AWS re:Invent

CMP423-R1

Hands-on deep learning inference with Amazon EC2 Inf1 instances

Wenming Ye

Sr. Solutions Architect

Amazon Web Services

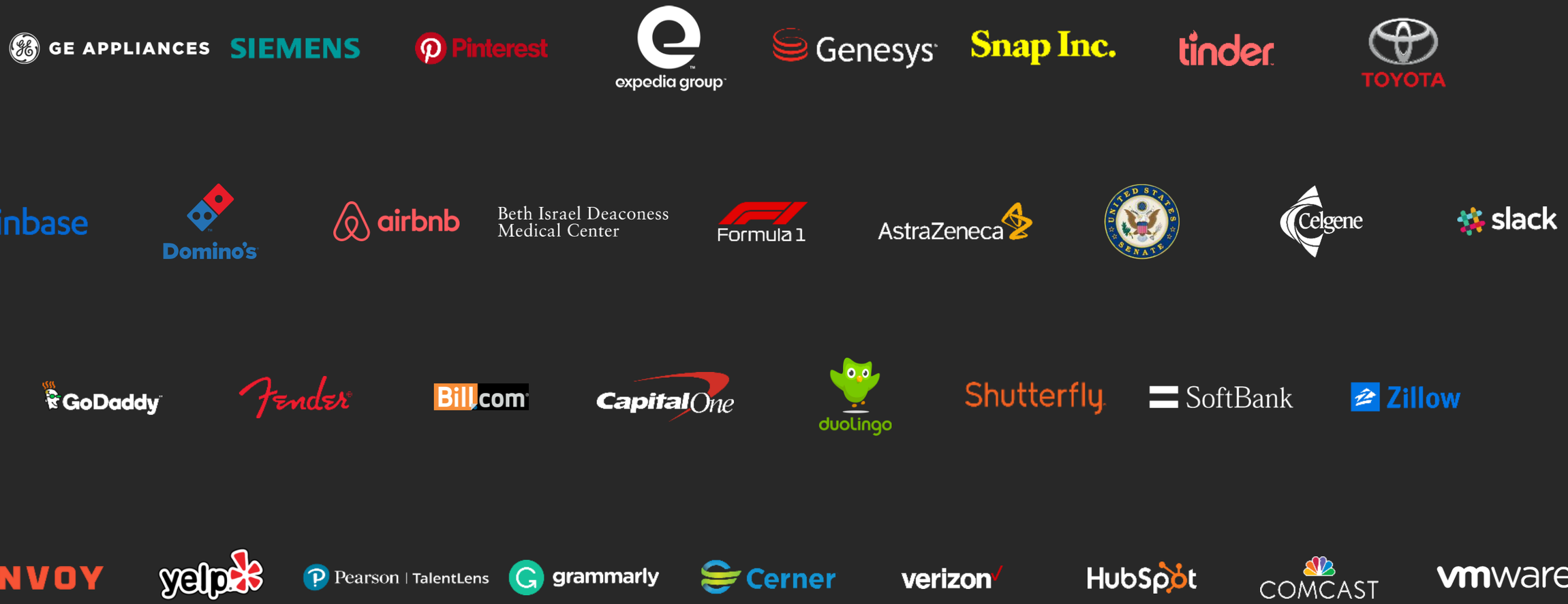
AWS re:Invent

aws

