
AWS
re:Invent
NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

AIM330

Optimizing deep learning models with compilation for faster inference

Robert Van Dusen

Sr. Product Manager
AWS Deep Learning

Lokesh Gupta

Software Development Manager
AWS Deep Learning

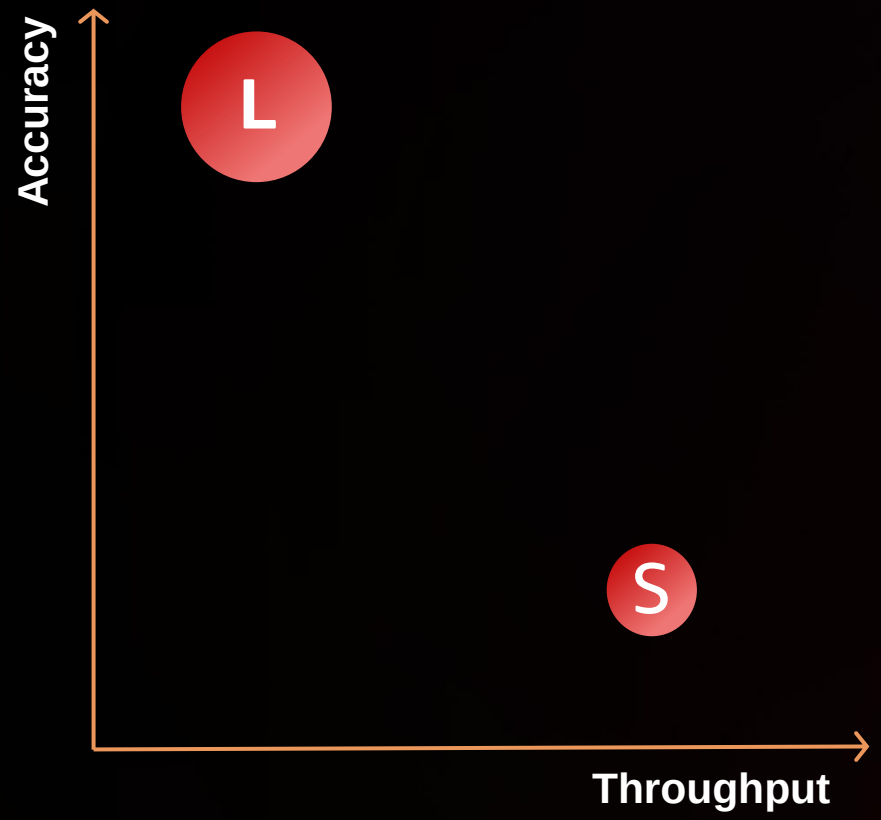
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

- Compiling models for inference with Amazon SageMaker Neo
- Benchmarks
- Demo
- Q&A

