
aws
INNOVATE
ONLINE CONFERENCE

aws INNOVATE
ONLINE CONFERENCE

分会场一：人工智能与机器学习

AI 基础架构资源的演进与选择—CPU, GPU, ARM 和 ASIC(AWS Inferentia)

王世帅，AWS 机器学习解决方案专家

Agenda

概览

GPU 实例

ARM 实例

机器学习专用推理芯片

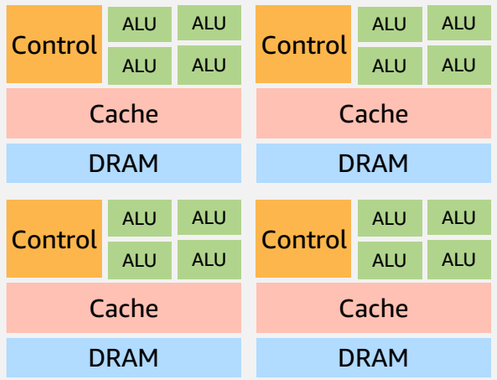
概览

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

aws INNOVATE
ONLINE CONFERENCE

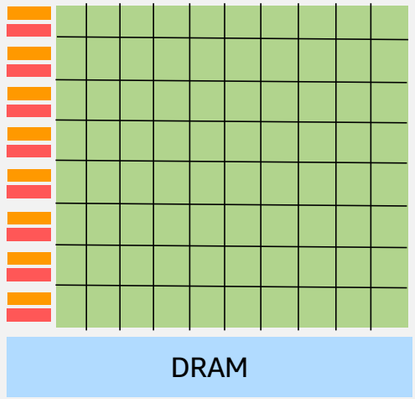
CPU、GPU、FPGA、ASIC 比较

CPU



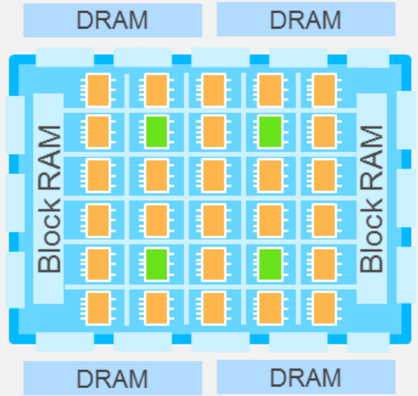
- 数10至数100个处理器核心
- 预定义的指令集和数据通路宽度
- 为通用计算而优化

GPU



- 数1,000个处理器核心
- 预定义的指令集和数据通路宽度
- 并行执行效率高

FPGA

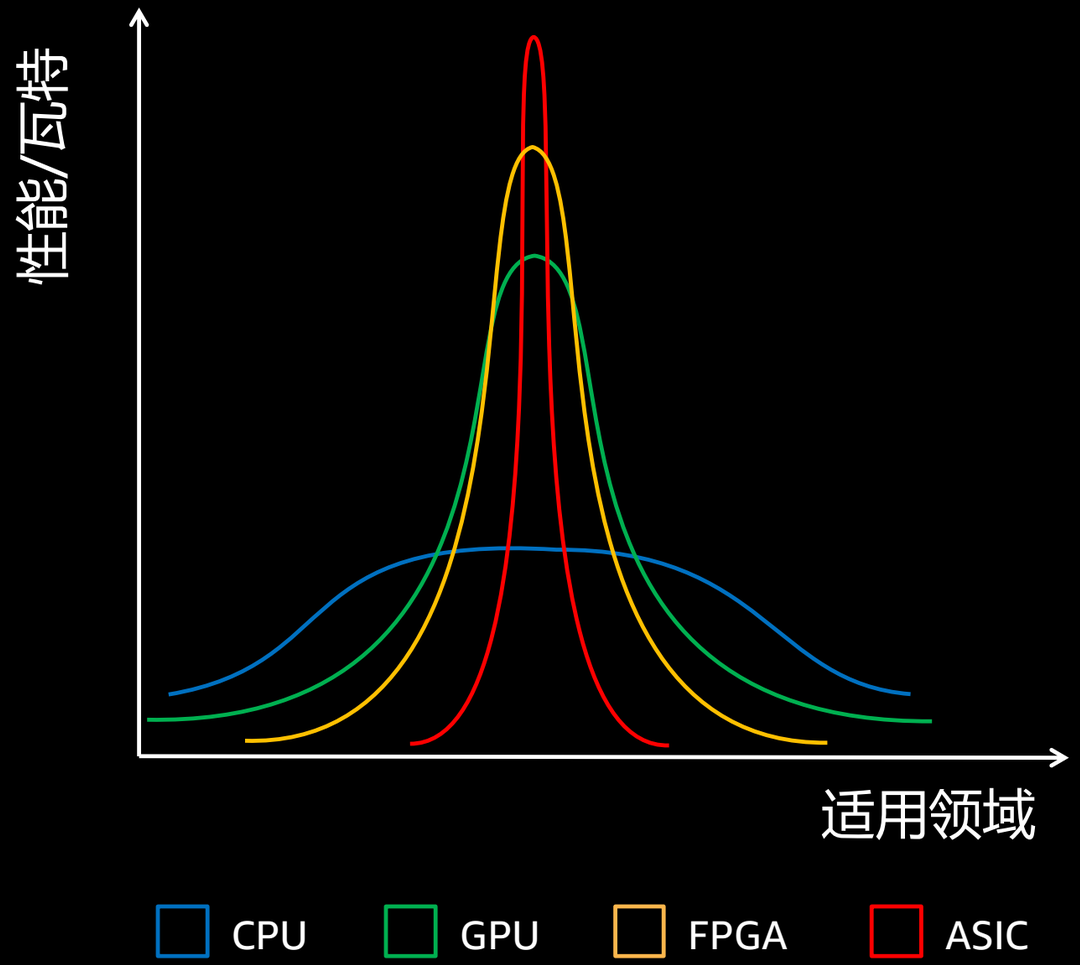


- 数百万个可编程数字逻辑单元
- 无预定义的指令集或数据通路宽度
- 提供硬件定时的速度和稳定性

ASIC

- 针对特定用途/功能的优化和定制设计
- 通过 API 提供预定义软件体验

使用定制芯片优化机器学习性能



示意性图表，仅供参考

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

为机器学习训练选择正确的实例类型

- 实例家族：
  - 优化时间/成本?
  - 模型复杂度
  - 模型能否使用 GPU 做训练?
- 实例大小：
  - 需要多少内存/ CPU 核心?
  - 模型能否利用多个 GPU?
- 实例个数：
  - 模型能否在多台机器上做分布式训练?