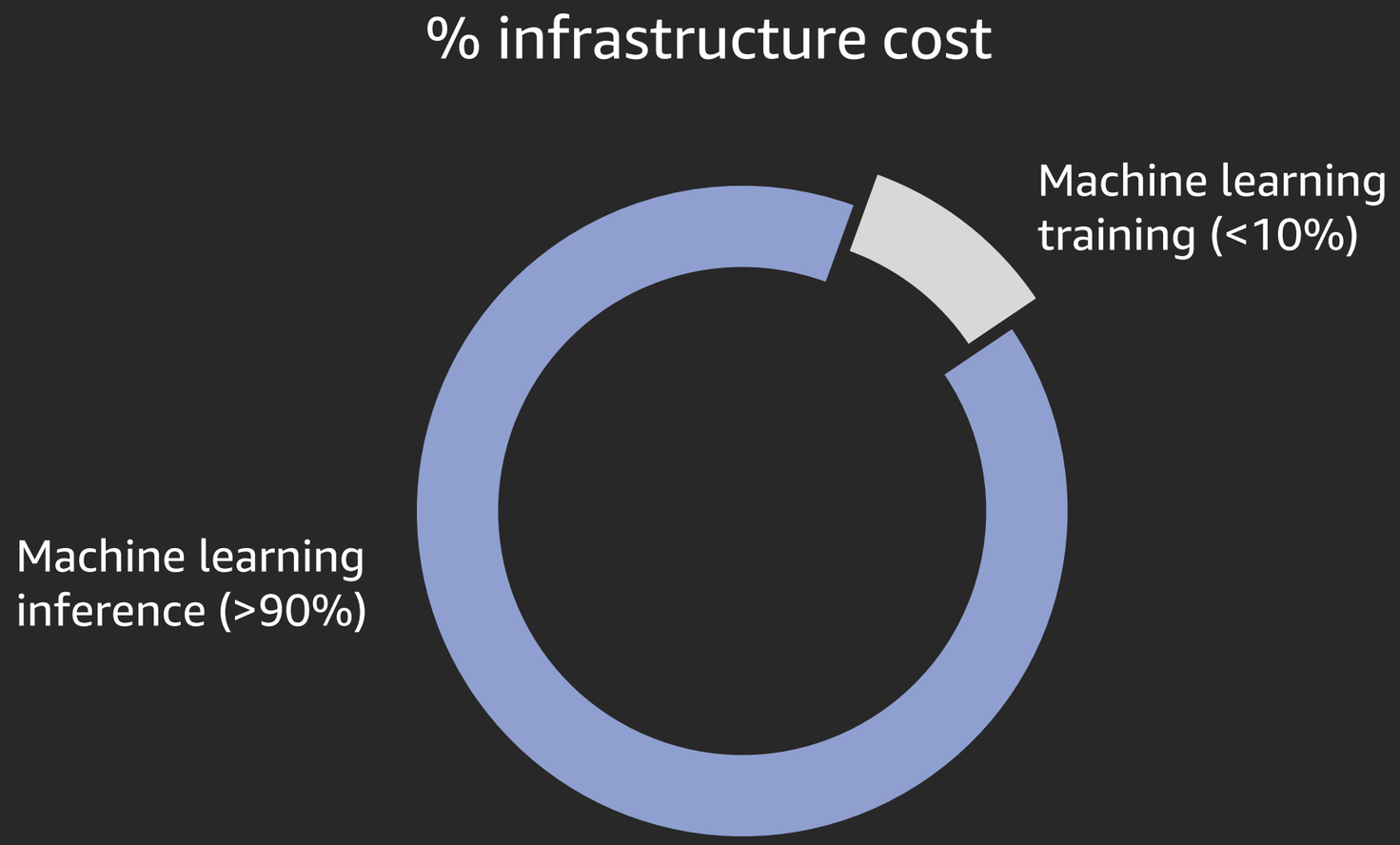
Agenda

- AWS Inf1 – Review from CMP324 [35 min.]
- Lab Logistics and Setup: [5 min.]
- Lab 1-4 [80 min.]
- Learning Resources and Summary

More machine learning happens on AWS than anywhere else



