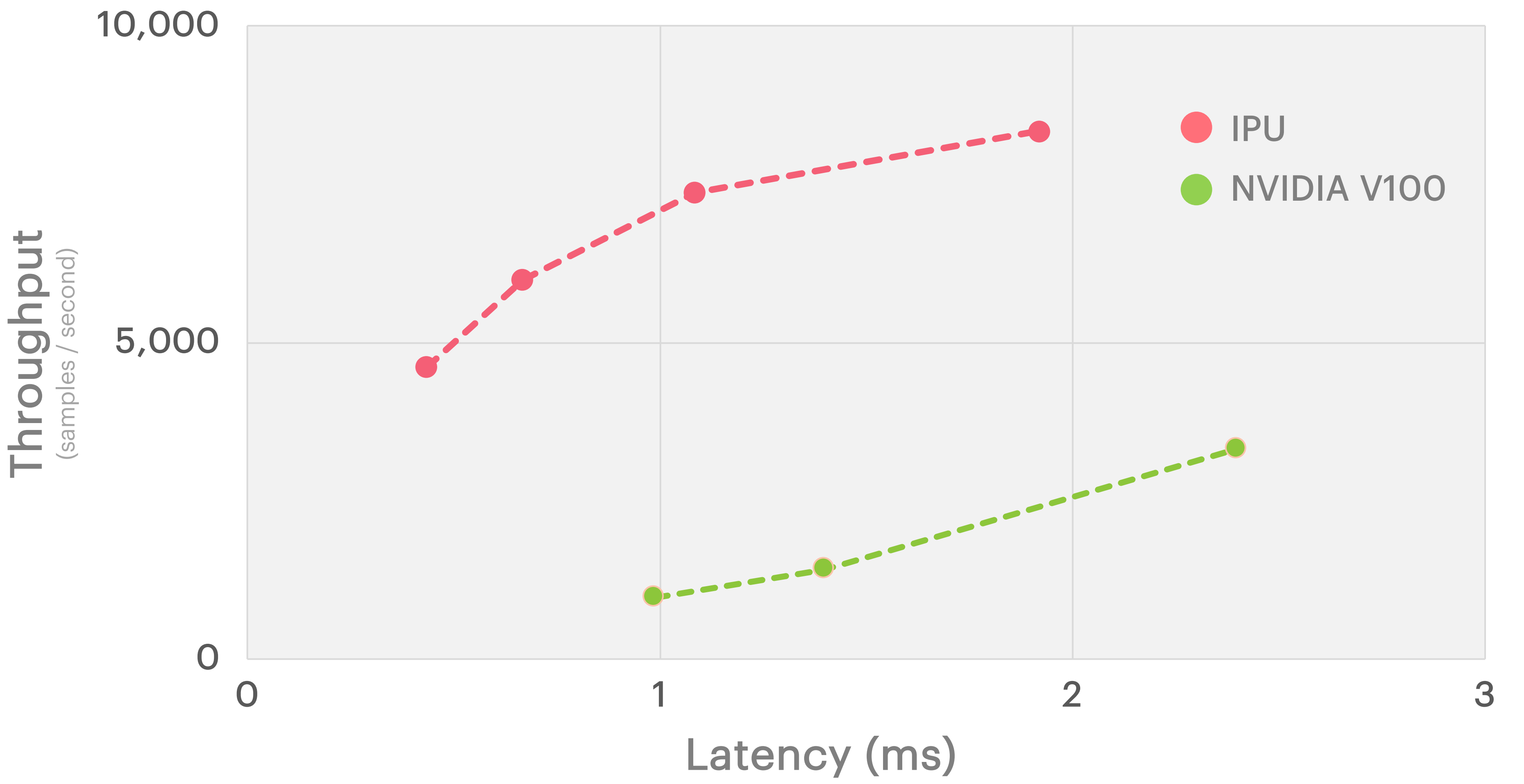
2x higher throughput at lowest latency



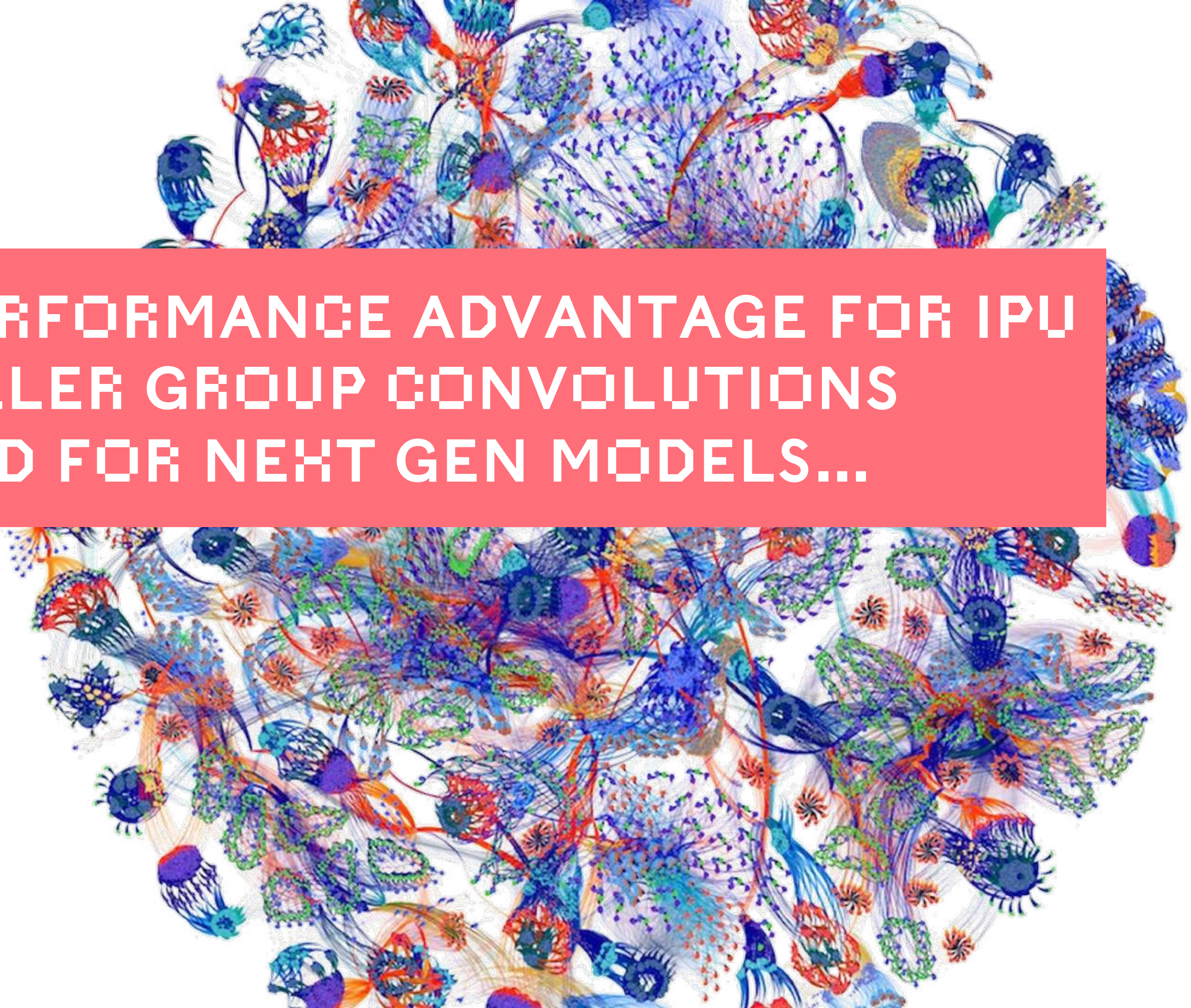
NOTES:
Graphcore results on one C2 Card using two IPUs, on SQuAD v1.1 data,
Graphcore C2 (SDK 1.0.194) using PopART @ 300W TDP
NVIDIA results for 1xV100 with TensorRT 7.0 using SQuAD v1.1 data, published 12 Mar 2020
https://developer.nvidia.com/deep-learning-performance-training-inference.

