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

SUS206

Sustainability and AWS silicon

Paul Mazurkiewicz (he/him)
Senior Principal Engineer
AWS Hardware Engineering

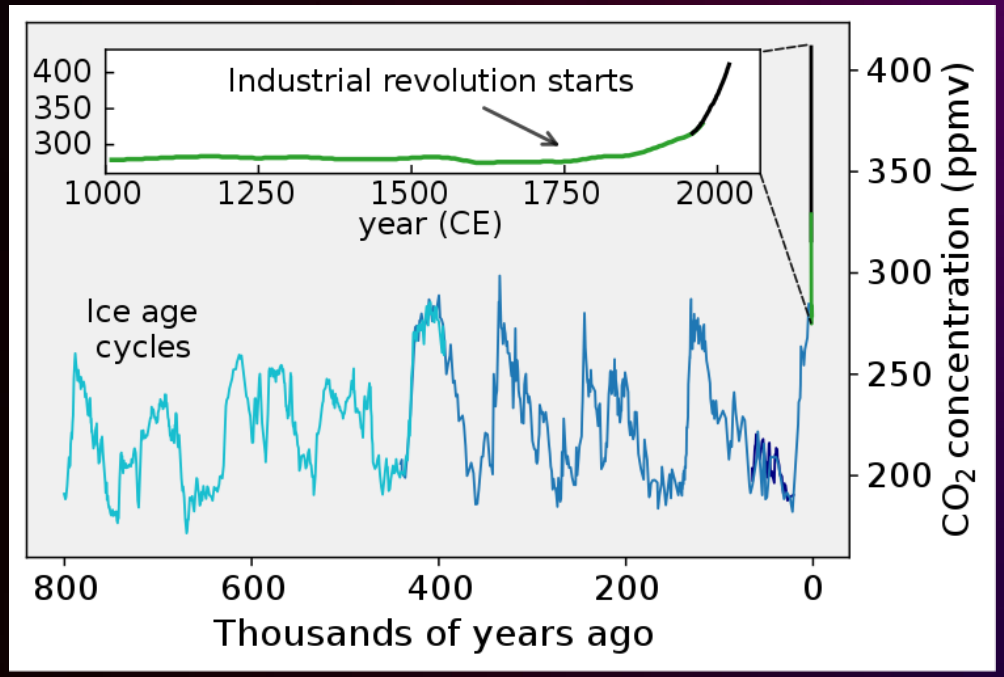
Kamran Khan (he/him)
Sr. Product Manager
Annapurna Labs, AWS

David Chaiken (he/him)
Chief Architect
Pinterest

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

What is contributing to climate change?

- Carbon dioxide and other gases absorb radiation from the sun
- Greenhouse gases regulate temperature
- Too much is problematic

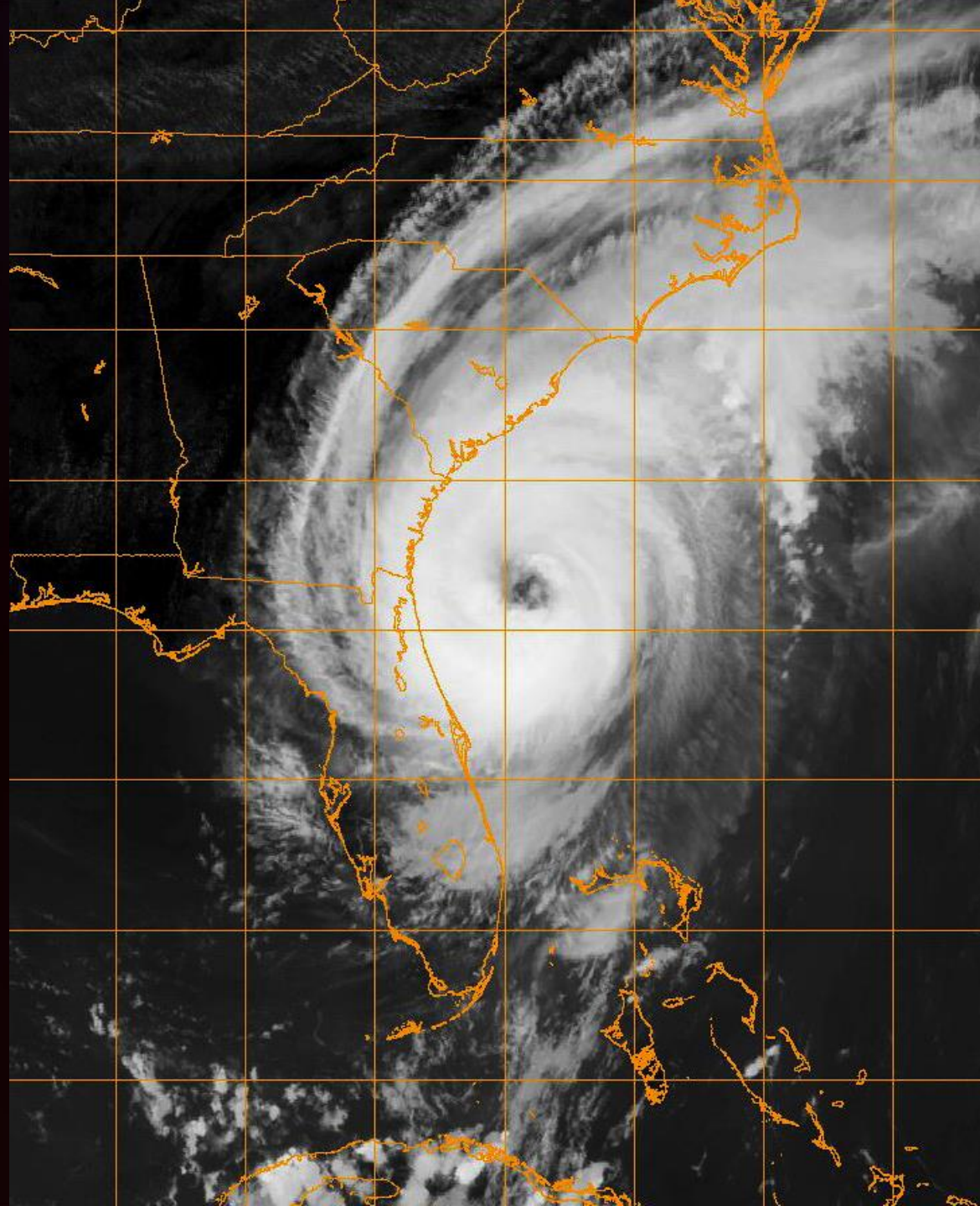


Carbon dioxide levels in Earth's atmosphere

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

What are the effects of climate change?