Inference accounts for the majority of machine learning infrastructure costs



AWS Inferentia

Optimizing ML performance with a custom chip

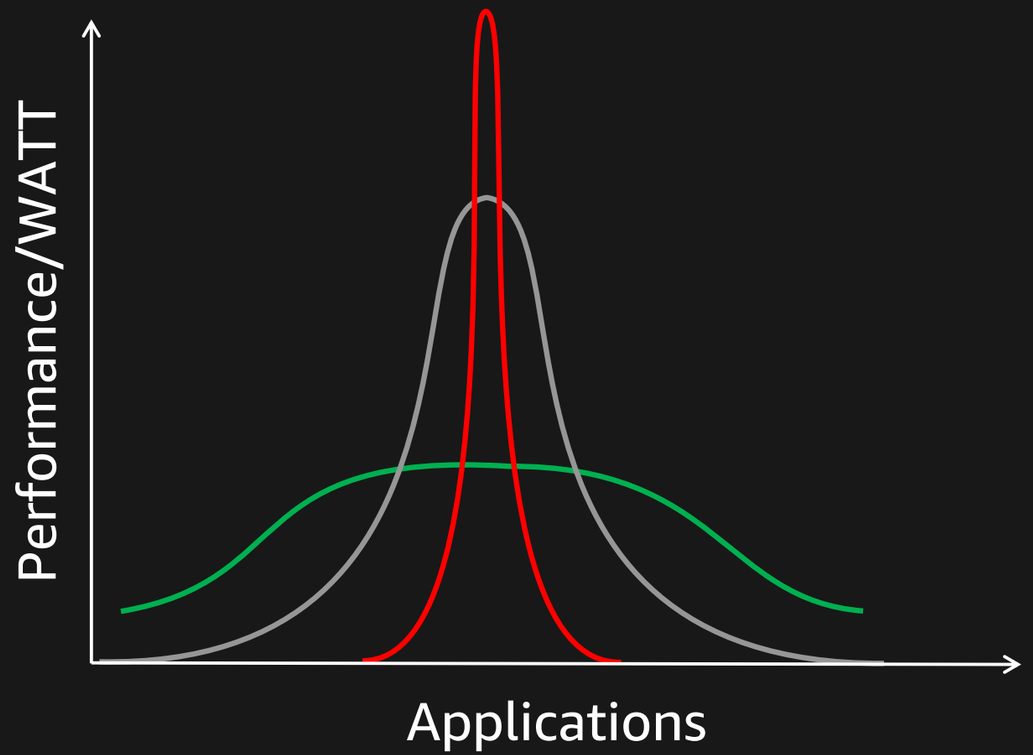
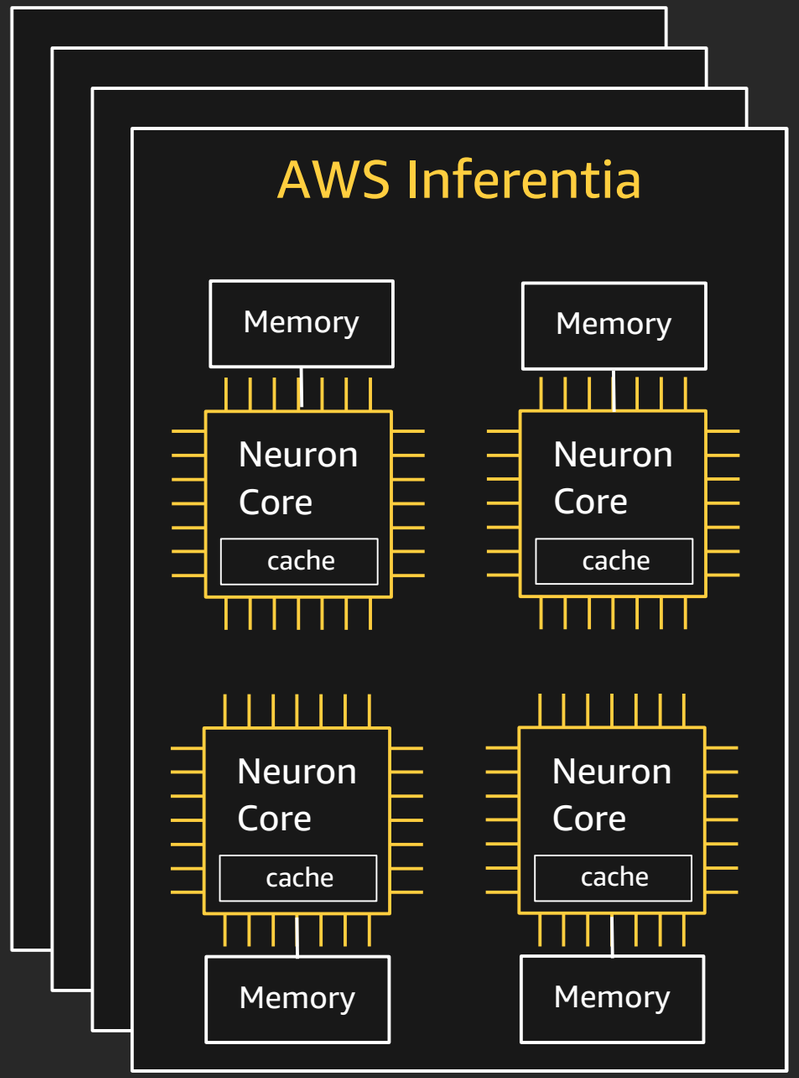


