



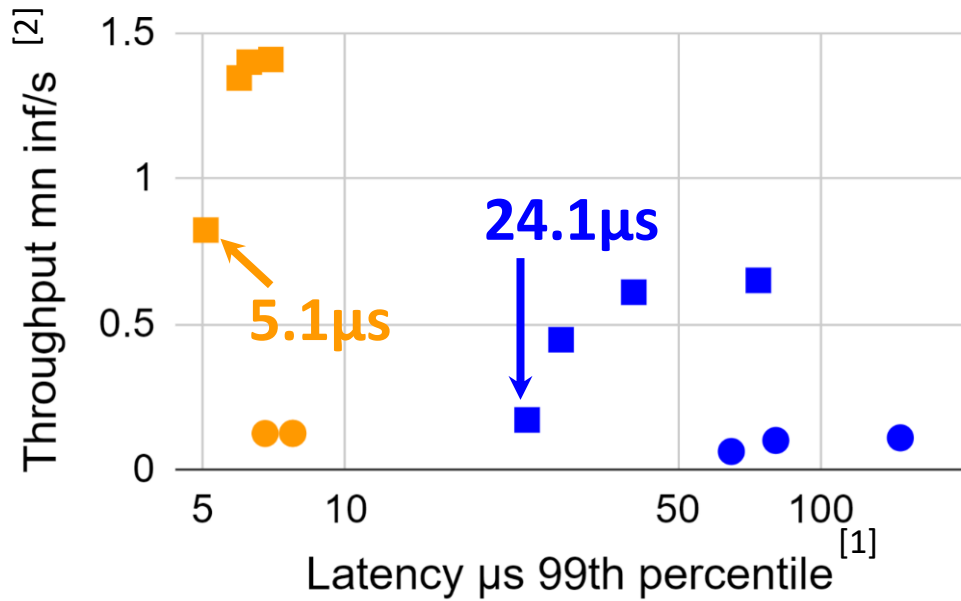
myrtle.ai

Fast Compute: How fast is fast?

