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

CMP226

How four customers reduced ML inference costs and drove innovation

Kamran Khan

Sr Product Manager
AWS

Jesús Hormigo



Kento Adachi



Zachary Schmidt



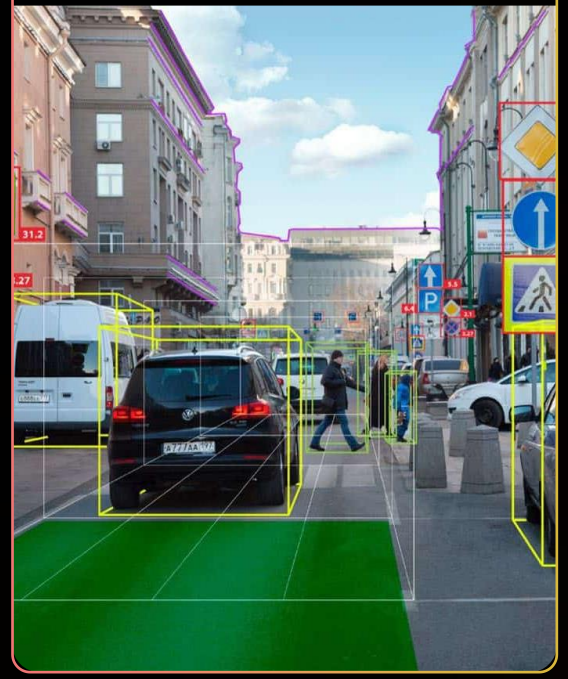
Alejandro Jaimes



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

AI is everywhere

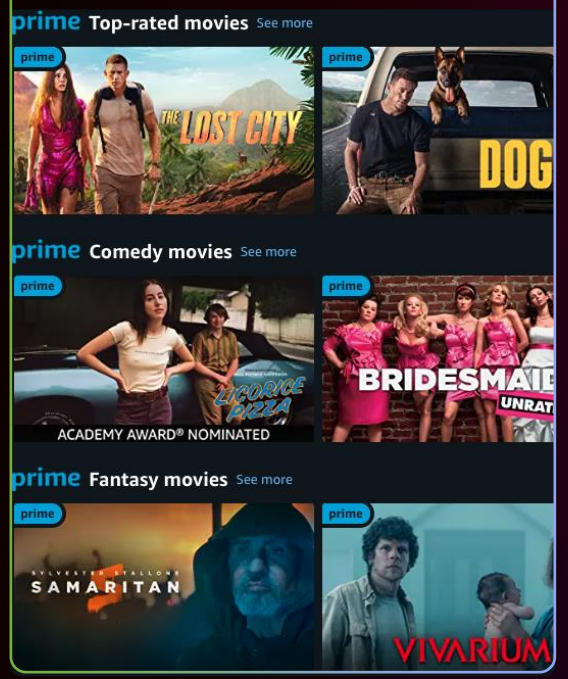
Vision



Language



Recommendation



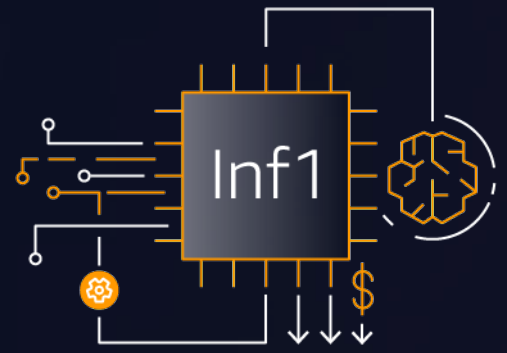
Robotics



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

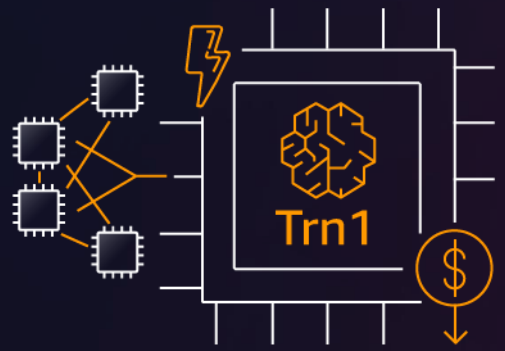
AWS chips optimized for deep learning

AWS Inferentia



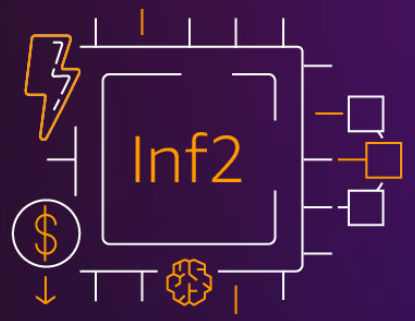
Lowest cost inference in the cloud for running deep learning models— up to 70% lower cost than GPU instances

AWS Trainium



The most cost-efficient high performance DL training instance

AWS Inferentia2



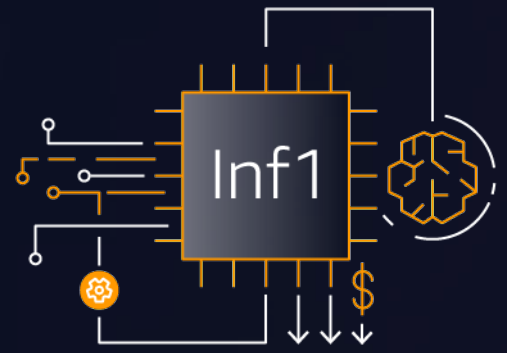
High-performance DL inference for today's largest models

In Preview Now!

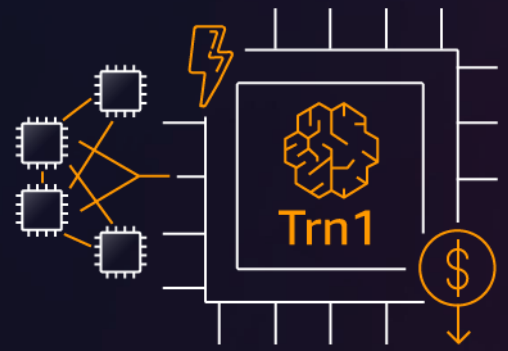
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS chips optimized for deep learning