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

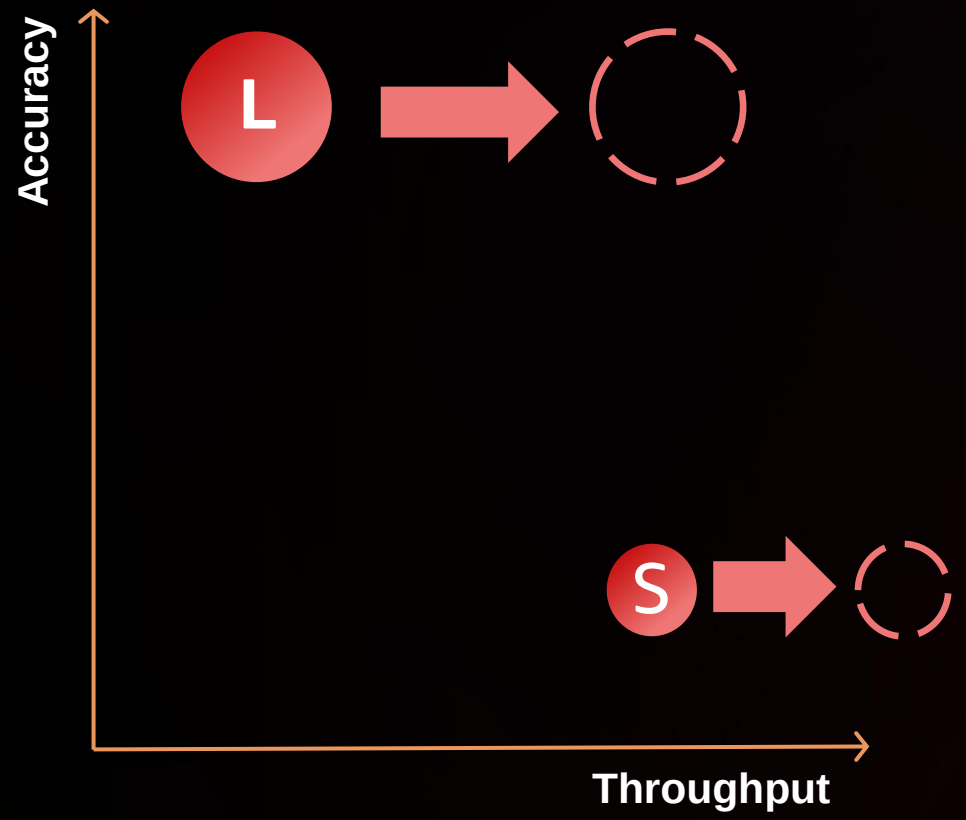
Compromise on performance for accuracy?



- Larger, complex models: better accuracy but worse latency/throughput
- At odds with business requirements (e.g., latency, throughput, budget)

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker Neo improves model performance



Amazon SageMaker Neo

- Optimize compute
- Reduce memory demands
- Decrease infrastructure costs
- Up to 25x latency improvement

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SageMaker Neo optimizes for inference

How Neo inference compiler works

Model parsing

Convert a TensorFlow, MXNet, PyTorch, or XGBoost model into a common format

Graph optimizations

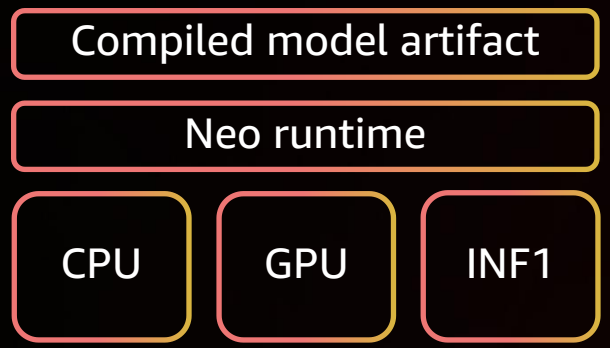
Detect patterns in the ML model structure to reduce the execution time