© 2024 Myrtle Software Ltd. All rights reserved

STAC Summits, Fall 2024

Myrtle.ai



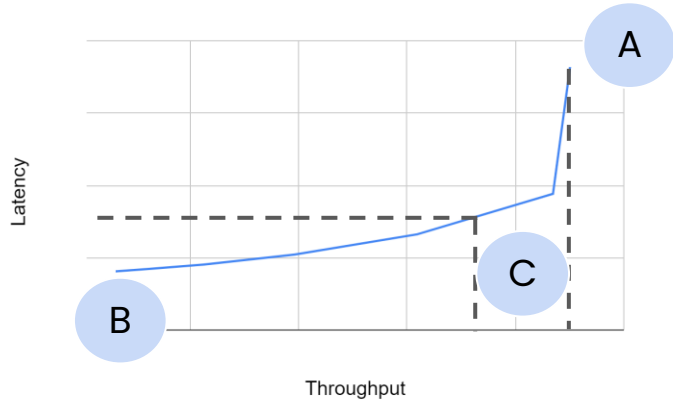
STAC-ML Inference

- Sumaco LSTM A
- Sumaco LSTM B
- Tacana LSTM A
- Tacana LSTM B

VOLLO

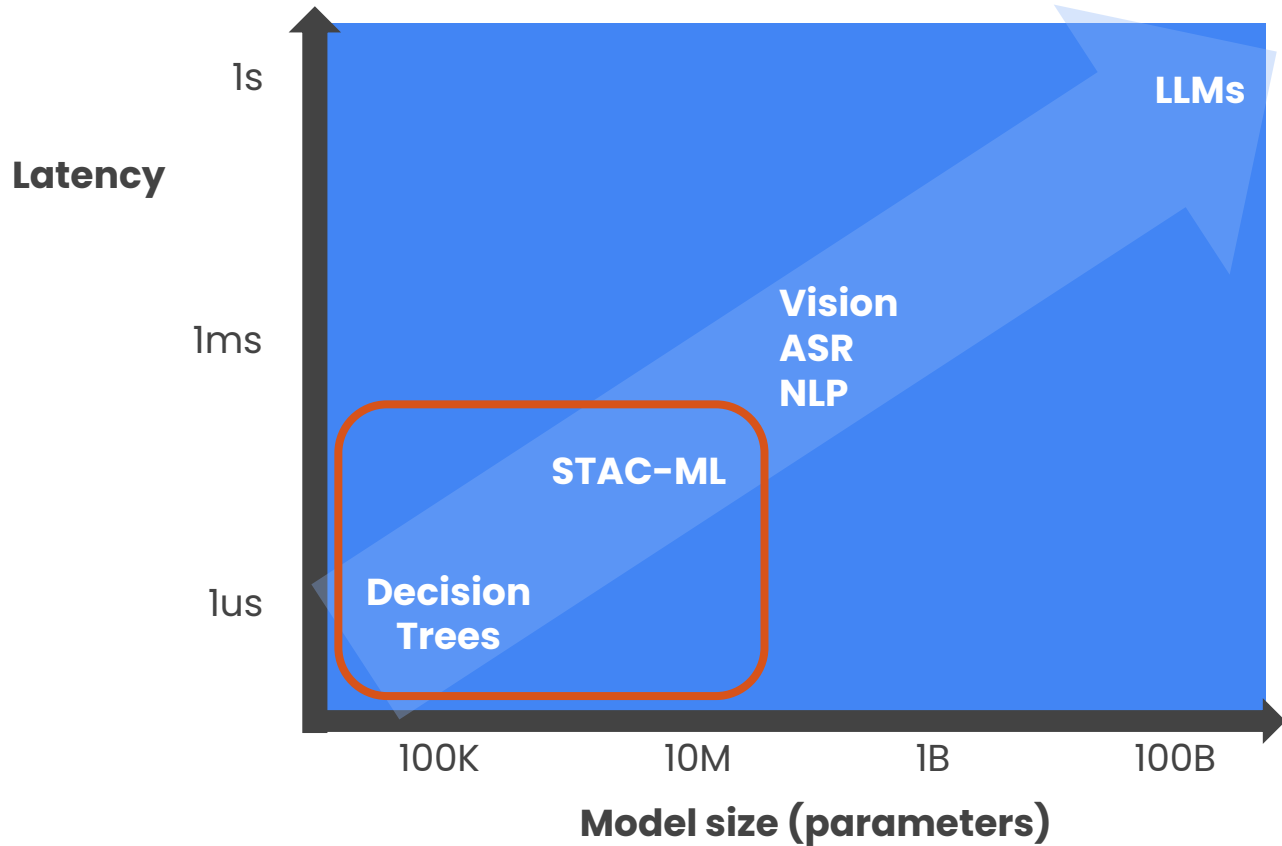
What do we mean by fast?

- **A** Throughput – Inferences per second
 - Can be increased with horizontal scaling
 - Relevant for ML model training, backtesting, etc.
- **B** Latency – Time to respond after receiving a request
 - Needs system to be designed around low latency
 - Small differences can be critical in latency-sensitive workloads
- **C** Latency-bounded throughput



Today we'll focus on **Latency**

Latency -vs- Model Size



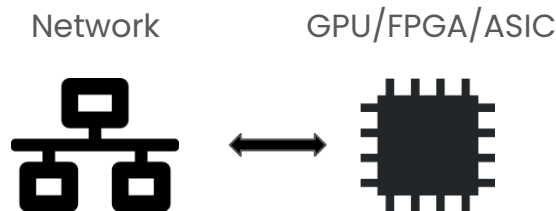
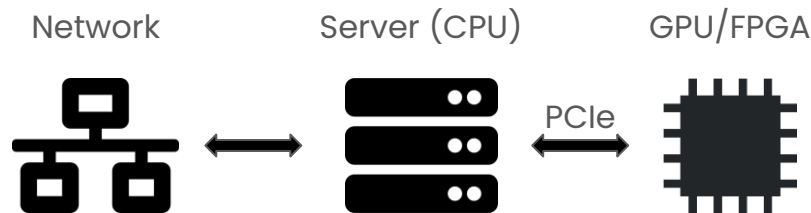
Inference platforms

Custom chips:

- **ASSP**
 - For ML, mostly focused on throughput (not latency)
- **ASIC**
 - Inflexible & long TTM
 - High NRE & risky

General purpose chips:

- **CPU**
- **GPU**
- **FPGA**

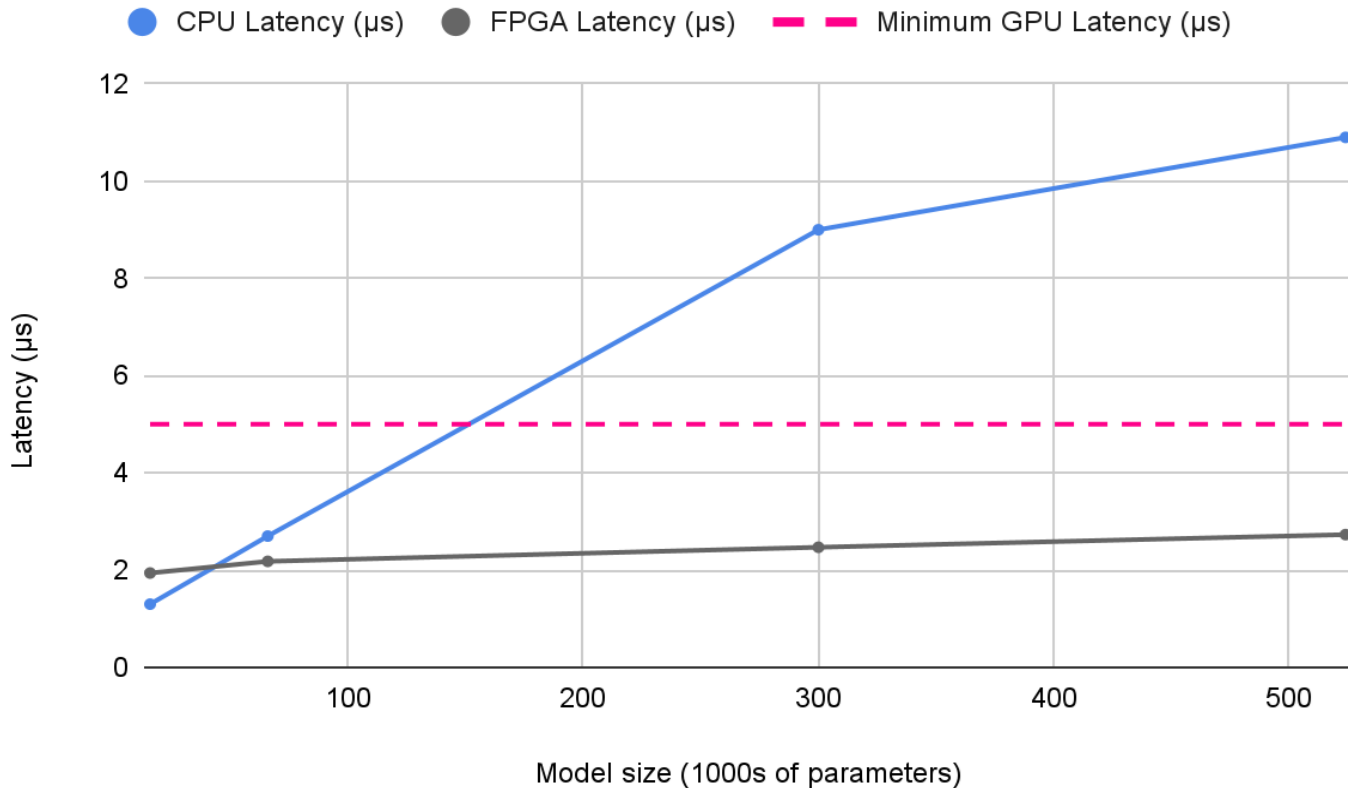


What affects latency in AI?

