Lower the Cost of Running ML Applications with New Amazon EC2 Inf1 Instances

Gadi Hutt, Senior Director of Business Development and Product, AWS

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

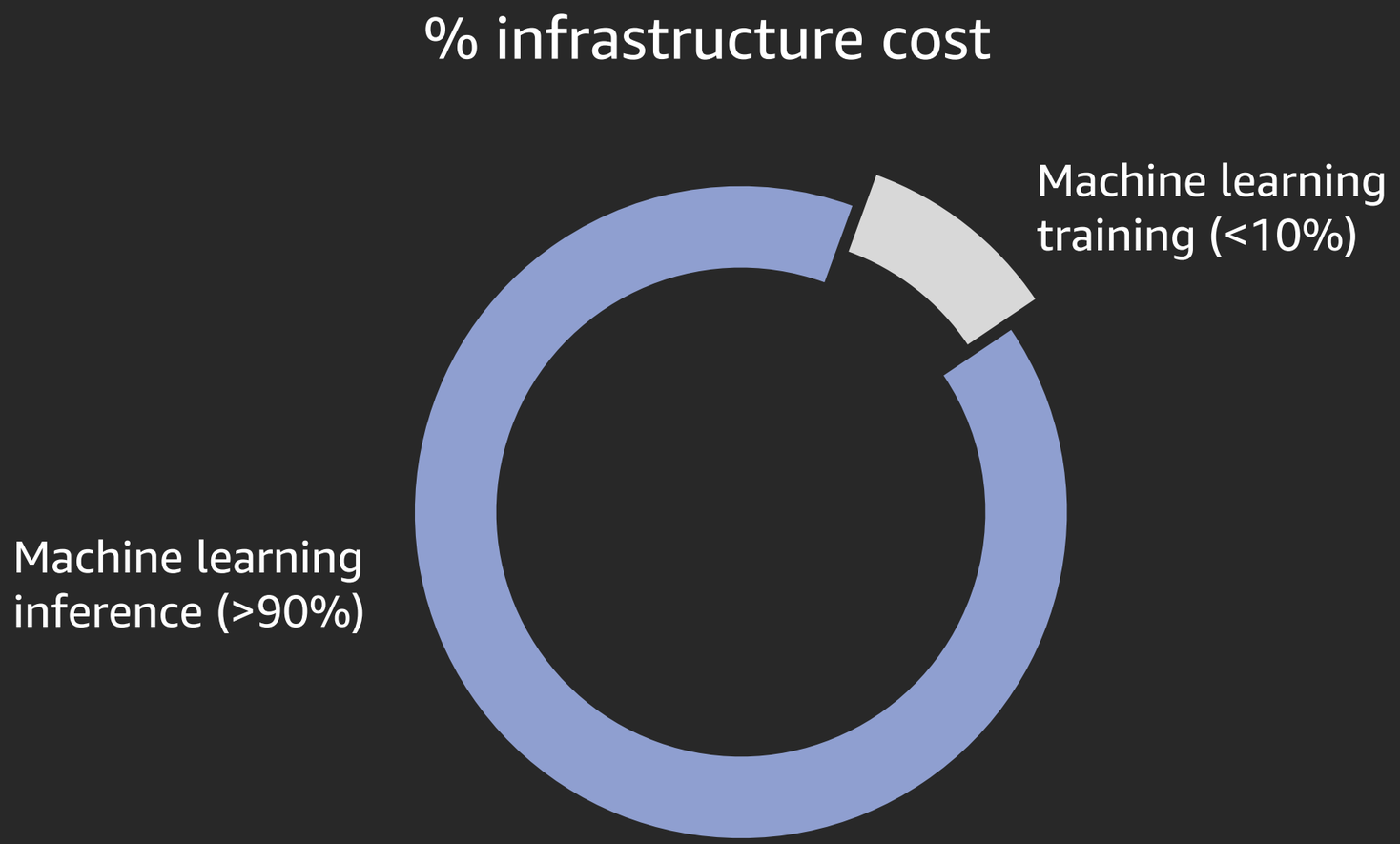
Agenda

- Machine learning at AWS
- AWS Inferentia overview (HW and SW)
- Amazon EC2 Inf1 instance
- Customer story: Alexa Text-to-Speech (TTS)

MACHINE LEARNING IS HAPPENING IN COMPANIES OF EVERY SIZE AND INDUSTRY

Tens of thousands customers have chosen AWS for their ML workloads | More than twice as many customers using ML than any other cloud provider

Machine learning Infrastructure costs



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

GPU based
EC2 G4 instances

Applications that require access to CUDA, CuDNN or TensorRT libraries

Amazon EC2 G4 instances based on NVIDIA T4 GPUs

CPU based
EC2 C5 instances

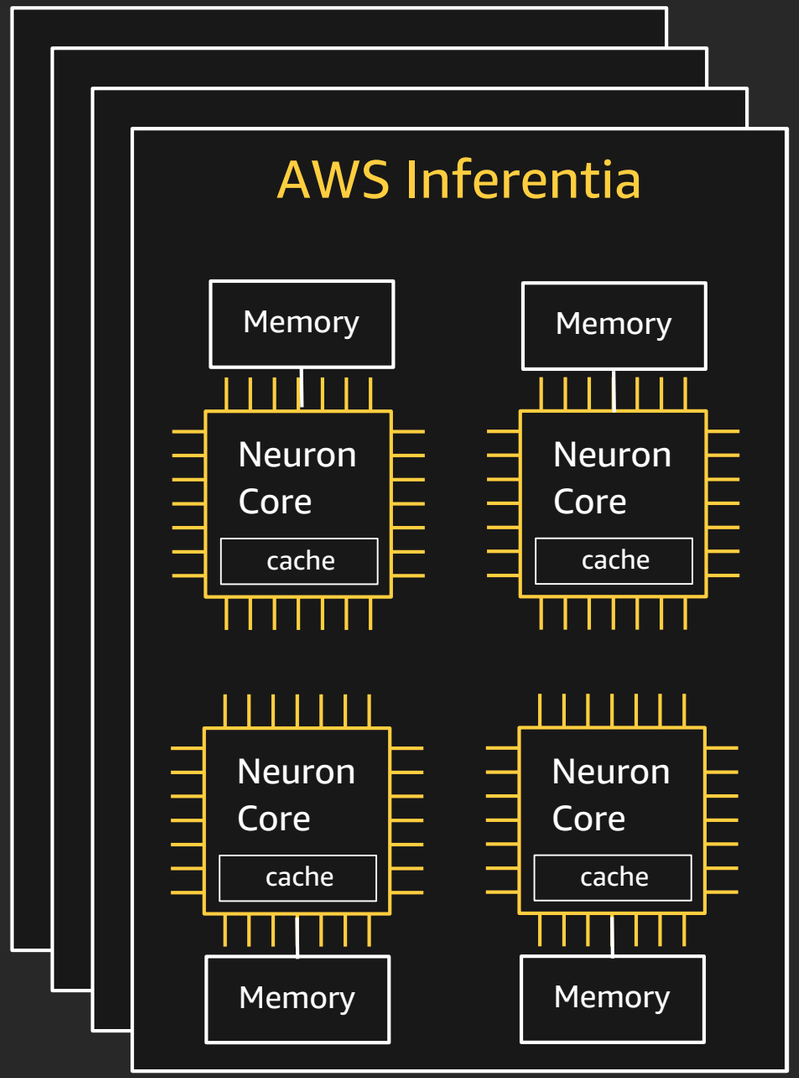
Low sensitivity to inference performance

