GRAPHCORE

HIGH-PERFORMANCE AI FOR FINANCIAL MODELLING

- Founded in 2016
- Technology: Intelligence Processor Unit (IPU)
- Team is over 400 globally
- Offices in UK, US, China, Norway
- Raised over $450M

SEQUOIA ATOMICO BMW i VENTURES DELL BOSCH SAMSUNG Merian GLOBAL INVESTORS
SOFINA Microsoft Amadeus Capital Partners pitango VENTURE CAPITAL draperesprit Foundation CAPITAL

“IPU’s novel architecture is already enabling us to explore new techniques that have been inefficient or simply not possible before.”

CITADEL

“We were able to train one of our proprietary probabilistic models In 4.5 minutes instead of 2 hours. That’s 26x faster. ”

carmot CAPITAL

IN FINANCE

MODELS ARE OFTEN OVERSIMPLIFIED TO ACHIEVE COMPETITIVE LATENCY OR FIGHT NOISE

IPU IS ENABLING ADVANCED MODELS TO RUN
UP TO 300x FASTER

DEEPER INSIGHTS AND BETTER DECISIONS
IN SHORTER TIME

IPU DELIVERS UNIQUE BENEFITS FOR FINANCIAL MODELING

Inference

- Make decisions with lower latency
- Use models of greater complexity
- Run backtesting faster

Training

- Iterate faster when experimenting with model architectures
- Use probabilistic models to avoid overfitting and account for noise
- Re-train continuously–“nowcasting”

FASTER AND SMARTER FUTURE

Inference

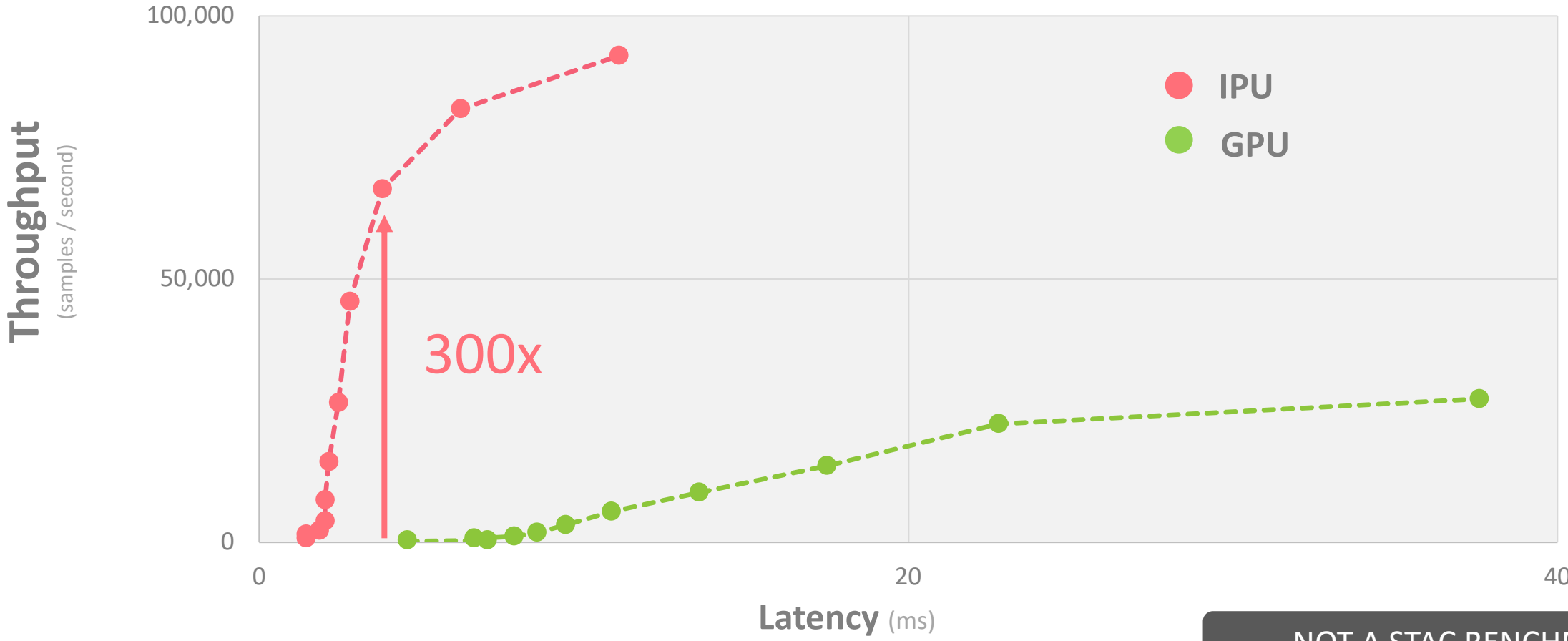
- **300x higher throughput** for LSTM at the same latency
- Efficient for **non-vectorizable models**