GPU 实例

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

aws INNOVATE
ONLINE CONFERENCE

NVIDIA GPU 矩阵

架构	Tesla	GeForce	AWS 实例系列	GPU 发布日期	制造工艺	最小实例类型
Kepler	K80	GTX780	Amazon EC2 P2	2014.11	28nm	p2.2xl 4C/61G
Maxwell	M60	GTX980	Amazon EC2 G3	2015.08	16nm	g3s.xl 4C/31G
Pascal	P4/P100	GTX1080	N/A	2016.09	16nm	N/A
Turing	T4	RTX2080	Amazon EC2 G4	2018.09	12nm	g4dn.xl 4C/16G 125G SSD
Volta	V100	N/A	Amazon EC2 P3	2017.06	12nm	p3.2xl 8C/61G

Amazon GPU 实例类型

P3/P3dn GPU 计算实例

- 可支持高达 8 个 NVIDIA V100 Tensor Core GPU
- 并可为机器学习和 HPC 应用提供高达 100Gbps 的网络吞吐量
- 可以实现最高 1 PetaFLOP 的混合精度性能

G4: GPU 计算实例

- 提供最新一代的 NVIDIA T4 GPU
- 高达100 Gbps 的网络吞吐量
- 高达1.8 TB 的本地 NVMe 存储
- 可在生产和图形密集型应用程序中部署机器学习模型

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

NVIDIA Tesla V100 关键规格

- 采用12纳米制造工艺，具有211亿个晶体管（比上一代 P100 GPU 多40%）
- 与 P100 GPU 相比，其能源效率提高了50%，在相同功率范围内有显著的性能提升
- 每个 GPU 具有专属640个 Tensor 内核，以提高混合精度性能
- 每个 GPU 具有125 TFLOPS 的混合精度性能
- 第二代 NVLink，通过6个 NVLink 端口以25GB/s 的速度实现 GPU 到 GPU 的通信，总计300GB/s
- 16GB GPU 内存，峰值内存带宽为900 GB /秒

用于计算的 GPU

矩阵运算，图像处理，物理仿真均可极大地受益于 GPU

Matrix A

a_{11}	a_{12}	a_{13}
a_{21}	a_{22}	a_{23}
a_{31}	a_{32}	a_{33}

×

Matrix B

b_{11}	b_{12}	b_{13}
b_{21}	b_{22}	b_{23}
b_{31}	b_{32}	b_{33}

=

Matrix C

$a_{11}.b_{11} + a_{12}.b_{21} + a_{13}.b_{31}$	$a_{11}.b_{12} + a_{12}.b_{22} + a_{13}.b_{32}$	$a_{11}.b_{13} + a_{12}.b_{23} + a_{13}.b_{33}$
$a_{21}.b_{11} + a_{22}.b_{21} + a_{23}.b_{31}$	$a_{21}.b_{12} + a_{22}.b_{22} + a_{23}.b_{32}$	$a_{21}.b_{13} + a_{22}.b_{23} + a_{23}.b_{33}$
$a_{31}.b_{11} + a_{32}.b_{21} + a_{33}.b_{31}$	$a_{31}.b_{12} + a_{32}.b_{22} + a_{33}.b_{32}$	$a_{31}.b_{13} + a_{32}.b_{23} + a_{33}.b_{33}$

数值精度:

- 半精度 (FP16) – 16位浮点数
- 单精度 (FP32) – 32位浮点数
- 双精度 (FP64) – 64位浮点数

新 Tensor 核心

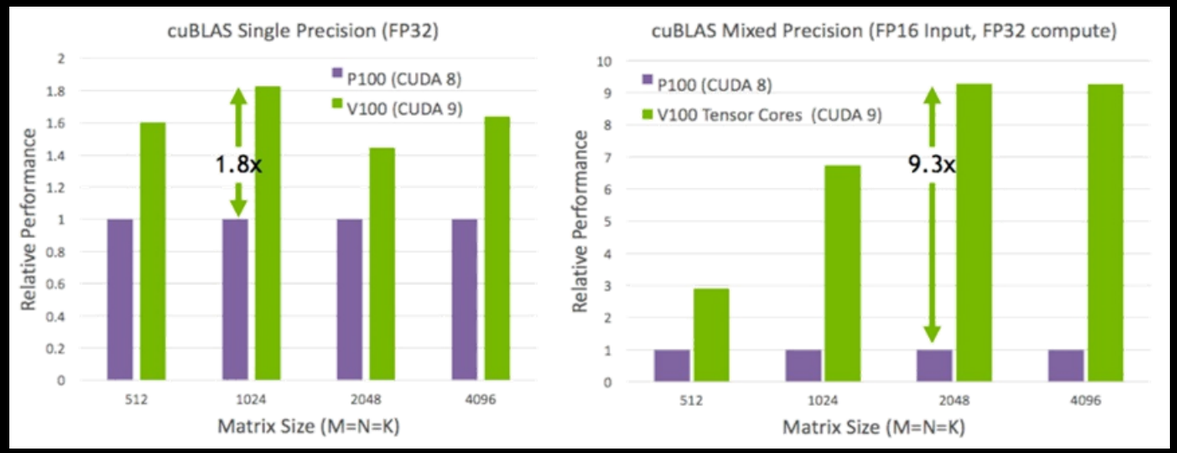
- 除了5,120个 CUDA 核心外，每个 Tesla V100 还具有640个 Tensor 核心，每个 GPU 具有 125 TFLOPS 的混合精度性能。
- 每个 Tensor 核心都提供一个 4x4x4 矩阵处理阵列，该阵列执行运算 D = A X B +C。

$$D = \underset{\text{(FP16)}}{A} \times \underset{\text{(FP16)}}{B} + \underset{\text{(FP16 or FP32)}}{C}$$

(FP16 or FP32)

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \quad B = \begin{bmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \\ b_{31} & b_{32} & b_{33} & b_{34} \\ b_{41} & b_{42} & b_{43} & b_{44} \end{bmatrix} \quad C = \begin{bmatrix} c_{11} & c_{12} & c_{13} & c_{14} \\ c_{21} & c_{22} & c_{23} & c_{24} \\ c_{31} & c_{32} & c_{33} & c_{34} \\ c_{41} & c_{42} & c_{43} & c_{44} \end{bmatrix}$$

Tesla V100 GPU



Source: NVIDIA

Amazon EC2 P3 实例矩阵

Instance size	GPUs	GPU memory	GPU peer to peer	vCPUs	CPU type	Memory (GB)	Network bandwidth	Amazon EBS bandwidth	Local instance storage
P3.2xlarge	1 x V100	16 GB/GPU	No	8	Broadwell	61	Up to 10 Gbps	1.7 Gbps	NA
P3.8xlarge	4 x V100	16 GB/GPU	NVLink	32	Broadwell	244	10 Gbps	7 Gbps	NA
P3.16xlarge	8 x V100	16 GB/GPU	NVLink	64	Broadwell	488	25 Gbps	14 Gbps	NA
P3dn.24xlarge	8 x V100	32 GB/GPU	NVLink	96	Skylake	768	100 Gbps	14 Gbps	2 TB NVMe