Intel Skylake CPUs
Support for AVX-512/ VNNI instruction set

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



Inf1 instances for ML inferencing

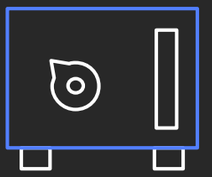
Object detection

Natural language processing

Personalization

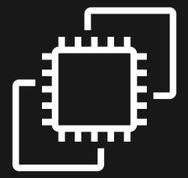
Speech recognition

Search

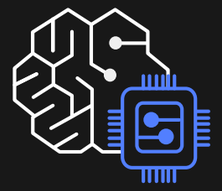


Fraud detection

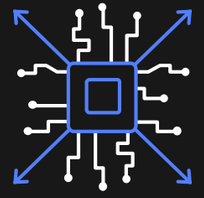
Inf1 instances are built from the ground up by AWS



AWS Nitro

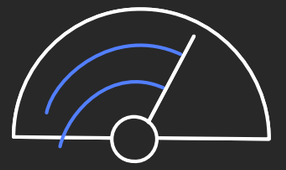


AWS Inferentia



AWS Custom 2nd Gen Intel Xeon Scalable Processors

AWS Neuron



High performance

Low cost

Inf1 instance sizes

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

- Available in 4 sizes
- Single and multi chip instances

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

- Amazon SageMaker, Amazon EKS, and Amazon ECS support coming soon

Best price/performance for ML inference in the cloud

Instance type	Throughput (Seq/Sec)	OD Price ($/Hr)	Cost-per-1M inferences	Throughput: Inf1 vs. G4	Cost-per-inference: Inf1 vs. G4
inf1.24xlarge	26,100	$7.619	$0.081	4.24x	54.1%
G4.12xlarge	6,156	$3.912	$0.1765		

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

EC2 Inf1 offers up to 4x higher throughput[1] and 54% lower cost per inference[1] than EC2 G4 instances

Highest price to performance in the cloud for ML inference

AWS Neuron

high-performance deep learning inference SDK

Neuron Compiler

Neuron Runtime

Profiling tools

Supports all major frameworks

TensorFlow mxnet PyTorch

github.com/aws/aws-neuron-sdk

AWS Neuron support forum

https://forums.aws.amazon.com/forum.jspa?forumID=355

AWS Neuron quick tour

Compile

Neuron Compiler

(NEFF output)

Deploy

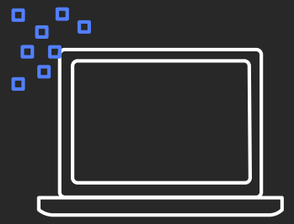
Neuron Runtime
(NRT)

Neuron Binary
(NEFF)

Profile

Neuron Tools

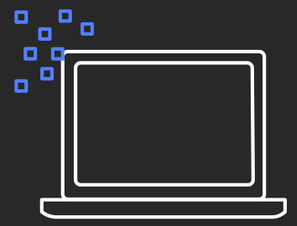
AWS Neuron Compiler highlights



Smart partitioning

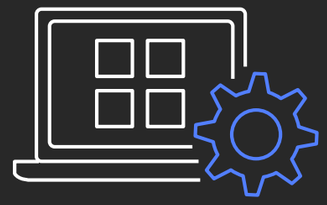
Automatically optimize neural-net compute

AWS Neuron Compiler highlights



Smart partitioning

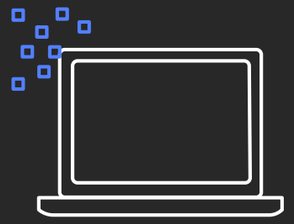
Automatically optimize neural-net compute



FP32 auto casting

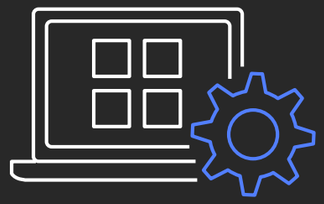
Ingest FP32 trained models, and Neuron auto casts to BF16