Training

- **26x faster sampling** with MCMC for alpha estimation and option pricing
- Up to **13x shorter time to train** for reinforcement learning

LSTM INFERENCE FOR TIME SERIES ANALYSIS

>300 higher throughput at lower latency
for feature generation and alpha estimation



NOT A STAC BENCHMARK

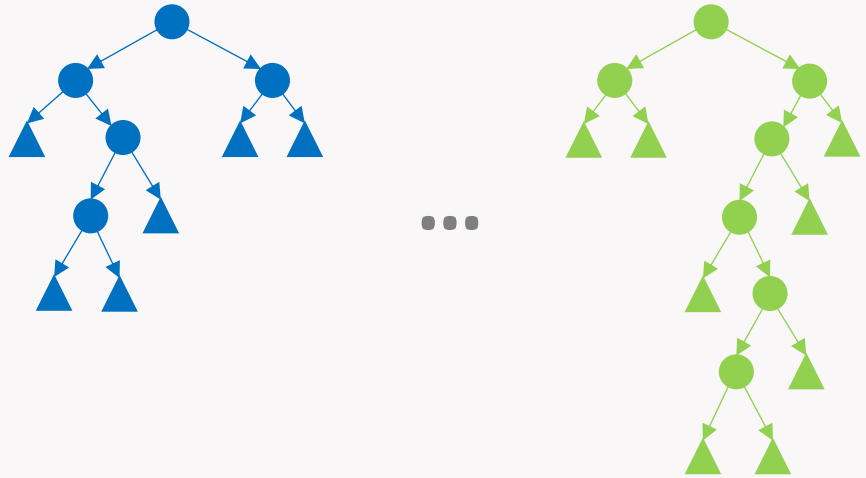
NOTES:
2 LSTM layers, each with 256 units, 200 time steps, 16 input dimensions, real data, mixed precision
IPU: Graphcore C2 using TensorFlow and PopNN (SDK 1.1.11) @ 300W TDP (Batch Sizes upto 1024)
GPU: using TensorFlow with optimizations @ 300W TDP (Batch Sizes upto 1024)