RESNET-50 : INFERENCE

Lowest Latency Comparison: 4.7x higher throughput | 2.2x lower latency



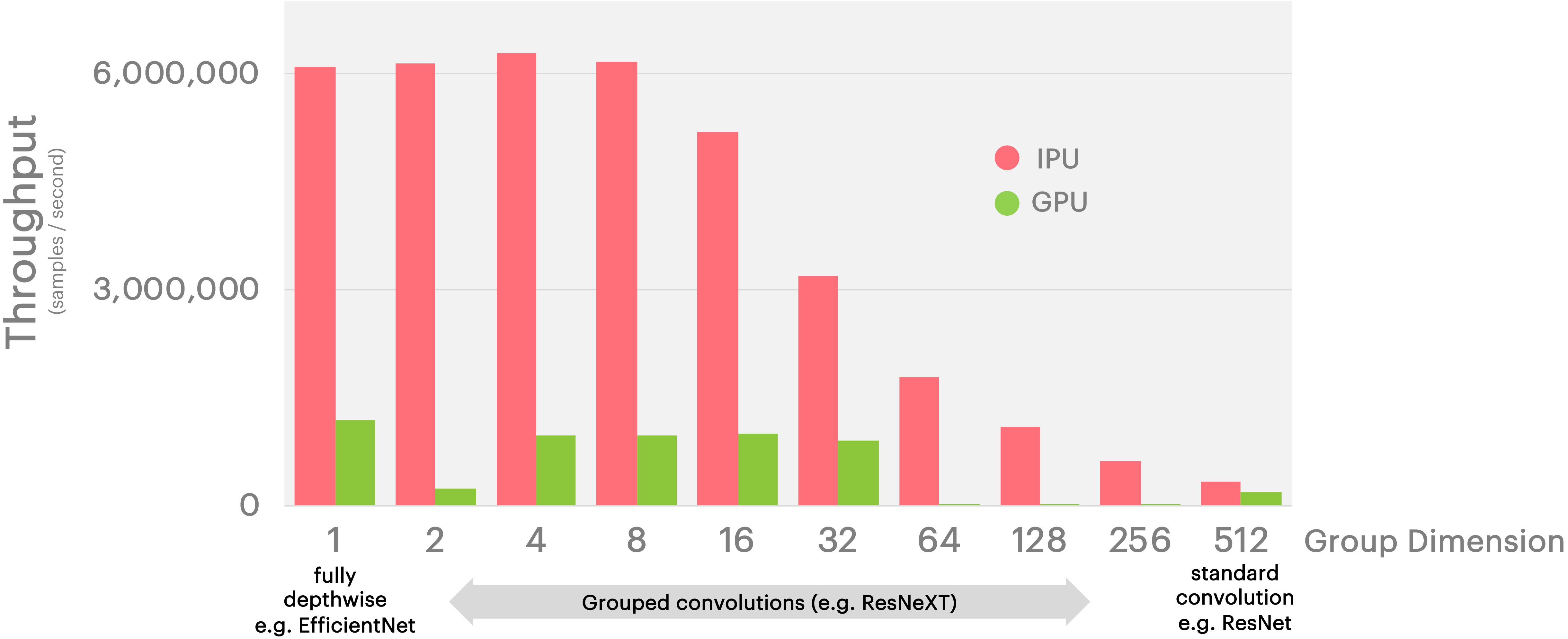
NOTES: ResNet-50 v1.5
Graphcore : C2 Platform using TensorFlow with synthetic data, (Batch Size 2,4,8,16), (SDK 1.1.11) @ 300W TDP
NVIDIA V100 with TensorRT 7.0 published 12 Mar 2020 https://developer.nvidia.com/deep-learning-performance-training-inference using synthetic data, (Batch Size 1,2,8) @ 300W TDP