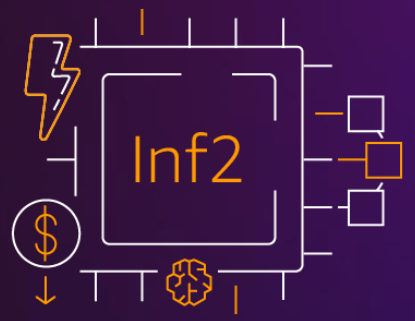
AWS Inferentia



AWS Trainium



AWS Inferentia2



AWS Neuron

Seamless Integration with ML frameworks like PyTorch and TensorFlow

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

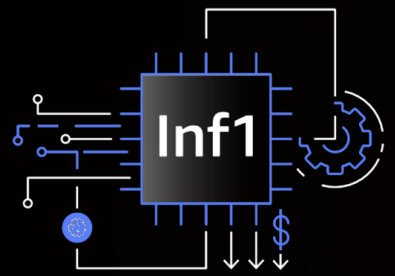
Amazon EC2 Inf1 Instances

INTRODUCED IN 2019, INF1 IS THE FIRST AWS-DESIGNED DL SILICON



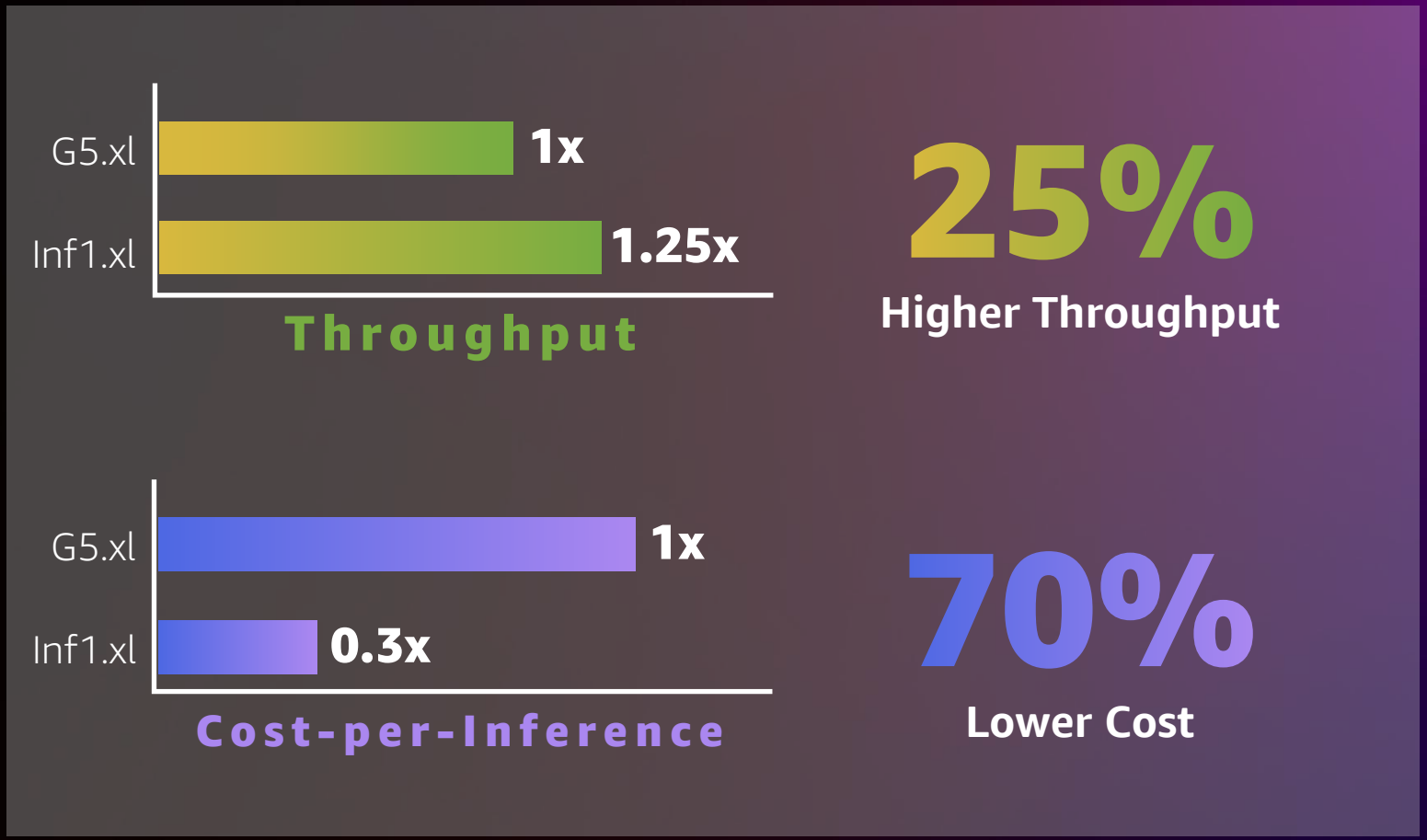
AWS Inferentia

High performance deep learning inference chip, custom designed by AWS



EC2 Inf1 Instances

Lowest Cost Inference in the cloud



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

[1]*Neuron v1.16, fp32, bert-base-uncased, sequence-length=128, PT 1.8, b6*

Amazon Search Saves 85% on Infrastructure Costs



The Amazon Search product search engine indexes billions of products, serves hundreds of millions of customers worldwide, and is one of the most heavily used services in the world. Since, these searches happen in real-time, they need the P99 latency to be under 10 milliseconds to ensure a delightful user experience. **Using AWS Inferentia-based Amazon EC2 Inf1 instances for their Transformer-based NLP models delivers latencies under 3ms and saves up to 85% in infrastructure costs** while maintaining strong throughput.

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

Customer momentum with AWS Inferentia

airbnb

AUTODESK

Finch COMPUTING

sprinklr

NTTPC

amazonrobotics

amazonadvertising

amazon alexa

SCREENING EAGLE

The Asahi Shimbun

CONDÉ NAST

Talroo™

Anthem

DISCO

SKYWATCH

Money Forward

Amazon Rekognition

DMP DIGITAL MEDIA PROFESSIONALS

prime video

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

More customer testimonials

Screening Eagle

Jesús Hormigo

Chief of Cloud & AI Officer



SCREENING EAGLE

aws

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

Screening Eagle Technologies



- We produce built-world technology
- Founded in 2019
- Mission: Protect the built world
- Headquarters in Switzerland
- 300 employees worldwide
- Hardware and software engineering production in-house

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

Protect the Built World

- Enable collection of the built world data
- Determine the condition of structures to secure quality materials and longevity of the built world
- Reduce embodied carbon in the built world by ensuring longevity of structures
- Use of AI to interpret large sets of captured data without the need of a specialist