- Hotter temperatures
- More severe storms
- Increased drought
- Warming, rising oceans
- Loss of species
- Food scarcity



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

Further and faster, together

September 19, 2019

Amazon and Global Optimism announced The Climate Pledge, a commitment to meet the goals of the Paris Agreement 10 years early – and achieve net-zero carbon by 2040. Amazon is a co-founder and first signatory of The Climate Pledge.



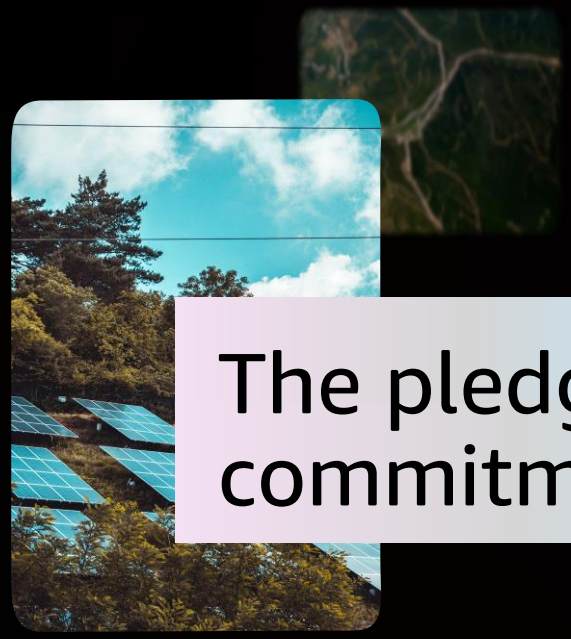




aws

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

Joining The Climate Pledge



The pledge commitments

01 **Regular reporting**

Measure and publicly report greenhouse gas emissions on a regular basis across scopes 1, 2, and 3

02 **Carbon elimination**

Implement decarbonization strategies through real business changes and innovations

03 **Credible offsets**

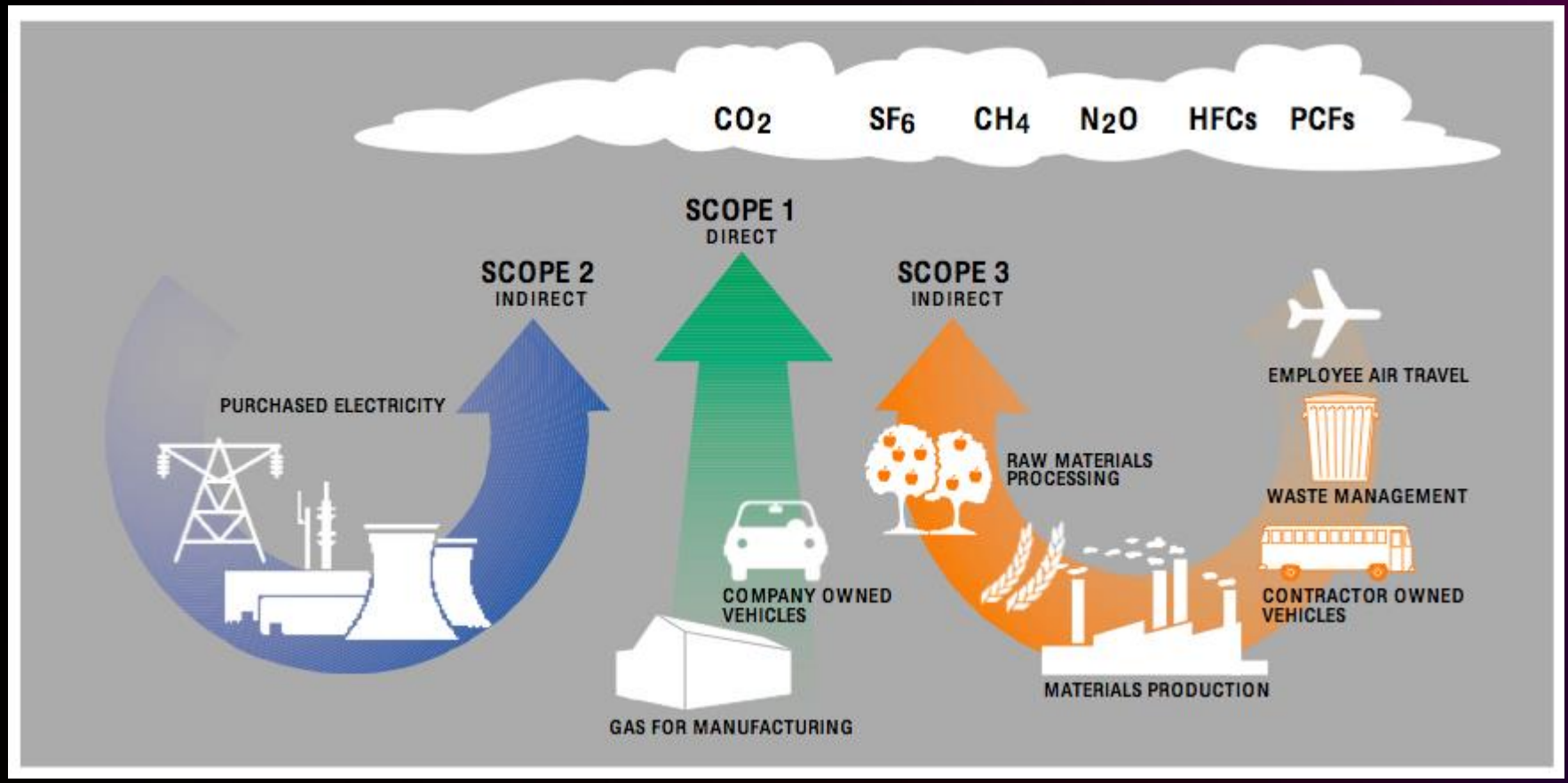
Take action to neutralize any remaining emissions with additional, quantifiable, real, permanent, and socially beneficial offsets to achieve net-zero annual carbon emissions by 2040

aws

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


With 375+ signatories
from around the world
NESPRESSO.
CARBON CURE.
asos
cabify
gsk
BEST BUY
Snap Inc.
itv
Deloitte.
CLIF
Mercedes-Benz
reckitt
Weyerhaeuser
Microsoft
Unilever
jetBlue
CANARY WHARF GROUP PLC
BROOKS
hotelbeds
PROSEGUR
JLL
CBRE
MAERSK
Infosys
verizon
SIEMENS
Ingenuity for life
vmware
HRS
amazon
P&G
SAP
Uber
Klarna.
Iceland
Johnson Controls
VISA
IBM
Coca-Cola EUROPEAN PARTNERS
McKinstry
For The Life Of Your Building
Alaska AIRLINES
CRANSWICK plc
Great British Taste
Telefónica
Lime
IPG
MACQUARIE
PEPSICO
BOOM
PHILIPS
Sainsbury's