AWS Neuron Compiler highlights



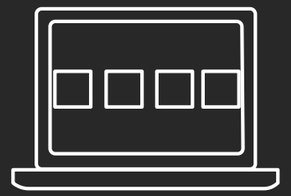
Smart partitioning

Automatically optimize neural-net compute



FP32 auto casting

Ingest FP32 trained models, and Neuron auto casts to BF16

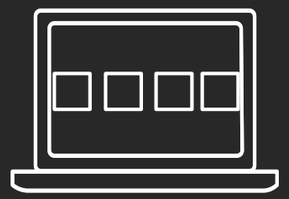


NeuronCore pipeline

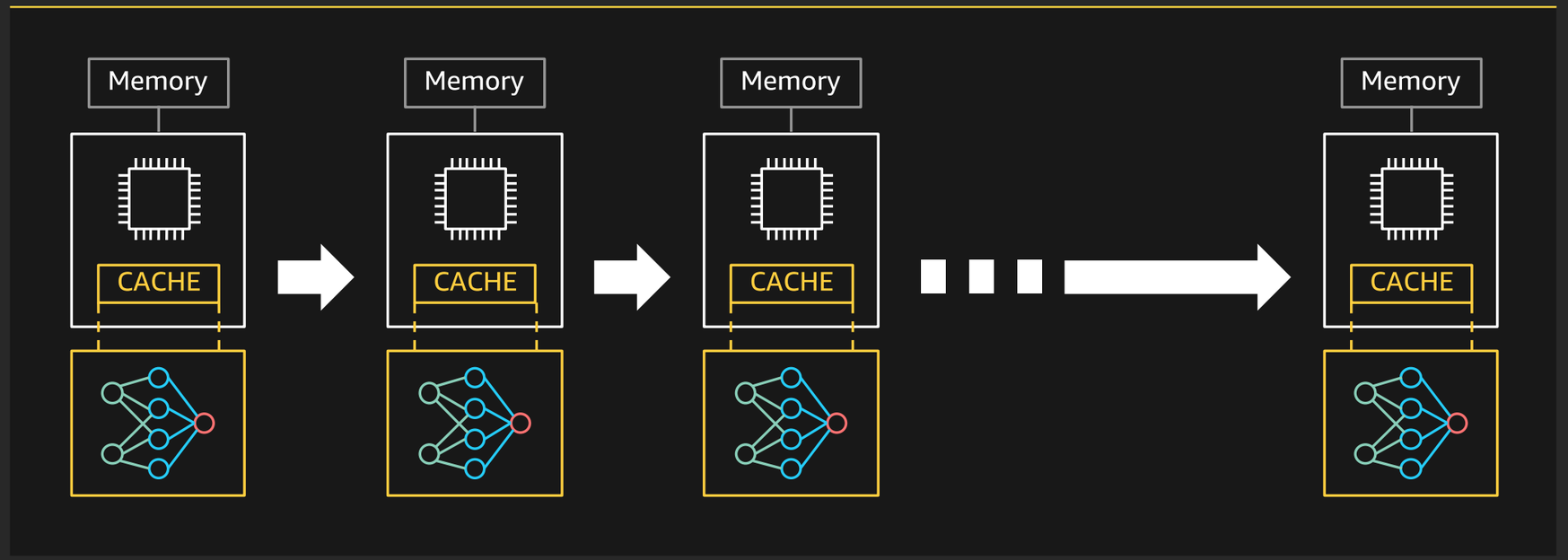
Super low-latency full bandwidth

NeuronCore Pipeline

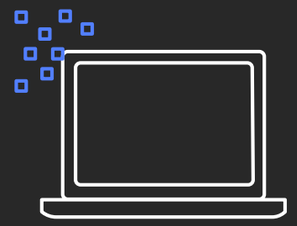
Enabling high-throughput for latency sensitive applications



NeuronCore pipeline

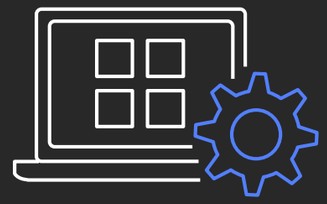


AWS Neuron Compiler highlights



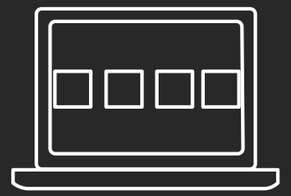
Smart partitioning

Automatically optimize neural-net compute



FP32 auto casting

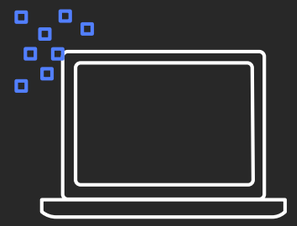
Ingest FP32 trained models, and Neuron auto casts to BF16