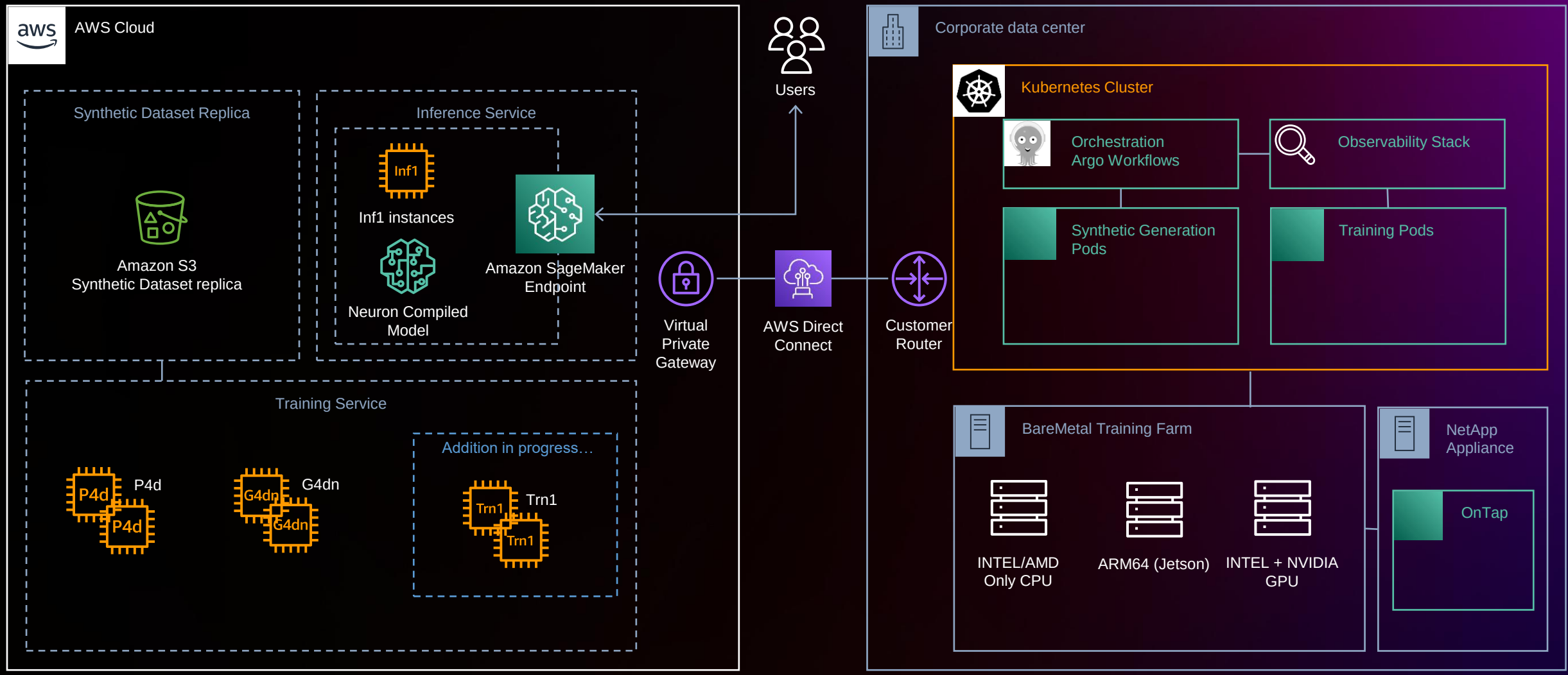
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Our use case

- Autonomous inspection data analysis (batch)
  - Batch processing of videos captured by drones and inspector cameras
  - Large file sizes up to 4K and 60 FPS
  - Return an annotated video / Parallel processing of frames
- Inferencing services (on-demand)
  - Near real-time processing of frames and short sets of data in a service
  - Individual frames (photos) and short time series data captured in-field
  - When inspector device can capture but cannot compute locally