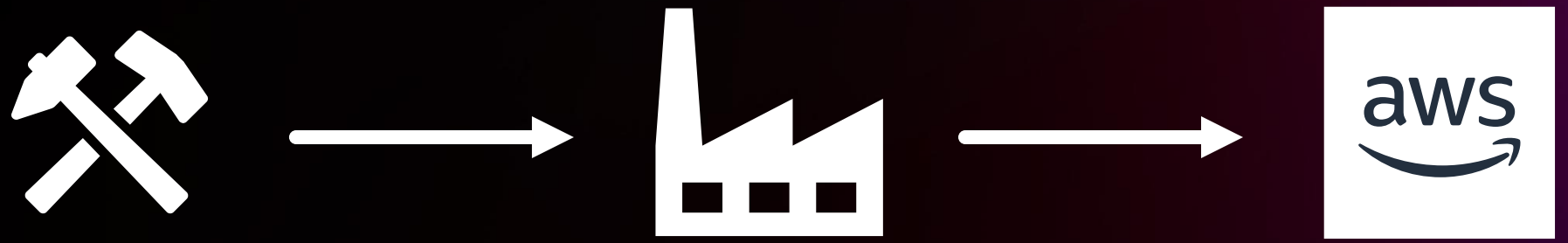
Types of carbon emissions



A carbon footprint is the total greenhouse gas emissions caused by an individual, event, organization, service, place, or product, expressed as carbon dioxide equivalent

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

The carbon footprint of hardware



Mining of raw materials

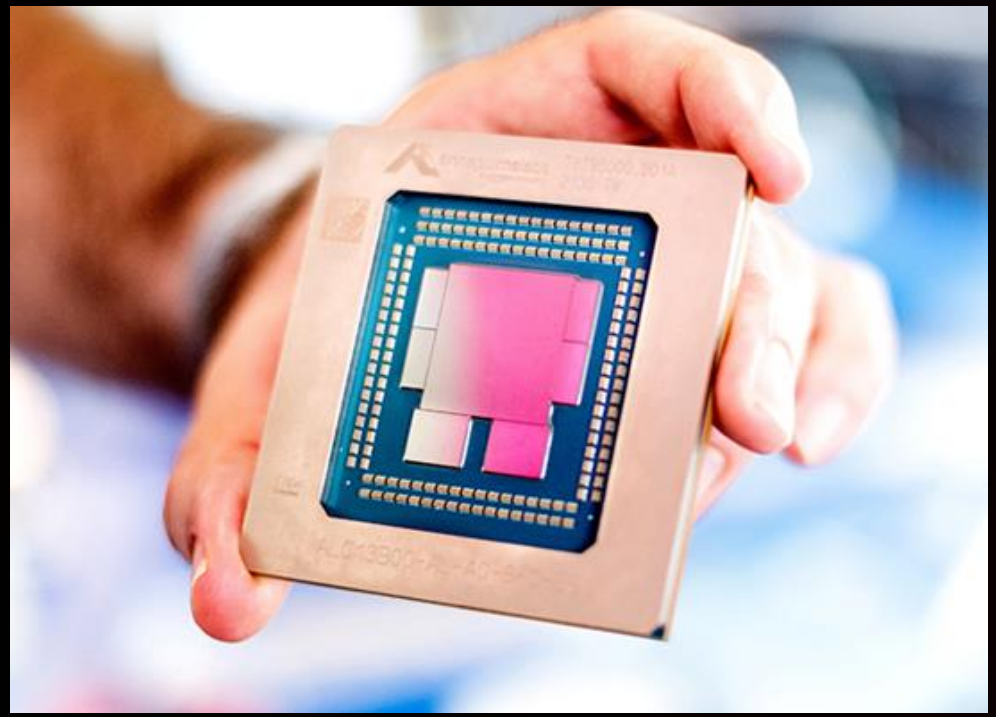
Factories

Data centers

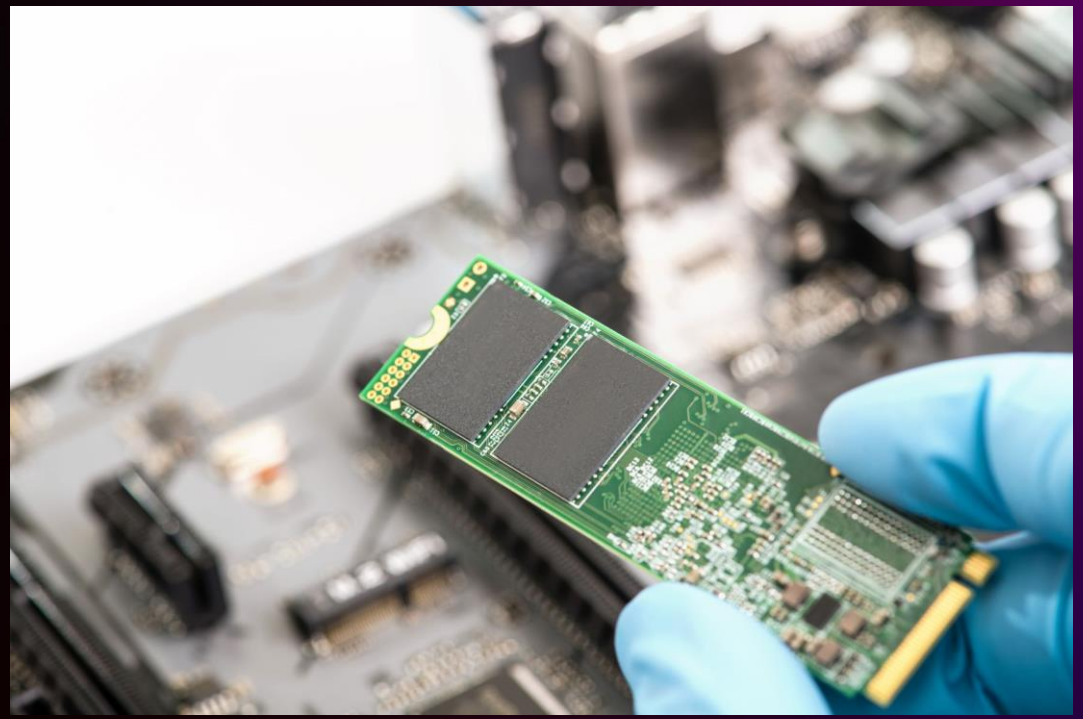
The embodied carbon of hardware is a result of the greenhouse gases emitted in the production process