适合大型模型以及更高的 batch sizes

96 Skylake vCPUs 支持 AVX-512指令集

100 Gbps 网络带宽适合大规模分布式训练以及快速获取训练数据

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

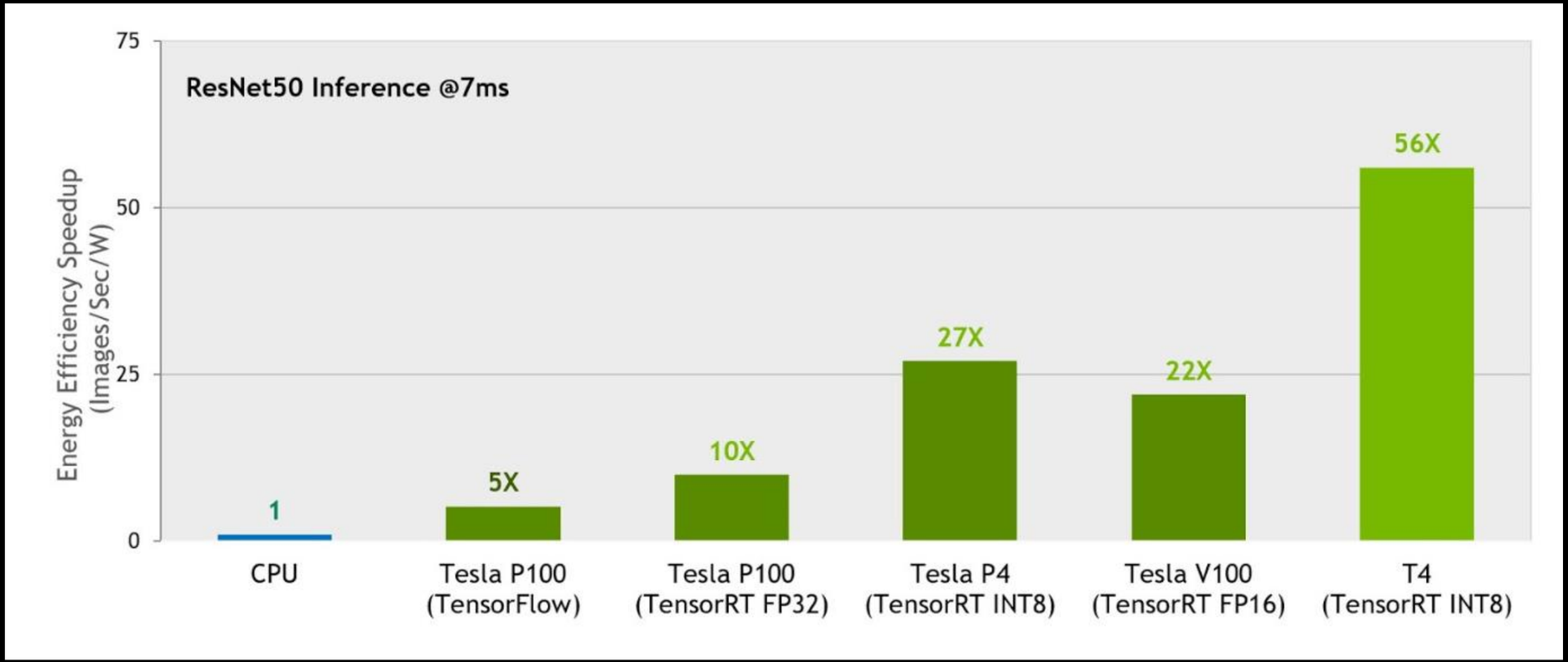
Amazon EC2 G4 实例矩阵

规格

GPU 架构	NVIDIA Turing
NVIDIA Turing Tensor 核心数量	320
NVIDIA CUDA® 核心数量	2560
单精度	8.1 TFLOPS
混合精度 (FP16/FP32)	65 TFLOPS
INT8	130 TOPS
INT4	260 TOPS
GPU 显存	16 GB GDDR6 300 GB/s
ECC	支持
互联带宽	32 GB / 秒
系统接口	x16 PCIe Gen3
外形尺寸	PCIe 半高卡
散热解决方案	被动式
计算 API	CUDA NVIDIA TensorRT™ ONNX

Instance Size	NVIDIA T4 Tensor Core GPUs	vCPUs	RAM	Local Storage
g4dn.xlarge	1	4	16 GiB	1 x 125 GB
g4dn.2xlarge	1	8	32 GiB	1 x 225 GB
g4dn.4xlarge	1	16	64 GiB	1 x 225 GB
g4dn.8xlarge	1	32	128 GiB	1 x 900 GB
g4dn.12xlarge	4	48	192 GiB	1 x 900 GB
g4dn.16xlarge	1	64	256 GiB	1 x 900 GB

Nvidia T4 相比 CPU 推理的能效比提高了50倍



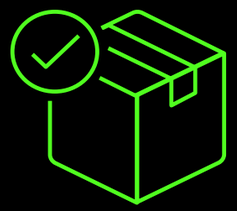
AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

Amazon Deep Learning AMI：深度学习框架镜像

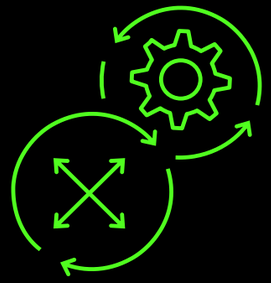
MASK R-CNN 训练测试

- 预配置了 TensorFlow、PyTorch、Apache MXNet、Chainer、Microsoft Cognitive Toolkit、Gluon、Horovod 和 Keras，让您可以快速地大规模部署和运行这些框架和工具。
- AWS Deep Learning AMI 内置的 AWS 优化的 TensorFlow、PyTorch、MXNet 框架，相比于开源框架有显著的性能提升。
- AWS Deep Learning AMI 不额外收费 – 您只为存储和运行应用程序时所用的 AWS 资源支付相关费用。

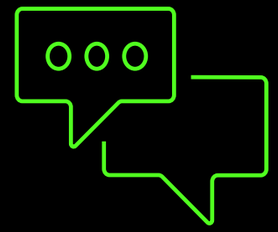
AWS Deep Learning Containers
针对深度学习环境的优化和可定制的容器