Chart for illustrative purposes

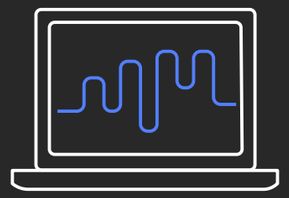
AWS Inferentia quick tour

AWS custom built: chip, software, and server

- 4 NeuronCores
- Up to 128 TOPS
- 2-stage memory hierarchy
  - Large on-chip cache and commodity DRAM
- Supports FP16, BF16, INT8 data types
- Fast chip-to-chip interconnect

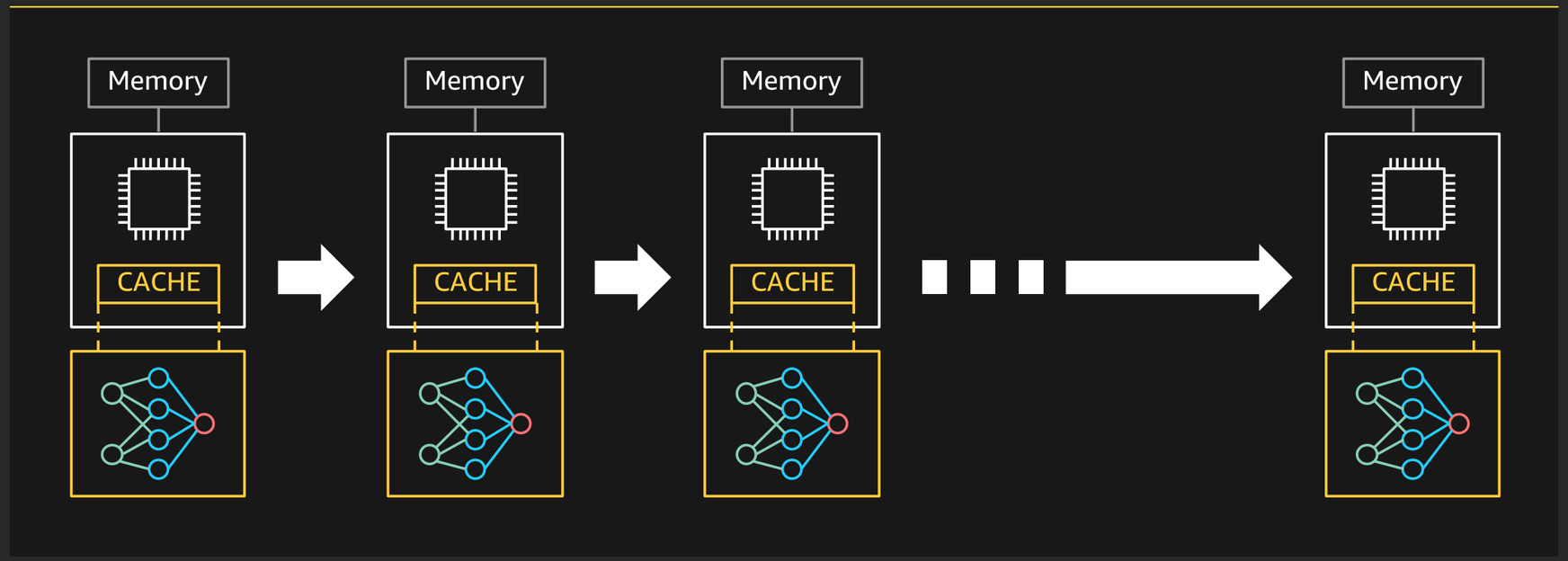


Low-latency inference at high load



NeuronCore pipeline

- Very low latency
- Full bandwidth due to on-chip cache



AWS Neuron

Introducing AWS Neuron

Software suite enabling high-performance deep learning inference on AWS Inferentia

Compiler

Runtime

Profiling and debugging tools

Supports all major frameworks

TensorFlow

mxnet

PyTorch

github.com/aws/aws-neuron-sdk

AWS Neuron support forum

AWS Neuron quick tour

Compile

Neuron Compiler
(NCC)

Deploy

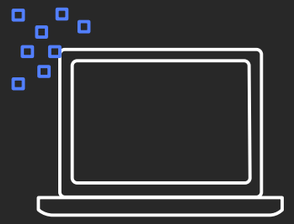
Neuron Runtime
(NRT)

Neuron Binary
(NEFF)

Profile

Neuron Tools

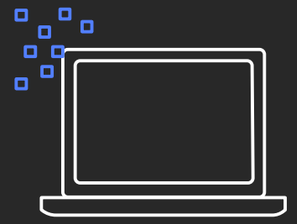
AWS Neuron highlights



Smart partitioning

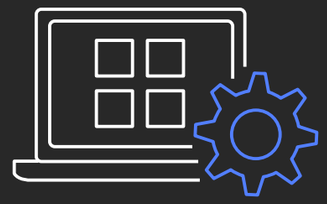
Automatically optimize neural-net compute

AWS Neuron highlights



Smart partitioning

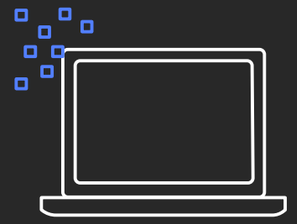
Automatically optimize neural-net compute



Auto FP32 casting

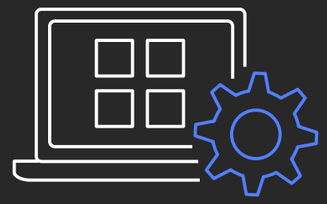
Ingest FP32 trained models, and Neuron auto casts to BF16

AWS Neuron highlights



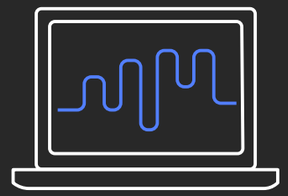
Smart partitioning

Automatically optimize neural-net compute



Auto FP32 casting

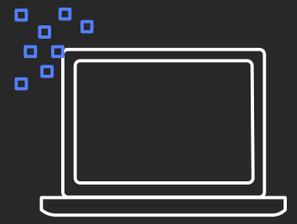
Ingest FP32 trained models, and Neuron auto casts to BF16