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

Embodied carbon in AWS hardware



Central processing unit



Solid-state drive

Silicon-based devices represent the highest embodied carbon in our hardware

aws

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

Carbon footprint of silicon

Quartz



Silicon crystal



Coal



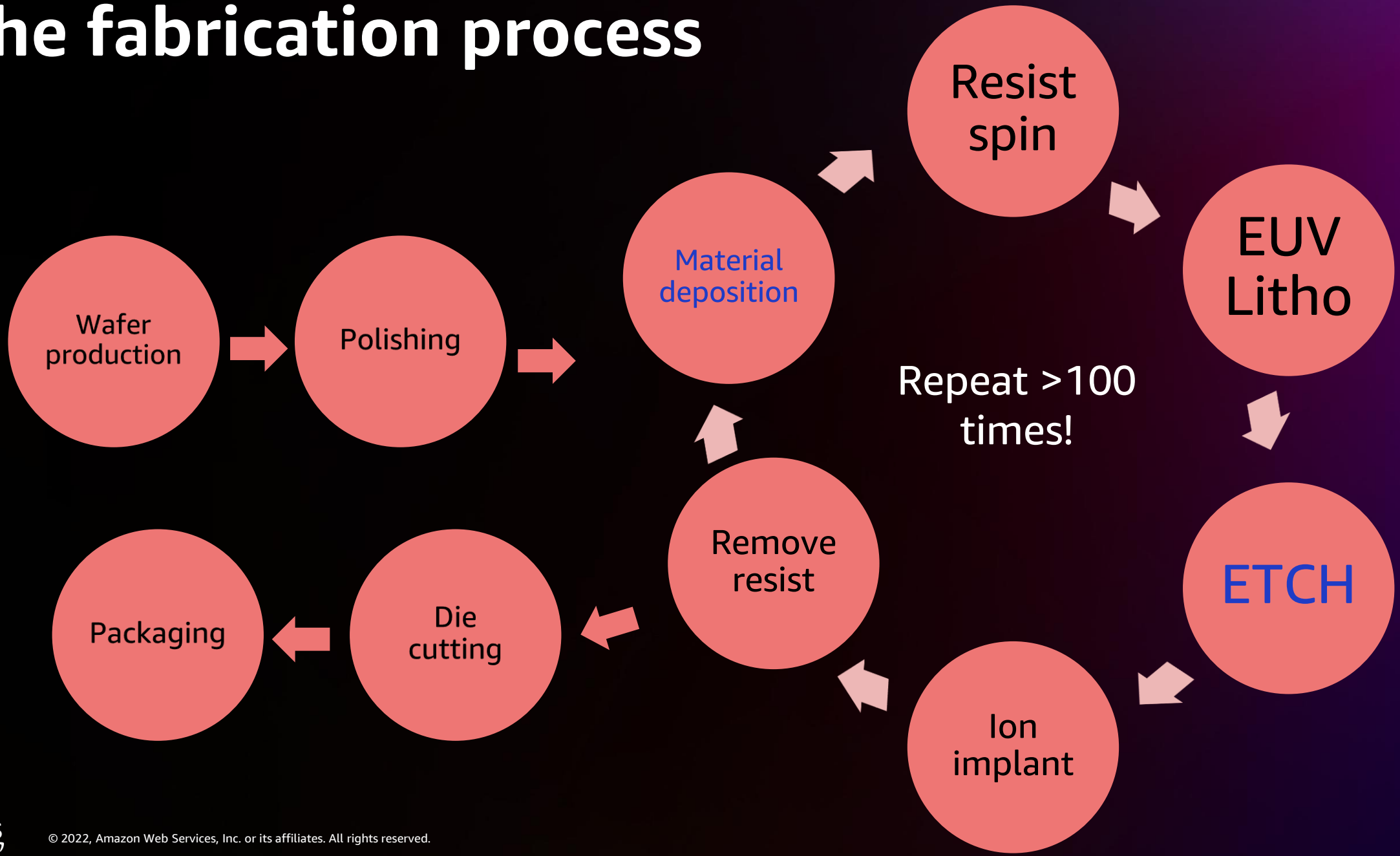
Silicon wafers



“Pure silicon crystal” by Dave Messina / CC BY 2.0
“Rock crystal” by Mauro Cateb / CC BY 2.0
Silicon wafers / CC BY-ND

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

The fabrication process



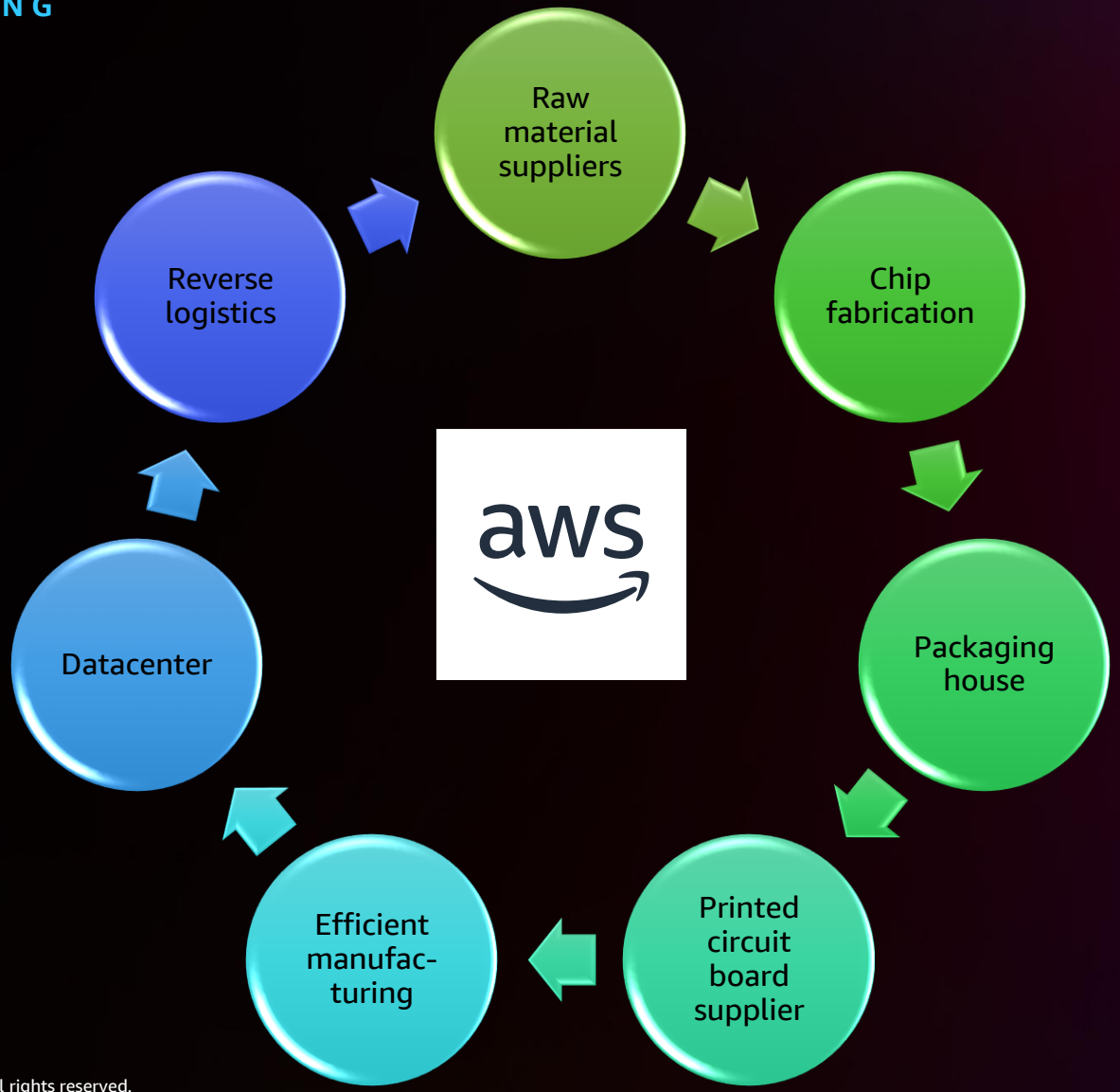
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The silicon supply chain