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

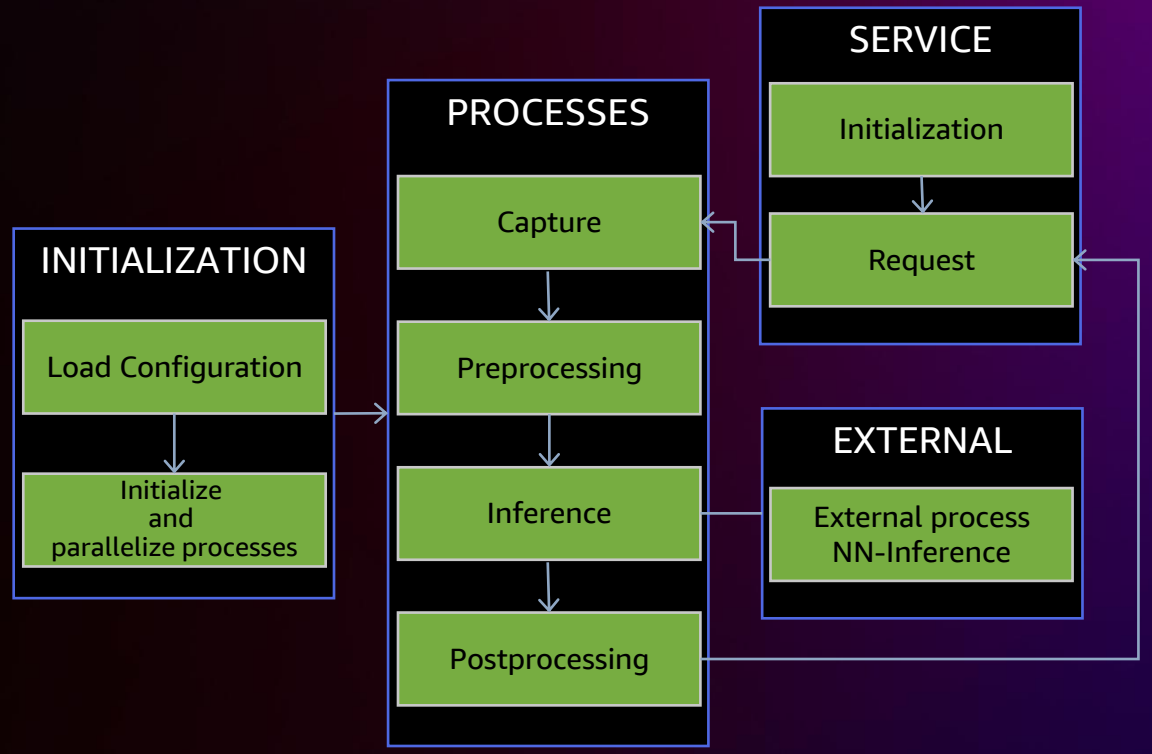
Our infrastructure



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

Screening Eagle's inference service

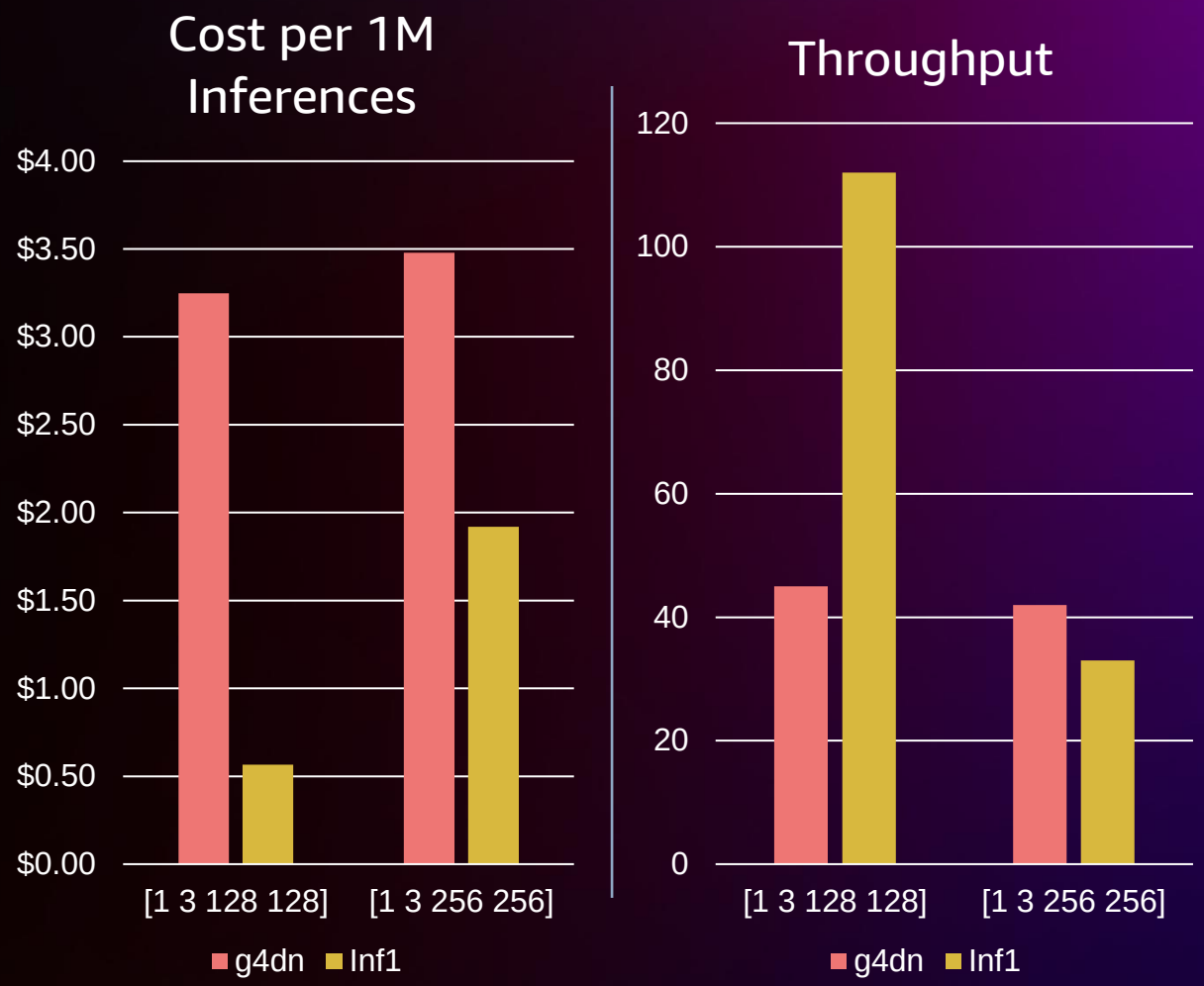
- REST API in a docker that uses the C++ inference-framework libs and Neuron based backend support
- Neuron cores are loaded to the max depending on the model size. Typically, inf1-xlarge runs 4 parallel inferences with or without batching
- Runs on AWS Inf1 instances as SageMaker endpoint



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

Reduced inference costs by 45%

- Transition from GPU to Inf1 was transparent without affecting the user experience
- Production cost was reduced ~45% vs GPU instances cost
- Latency improved in smaller inputs and remained equal in larger ones
- Achieved a cost per 1M inferences of $1.91 (on-demand inf1)



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

Thank you!

Jesús Hormigo
Chief of Cloud and AI Officer
jh@screeningeagle.com
@jesushormigo

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

Getting off the hamster wheel

Deploying a Modern ML Pipeline with AWS SageMaker, Neuron, and Inferentia

Zack Schmidt (he/him)

Data Science Lead

Actuate

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

Agenda

Where we were

Cost savings from new tech

Automation design and implementation

Where we are today

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

actuate

Cloud AI SaaS

Turns any camera into a real-time threat-detecting smart camera

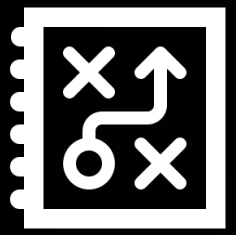
Constant retraining and frequent deployments



aws

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

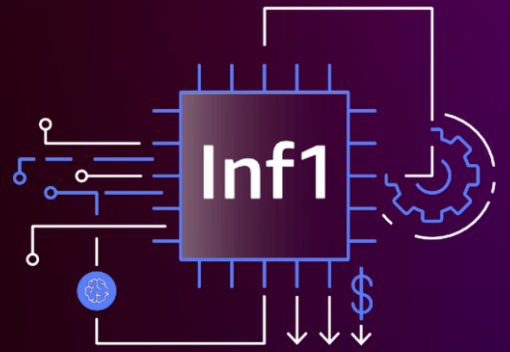
Migration to AWS Inferentia was fast



Moved ML cloud inference from standard GPU machine to Inferentia machine



Compiled existing PyTorch YOLO model using Neuron



Inference costs initially slashed by 70%

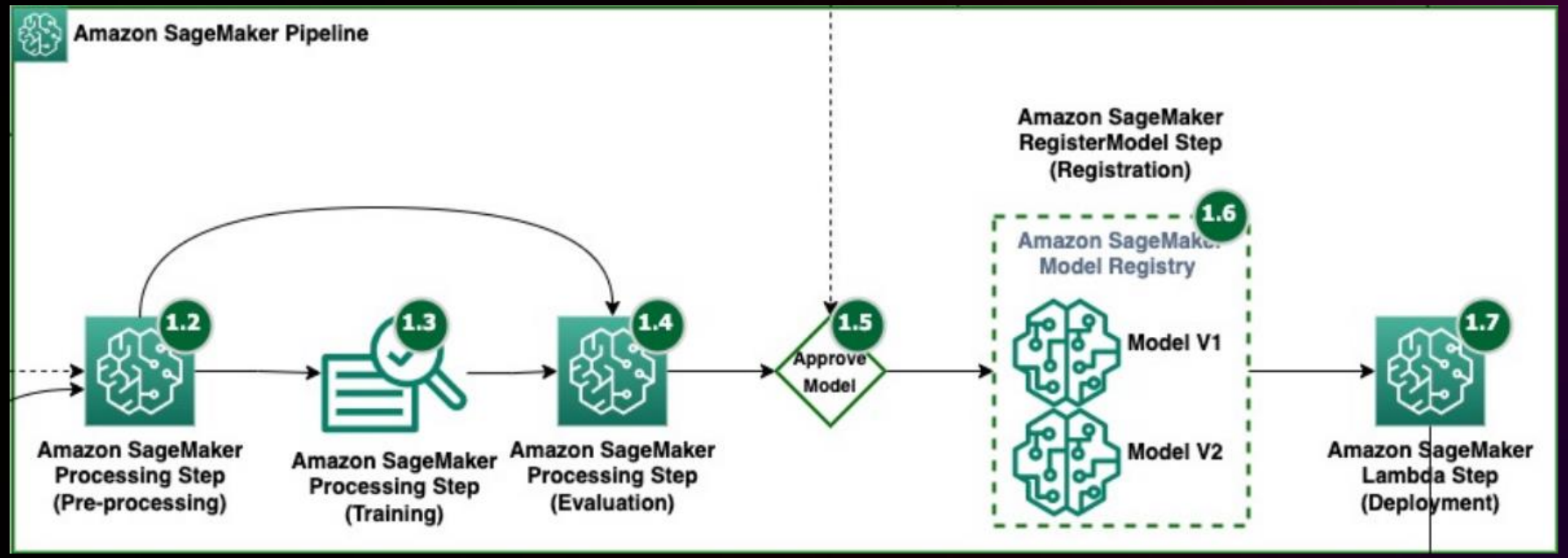
aws

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

Deployment time reduced to 4 minutes
w/SageMaker

Built YOLO and Neuron integration

Labeled data to Inferentia-deployable model: ~4 minutes of dev time



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

With AWS Inferentia we can focus on features

ML inference costs now slashed by 91%

Dev time to click “deploy” shortened from 4 weeks → 4 minutes

Focus shifted to UX improvements and fundamental AI research

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

Thank you!