NeuronCore pipeline

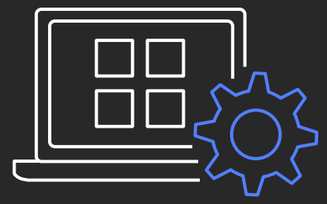
Very low latency
Full bandwidth

AWS Neuron highlights



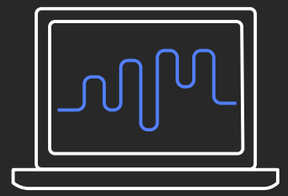
Smart partitioning

Automatically optimize neural-net compute



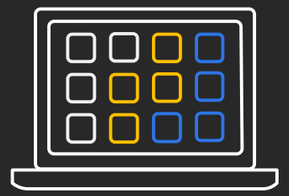
Auto FP32 casting

Ingest FP32 trained models, and Neuron auto casts to BF16



NeuronCore pipeline

Very low latency
Full bandwidth

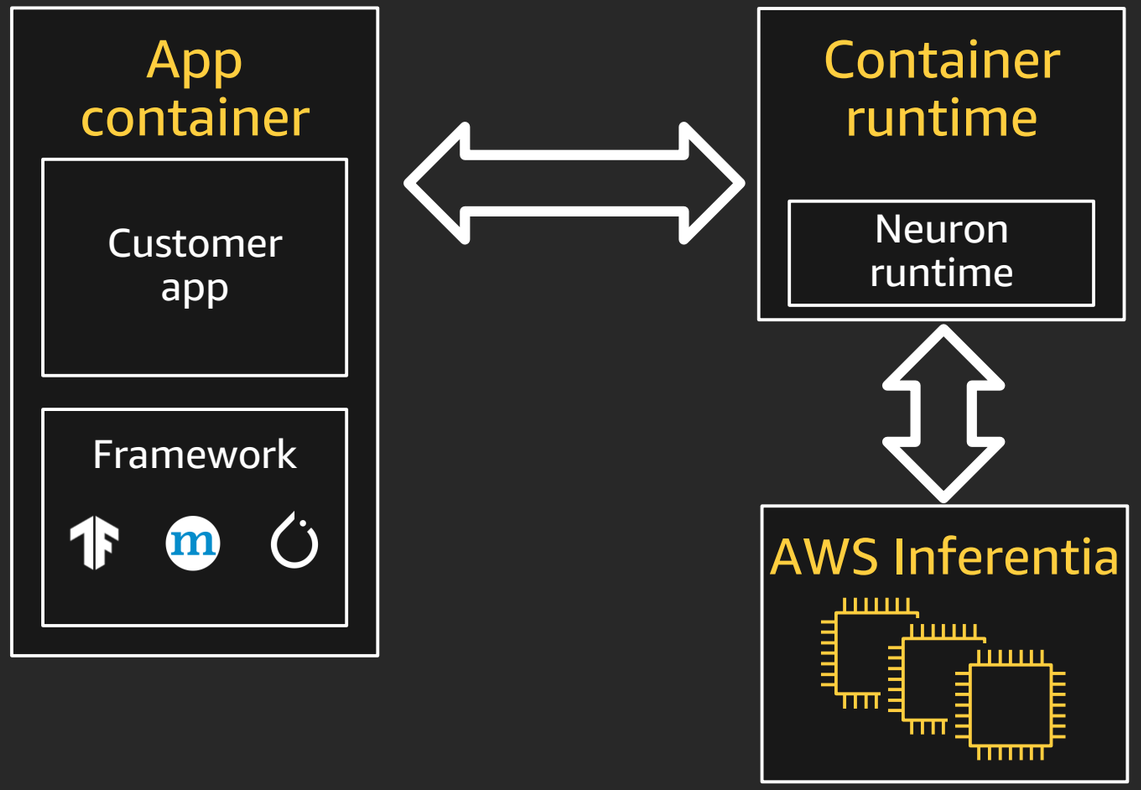


NeuronCore Groups

Concurrently run multiple models

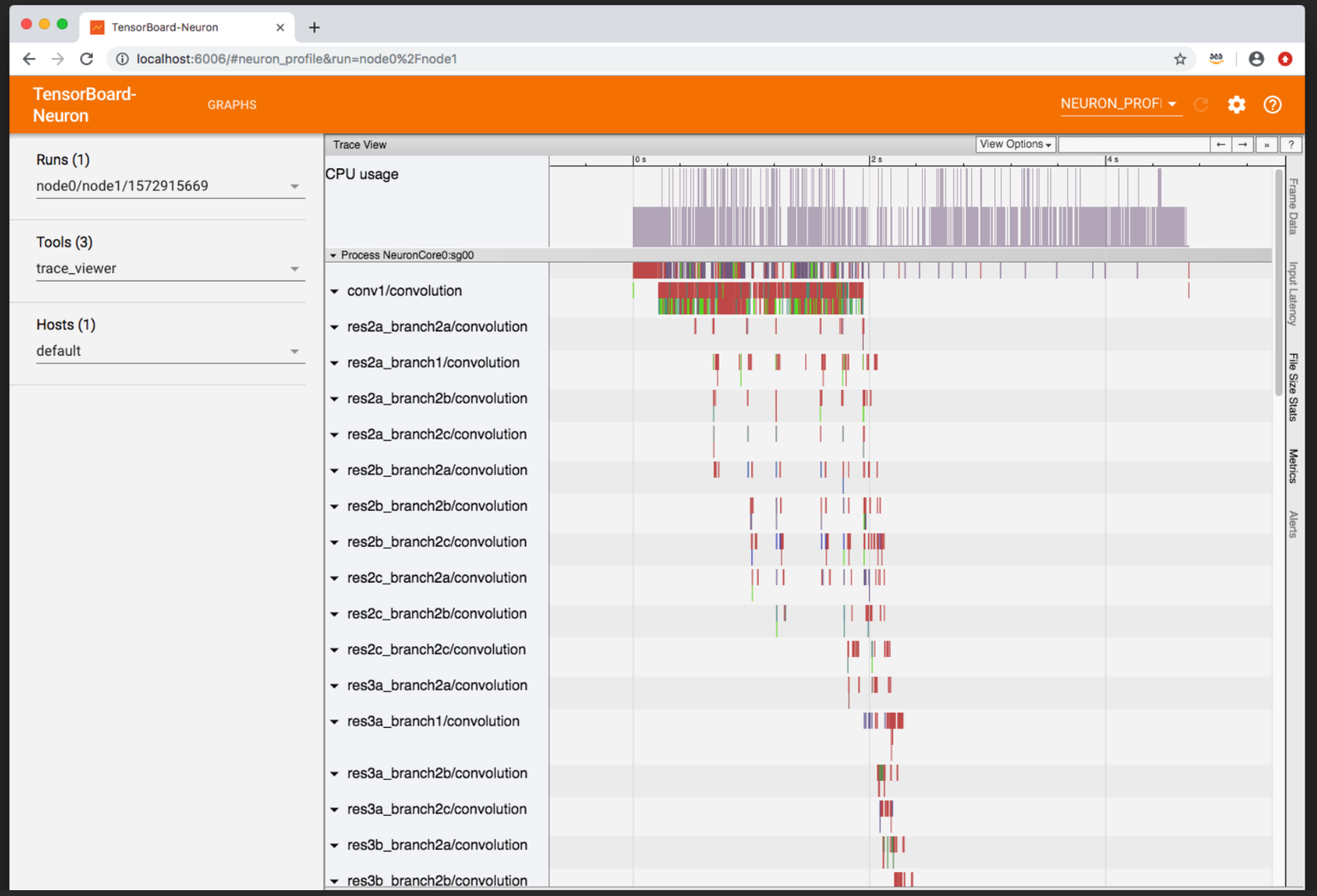
Neuron runtime

Deploy



Profiling with Neuron tools

Profile



Amazon EC2 Inf1 instances

Introducing Inf1 instances for ML inferencing

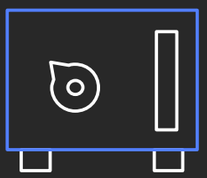
Object detection

Natural language processing

Personalization

Speech recognition

Image processing



Fraud detection

ML inference deployment options on Amazon EC2

Custom chip
EC2 Inf1 instances

Featuring AWS Inferentia

- Applications that leverage common ML frameworks
- Powered by AWS Inferentia
- Best price/performance for ML inferencing in the cloud
  - Up to 40% lower inference than EC2 G4 instances

Available today

GPU based
EC2 G4 instances

- Applications that require access to CUDA, CuDNN or TensorRT libraries
- Amazon EC2 G4 instances based on NVIDIA T4 GPUs

Launched

CPU based
EC2 C5 instances

- Low sensitivity to inference performance
- Intel Skylake CPUs Support for AVX-512/ VNNI instruction set

Launched

Inf1 instance sizes

Instance size	vCPUs	Memory (GIB)	Storage	Inferentia Chips	NeuronCore Pipeline Mode	Network B/W	EBS B/W
inf1.xlarge	4	8	EBS only	1	N/A	Up to 25 Gbps	Up to 3.5 Gbps
inf1.2xlarge	8	16	EBS only	1	N/A	Up to 25 Gbps	Up to 3.5 Gbps
inf1.6xlarge	24	48	EBS only	4	Yes	25 Gbps	3.5 Gbps