预打包的 Docker 容器映像
包含 AWS 优化的
TensorFlow 和 MXNet 框架



无需调整即可获得最
佳性能和可扩展性



与 Amazon EKS,
Amazon ECS, Amazon
EC2, Sagemaker 配合使用

关键特性

可扩展的容器镜像

支持
TensorFlow mxnet
PYTORCH HOROVOD

支持单节点和多节点的训练和推理

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

机器学习推理与训练

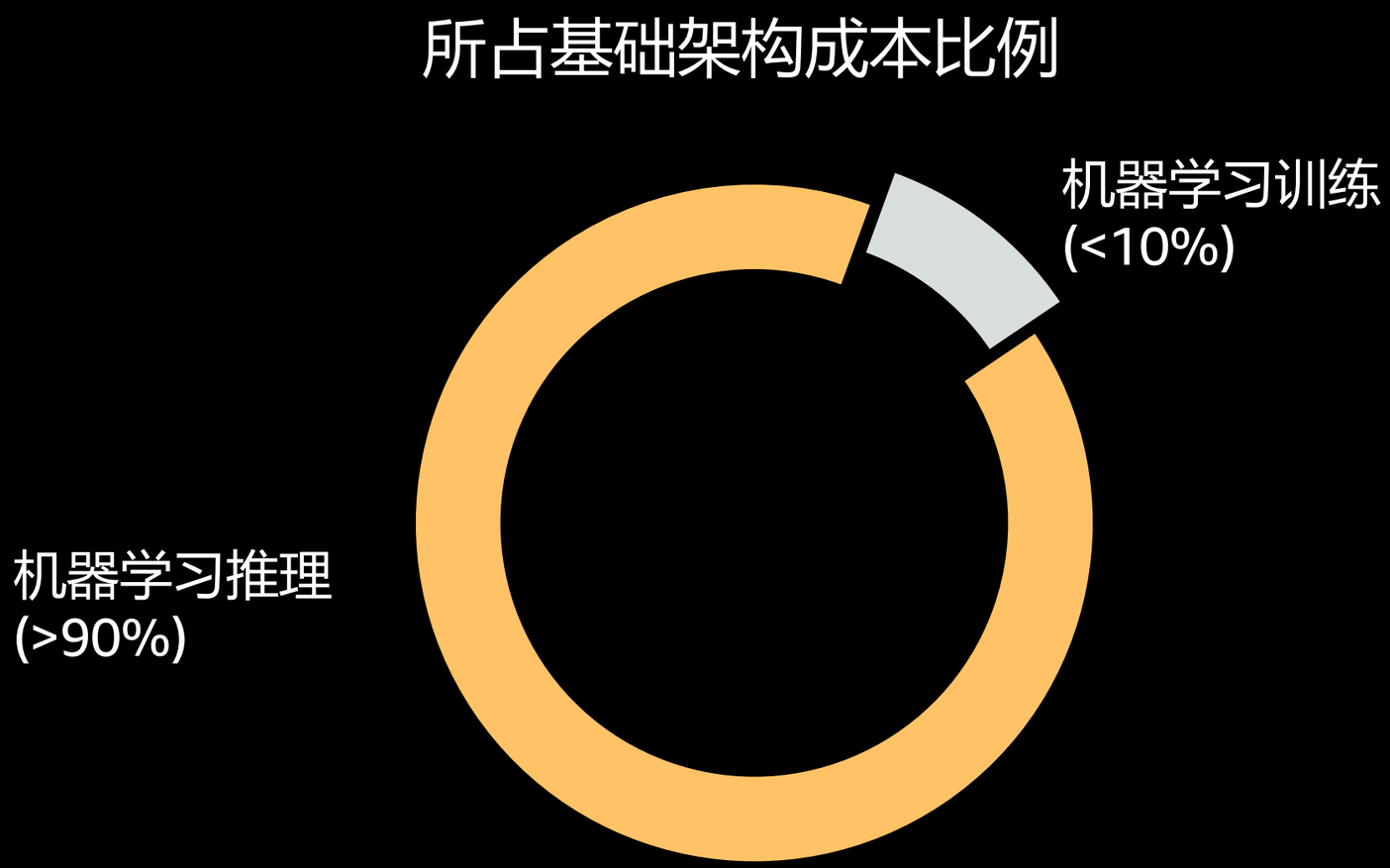
训练

- 需要高并行度和大批量处理能力，以实现更高的吞吐量
- 计算/内存密集型
- 独立，未集成到应用程序堆栈中
- 在云端运行
- 通常运行频率较低（训练一次，不经常重复）

推理

- 通常在单个输入上实时运行
- 非计算/内存密集型
- 集成到应用程序堆栈工作流程中
- 在边缘和云中的不同设备上运行
- 一直运行

“推理”成本在机器学习基础架构成本中占很大比例



AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

ARM 实例

aws INNOVATE
ONLINE CONFERENCE

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

处理器的最广泛选择

Intel® Xeon Scalable processors

AMD EPYC processors

Graviton processors

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

AWS Graviton2 处理器

AWS Graviton2 处理器

AWS Graviton2 处理器

4X	5X	7X
更多的计算核心	更快的内存读写	性能提升

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

AWS Graviton2 支持6款实例类型

通用型	计算优化型	内存优化型
4GB DRAM/vCPU	2GB DRAM/vCPU	8GB DRAM/vCPU
M6g	**C6g**	**R6g**
M6gd	**C6gd**	**R6gd**
M6g 目前提供预览	Coming in 2020	

全部具有增强网络功能，EBS，部分具有本地 NVMe SSD

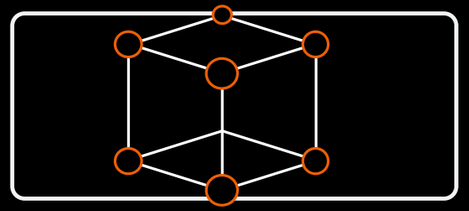
AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