EFFICIENT FOR NON-VECTORIZED MODELS

Benefit from thread independency

Heterogeneous Experts

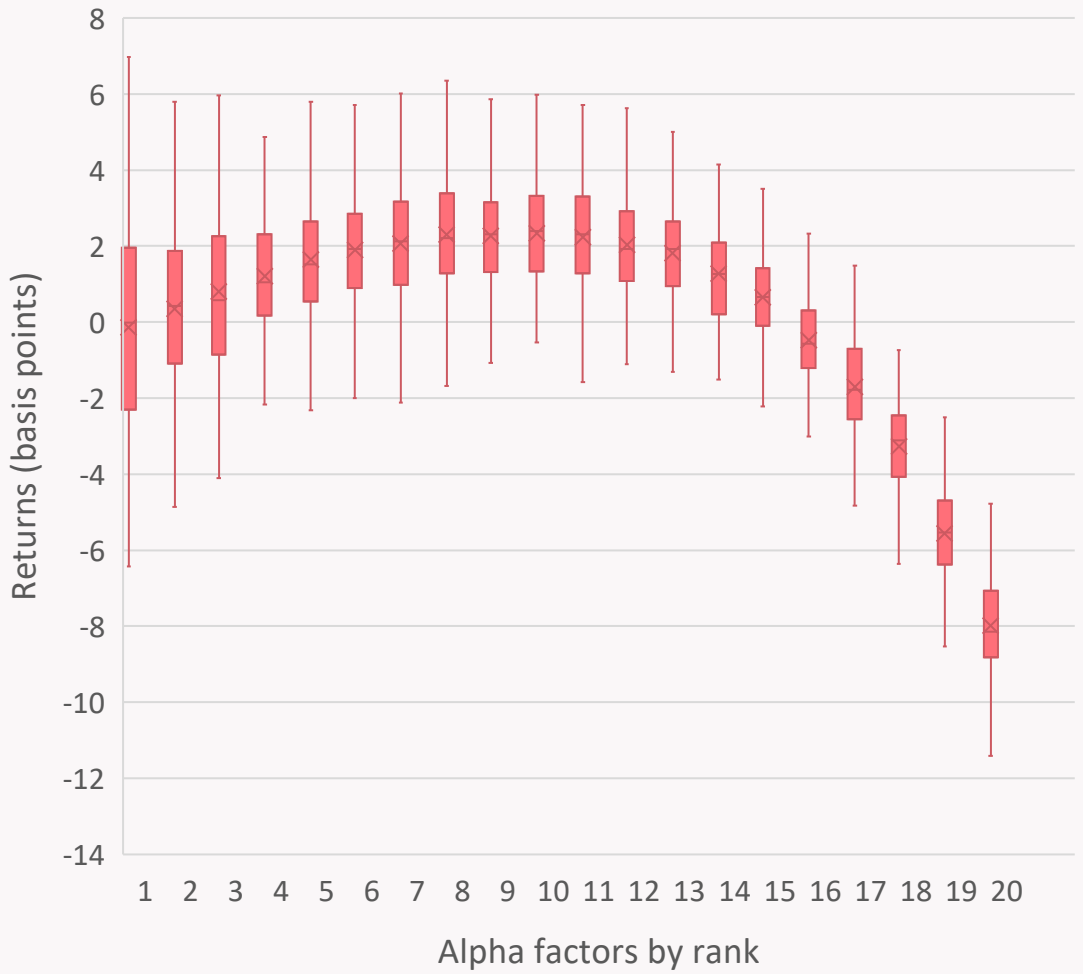
Sparse Random Forests



FASTER TIME-TO-INSIGHT

Feature generation and prediction done in one step on IPU

ALPHA ESTIMATION WITH MCMC



Markov Chain Monte Carlo (MCMC):

predictions are distributions,

not just point values

This helps to:

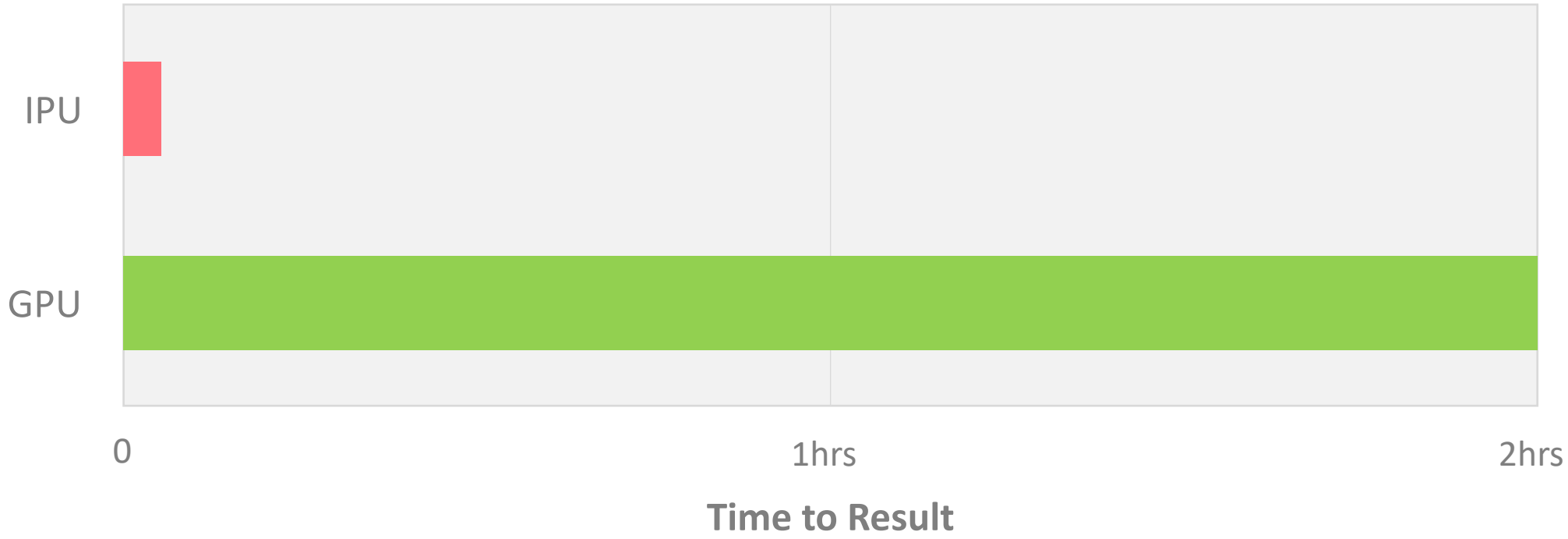
- Avoid overfitting
- Deal with noisy data
- Handle non-linearities