Tensor optimizations

Detect patterns in the shape of input data to allocate memory efficiently

Code generation

Use a low-level compiler to generate machine code for each target

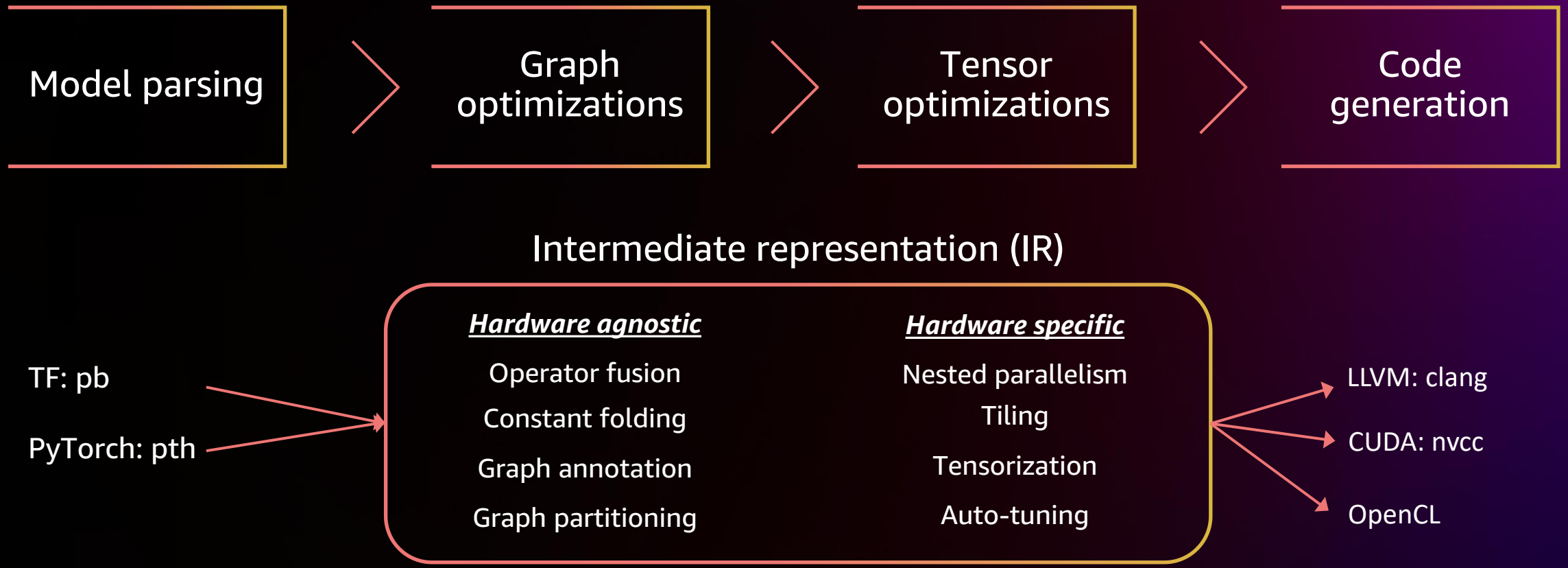


- Optimized model deployed to target instance with the Neo runtime
- Supported hardware includes CPU, GPU, INF1, and edge devices