CRADLE-TO-CRADLE THINKING



Sustainable energy





Sustainable materials

CC BY 3.0

aws

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

Reducing hardware carbon footprint

Focusing on renewable energy in our supply chain

Energy efficiency is integrated into our chip design process

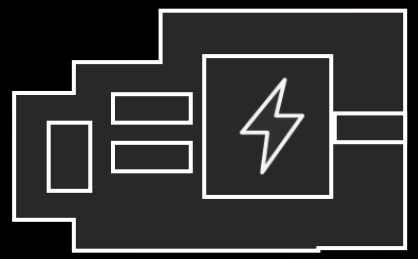


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS silicon design

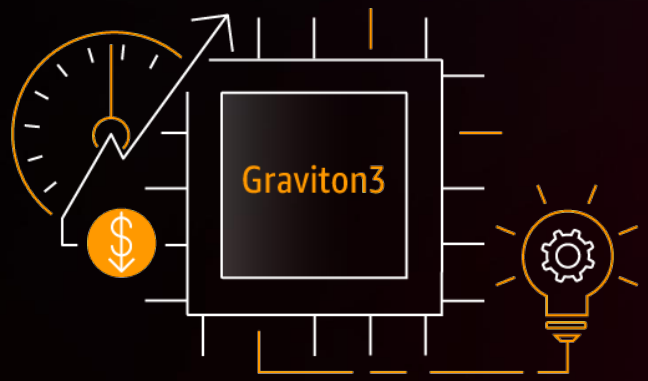
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Journey of silicon innovation at AWS



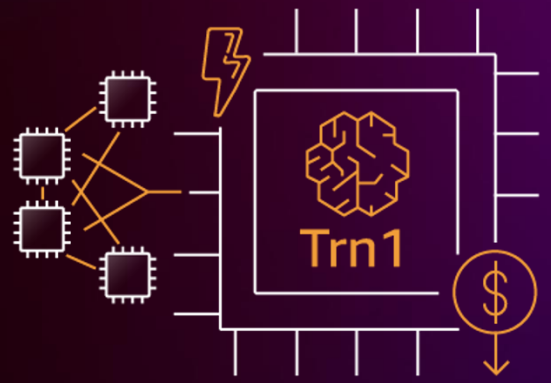
AWS Nitro System

Hypervisor, Nitro Cards, network, storage, SSD, and security



AWS Graviton

Powerful and efficient, modern applications



AWS Inferentia and AWS Trainium

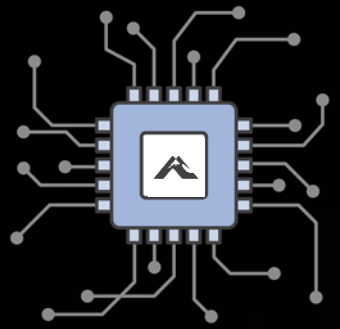
Machine learning acceleration

aws

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

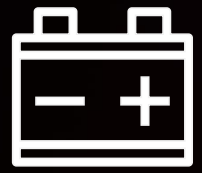
Graviton3-based seventh-generation EC2 instances

Graviton3



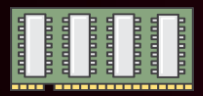
Industry-leading performance

Efficiency



Use up to 60% less energy for the same performance than comparable EC2 instances

DDR5



First DDR5 system in Amazon EC2

Best price performance



Best price performance in Amazon EC2

aws

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

What customers are saying about Amazon EC2 C7g instances



"We found Graviton3-based C7g instances deliver 20-80% higher performance vs. Graviton2 based C6g instances, while also reducing tail latencies by as much as 35%."



"We were able to run 30% fewer instances of C7g than C6g serving the same workload, and with 30% reduced latency."



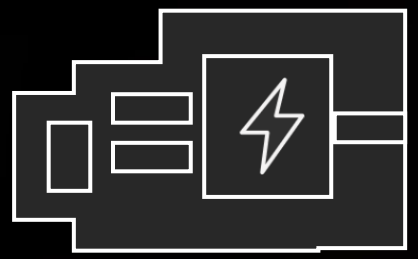
"They are suitable for even the most demanding latency sensitive workloads while providing significant price performance benefits."



"We have now found Graviton3 C7g instances to be 40% faster than the Graviton2 C6gn instances for those same simulations."