NeuronCore pipeline

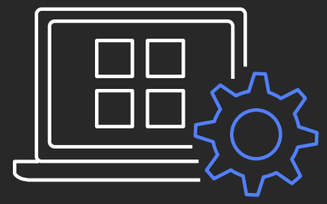
Super low-latency full bandwidth

AWS Neuron Compiler highlights



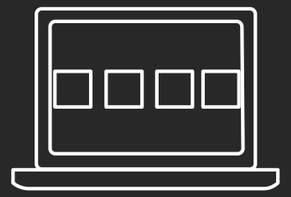
Smart partitioning

Automatically optimize neural-net compute



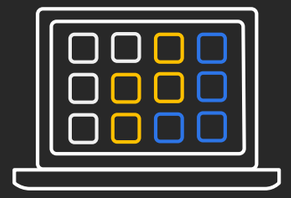
FP32 auto casting

Ingest FP32 trained models, and Neuron auto casts to BF16



NeuronCore pipeline

Super low-latency full bandwidth

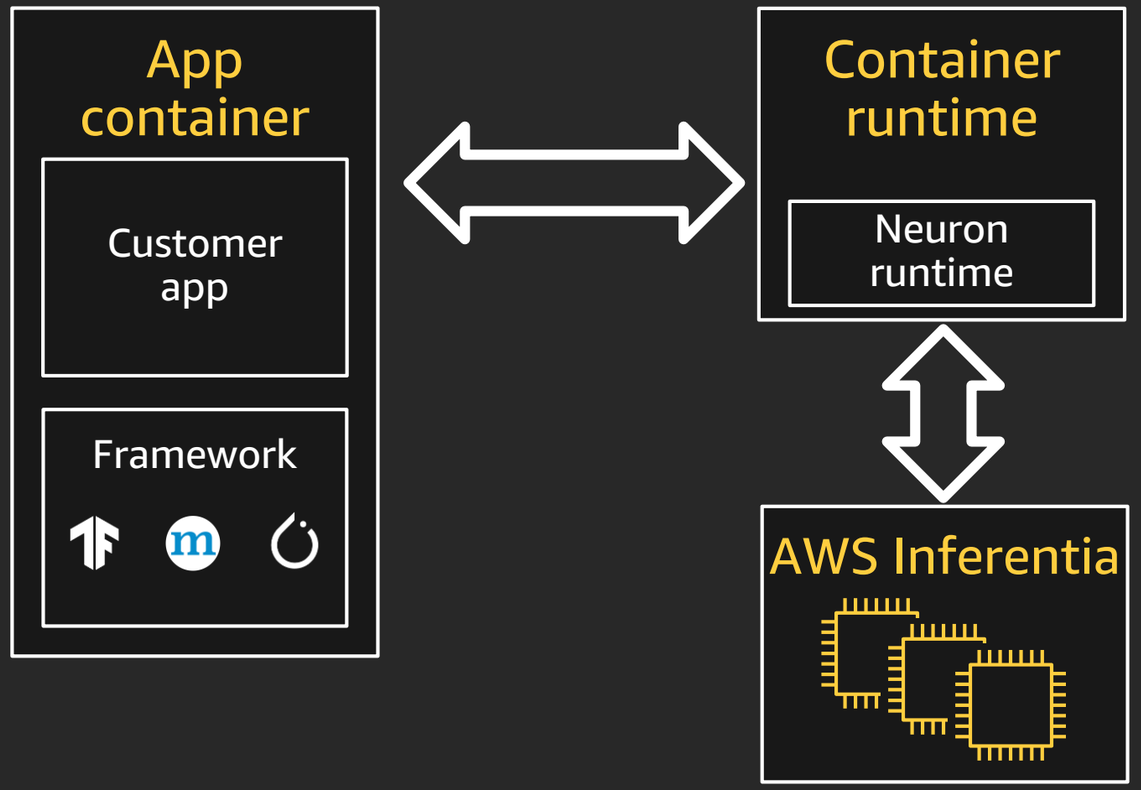


NeuronCore Groups

Concurrently run multiple models