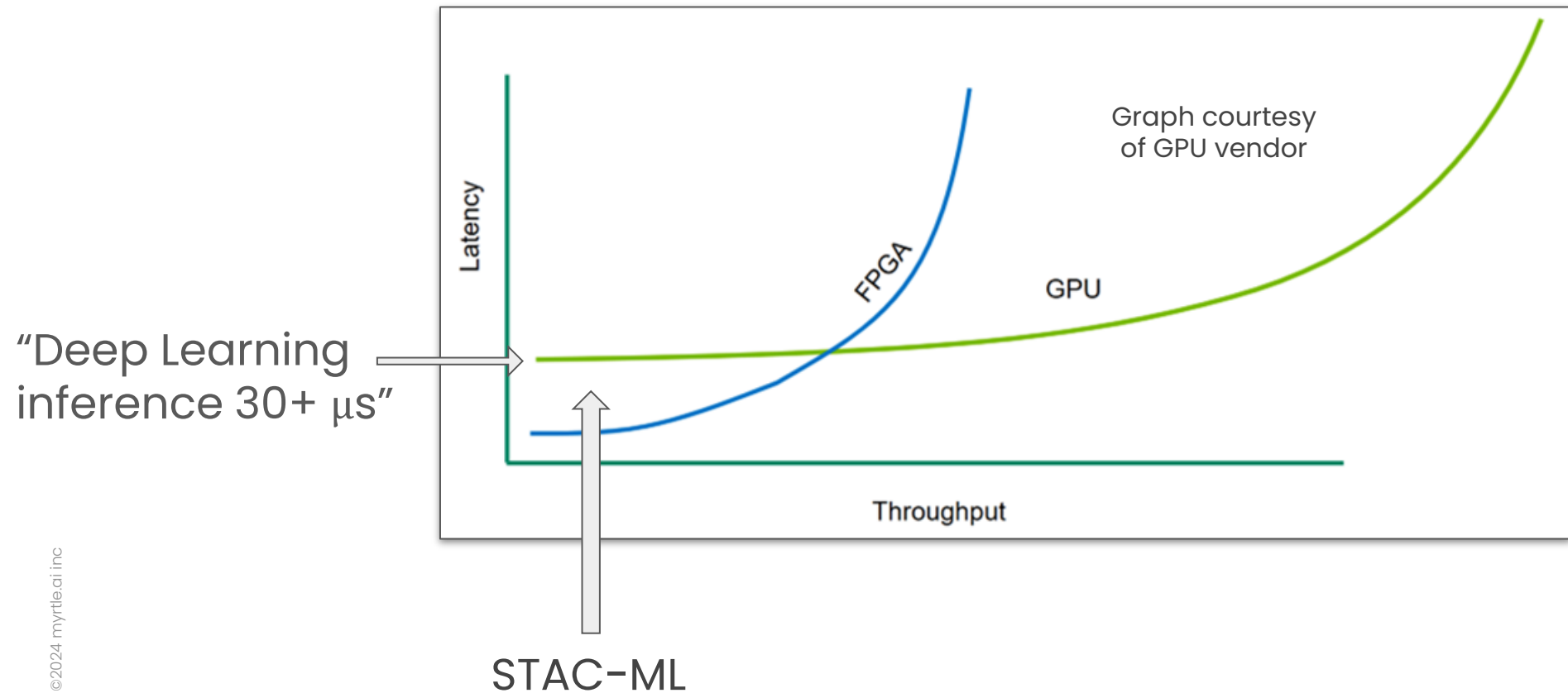
- **IO**
 - Network \leftrightarrow CPU (if using one) $0.5\mu\text{s}$
 - CPU \leftrightarrow accelerator (if using one) $0.5\mu\text{s}$
- **Memory**
 - Loading model parameters for compute $1-10\mu\text{s}+$
- **Compute**
 - Startup $5-10\mu\text{s}$ on GPU, $<1\mu\text{s}$ on CPU/FPGA
 - The actual work!

Overheads are substantial in low latency ML inference

Latency overheads critical for small models



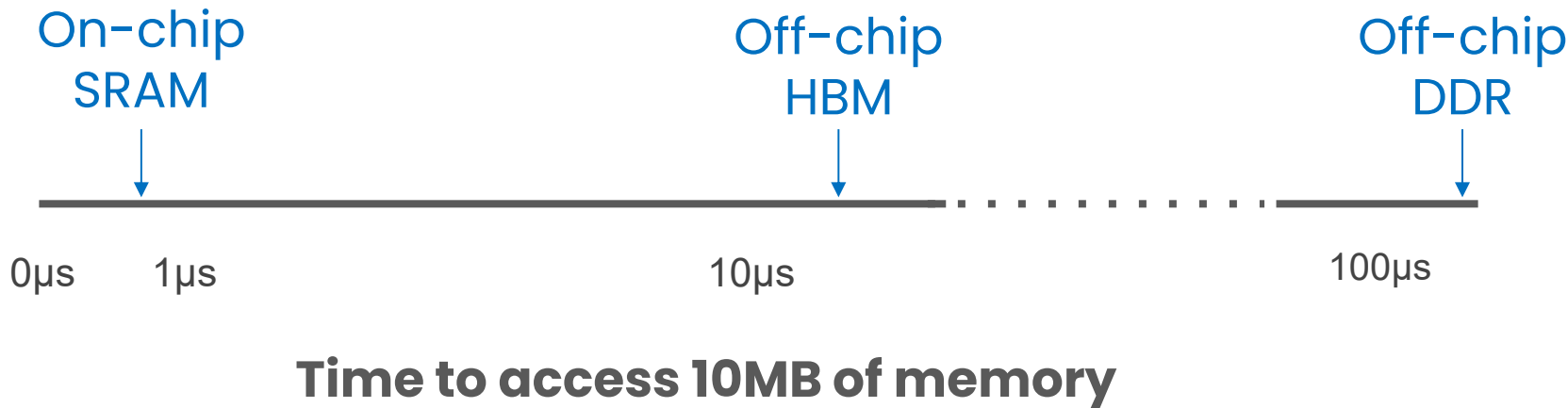
Latency/throughput trade-off for larger models