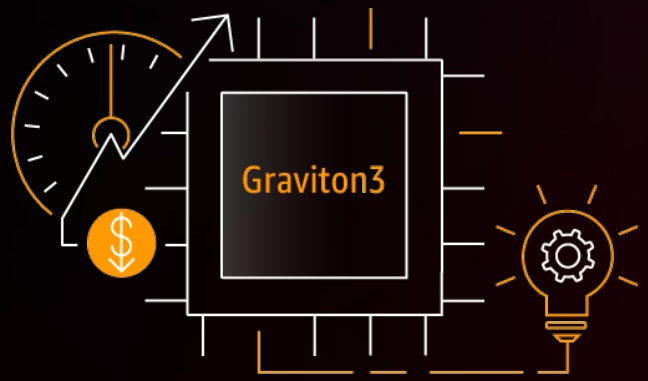
aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Journey of silicon innovation at AWS



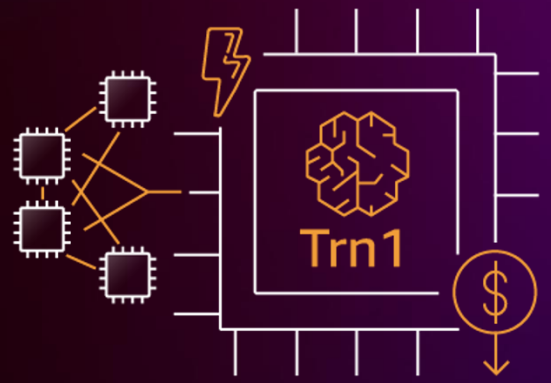
AWS Nitro System

Hypervisor, Nitro Cards, network, storage, SSD, and security



AWS Graviton

Powerful and efficient, modern applications



AWS Inferentia and AWS Trainium

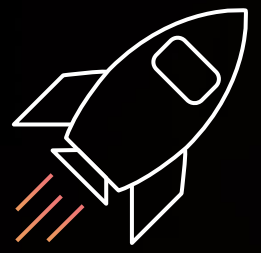
Machine learning acceleration

aws

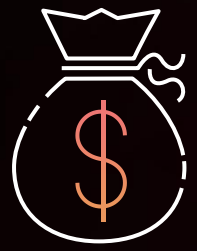
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Inferentia

HIGH PERFORMANCE AND LOWEST COST INFERENCE IN THE CLOUD



Up to 2.3x higher throughput



Up to 70% lower cost per inference than comparable GPU-based Amazon EC2 Instances



Support for popular ML frameworks including TensorFlow, PyTorch, and MXNet

aws

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

Customer momentum with Inferentia

qualtrics XM

airbnb

Finch COMPUTING

sprinklr

AUTODESK

amazon robotics

amazon advertising

amazon alexa

SCREENING EAGLE

NTTPC

CONDÉ NAST

Talroo™

Anthem

DISCO

SKYWATCH

Money Forward



Amazon Rekognition

DMP DIGITAL MEDIA PROFESSIONALS

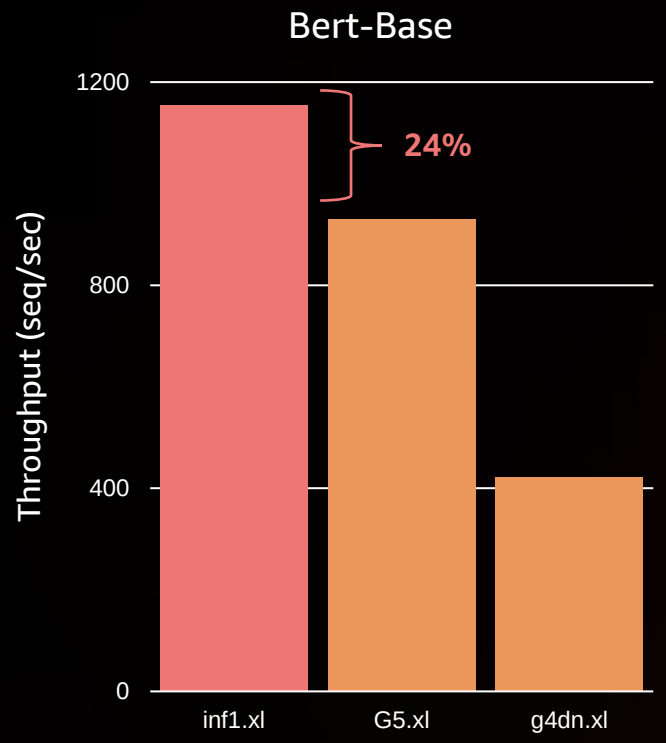
prime video

aws

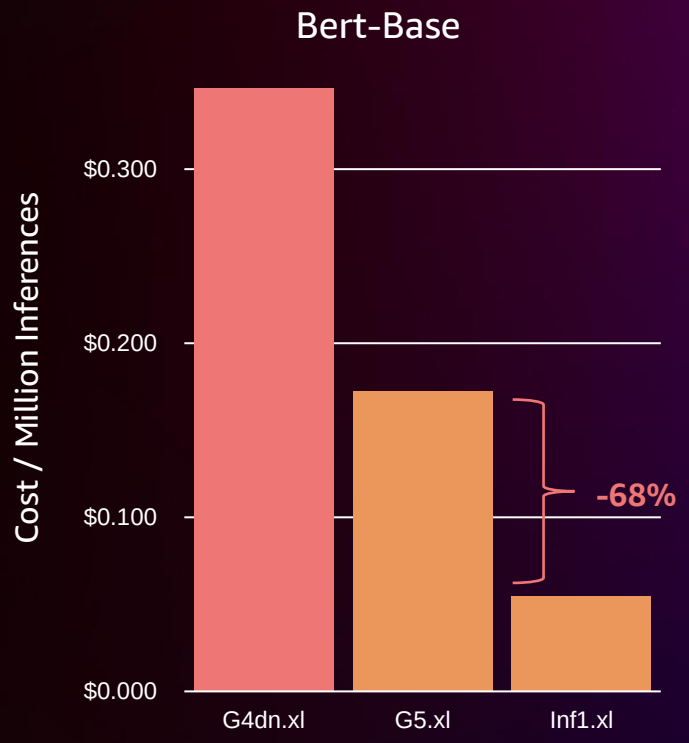
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Inf1: highest performance and lowest cost

Inf1 delivers 2.74x higher throughput than G4 and 24% higher throughput than G5



Inf1 delivers 84% lower cost than G4 and 68% lower cost than G5



Bert-Base numbers derived from Nvidia Performance Page; PyTorch 1.9, seq =128, FP16