HUGE PERFORMANCE ADVANTAGE FOR IPU ON SMALLER GROUP CONVOLUTIONS REQUIRED FOR NEXT GEN MODELS…

GROUP CONVOLUTION KERNELS

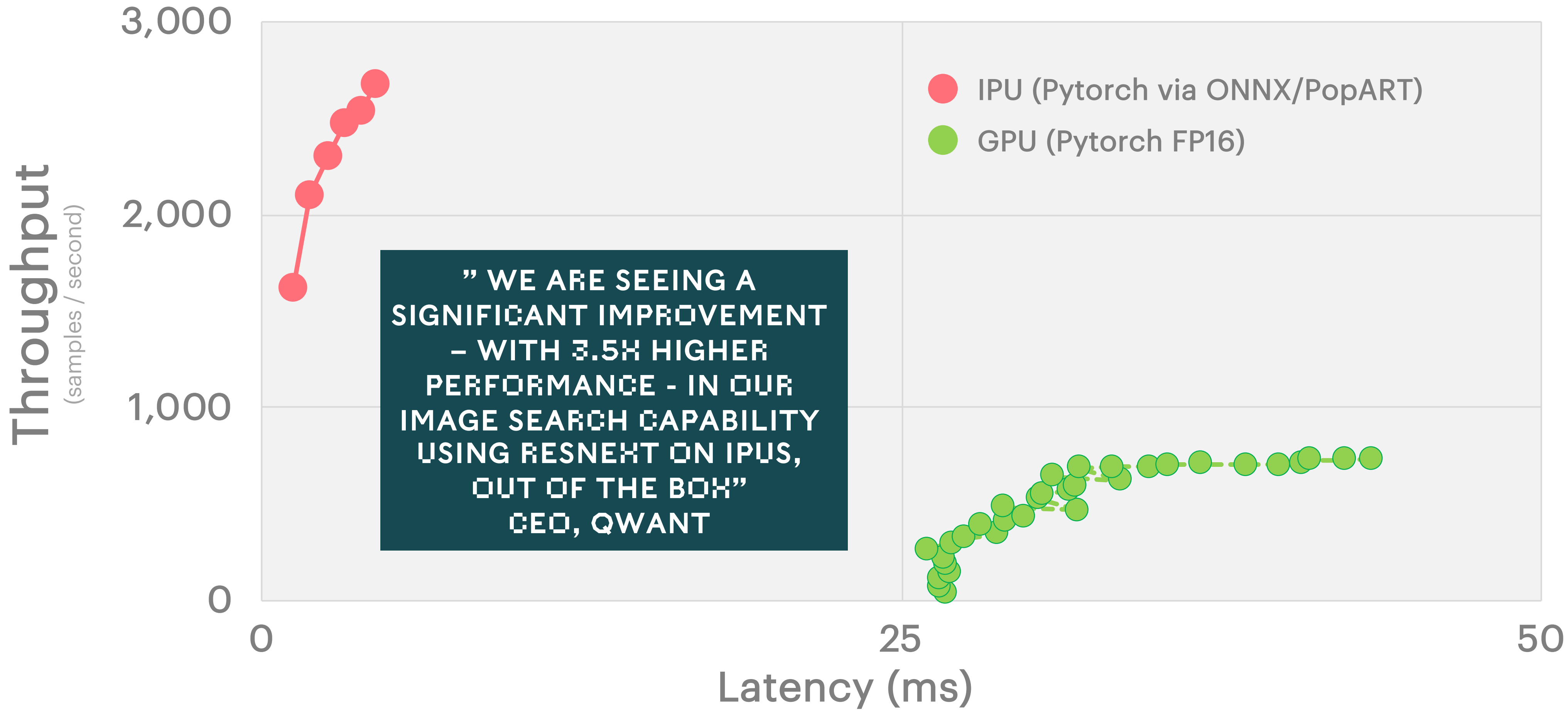
from 4x to >100x throughput for depthwise/grouped convolutions



NOTES:
Results averaged over 10,000 iterations. Filter/kernel size 3x3, field size 7x7, number of filters 512, batch size 32
Repeated for varying group dimensions from standard convolution (512) to full depth-wise (1)
Same code on IPU (Graphcore C2 – SDK 1.1.11) and GPU using TensorFlow (measured Apr'20) | Forward pass only | Both @ 300W TDP

RESNEXT-101 : INFERENCE

Lowest Latency Comparison: 6x higher throughput | 22x lower latency
Highest Throughput Comparison: 3.7x higher throughput | 10x lower latency

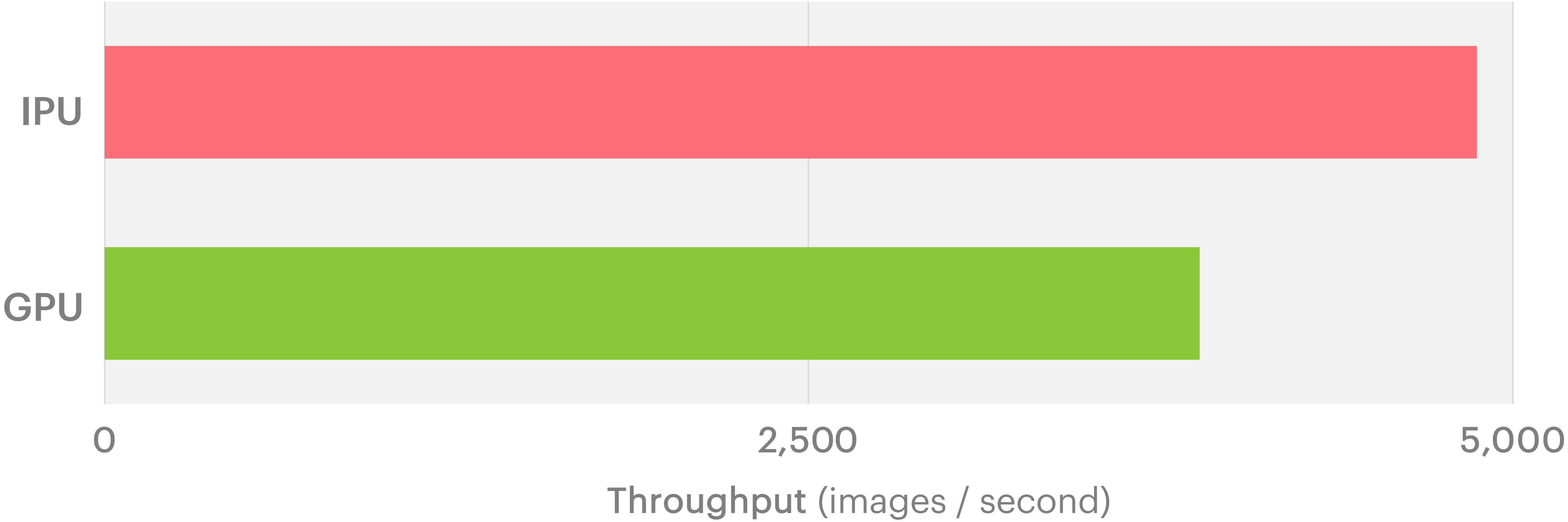


” WE ARE SEEING A SIGNIFICANT IMPROVEMENT – WITH 3.5X HIGHER PERFORMANCE - IN OUR IMAGE SEARCH CAPABILITY USING RESNEXT ON IPUS, OUT OF THE BOX”
CEO, QWANT

NOTES:
ResNext-101_32x4d | Real data (COCO) IPU: Graphcore C2 (SDK 1.1.11) using Pytorch (via ONNX/PopART) Batch Size 2-12 @ 300W TDP
GPU using Pytorch FP16 Batch Size 1-32 @ 300W TDP