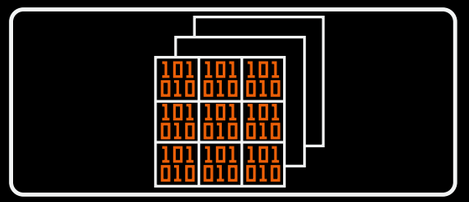
广泛的应用场景

Web 和游戏服务器

开源数据库

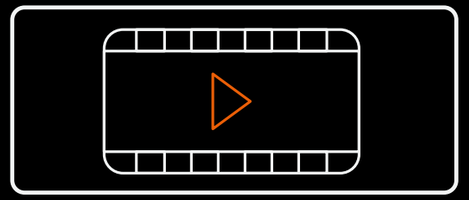
高性能计算

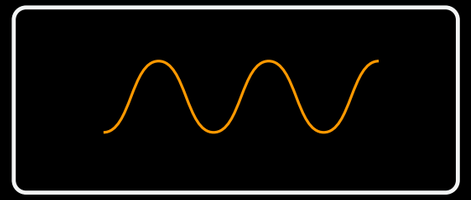




内存缓存

媒体编码





EDA

数据分析

微服务

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

场景举例：机器学习

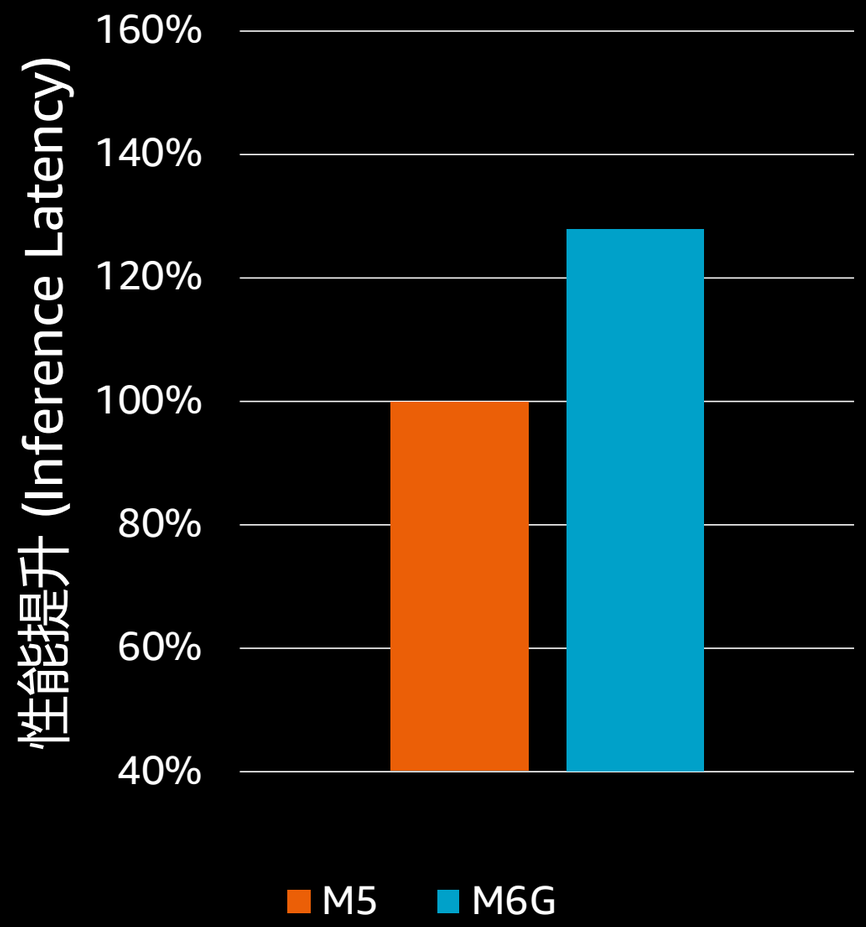
BERT: 目前最好的文本编码器

- 特征表示（编码）对于机器学习至关重要
- 深度神经网络用于文本特征表示

Graviton2 支持 fp16 和 int8

M6g 相较于 M5 的优势：

- M5 AVX-512 仅限于 FP32
- 使用 FP16， M6g 在推理方面相较于 CPU 表现更好



BERT 分类器，使用 TVM，序列长度64
Batch size 1; xlarge 型号实例

机器学习推理专用芯片

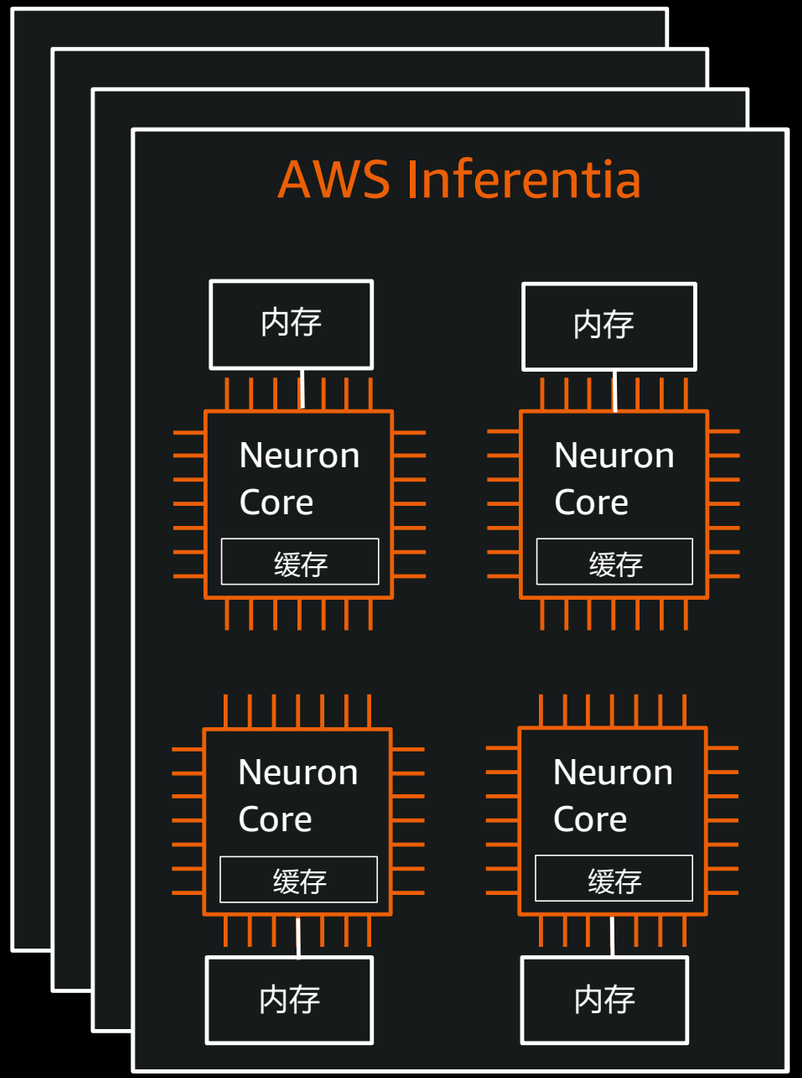
AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

aws INNOVATE
ONLINE CONFERENCE

AWS Inferentia 简介

AWS 独家定制：从芯片、软件到服务器

- 4 个 NeuronCores
- 最高 128 TOPS
- 2 级内存
  - 大容量片上缓存和 DRAM 内存
- 支持 FP16、BF16、INT8 数据类型
- 芯片间高速互联



AWS Neuron 简介

支持在 AWS Inferentia 上实现高性能深度学习推理的软件开发套件

编译器

运行时

性能分析和调试工具

支持所有主流框架

TensorFlow

mxnet

PyTorch

github.com/aws/aws-neuron-sdk

AWS Neuron
支持论坛

AWS Neuron 概览

编译

Neuron 编译器
(NCC)