MCMC has long been considered too computationally expensive

MCMC PROBABILISTIC MODEL : TRAINING

Customer implementation

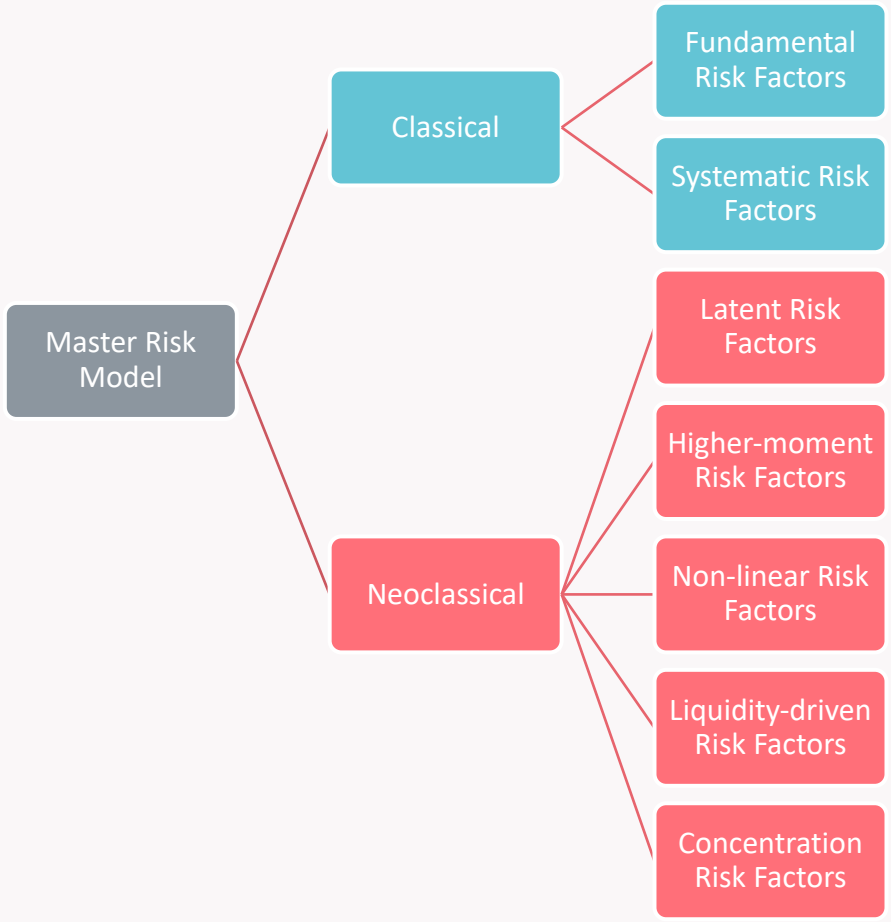
26x higher throughput



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

NOT A STAC BENCHMARK

RISK ESTIMATION AND PORTFOLIO GENERATION



IPU improves performance of complex models to:

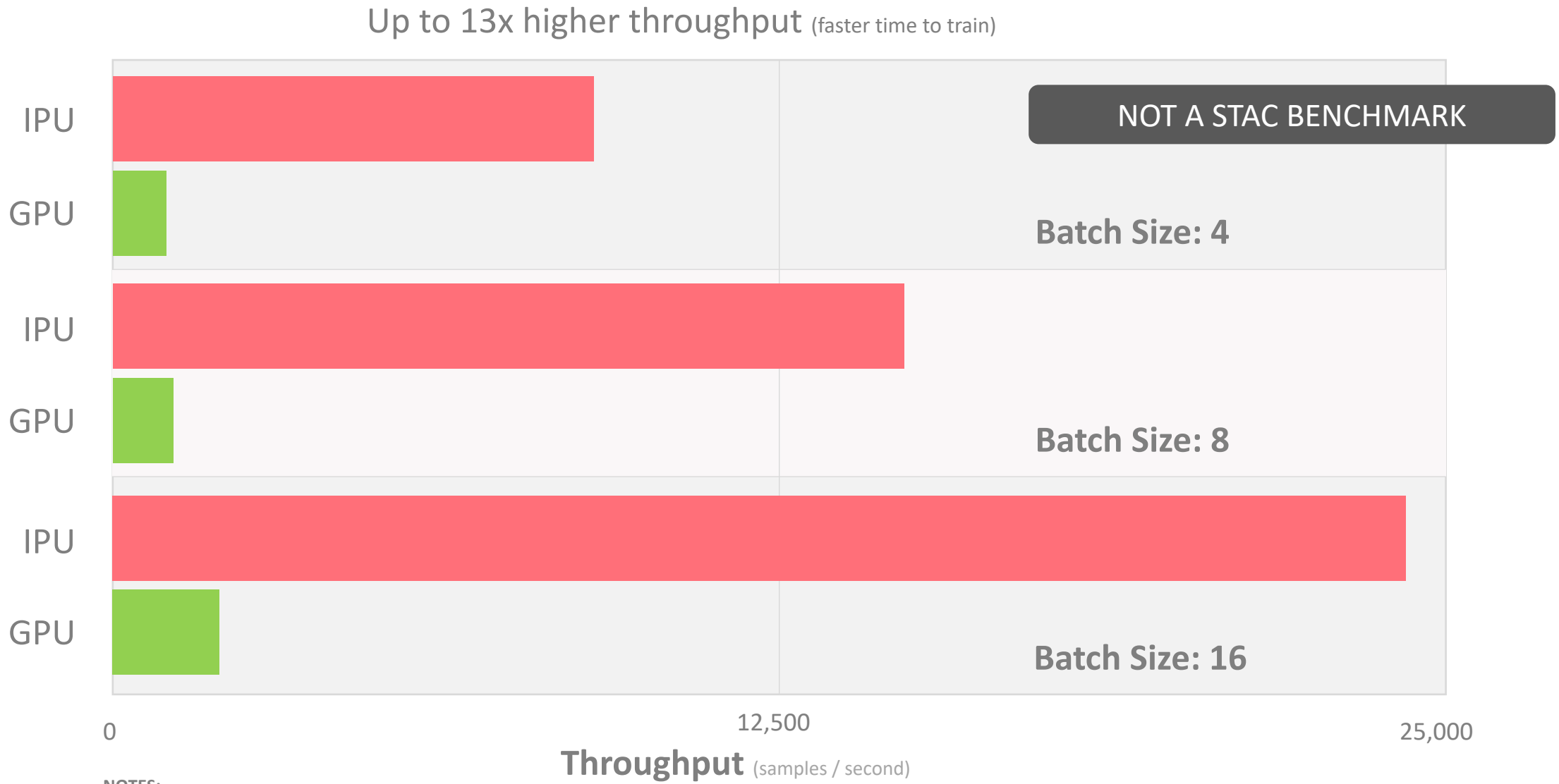
- Estimate full risk profile of a risk factor set
- Find statistical and latent risk factors

4.3x faster training of Variational Autoencoder:

- Identify latent risk factors
- Capture non-linear relationships between the channels more efficiently than SVD