RESNEXT-50 : TRAINING

1.25x higher throughput



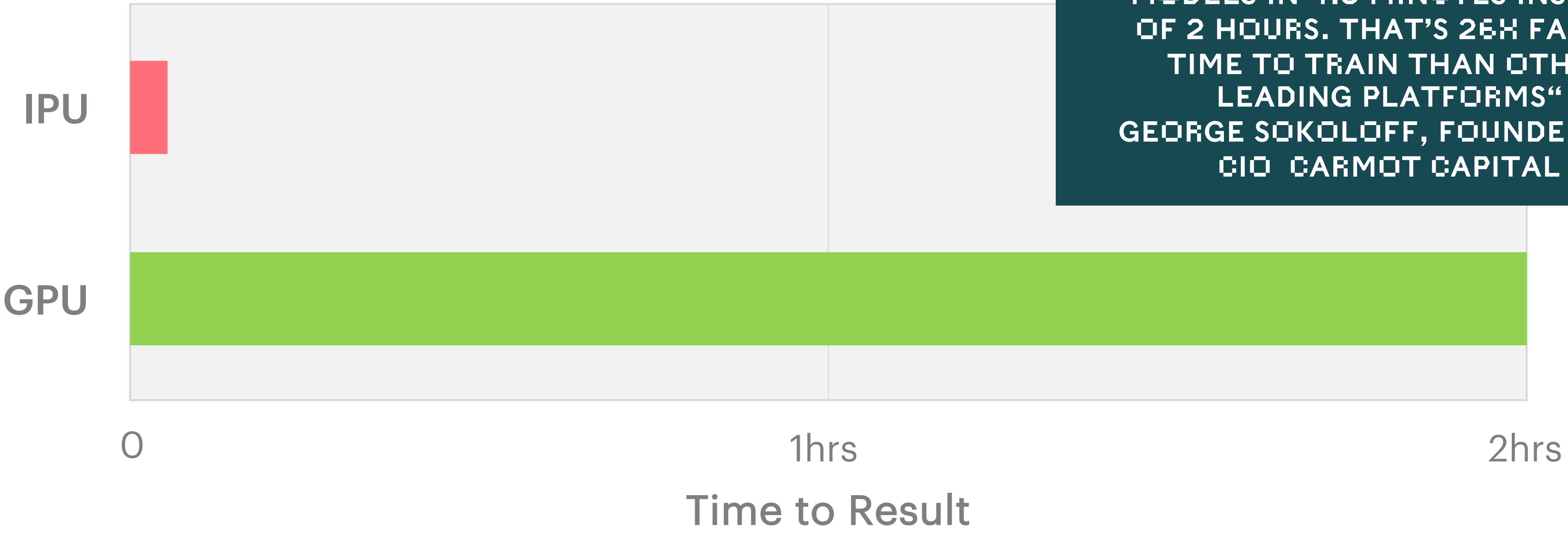
NOTES:
ResNext-50 | Real data (ImageNet dataset)
IPU: DSS8440 (SDK 1.1.11) using Tensorflow 16IPU pipelined. 4 shards, 4 replicas, Batch Size per processor 3
GPU: chassis comparison (power equivalent) using Pytorch (FP16) Batch Size 1024

IPU DELIVERS MASSIVE PERFORMANCE ADVANTAGE ON DIFFICULT MACHINE LEARNING PROBLEMS, e.g. FINANCE...

MCMC PROBABILISTIC MODEL : TRAINING

Customer implementation

26x faster time to result with 50% power



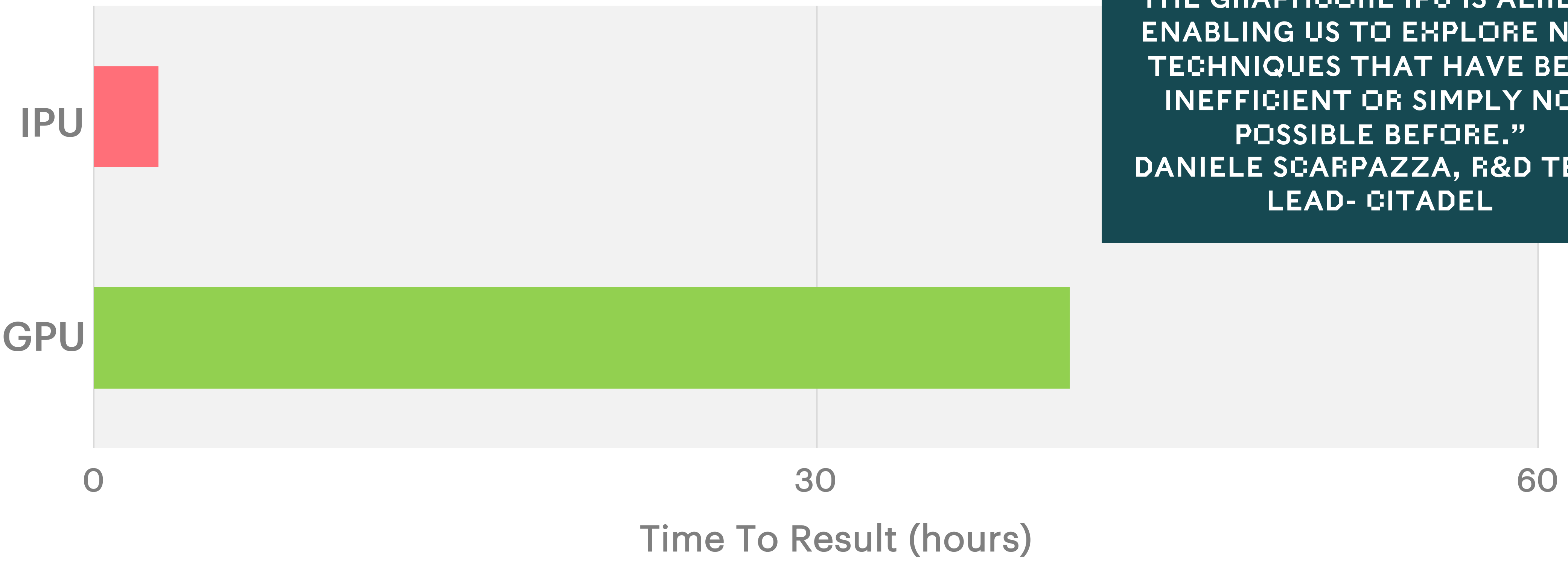
“WE WERE ABLE TO TRAIN ONE OF OUR PROPRIETARY PROBABILISTIC MODELS IN 4.5 MINUTES INSTEAD OF 2 HOURS. THAT’S 26X FASTER TIME TO TRAIN THAN OTHER LEADING PLATFORMS“
GEORGE SOKOLOFF, FOUNDER AND CIO CARMOT CAPITAL

NOTES:
Graphcore customer Markov Chain Monte Carlo Probability model (summary data shared with customer's permission)
IPU: Graphcore GC2 @ 150W TDP
GPU @ 300W TDP

MCMC PROBABILISTIC MODEL : TRAINING

TensorFlow Probability model - representative finance workload for alpha estimation