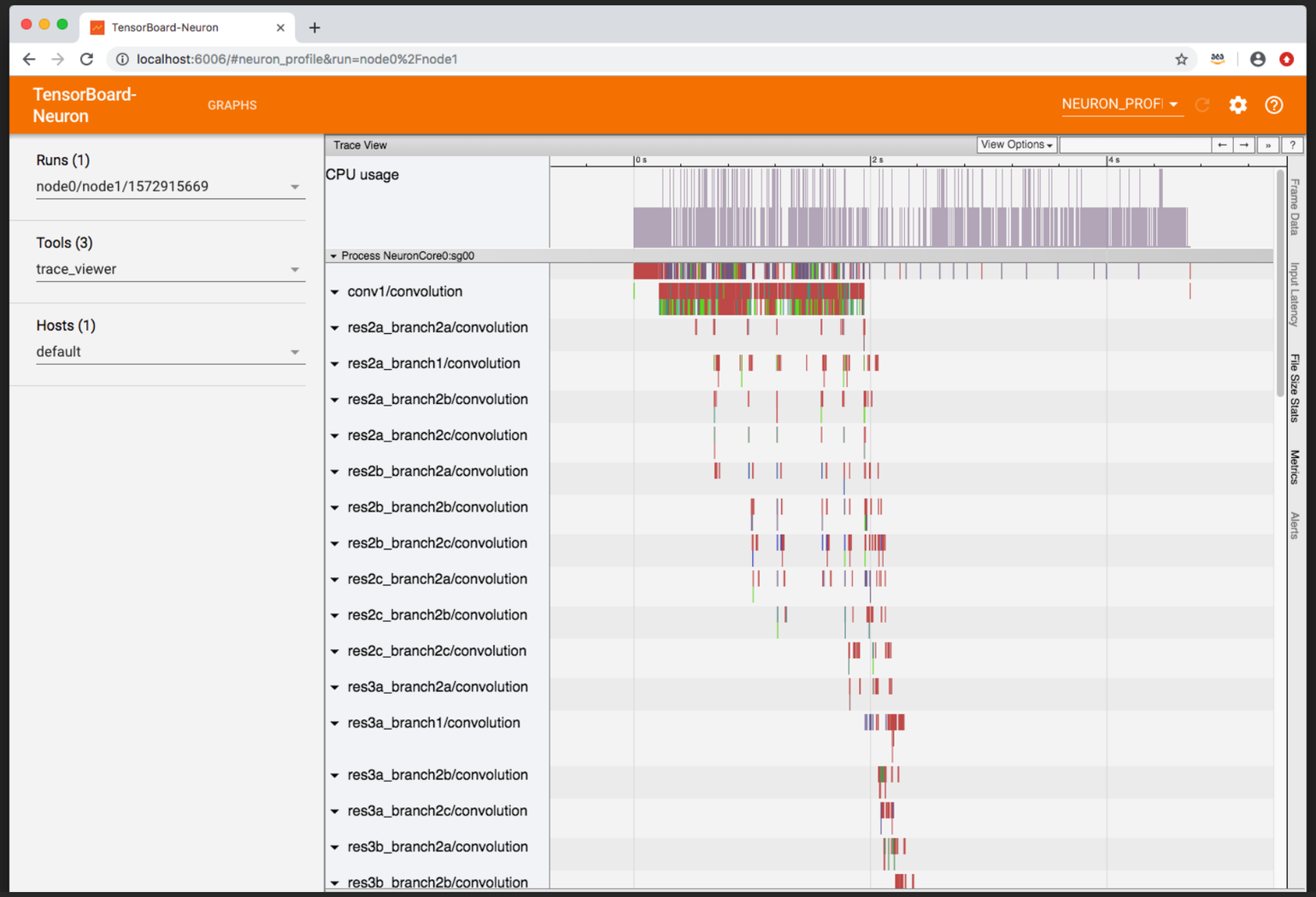
Neuron runtime

Deploy

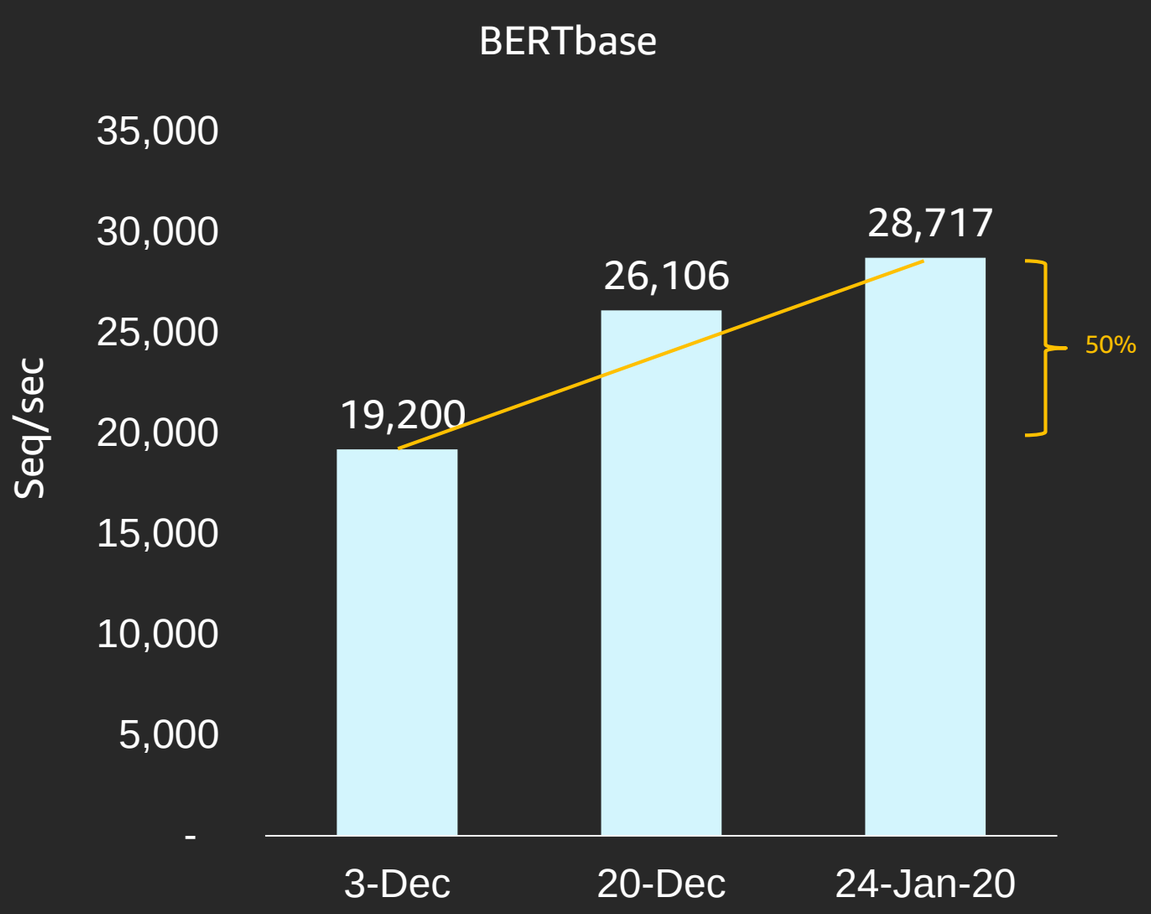


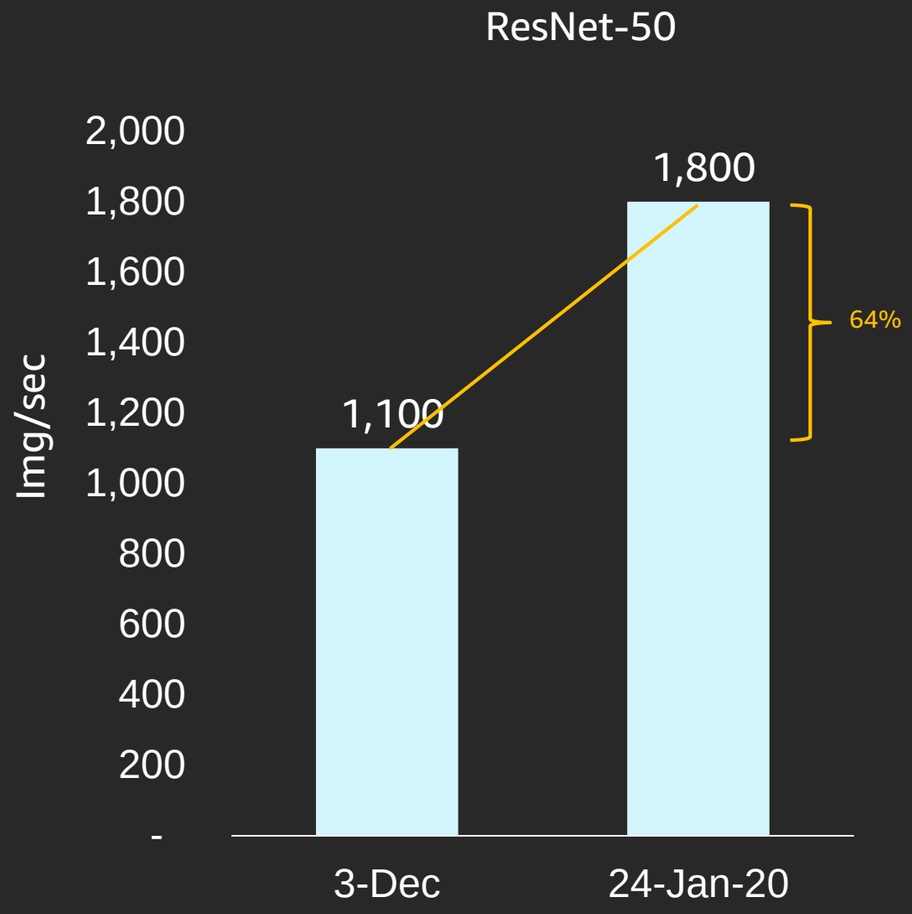
Profiling with Neuron tools

Profile



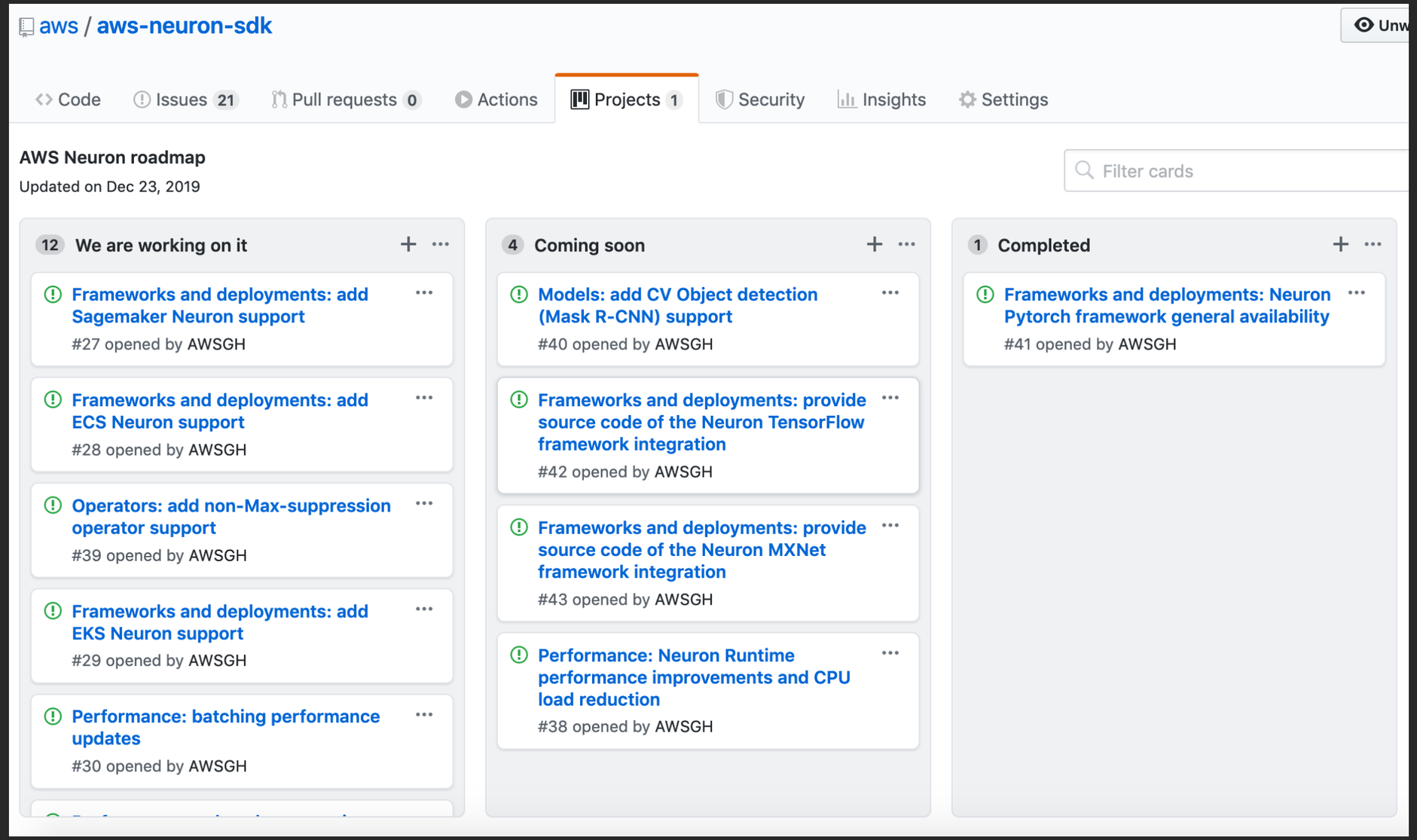
Neuron Performance since Dec-4 Launch





- Results based on running BERT-base and Resnet-50 model end-to-end with TensorFlow
- BERTbase measured with inf1.24xlarge / ResNet50 measured with inf1.1xlarge

AWS Neuron Roadmap



Alexa usage of Inf1 instances

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

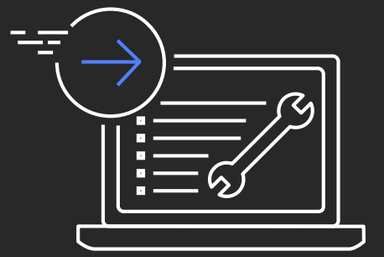
Neural text-to-speech challenges

- Low-latency requirement for dialog system
- High throughput requirement implied by streaming of the output speech
- Context generation is a sequence-to-sequence auto-regressive model
- Memory-bandwidth bound inference
- High temporal density of speech production model resulting with 90GFLOPs to generate 1s of output is Compute-bound inference
- Using EC2 GPU instances to meet requirements results in high operational cost