部署

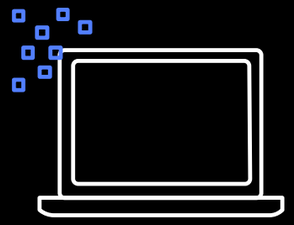
Neuron 运行时
(NRT)

Neuron Binary
(NEFF)

分析

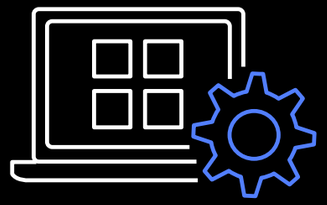
Neuron 工具

Neuron 特色



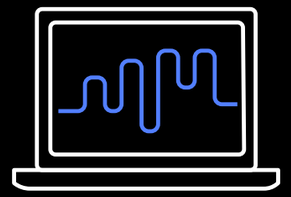
智能分区

自动优化神经网络的计算



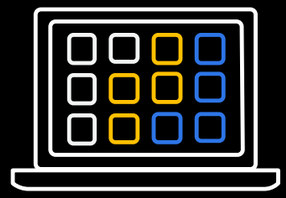
自动 FP32 转换

输入 FP32 训练的模型，Neuron 可自动转为 BF16



NeuronCore Pipeline

超低延迟全带宽

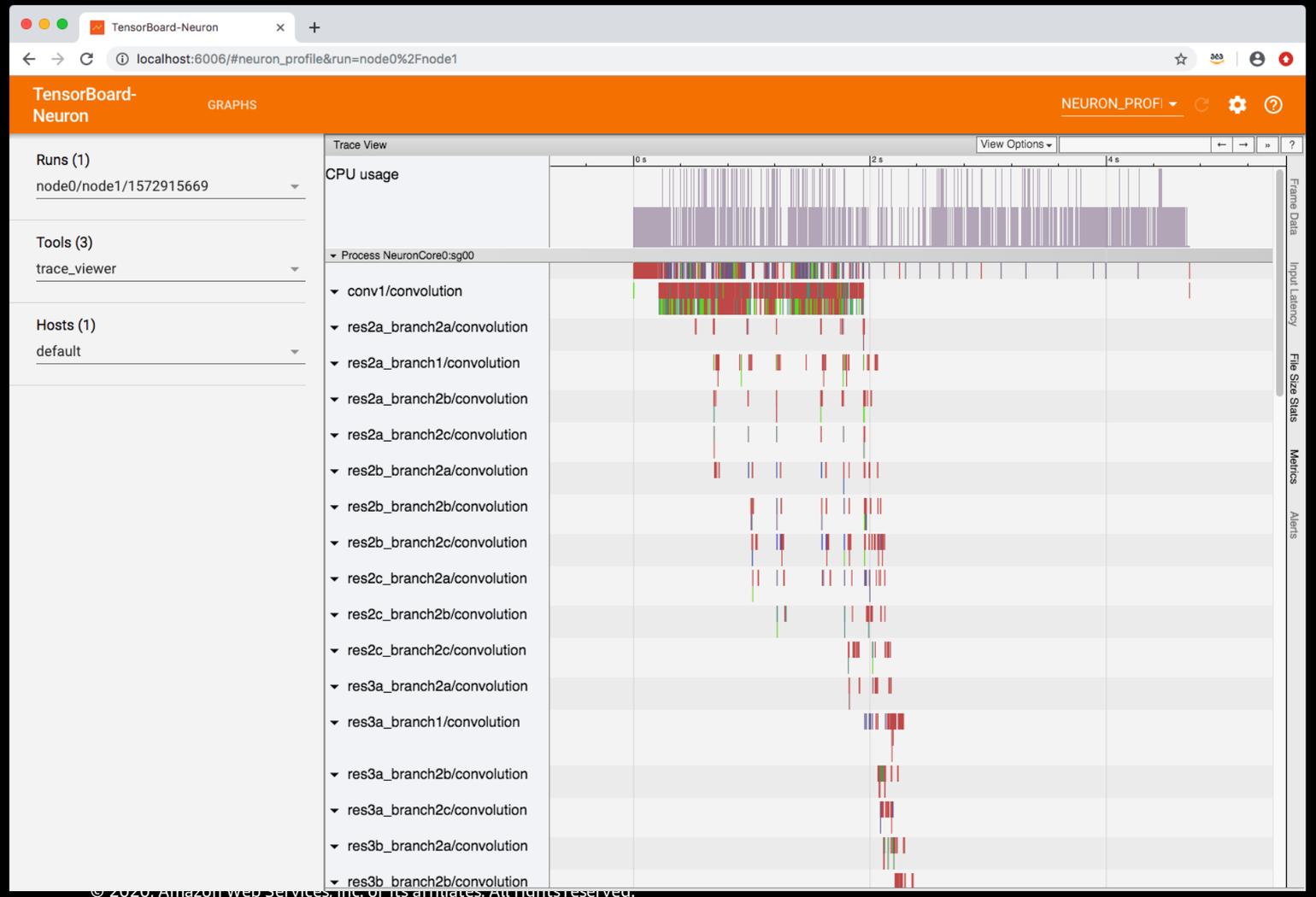


NeuronCore Group

并行运行多个模型

使用 Neuron工具进行性能分析

Profile

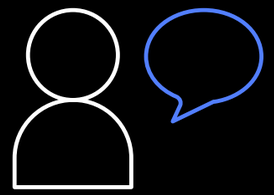


专用于机器学习推理的 Amazon EC2 Inf1 实例

对象检测

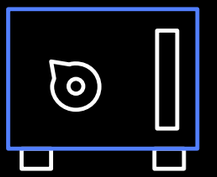
自然语言处理

个性化推荐



语音识别

图像处理



欺诈检测

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

Amazon EC2 Inf1 实例类型一览

实例规模	vCPU	内存 (GiB)	存储	AWS Inferentia 芯片	AWS Inferentia 芯片间互联	网络带宽	Amazon EBS 带宽
inf1.xlarge	4	8	仅 EBS	1	N/A	高达 25 Gbps	高达 3.5 Gbps
inf1.2xlarge	8	16	仅 only	1	N/A	高达 25 Gbps	高达 3.5 Gbps
inf1.6xlarge	24	48	仅 only	4	有	25 Gbps	3.5 Gbps
inf1.24xlarge	96	192	仅 only	16	有	100 Gbps	14 Gbps