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

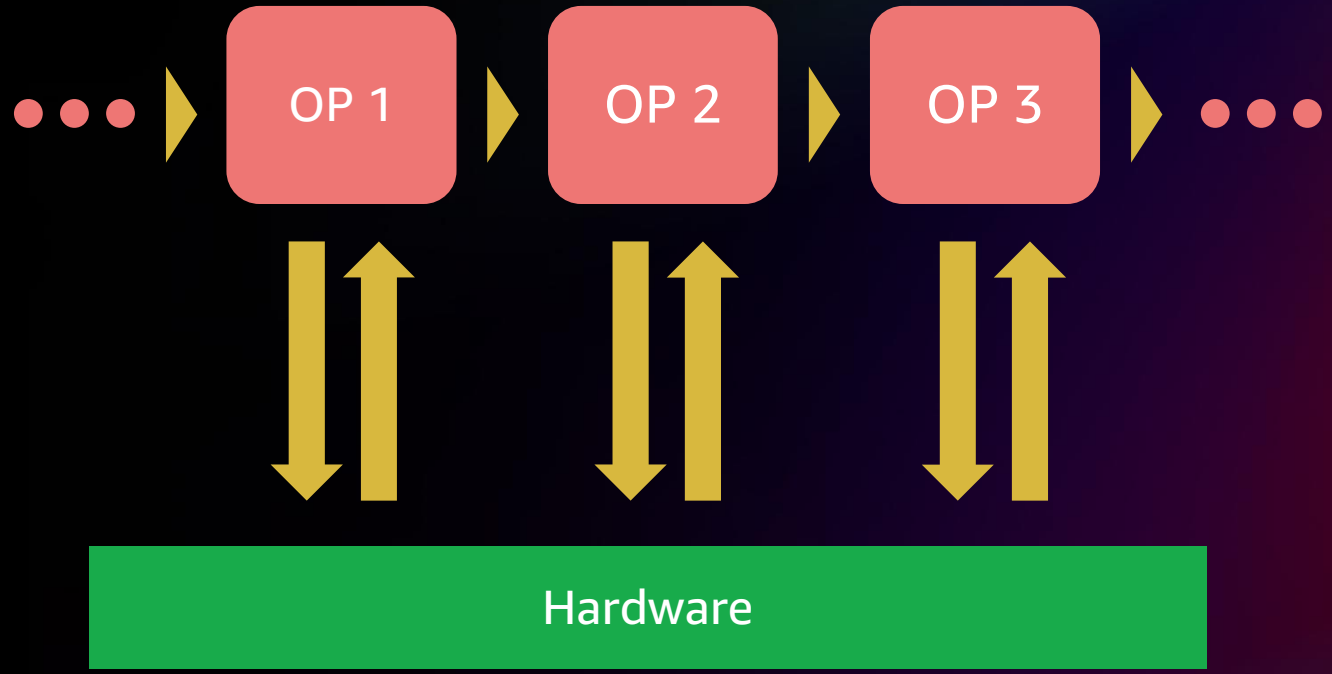
SageMaker Neo optimizes for inference

How Neo inference compiler works



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Sequential operator execution is inefficient



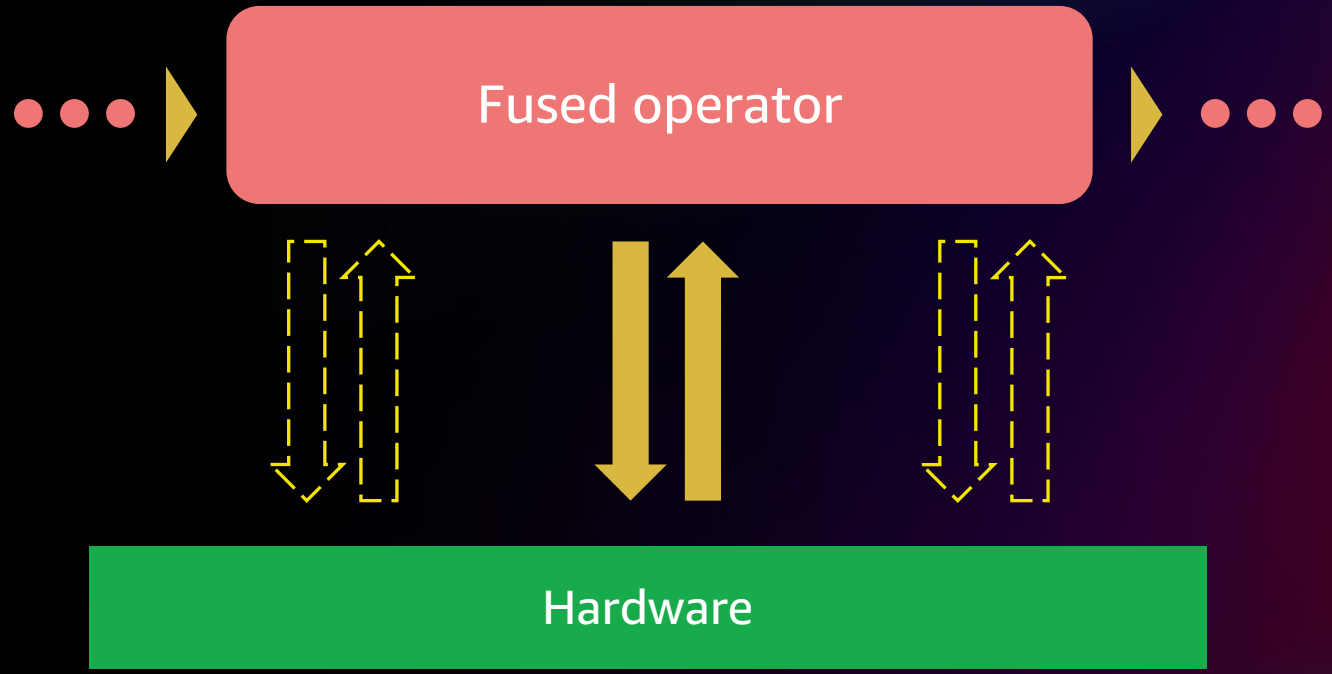
Operators in the computation graph

Kernel launches

Hardware

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Operator fusion merges sequential operators