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



Money Forward

Kento Adachi

Technical lead, CTO office,
Money Forward, Inc.

aws

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

Money Forward, Inc.



- Based on the vision of ‘Becoming the financial platform for all', Money Forward offer various services that solve financial issues for corporates, individuals, and financial institutions.
- On our 10th anniversary, we are providing one of Japan`s Largest User Base and Product Lineup in SaaS/Fintech Field.
- Our team(CTO Office AI Forward Div.) is primarily responsible for developing systems to utilize machine learning models in our services. I am the tech lead for that team.

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

Chatbot service HiTTO

- We provide an in-house DX-specific chatbot service at our group company, HiTTO Co.
- This service is a chatbot service that provides real-time answers to employee questions
- This time we needed to migrate the entire system for this service from a managed environment to an AWS environment in a **short period of two months**.



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

ML use-cases in HiTTO

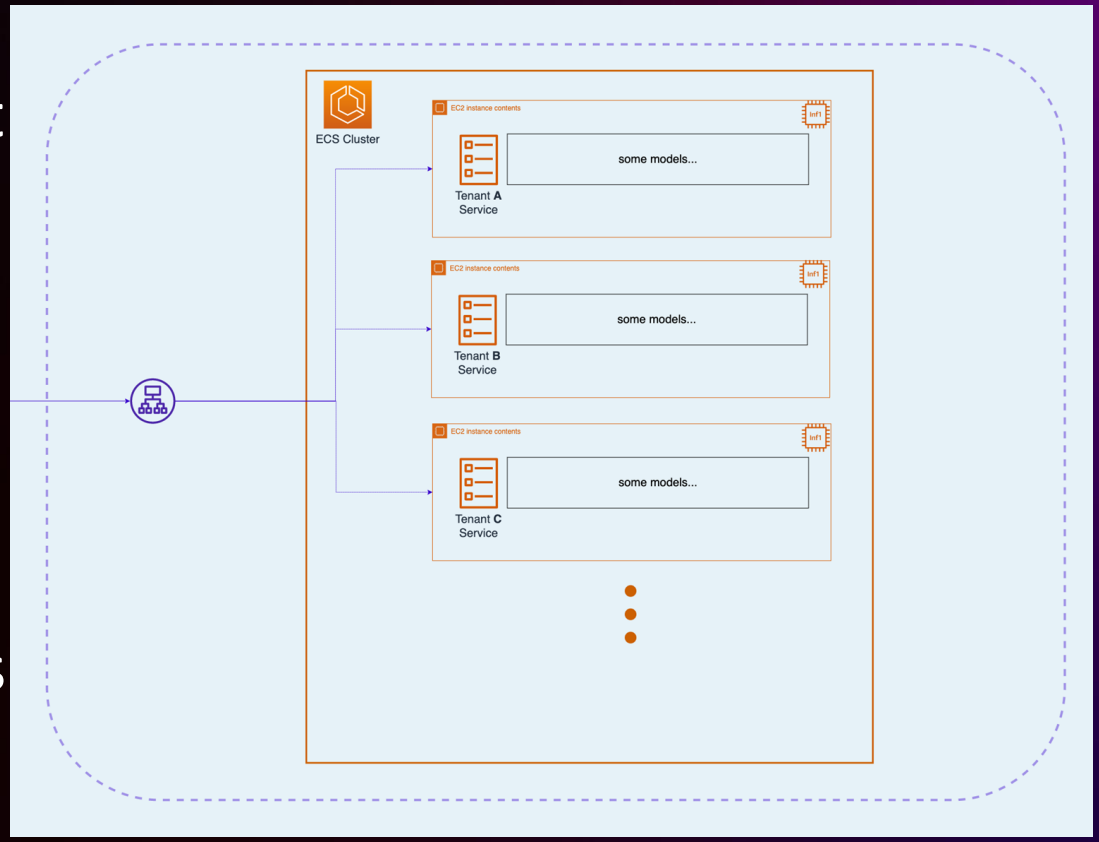
- Deep learning models of natural language processing
  - Since real-time performance is important for this service, very fast inference processing was required.
  - Since a model is prepared for each client company, it was important to be able to perform inference on many models simultaneously.

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

ECS on EC2 configuration with Inferentia

In this project, it is necessary to create a model and provide an inference environment for each customer who uses the system, so the number of models will be huge, existing for the number of customers. In practice, we have chosen to use ECS on EC2, a method of arranging a large number of services.

Although there are some maintenance costs due to the unavailability of AWS Fargate, this approach was chosen because the development team had more knowledge and skills in ECS than in SageMaker.



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

Why Inferentia?

There were three main reasons for adopting Inferentia

- Inference execution speed
  - In our trial runs, we found that the inference processing on the AWS Inferentia chip was very fast.
  - The speed of the AWS Inferentia chip was essential for us, as we needed to keep latency as low as possible due to the nature of the service.
- Reduced number of Compute Instances
  - The AWS Inferentia chip has four Neuron cores, making it very easy to serve multiple models on a single unit.
  - The ability to serve multiple models on a single unit has led to a reduction in the overall number of instances = a significant cost reduction.
- Ease of implementation
  - The fact that AWS Inferentia was so easy to use was the biggest reason for our adoption.
  - The AWS Inferentia chip was very easy to use, as all we had to do was compile an existing model, and very little advance preparation was required, which was very important to us, as we had a limited development time.

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

We migrated all systems in just two months

- We succeeded in completely transitioning the environment in two months
- Cost reduction of more than 50%
- Up to 90%+ processing time reduction

Inferentia was easy to implement, reduced cost, fast,
and was an integral part of our environment migration project!

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

Thank you!