- 提供 4 种规格
- 单芯片和多芯片实例

- 第 2 代 Intel Xeon 可扩展处理器
- 最高 100 Gbps 网络带宽

- 将支持应用于 Amazon SageMaker、Amazon EKS 和 Amazon ECS

立即尝试 Inf1 实例！

Inf1 实例目前已在 US-East-1 (N. Virginia) 和 US-West-2 (Oregon) 上线
更多区域 即将提供！

EC2 Inf1 实例支持按需、预留和 Spot 购买选项，此外也可包含在 Savings Plan 中

实例规模	按需	1 年标准预留实例（40% 折扣）	3 年标准预留实例（60% 折扣）
inf1.xlarge	$ 0.368/小时	$ 0.221/小时	$ 0.147/小时
inf1.2xlarge	$ 0.584/小时	$ 0.351/小时	$ 0.234/小时
inf1.6xlarge	$ 1.905/小时	$ 1.143/小时	$ 0.762/小时
inf1.24xlarge	$ 7.619/小时	$ 4.572/小时	$ 3.048/小时

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

Amazon EC2 上的机器学习推理部署选项

定制芯片
EC2 Inf1 实例

特色 AWS Inferentia

- 可使用通用机器学习框架的应用程序
- 由 AWS Inferentia 驱动
- 机器学习推理最佳性价比的实例
  相比 EC2 G4 实例，推理成本最多可降低 40%

已正式发布！

基于 GPU
EC2 G4 实例

- 需要访问 CUDA、CuDNN 或 TensorRT 库的应用程序
- 基于 NVIDIA T4 GPU 的 Amazon EC2 G4 实例

已全面发布！

基于 CPU
EC2 C5 实例

- 对推理性能要求较低的应用程序
- 支持 AVX-512/VNNI 指令集的 Intel Skylake CPU

已全面发布！

Alexa 服务概述

语音命令

文字到语音 (TTS)

Alexa 云服务

识别

识别结果

自动语音识别

识别结果

意图

自然语言理解

意图

“播放”指令

Alexa 技能

文字

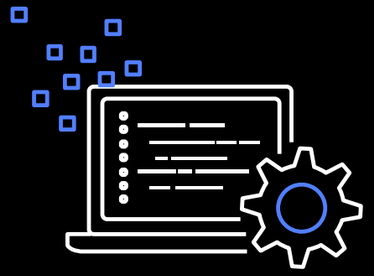
输出语音

文字到语音

应用神经网络文字到语音转换的挑战

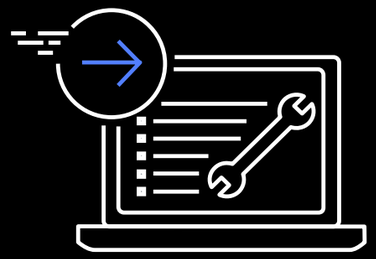
- 输出语音的延迟应尽可能保持最小，吞吐量则应尽可能最大
- 上下文的生成是一种序列到序列模型，推理性能受制于内存带宽
  由于需要处理整条句子，导致出现延迟
- 高密度运算，受制于计算能力，比如需要 90GFLOP 才能生成 1 秒时长的输出语音，需要非常高的吞吐量
- 目前很多用户依然在使用基于 GPU 的 EC2 P3 和 G4 实例进行推理
- 单个推理的成本较高，同时也在探寻低成本，低延迟并且高吞吐率的方法