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

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

AWS Inferentia enables sustainable ML inference



54%

Inferentia instances consume less power than G4dn instances

Document understanding



Target throughput:
1,000,000 inferences per hour

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

AWS Inferentia enables sustainable ML inference

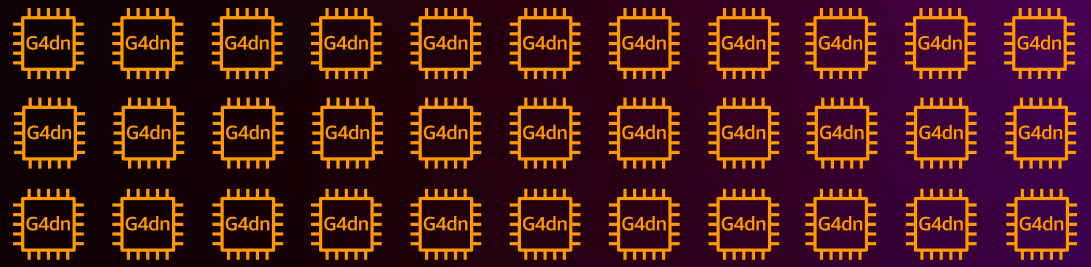
Document understanding



Target throughput:
1,000,000 inferences per hour

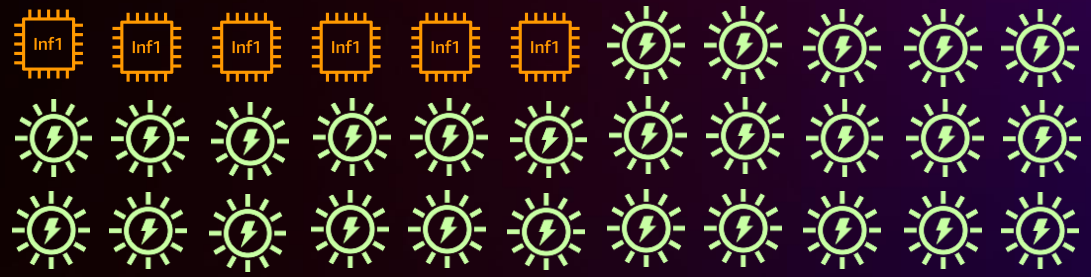
33

G4dn Instances



6

Inf1 Instances



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

AWS Inferentia enables sustainable ML inference

Document understanding



Target throughput:
1,000,000 inferences per hour

82% fewer instances

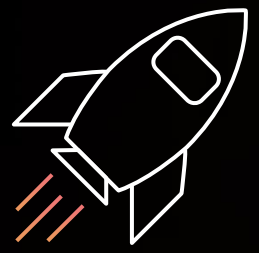
92% less energy

90% lower cost

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

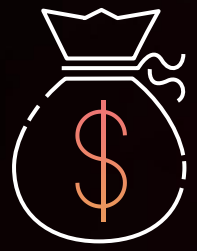
AWS Trainium

HIGH PERFORMANCE MACHINE LEARNING TRAINING CHIP, DESIGNED BY AWS



HIGH PERFORMANCE

Offers the highest performance in Amazon EC2 for ML training



LOWEST COST

Will provide the best price performance for training ML models in the cloud



FRAMEWORK SUPPORT

Integrated with leading ML frameworks, such as TensorFlow and PyTorch

aws

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

Customer momentum with AWS Trainium

Magic

Helixon
华深智药

PyTorch

amazon

sprinklr

Money Forward

CACTUS

ANTHROP\C

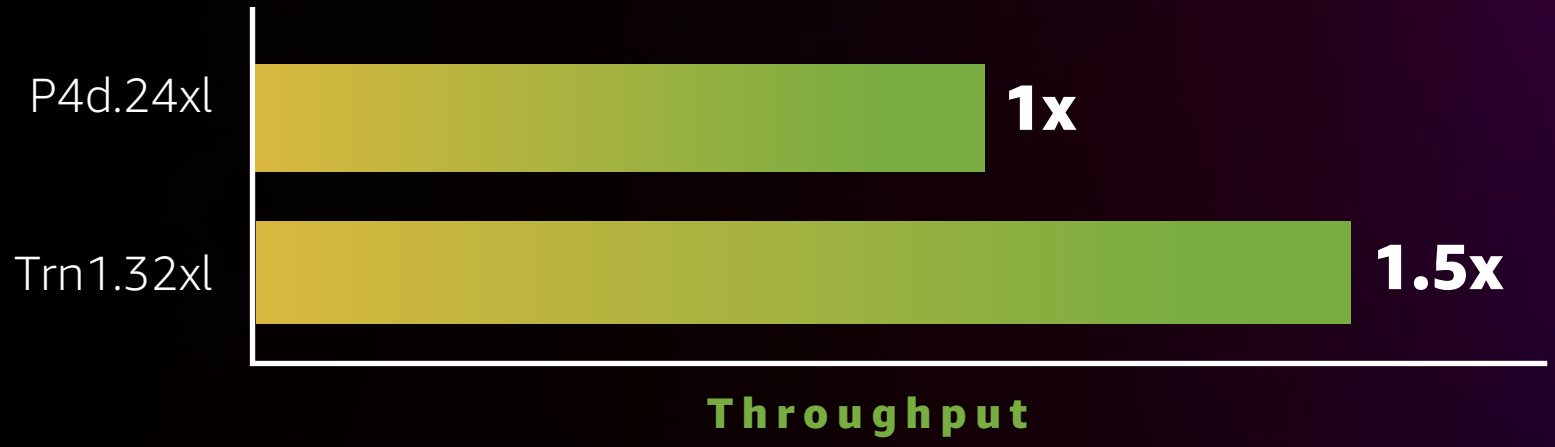
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Best cost to train with Trn1