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

Dataminr's AI Platform

Alex Jaimes

Chief Scientist and
SVP of AI





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

Dataminr's Real-Time Information Discovery Platform

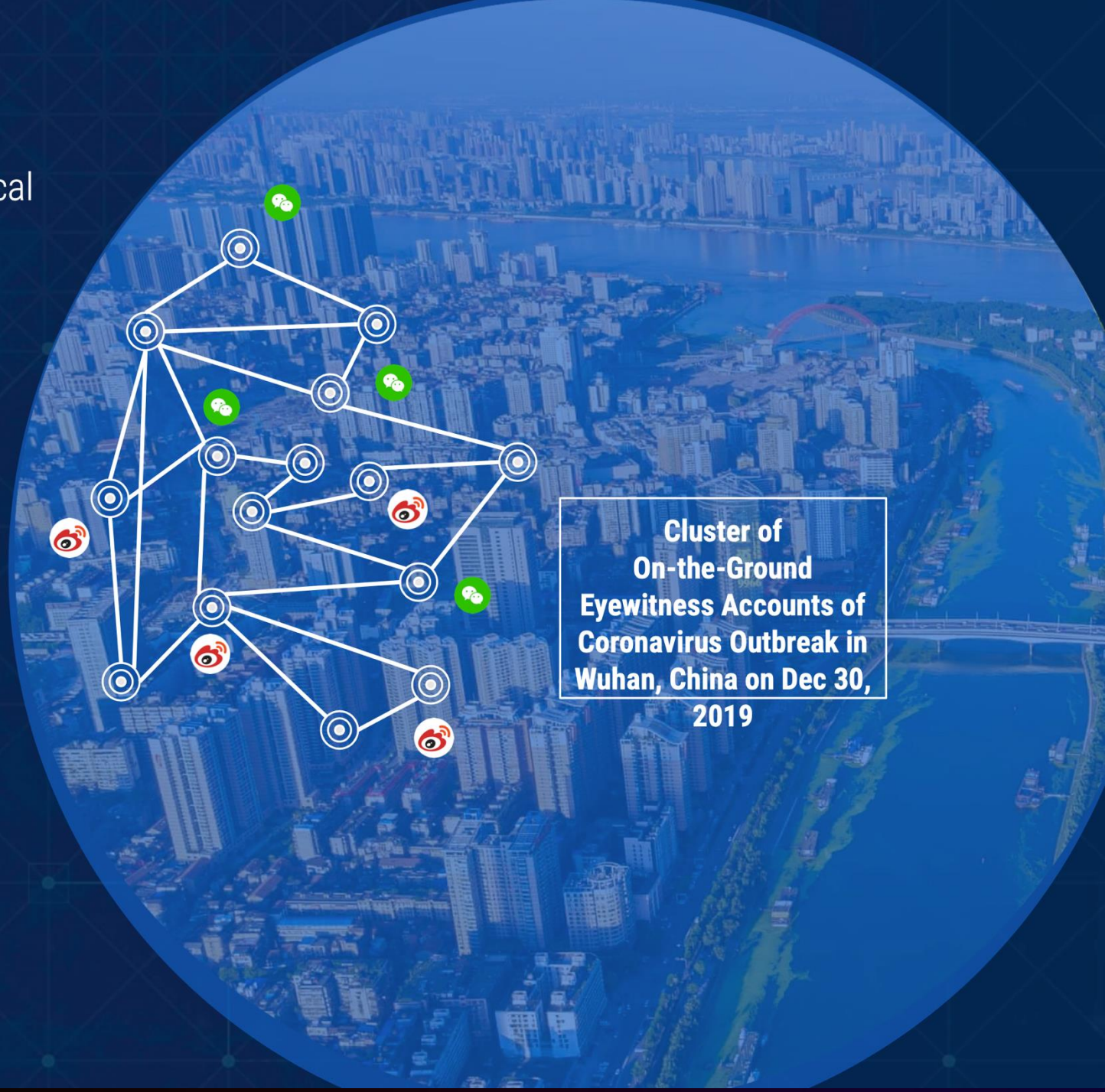
Detects Digital Patterns of Emerging Events and Critical Information from Within Billions of Daily Public Data Signals

Customers: Businesses, public sector agencies, NGOs, and newsrooms rely on Dataminr to get the earliest warnings about high impact events and critical information so they can know first and act faster

Data Scope: Integrates 300,000+ public data sources, in 100+ languages, from within global and regional social media platforms, blogs, web forums, deep and dark web, audio transmissions, and public IoT sensors

AI-Driven Advantage: 10+ Year historical event and data archive, including trillions of data historical data units

Multi-Modal Fusion AI: AI platform cross-correlates data formats spanning text, image, video, sound, and machine generated sensor data



Helping your customers mitigate threats

The New Composition of Corporate Risk

Corporate Security

CSO

Asset Protection
Executive Protection
Travel Safety
Employee Safety/Field Worker Safety
Loss Prevention & Supply Chain
Event/Venue Security
Route Safety

Cyber Risk

CISO

Vulnerability Management
Vendor & Third Party Risk
First party fraud detection
Physical/Cyber Convergence

Reputational Risk

CMO, CCO

Brand & Executive Reputation
3rd Party Brand Risk
Corporate Response
Crisis Communications
Digital Risk & Fraud
Media & Commentary
Misinformation Campaigns

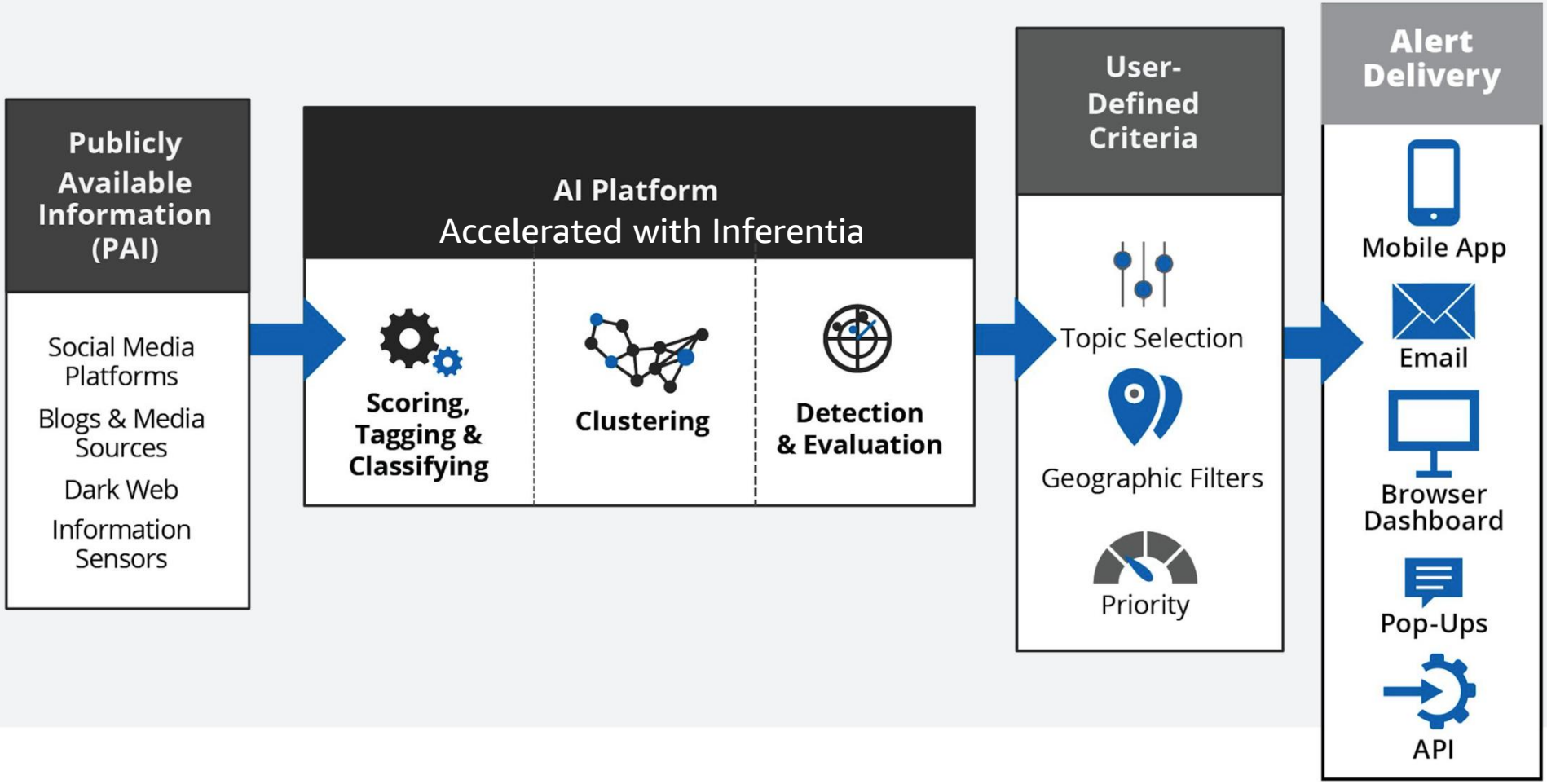
Business Intelligence

COO, CFO, GC, CRO

Financial Services
Supply Chains
Business Logistics
Delivery Disruption
Site Selection
Competitive Intelligence

How Dataminr Works

Dataminr processes enormous amounts of data – billions of messages, posts and data artifacts etc. – to deliver critical information to our customers in real-time.