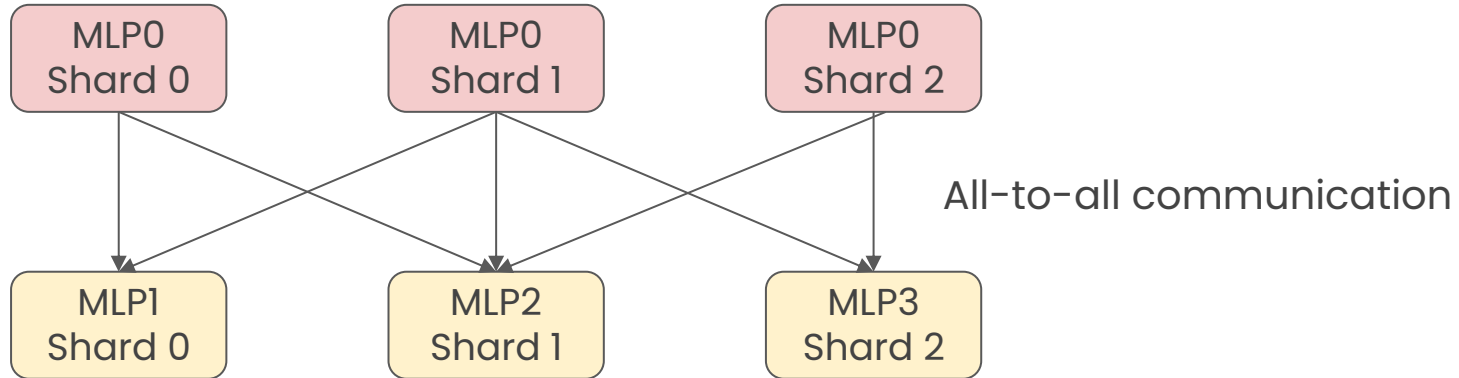
Memory efficiency is key for latency optimization

- Limited on-chip, fast-access memory for weights
- Even small ML models often have MBs of weights
- Going off-chip to fetch weights can blow latency budget, limiting model size
- Techniques to reduce model size such as quantisation and sparsity can be hugely beneficial



Compute latency depends on data flow

- Low latency requires parallelisation at low batch sizes
- Vector, SIMD, and Matrix multiplication instructions all good for parallelisation within a compute unit
- Sharding across compute units gives further parallelisation
- Requires all-to-all communication, a data flow problem
- **Synchronization can cost latency on CPU or GPU, particularly for small models**