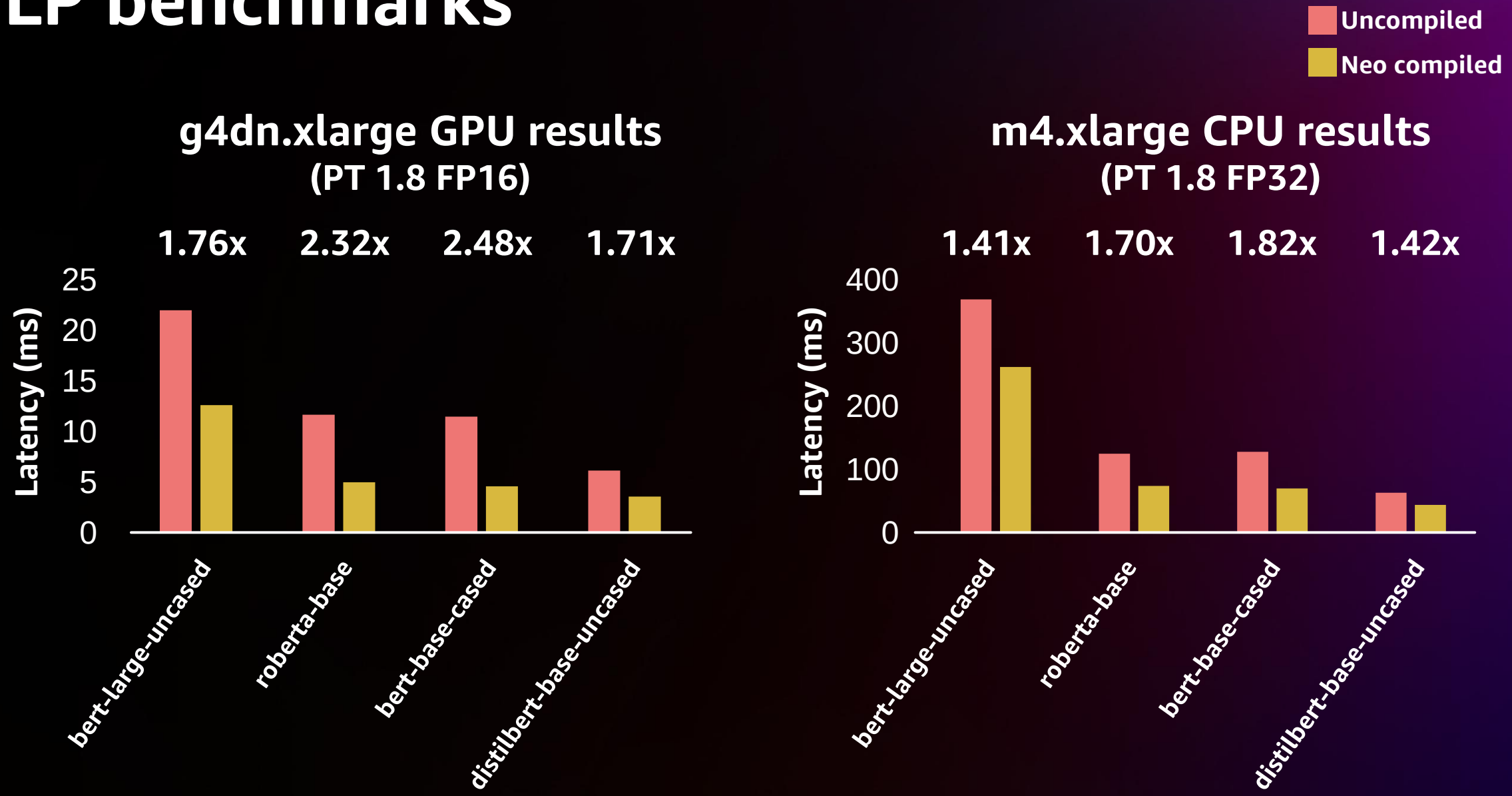
Operators in the computation graph

Kernel launches

Hardware

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

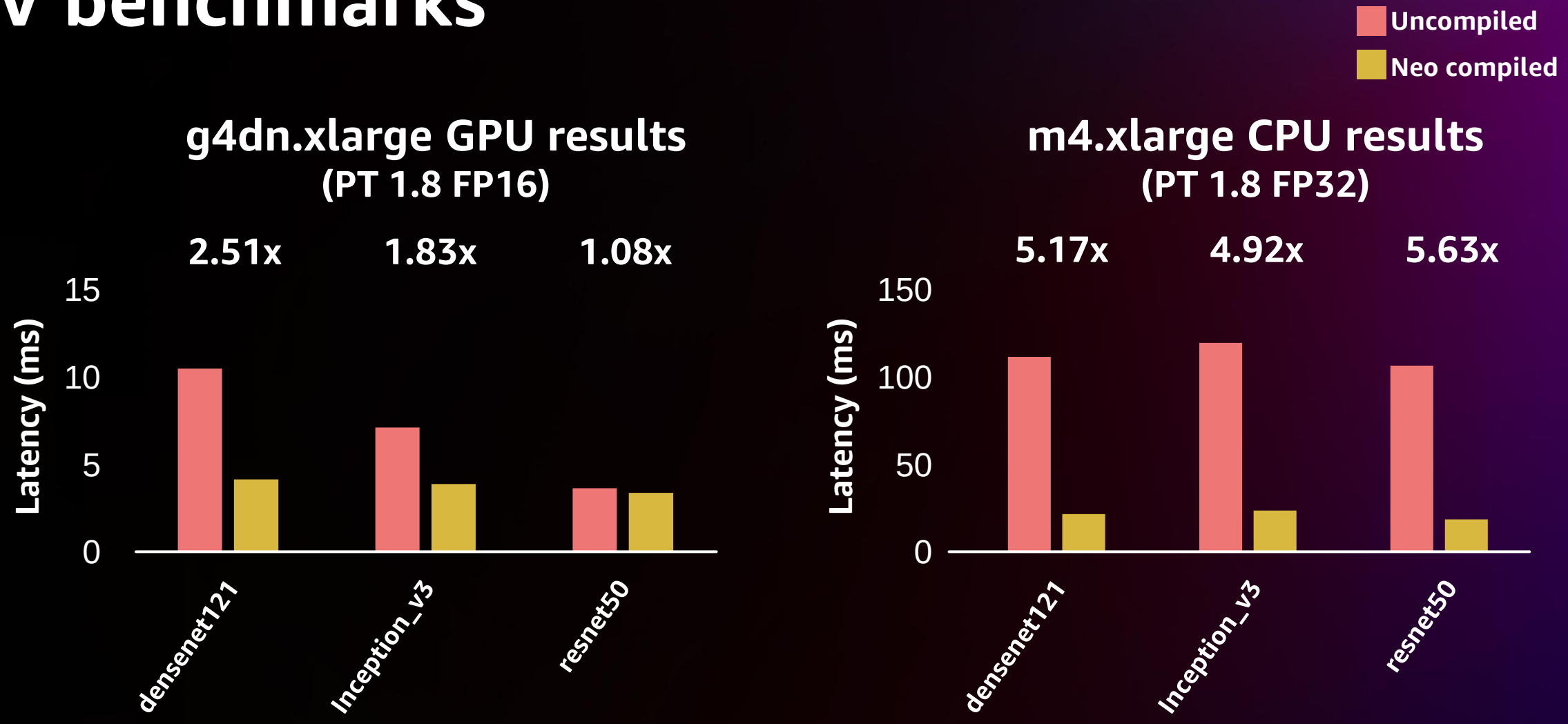
NLP benchmarks



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Note: p95 results, seq_len=128, batch_size=1, models from Hugging Face

CV benchmarks



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Note: p95 results, batch_size=1

Best use cases for compilation

- Static input shape
- Models with dynamic tensors or operators (e.g., mask-rcnn)
- Test models for improved performance

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Resources

- Amazon SageMaker Neo documentation



- Sample notebooks



- Customer examples and blog posts



aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Robert Van Dusen
robdusen@amazon.com

Lokesh Gupta
lokeshgu@amazon.com



Please complete the session survey in the **mobile app**

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.