inf1.24xlarge	96	192	EBS only	16	Yes	100 Gbps	14 Gbps

- Available in 4 sizes
- Single and multi-chip instances

- AWS 2nd Gen Intel Xeon Scalable Processors
- Up to 100 Gbps networking bandwidth

- Available to use with Amazon SageMaker, Amazon Elastic Kubernetes Service, and Amazon Elastic Container Service (coming soon)

Best price/performance for ML inference in the cloud

Instance type	Throughput (Seq/Sec)	OD Price ($/Hr)	Cost per inference	Throughput: Inf1 vs. G4	Cost-per-inference: Inf1 vs. G4
inf1.24xlarge	19,200	$7.619	$0.0003968	3.28x	40.7%
G4.12xlarge	5,846	$3.912	$0.0006692		

Results based on running BERT-base model end-to-end with TensorFlow

EC2 Inf1 offers up to 3x higher throughput and 40% lower cost per inference than EC2 G4 instances

Highest price-to-performance in the cloud for ML inference

Try Inf1 instances today

Inf1 instances are now available in the US-East-1 (N. Virginia) and US-West-2 (Oregon) Regions
More Regions coming soon!

EC2 Inf1 instances support on demand, reserved and spot purchasing options; also available as part of Savings Plan

Instance size	On Demand	1-Year Standard RI (40% discount)	3-Year Standard RI (60% discount)
inf1.xlarge	$ 0.368/Hr	$ 0.221/Hr	$ 0.147/Hr
inf1.2xlarge	$ 0.584/Hr	$ 0.351/Hr	$ 0.234/Hr
inf1.6xlarge	$ 1.905/Hr	$ 1.143/Hr	$ 0.762/Hr
inf1.24xlarge	$ 7.619/Hr	$ 4.572/Hr	$ 3.048/Hr

Alexa use case

Neural text-to-speech challenges

Low latency requirement for dialog system
High throughput requirement implied by streaming of output speech

Context generation is a sequence-to-sequence auto-regressive model
Inference-bound memory bandwidth

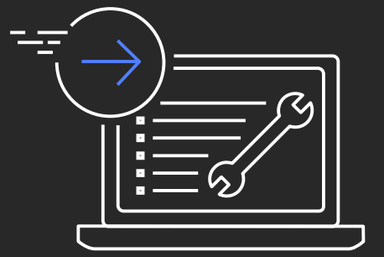
High temporal density of speech production model resulting in 90 GFLOPs to generate 1 second of output
Compute-bound inference