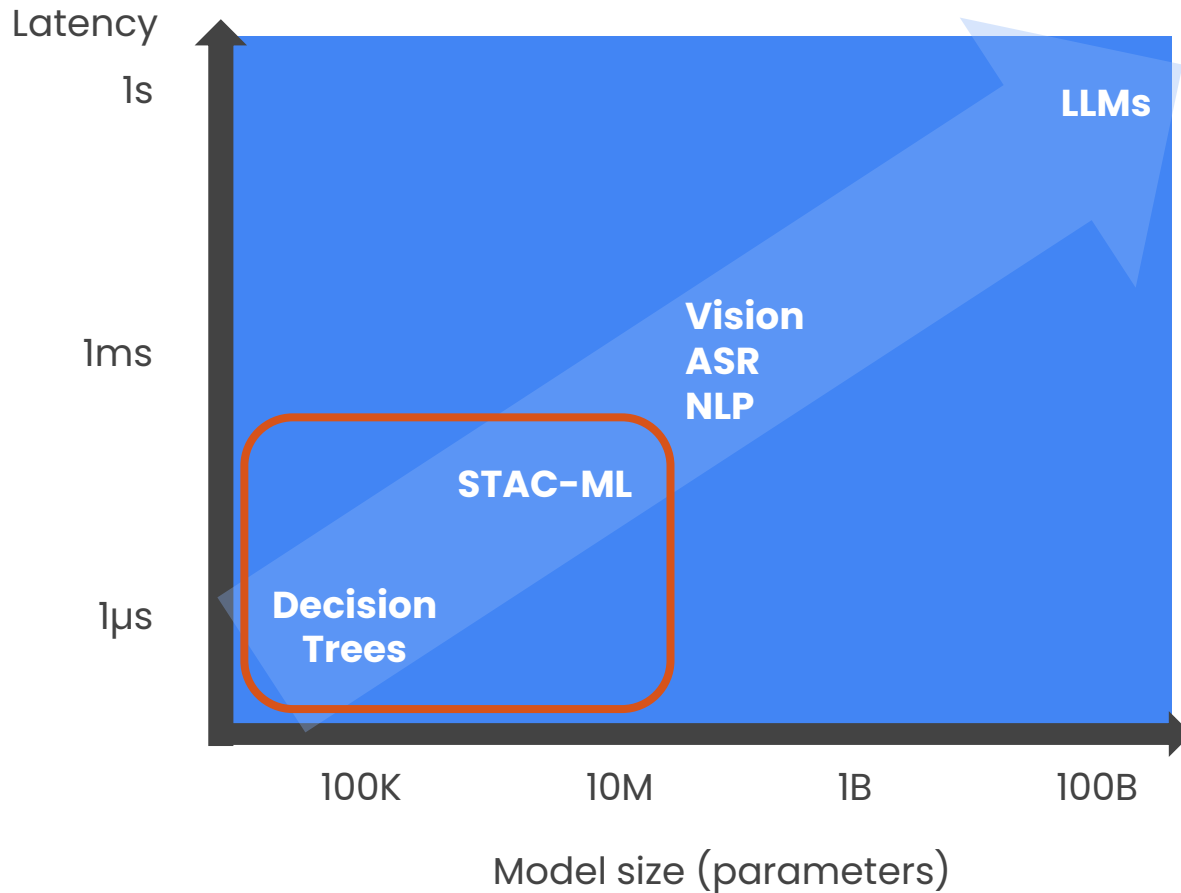




Conclusion

FPGA: optimum choice for latency-optimization

...but how easy is it for an ML team to run their model on an FPGA?



VOLLO – FPGA Overlay for Finance

✓ Faster, more intelligent decisions

Unrivalled latency

High throughput and power-efficiency

- For co-located servers



Simple to program

- Design entirely in PyTorch or TensorFlow

Quick to evaluate, develop & deploy

- Virtual Machine for performance estimation & fast model iteration
- Deploy in standard PCIe card or RTL integrations