15.2x faster time to train



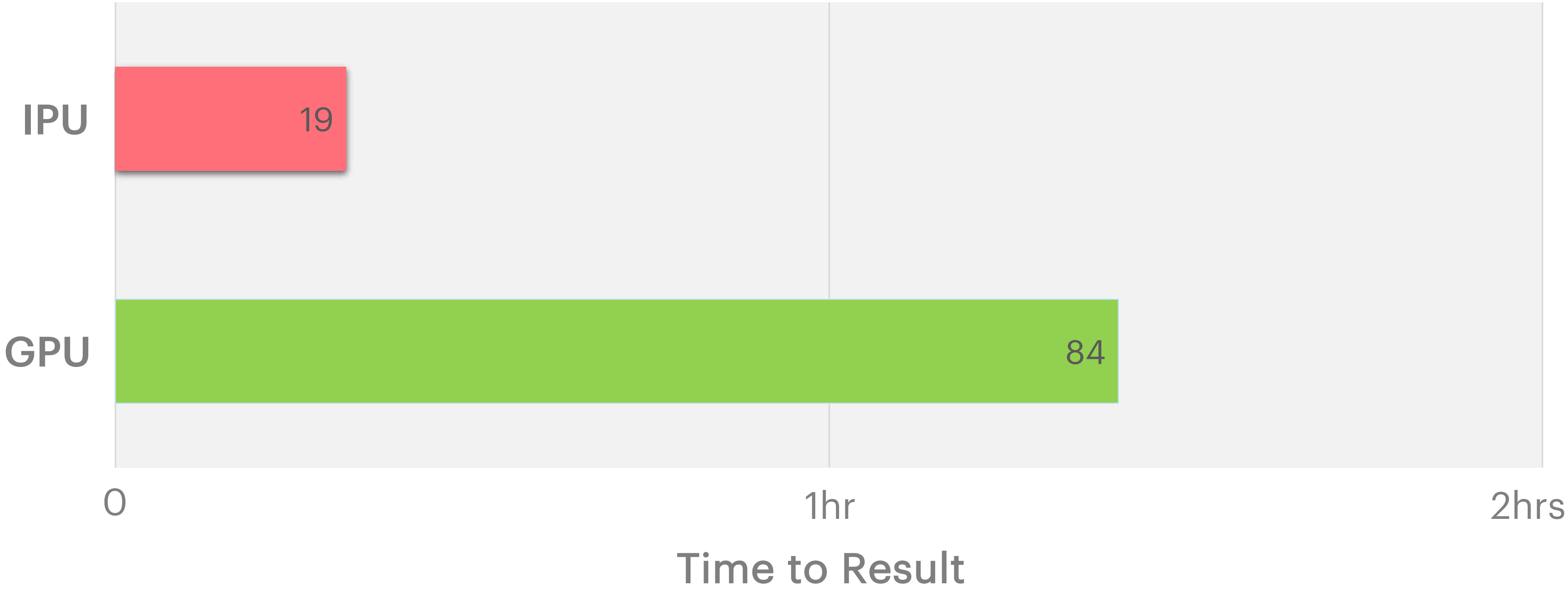
> “THE GRAPHCORE IPU IS ALREADY ENABLING US TO EXPLORE NEW TECHNIQUES THAT HAVE BEEN INEFFICIENT OR SIMPLY NOT POSSIBLE BEFORE.”
> DANIELE SCARPAZZA, R&D TEAM LEAD- CITADEL

NOTES:
Markov Chain Monte Carlo – Probabilistic model with TensorFlow Probability, representative of workload used by Carmot Capital
Neural network with 3 fully-connected layers (num units in 1st layer=40, #dimensions in training set =22, #leapfrog steps=1000, calcs in sliding window=200)
IPU: C2 card (300W TDP) results (SDK 1.1.11) – 800 samples
GPU (300W TDP) - 800 samples

VAE PROBABILISTIC MODEL : TRAINING

TensorFlow Variational Autoencoder model – MCMC & VI combination

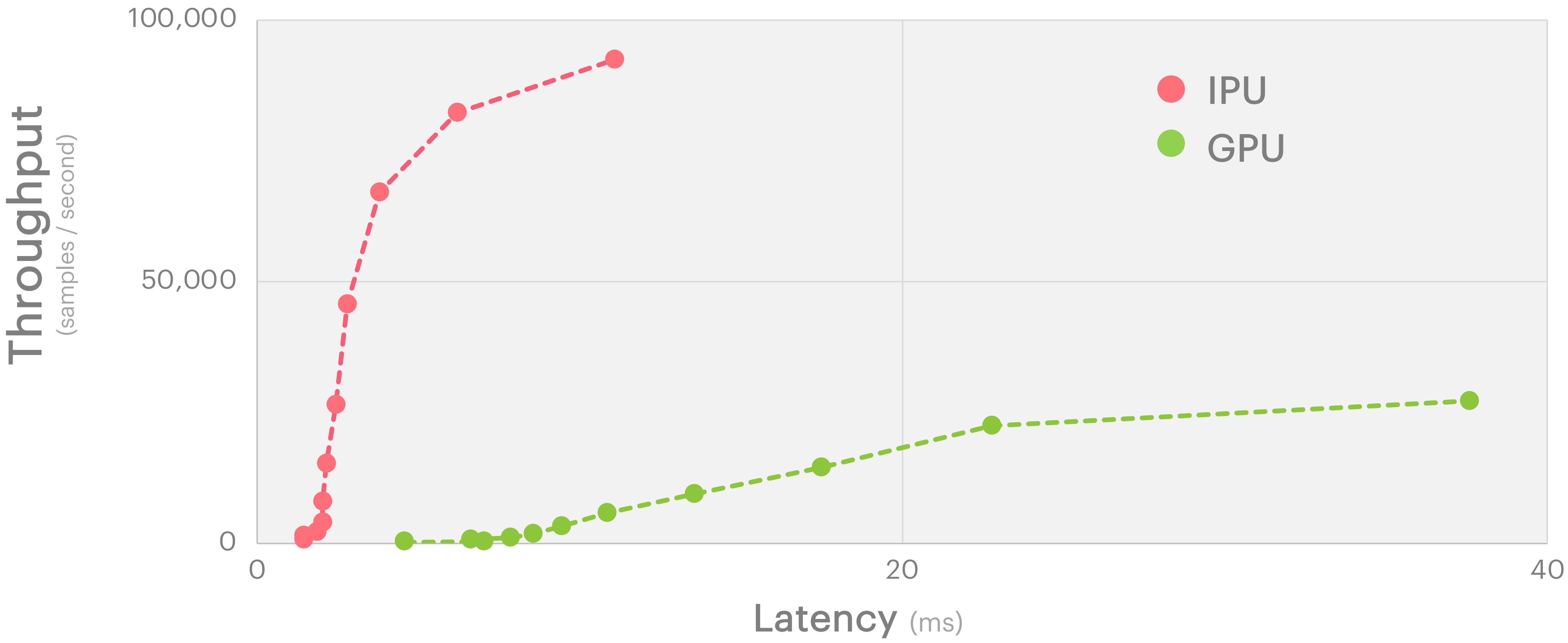
4.3x faster time to train with 50% power



NOTES:
Contrastive Divergence VAE - example variational autoencoder model using MCMC & Variational Inference, based on ICML paper (https://arxiv.org/pdf/1905.04062.pdf).
Scalar control variate used in place of a vector.
IPU: Running on single Graphcore GC2 @ 150W TDP (SDK 1.1.11) | GPU @ 300W TDP
Both using TensorFlow, real data MNIST, Batch size 100 (as in ICML paper).