At global scale, time is our critical resource

- **Inferentia** enables us to do more, faster.
- **3 key benefits** across our portfolio of AI capabilities:
  1. **Growth** – to apply more deep models to more input data
  2. **Acceleration** – lowering latency
  3. **Accuracy** – deeper models, with the same time constraints

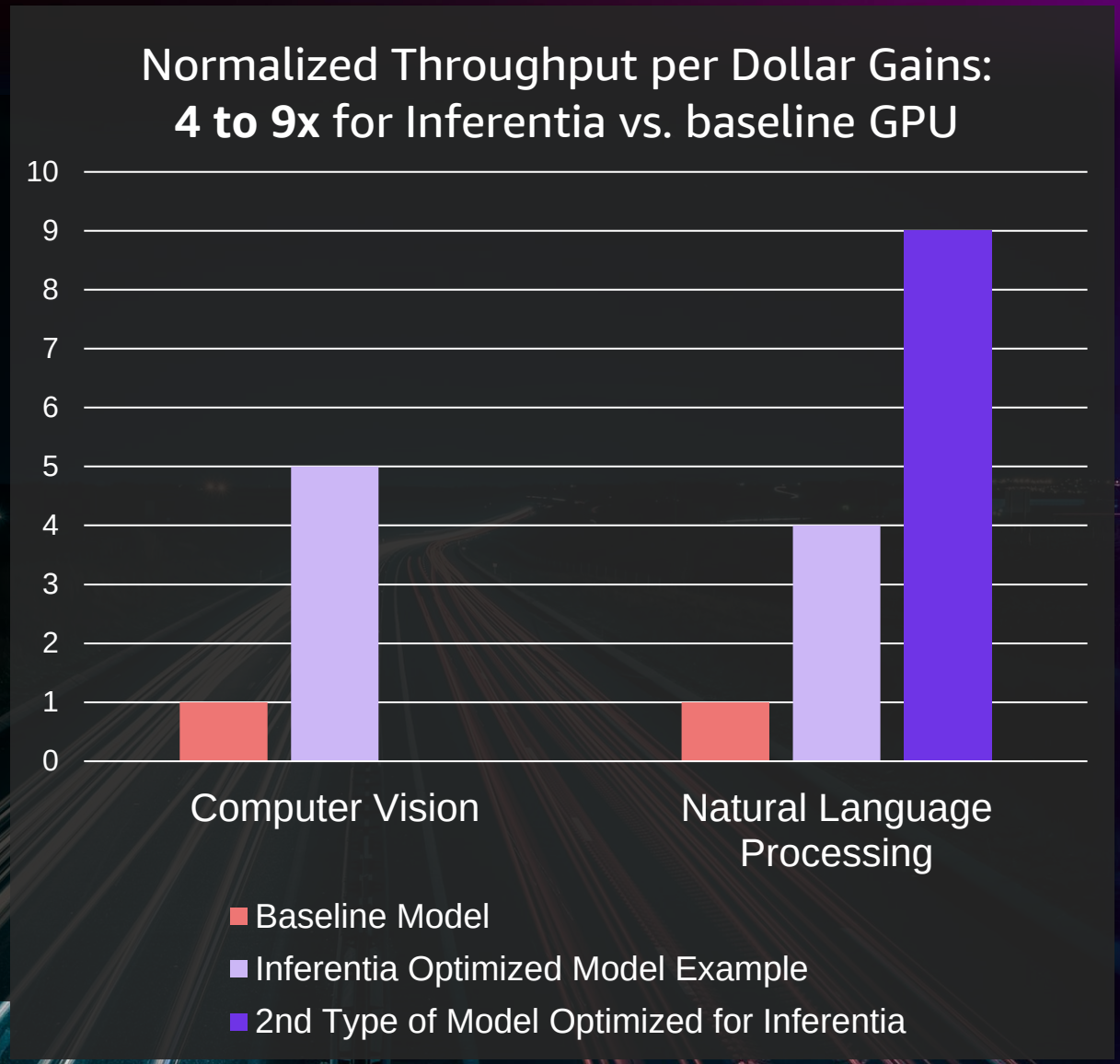


aws

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

Production results with AWS Inferentia

- In **production** throughput per dollar, we've achieved:
- **Vision**: **5x increase**
- **NLP** Tasks: **4x to 9x increase**
  - Coupled with more language coverage
  - Able to use the most innovative models
- We are building an MLOps service to compile models for Inf1 automatically, to leverage Inf1 across all our models



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

Thank you!

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

AWS Trainium

HIGH PERFORMANCE AND LOWEST COST TRAINING



HIGH PERFORMANCE

Highest performance in Amazon EC2 for ML training



LOWEST COST

Lowest cost-to-train ML models in Amazon EC2



FRAMEWORK SUPPORT

Integrated with leading ML frameworks and orchestration libraries

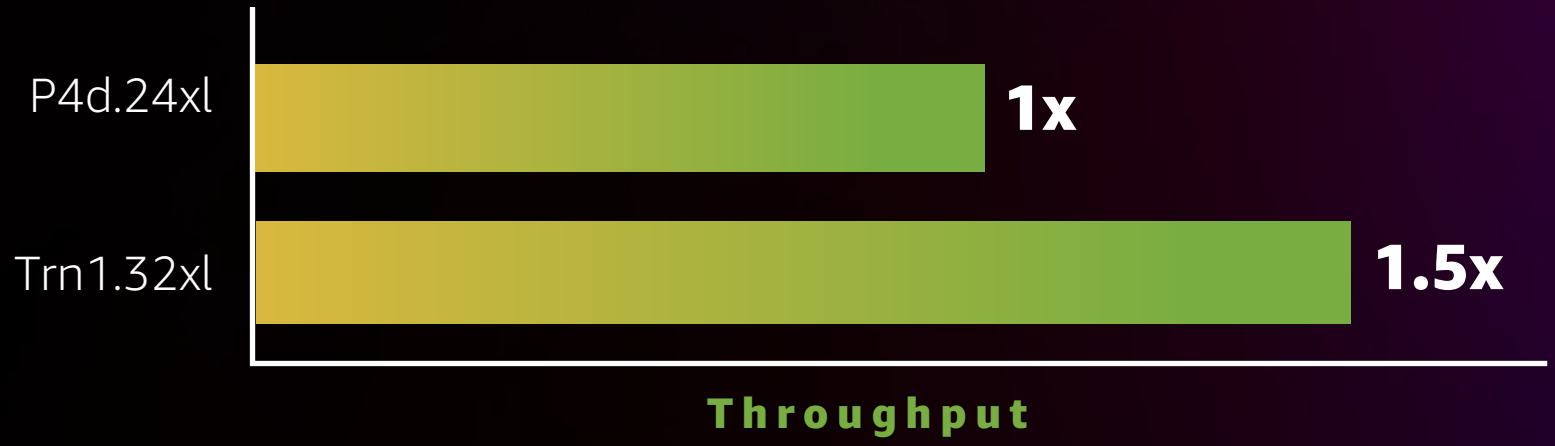
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Best cost to train with Trn1