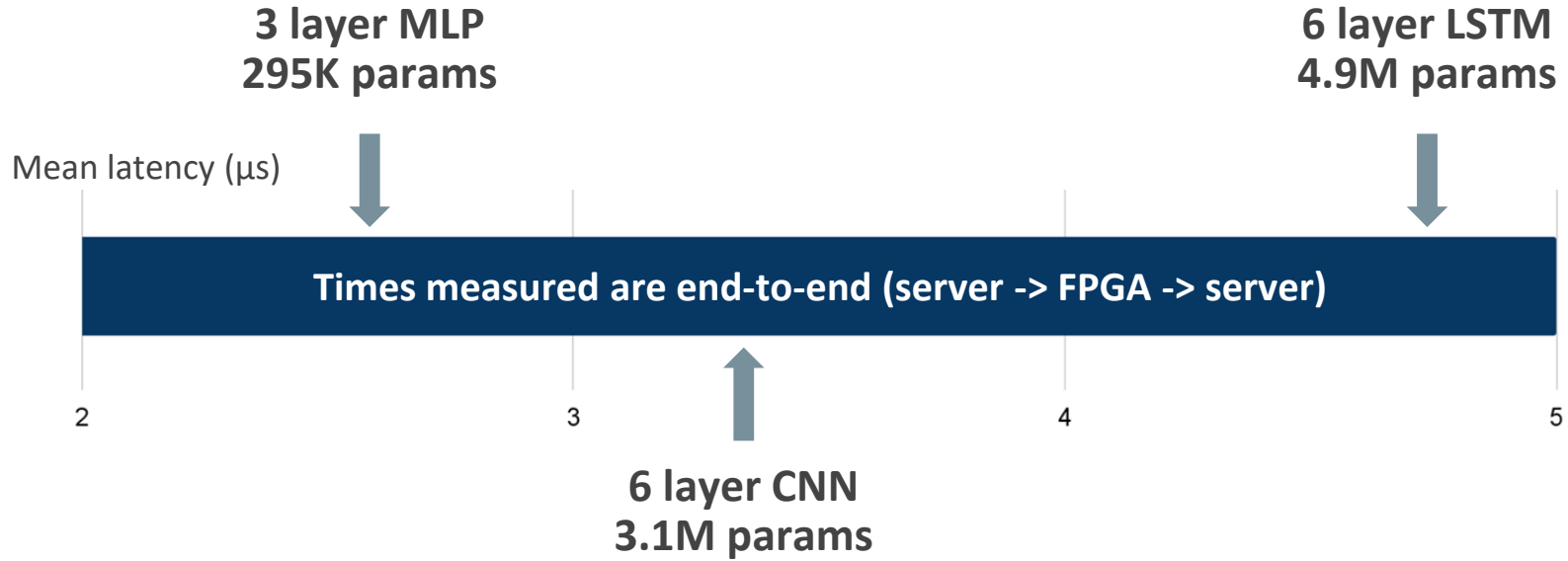
Proven

- In production today with leading innovators



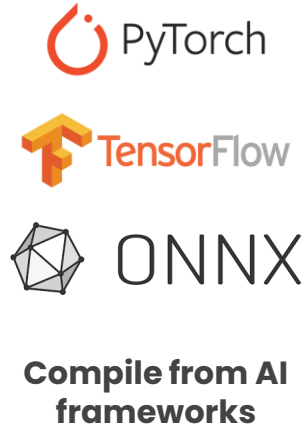
Sub-5 μ s round-trip latencies

vollo.myrtle.ai freely available repositories with examples



VOLLO SDK

Deploy ML models straight to FPGA



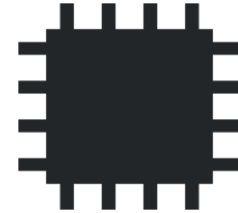
**VOLLO
Compiler**



Free-to-use with
bit-accurate
& cycle-level
simulators that show
real performance



**VOLLO
FPGA**



We sell annual
licenses for PCIe
accelerators or
encrypted netlists

Software flow for ML engineers to design for and infer in lowest latency



VOLLO trees

- Supports
 - Decision Tree models
 - Random Forests
 - Gradient-boosting models including
 - XGBoost
 - LightGBM
 - CatBoost
- Benchmark latencies 1.7 μ s \rightarrow 2.5 μ s
- Compiler imports ONNX models

model	num trees	max depth	input features	fully populated	mean latency (us)	99th percentile latency (us)
single-decision-t1-d1-f32	1	1	32	No	1.7	1.9
example-t1000-d5-f128	1000	5	128	No	1.9	2.1
example-t1000-d5-f128-full	1000	5	128	Yes	1.9	2.1
example-t1000-d8-f128	1000	8	128	No	1.9	2.1
example-t1000-d8-f128-full	1000	8	128	Yes	1.9	2.1
example-t512-d8-f512	512	8	512	No	1.9	2.1
example-t1024-d8-f1024	1024	8	1024	No	2.0	2.2
example-t4096-d8-f1024-full	4096	8	1024	Yes	2.5	2.6



ONNX

NOT STAC BENCHMARKS

Demand-Driven Roadmap

✓ Available

📅 On target



10x bigger models, up to 64M params per board

Evaluate Vollo for your models today. No cost, no security risk

Cycle-level & bit-accurate simulators available from:

vollo.myrtle.ai

or contact us at:

vollo@myrtle.ai