LSTM : INFERENCE

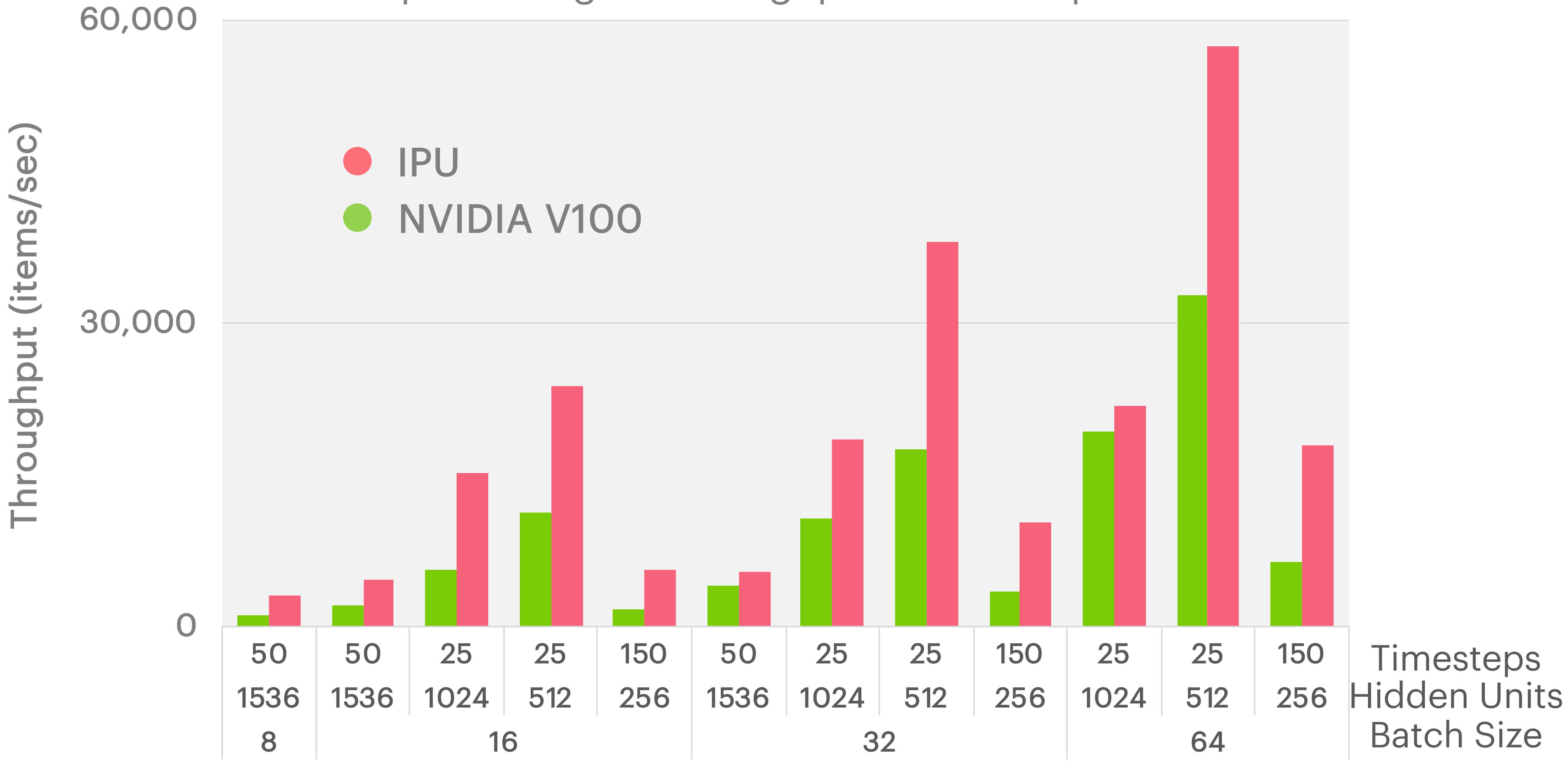
>300 higher throughput at lower latency



NOTES:
2 LSTM layers, each with 256 units, 200 time steps, 16 input dimensions, real data, mixed precision
IPU: Graphcore C2 using TensorFlow and PopNN (SDK 1.1.11) @ 300W TDP (Batch Sizes upto 1024)
GPU: using TensorFlow with optimizations @ 300W TDP (Batch Sizes upto 1024)