Alexa TTS migration to EC2 Inf1—Ease of integration

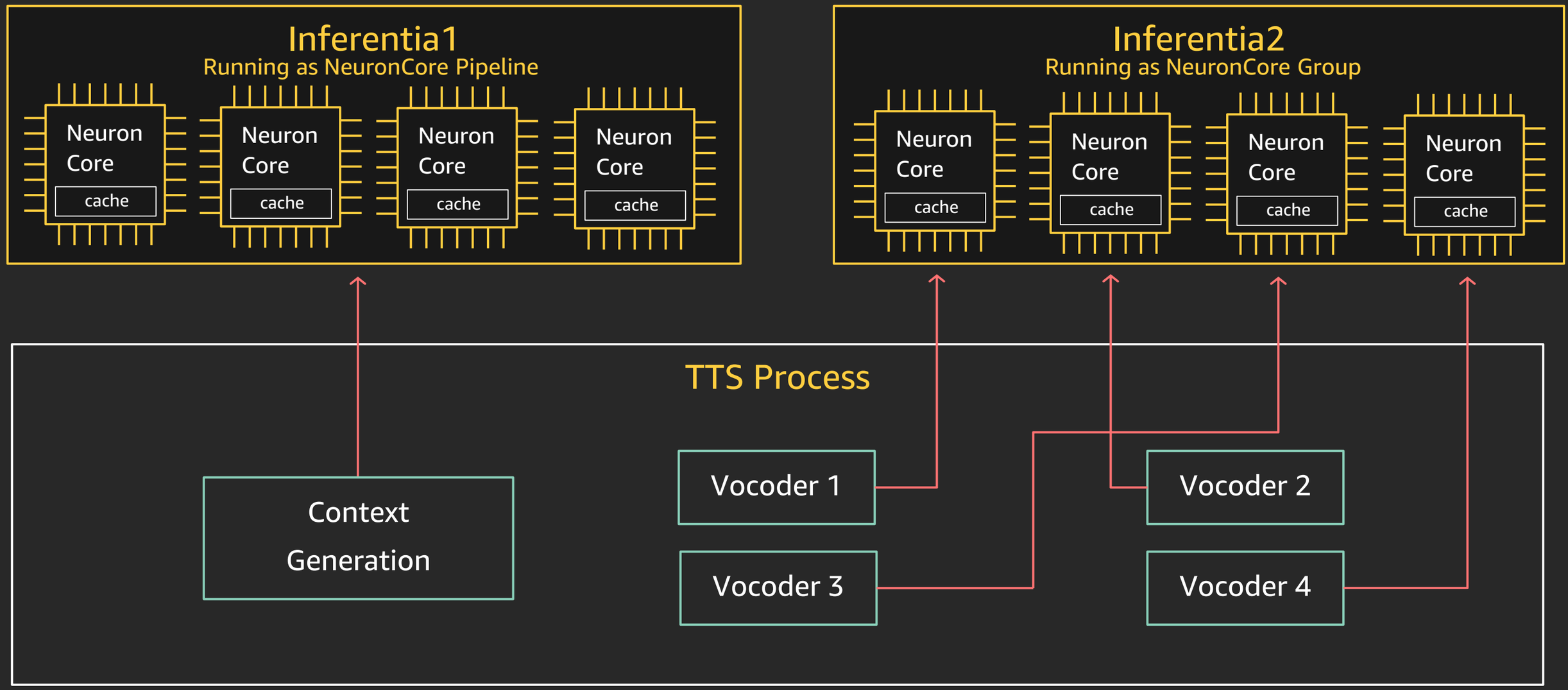
Alexa TTS uses MXNet that is supported natively by AWS Neuron

Support for C and Python APIs



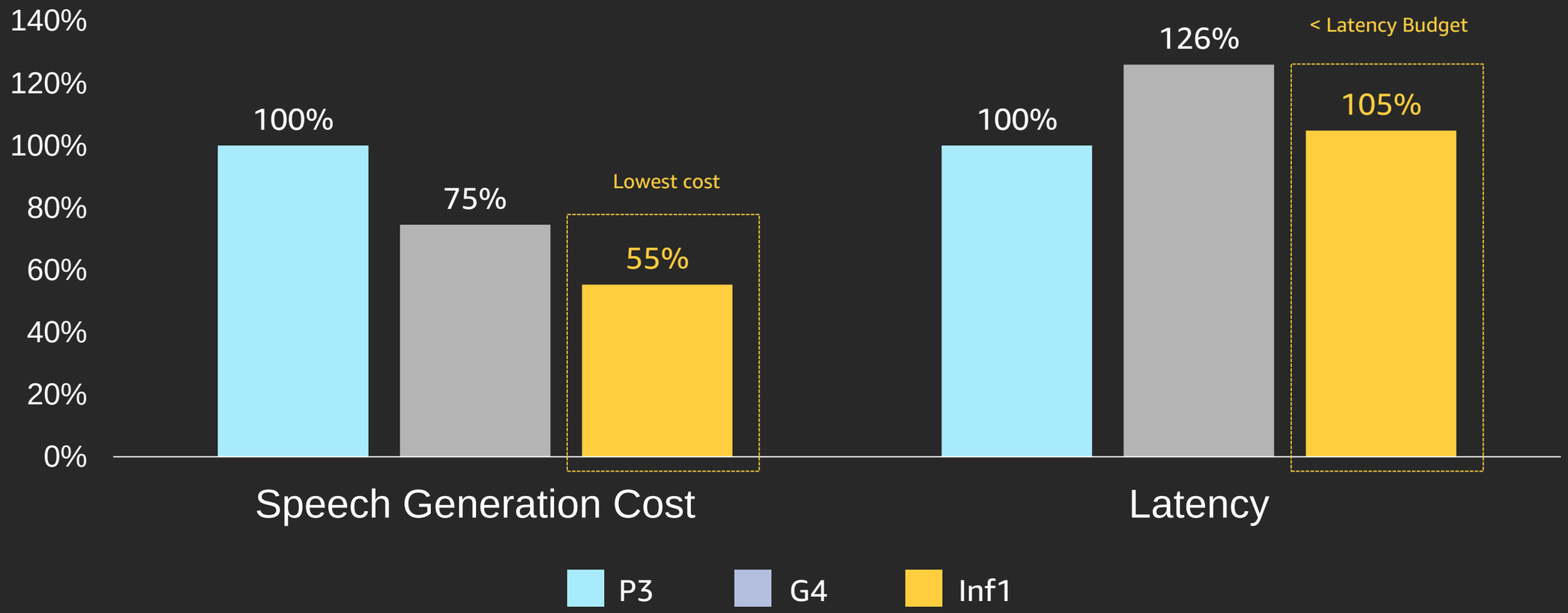
Options to migrate original FP32 models to FP16 or Bfloat16