Using EC2 GPU instances to meet requirements results in high operational cost

Alexa TTS migration to EC2 Inf1: Ease of integration

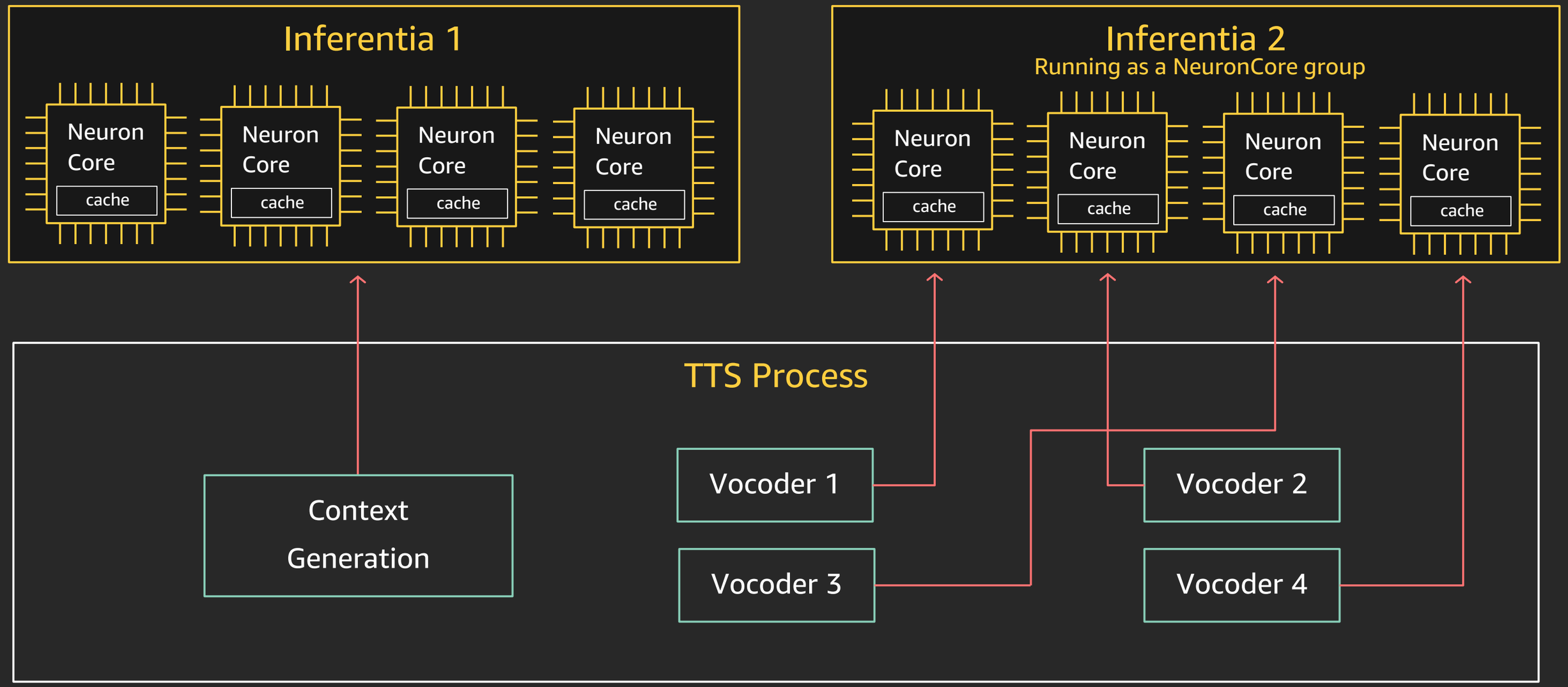
Alexa TTS uses MXNet supported natively by AWS Neuron