LSTM : TRAINING

up to 3x higher throughput with 50% power

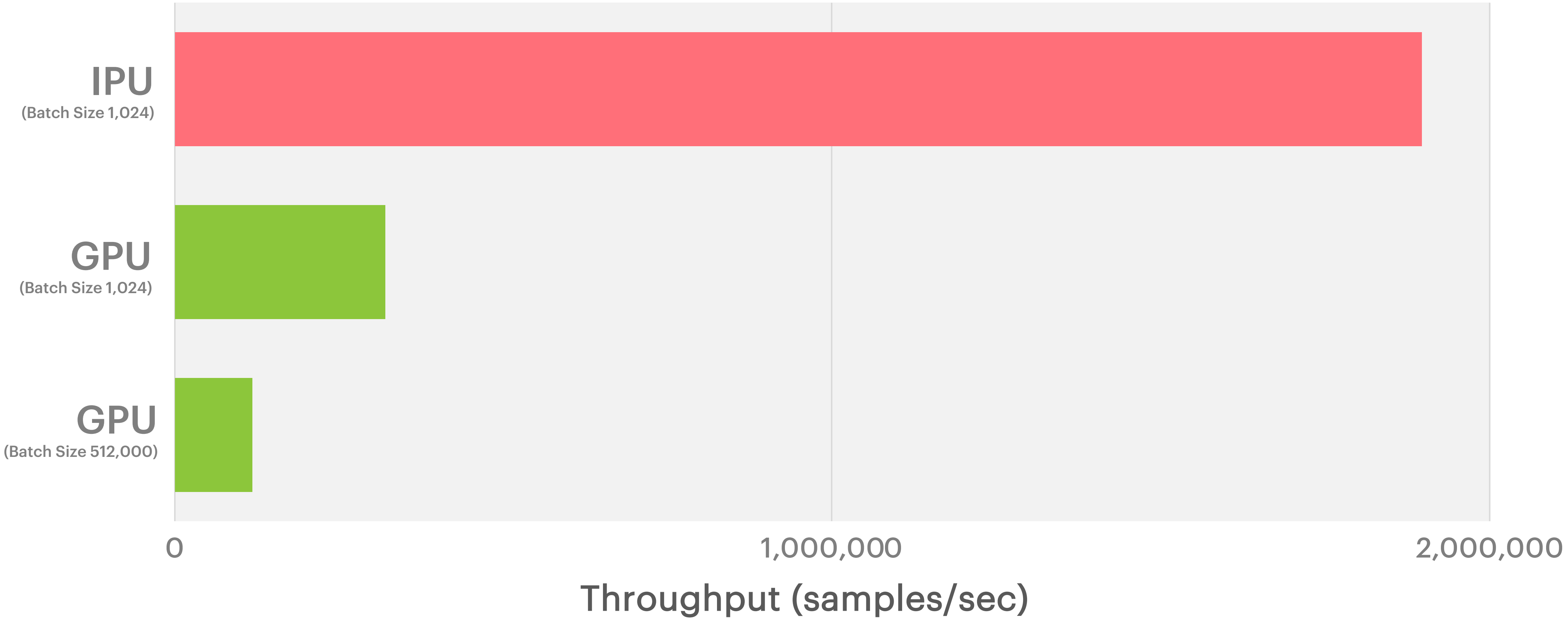


NOTES:
LSTM Single Layer benchmark vs V100 Deepbench resuts
IPU: Graphcore IPU using TensorFlow @ 150W TDP (SDK 1.1.11)
GPU: V100 DeepBench results @ 300W TDP (https://github.com/baidu-research/DeepBench/blob/master/results/train/DeepBench_NV_V100.xlsx)

TIME SERIES ANALYSIS : TRAINING

SALES FORECASTING MODEL | Multi-Layer Perceptron (MLP) + Embedding

>5.9x higher throughput (faster time to train)

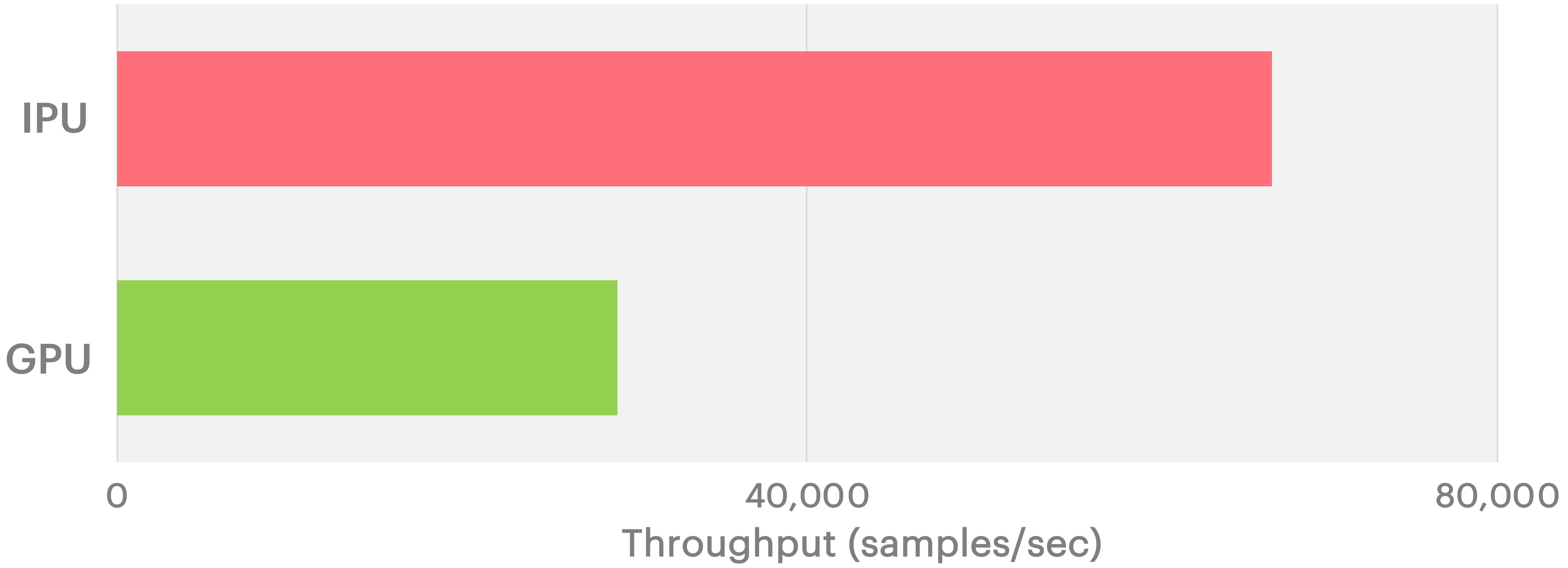


NOTES:
Multi-Layer Perceptron (MLP) + Embeddings model for forecasting, Real data
IPU: Graphcore C2 (SDK 1.1.11) using TensorFlow @ 300W TDP
GPU using TensorFlow @ 300W TDP