Alexa TTS 迁移至 Amazon EC2 Inf1 — 易于迁移



Alexa TTS 所使用的 MXNet 已被 AWS Neuron 原生支持

支持 C 和 Python API

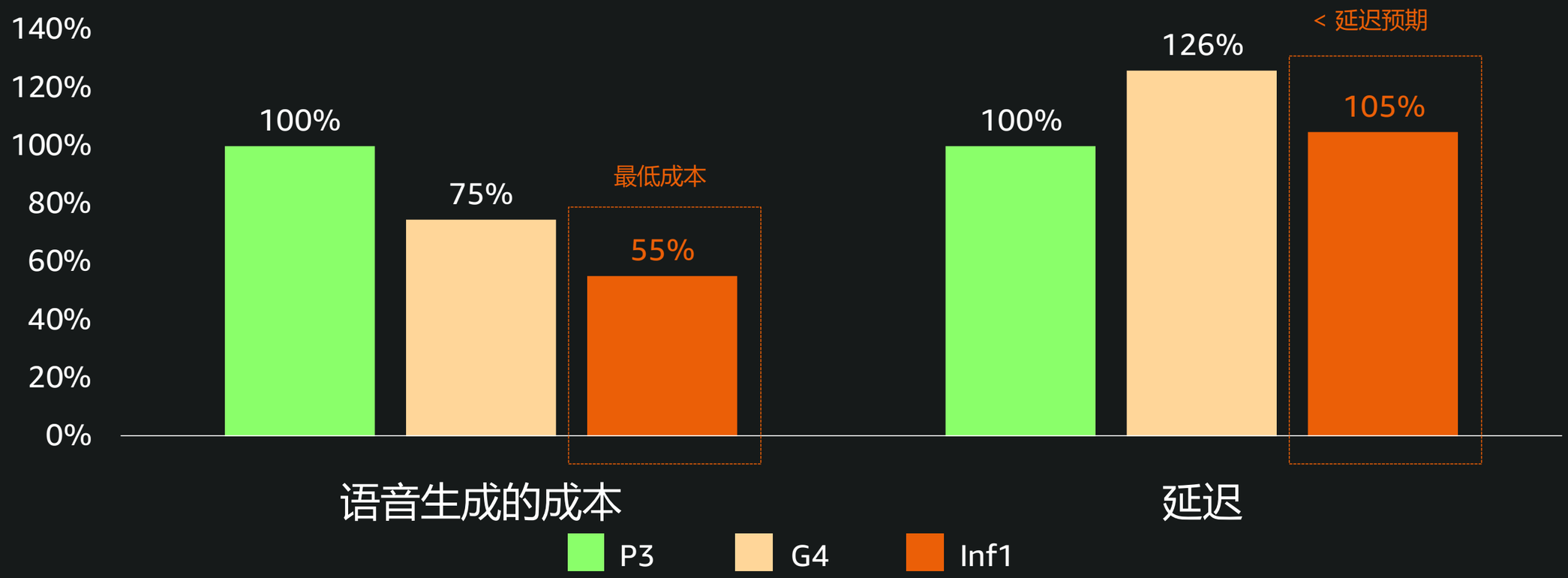


提供选项：
将原始的 FP32 模型迁移为 FP16 或 Bfloat16

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

Alexa TTS 迁移至 Amazon EC2 Inf1

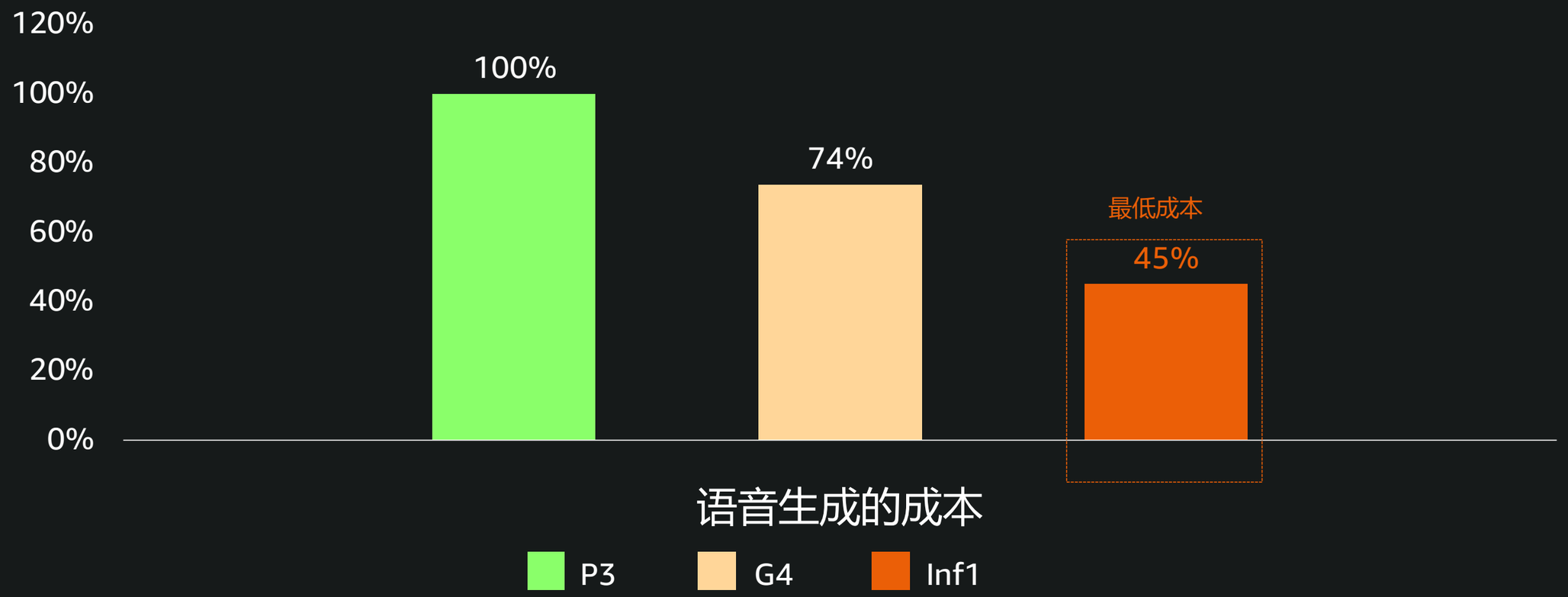
在相同延迟相当的情况下，Alexa 常规语音生成的推理成本可降至最低



AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

Alexa TTS 迁移至 Amazon EC2 Inf1

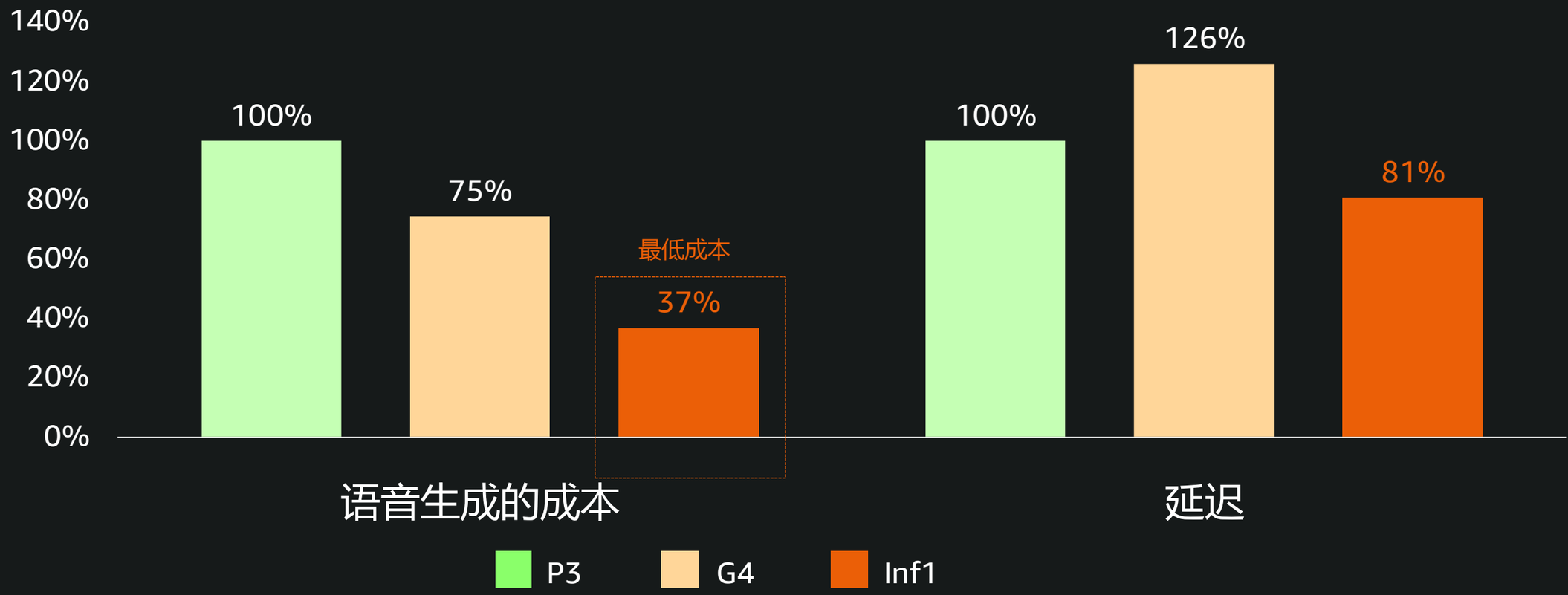
Alexa 长文本语音生成的推理（如图书、新闻）获益更大



AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

Alexa TTS 迁移至 Amazon EC2 Inf1

通过进一步软件优化，Alexa 将能全面发挥 Inf1 的优势进而获得更大收益



AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

感谢参加 AWS INNOVATE 2020 在线技术大会

我们希望您在这里找到感兴趣的内容！

也请帮助我们完成**投票打分**和**反馈问卷**。

欲获取关于 AWS 的更多信息和技术内容，可以通过以下方式找到我们：

微信订阅号：AWS 云计算（awschina）

微信服务号：AWS Builder 俱乐部（amazonaws）

新浪微博：https://www.weibo.com/amazonaws/

抖音：亚马逊云计算（抖音号：266052872）

视频中心：http://aws.amazon.bokecc.com/

博客：https://aws.amazon.com/cn/blogs/china/

更多线上活动 ： https://aws.amazon.com/cn/about-aws/events/webinar/

AWS 中国账户注册

AWS 全球账户注册

AWS INNOVATE ONLINE CONFERENCE

AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营