Support for C and Python APIs



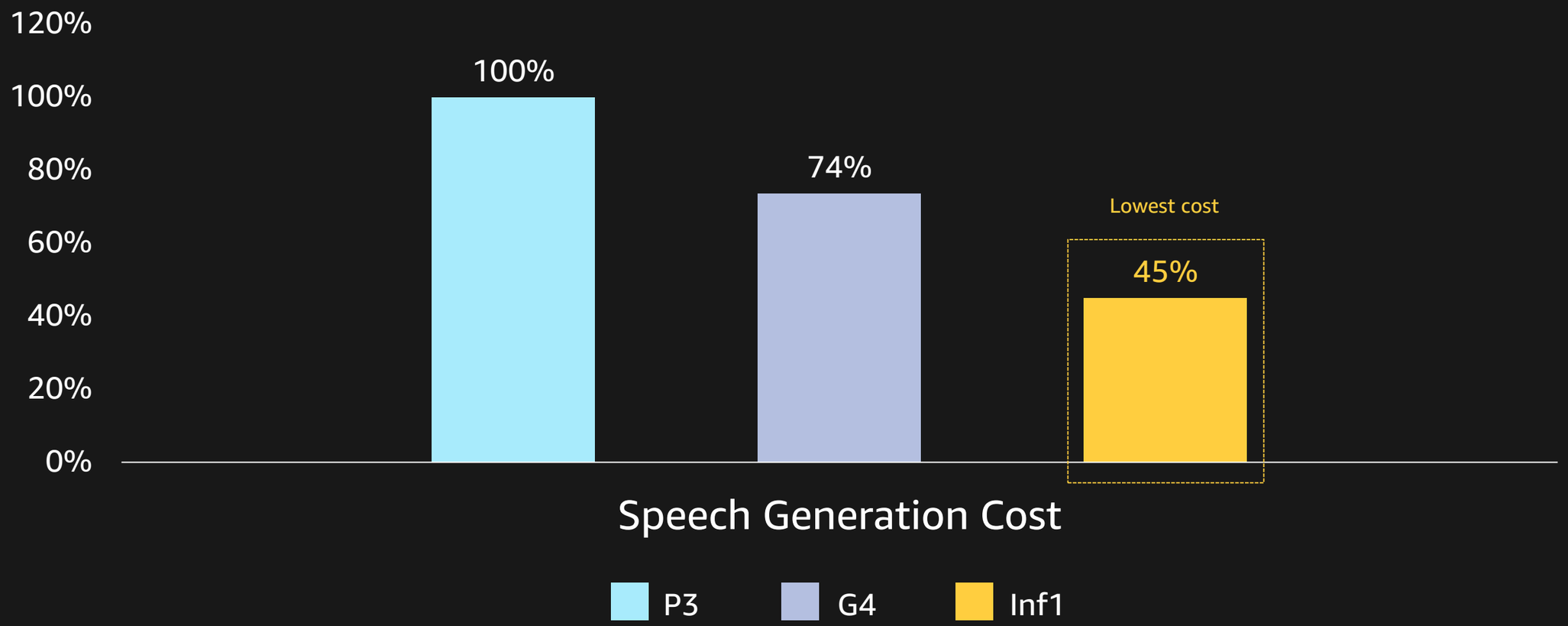
Options to migrate original FP32 models to FP16 or Bfloat16

Alexa TTS migration to EC2 Inf1: Architecture



Alexa TTS migration to EC2 Inf1: Long text

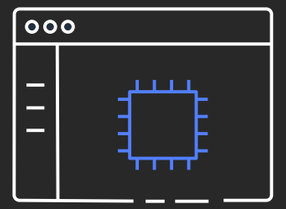
Alexa long text traffic (ex: books, news) has even higher gains



Getting started

Github
Tutorials/drivers

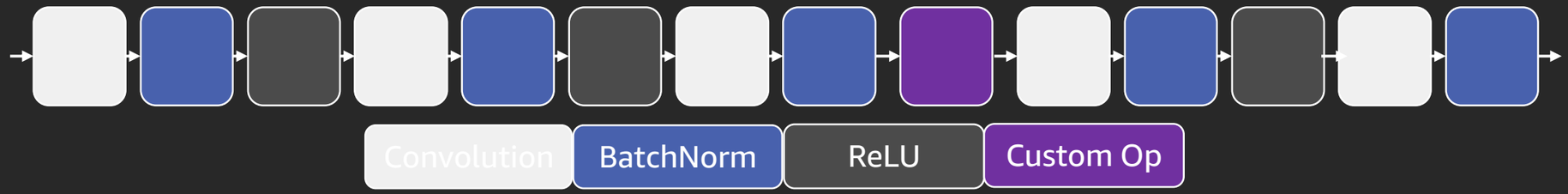
AWS Neuron
developer forum



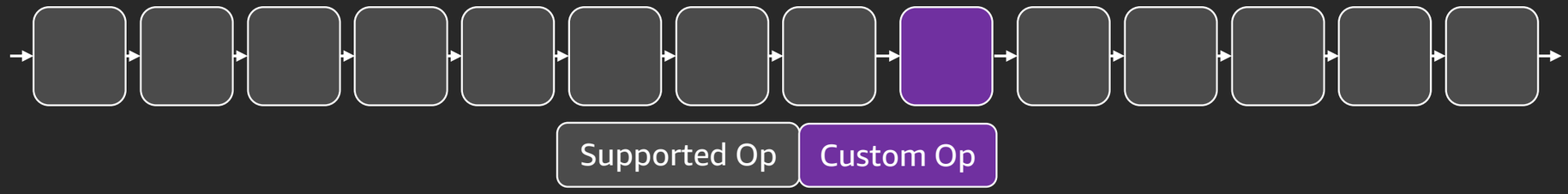
EC2 Inf1
landing page

Subgraph Compilation – under the hood

- We start with a normal computation graph for an MXNet model



- We search for subgraphs of supported operators



Hands-on labs

1. Set up development instance on a **c5d.4xlarge**

2. Deploy and model-serve on Inf1.2xlarge instance

3. Load test run

4. Debugging and profiling

Full URL:

https://github.com/awshlabs/reinvent19Inf1Lab

Questions? Use Forum at:
https://forums.aws.amazon.com/forum.jspa?forumID=355

Thank you!

Wenming Ye

Twitter @wenmingye