NOT A STAC BENCHMARK

PORTFOLIO REBALANCING WITH REINFORCEMENT LEARNING

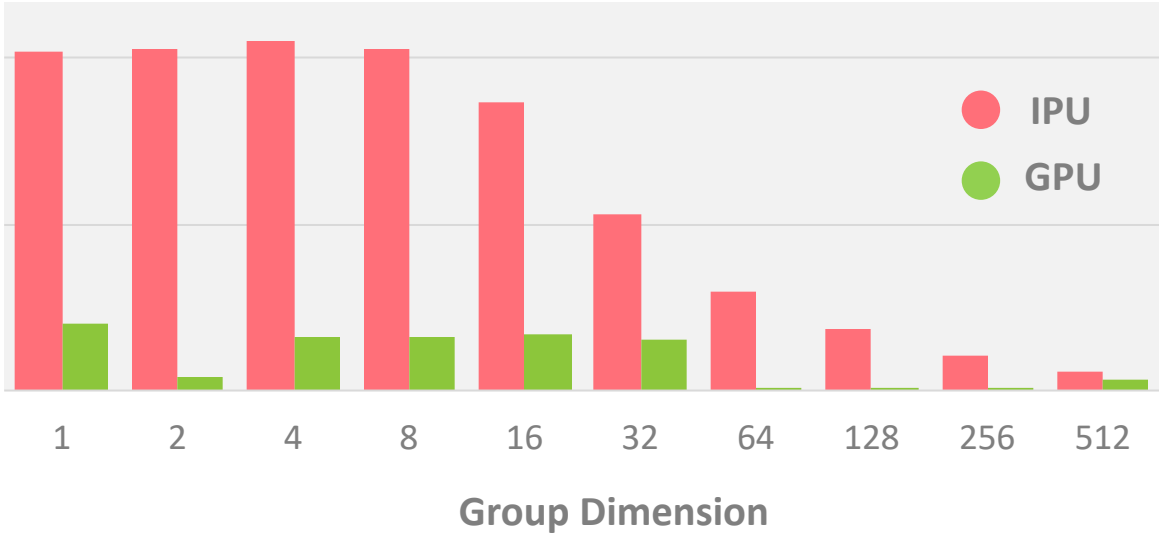


NOTES:
Reinforcement policy model training | representative of large-scale reinforcement learning systems using LSTM
IPU: Graphcore C2 using TensorFlow @ 300W TDP
GPU: using TensorFlow @ 300W TDP

EFFICIENT FOR ALTERNATIVE DATA

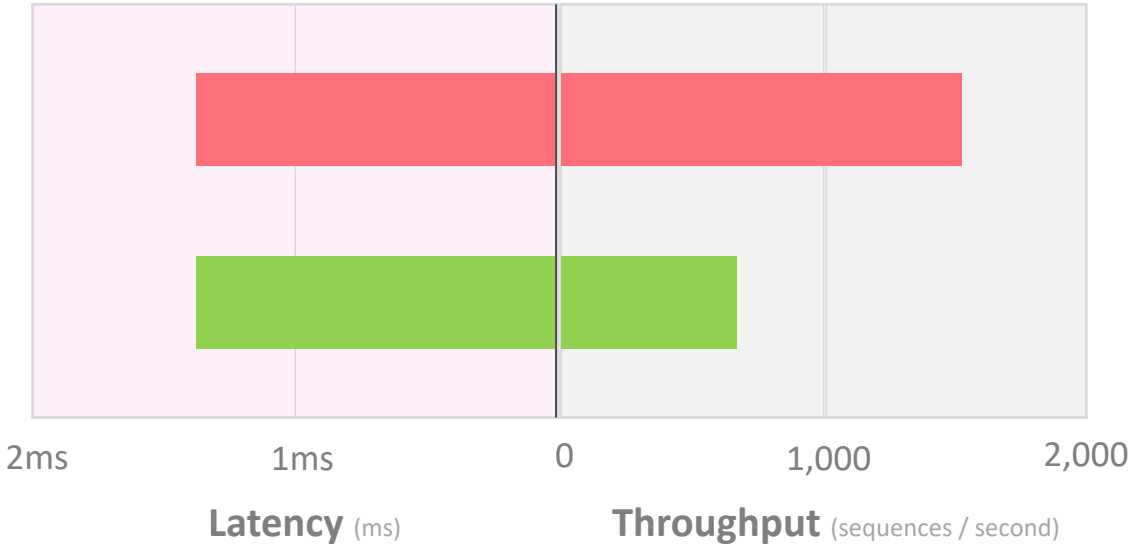
COMPUTER VISION

Up to 100x throughput for grouped convolutions



NLP

2x throughput at lowest latency for BERT-base



NOT STAC BENCHMARKS


OUR IPU LETS INNOVATORS CREATE THE NEXT
BREAKTHROUGHS IN MACHINE INTELLIGENCE
THANK YOU
Alexander Tsyplikhin
alext@graphcore.ai
13