DENSE AUTOENCODER : TRAINING

for content recommendation and ranking

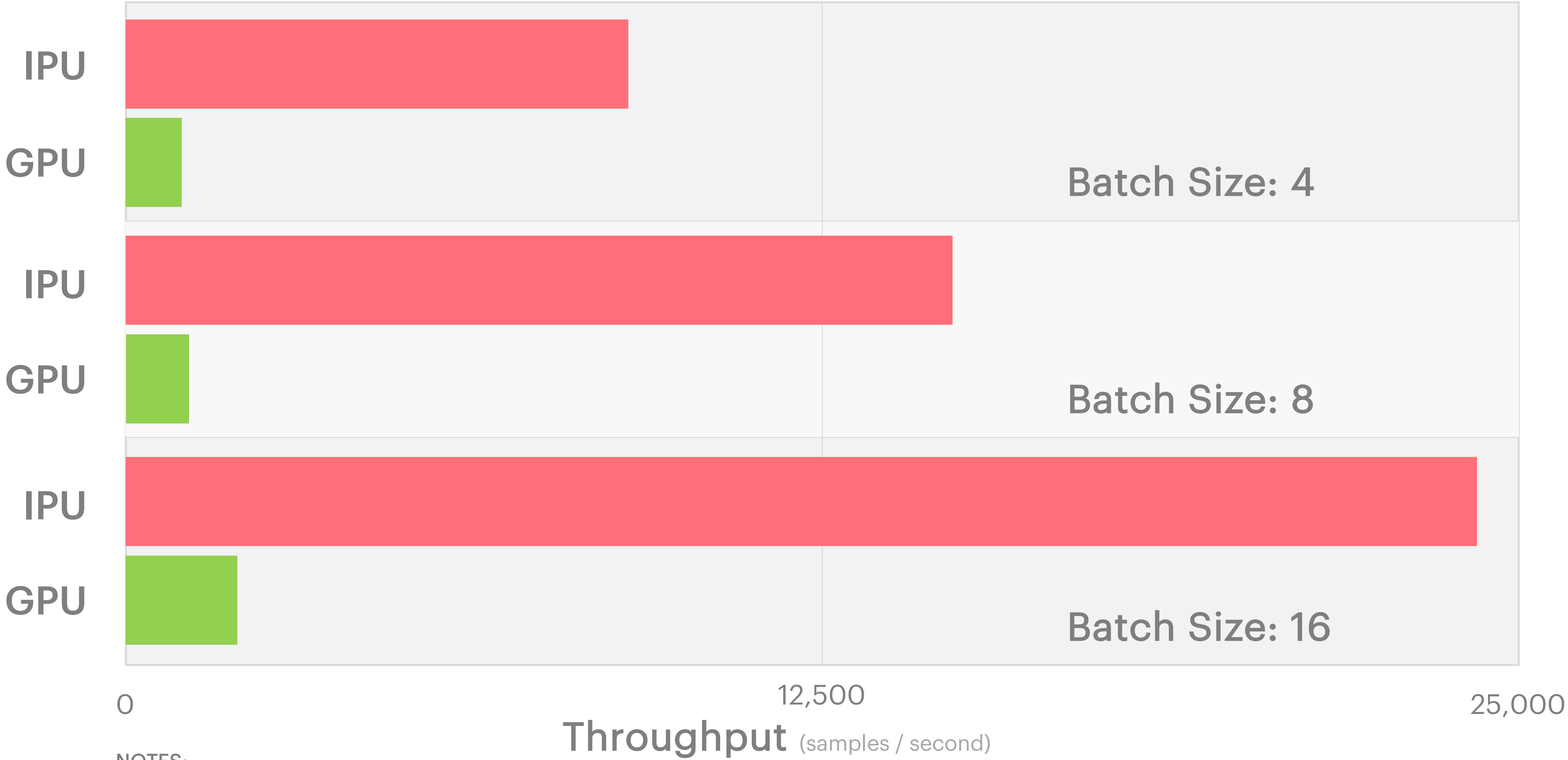
2.3x higher throughput (faster time to train)



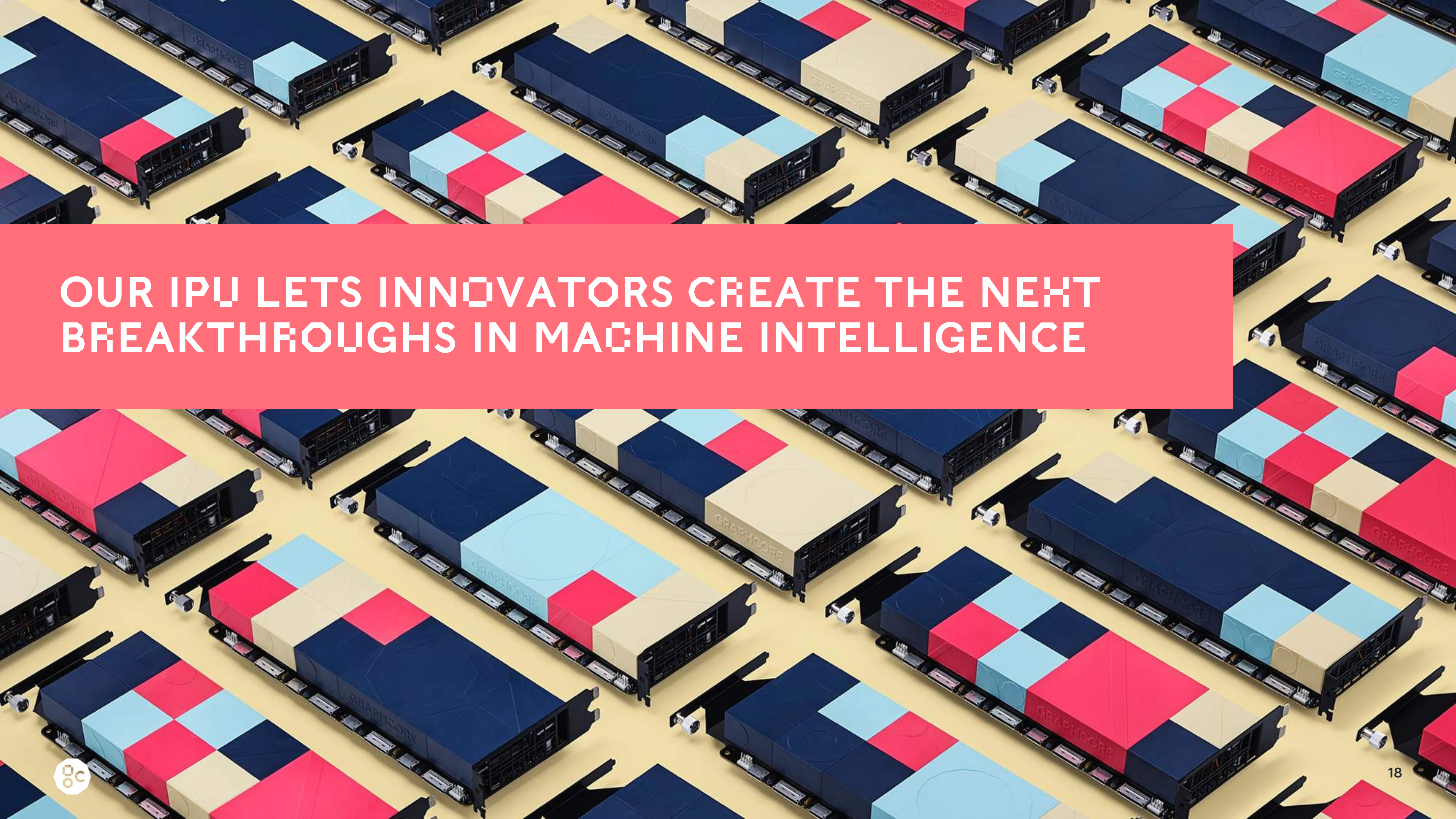
NOTES:
Deep autoencoder with 6 fully-connected layers and constrained decoder, BS 64 | Content recommendation | Training using open source Netflix 3m data-samples
IPU: Graphcore C2 (SDK 1.1.11) @ 300W TDP
GPU @ 300W TDP

REINFORCEMENT LEARNING POLICY: TRAINING

up to 13x higher throughput (faster time to train)



NOTES:
Reinforcement policy model training | representative of large-scale reinforcement learning systems using LSTM
IPU: Graphcore C2 using TensorFlow @ 300W TDP
GPU: using TensorFlow @ 300W TDP



OUR IPU LETS INNOVATORS CREATE THE NEXT BREAKTHROUGHS IN MACHINE INTELLIGENCE