Alexa TTS migration to EC2 Inf1—Architecture



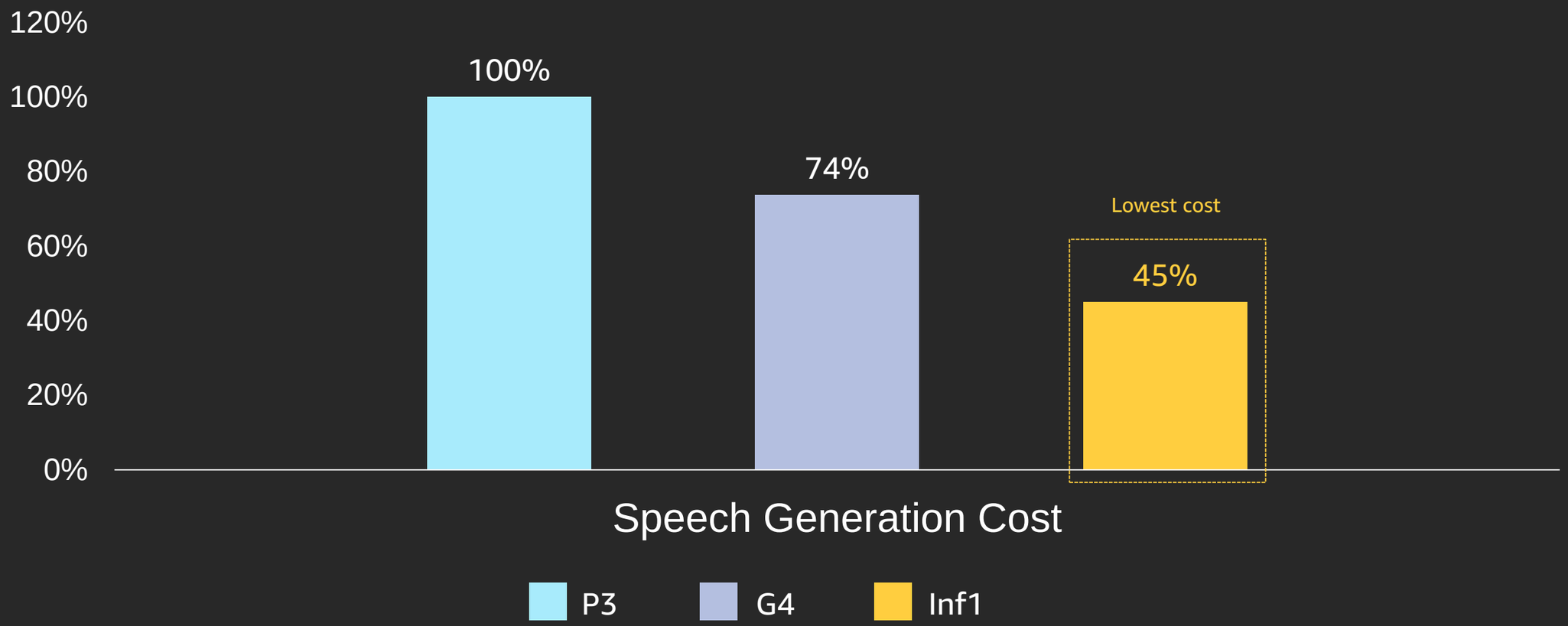
Alexa TTS migration to EC2 Inf1—Gains

Alexa generic traffic is able to get the lowest cost at the same latency budget



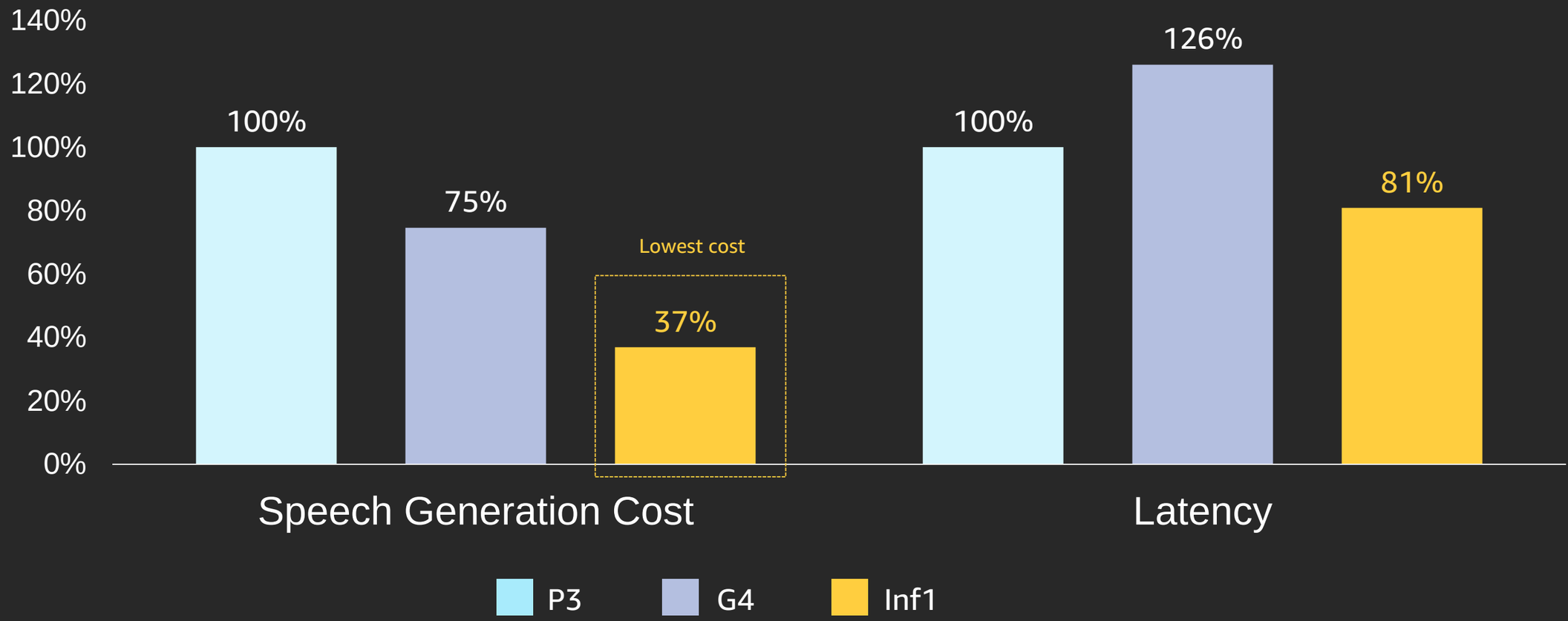
Alexa TTS migration to EC2 Inf1—Long texts

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



Alexa TTS migration to EC2 Inf1—Projected gains

With software optimizations Alexa can take full advantage of Inf1 for significant gains



Try Inf1 instances today!

Inf1 instances are available in US-East-1 (N. Virginia) and US-West-2 (Oregon) today

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	1YR Standard RI (40% discount)	3YR 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

Thank you!

© 2020, Amazon Web Services, Inc. or its Affiliates. All rights reserved.