AMAZON EC2 TRN1 INSTANCES DELIVER THE BEST TRAINING PRICE-PERFORMANCE

1.5x

Higher throughput



50%

Lower cost



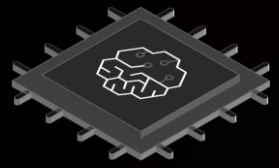
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

* Measured on PyTorch BERT-Large

Amazon EC2 Inf2 Instances

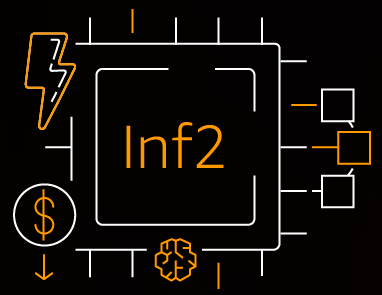
NEW

IN PREVIEW: THE MOST COST-EFFICIENT DL INFERENCE INSTANCE



AWS Inferentia2

High-performance machine learning inference chip, purpose-built by AWS



EC2 Inf2 Instances

Fastest and lowest cost inference in the cloud

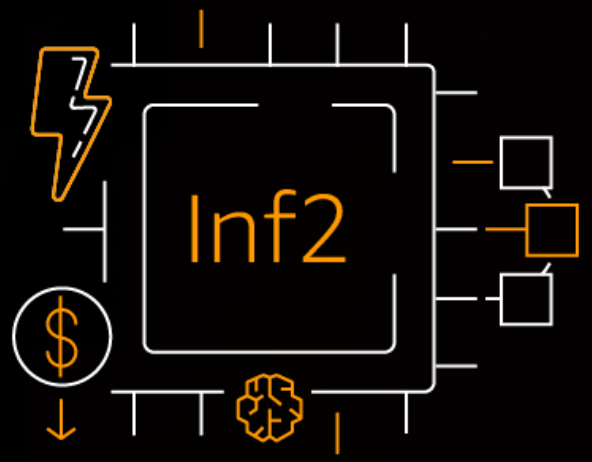
Inf2	
BF16/FP16 2.3 Peta FLOPS	INT8 4.6 Peta OPS
AGGREGATE ACCELERATOR MEMORY 384 GB	NETWORK CONNECTIVITY 100 Gbps
NEURON-CORE V2 NEURON-LINK V2	PYTORCH, TENSORFLOW, SUPPORTED

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

Amazon EC2 Inf2 Instances

NEW

IN PREVIEW: THE MOST COST-EFFICIENT DL INFERENCE INSTANCE



Up to 4x higher throughput and 10x lower latency (vs. Inf1)

10 TB/s aggregated memory bandwidth (15x vs. Inf1)

Deploy 175B parameter models in a single server

Support for dynamic input shapes and control-flow

Instance size	vCPUs	Inferentia2 chips	Accelerator memory	NeuronLink	Instance memory	Instance networking
Inf2.xlarge	4	1	32 GB	N/A	16 GB	Up to 15 Gbps
Inf2.8xlarge	32	1	32 GB	N/A	128 GB	Up to 25 Gbps
Inf2.24xlarge	96	6	192 GB	Yes	384 GB	50 Gbps
Inf2.48xlarge	192	12	384 GB	Yes	768 GB	100 Gbps

aws

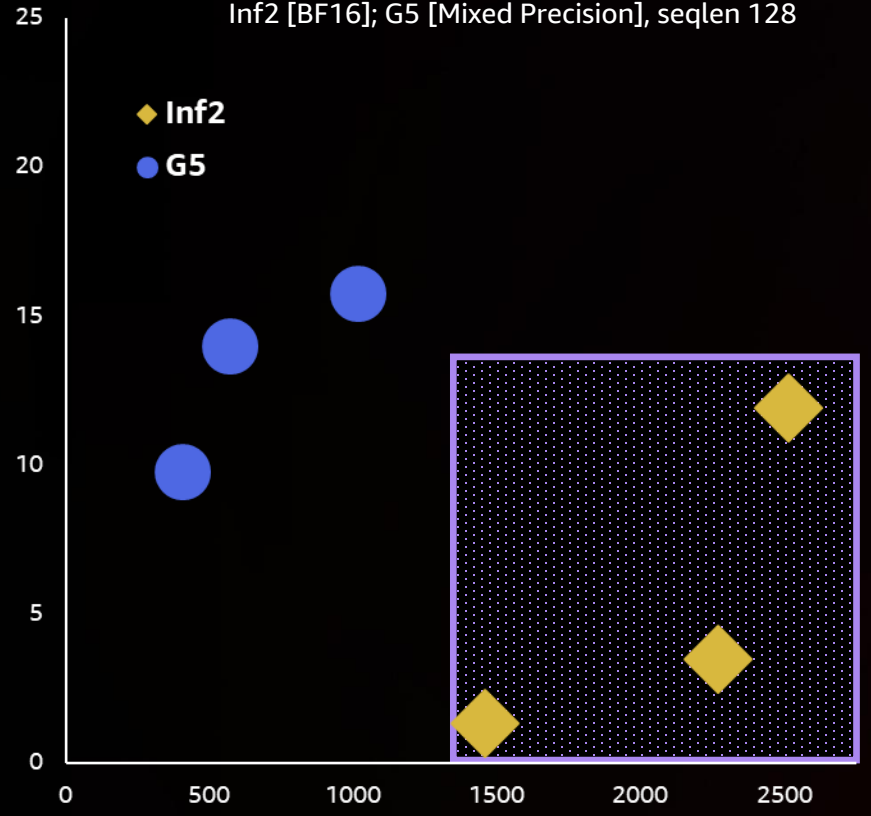
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Inf2 performance vs. GPUs

AMAZON EC2 INF2 INSTANCES DELIVER HIGH-PERFORMANCE AND LOW-LATENCY INFERENCE

RoBERTa-Base Benchmarks

Inf2 [BF16]; G5 [Mixed Precision], seqlen 128



Inf2 Performance vs. G5

Inf2 [BF16]; G5 [Mixed Precision], seqlen 128

Benchmark	Throughput	Latency
BERT-base	2.6x	6.7x
BERT-large	2.2x	5.0x
RoBERTa-base	2.6x	8.1x

up to 2.6x higher throughput

up to 8.1x lower latency

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

Thank you!

Jesús Hormigo
SCREENING EAGLE

Kento Adachi
Money Forward

Zachary Schmidt
actuate

Alejandro Jaimes
Dataminr®



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

aws

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