AMAZON EC2 TRN1 INSTANCES DELIVER THE BEST TRAINING PRICE-PERFORMANCE

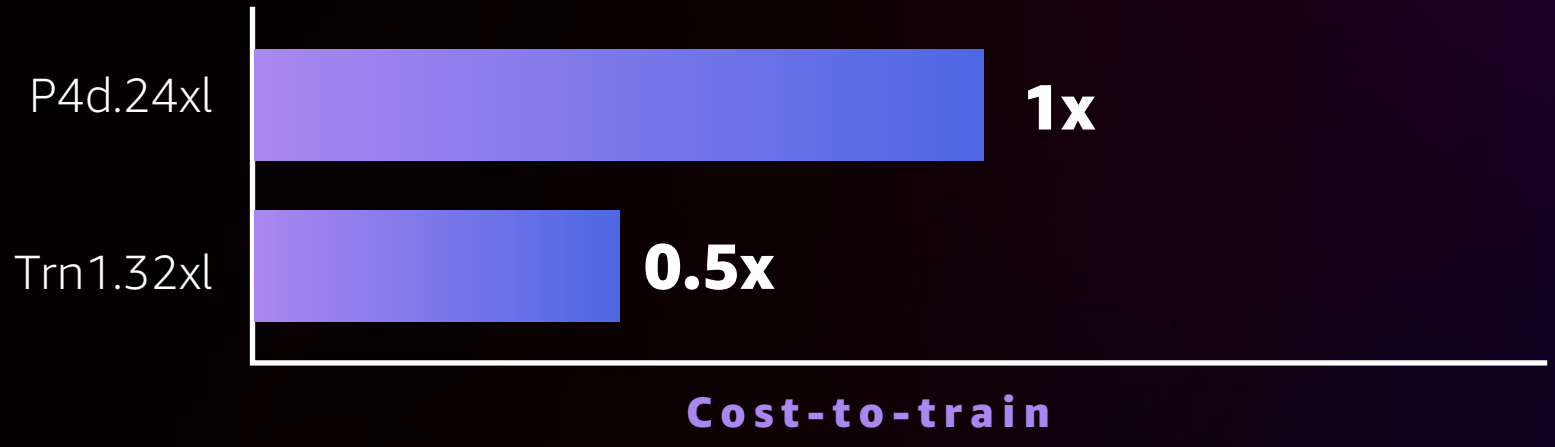
1.5x

Higher throughput



50%

Lower cost



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

* Measured on PyTorch Bert-Large

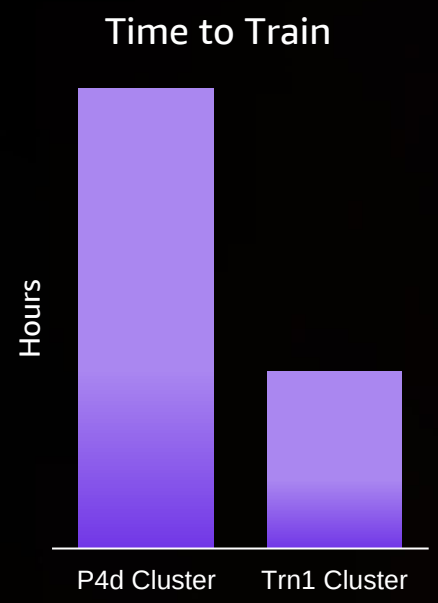
Trainium: High performance, less power

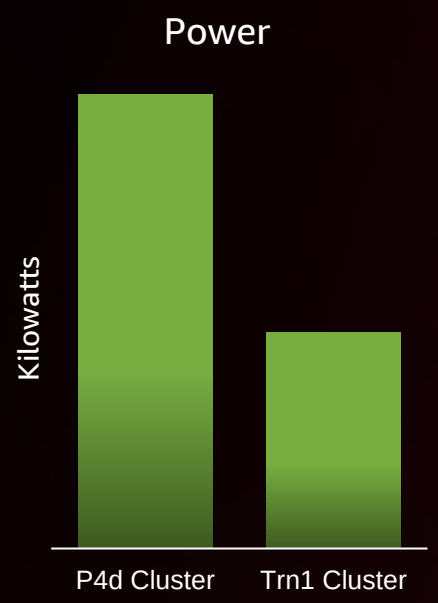
TRAINING BERT-LARGE WITH AWS TRAINIUM

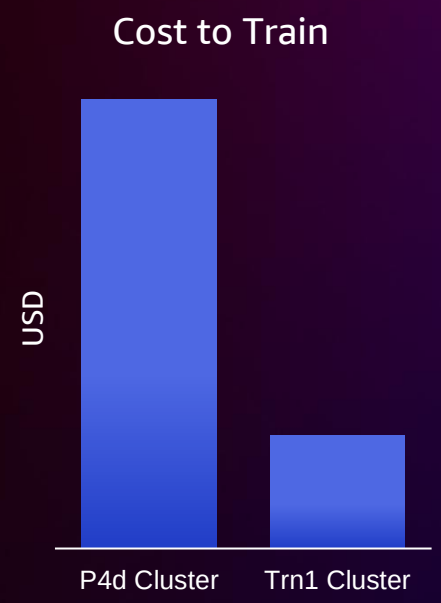
2.6x
Faster Training

52%
Less Energy

75%
Lower Cost







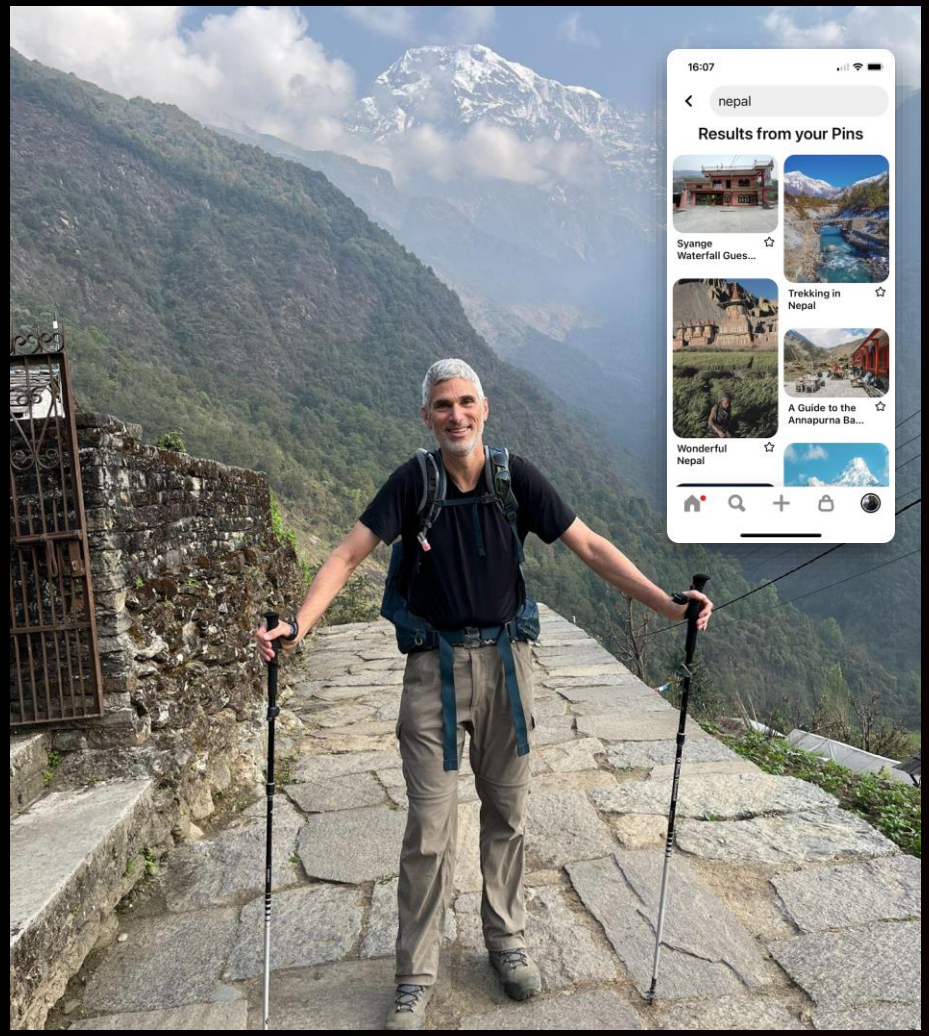
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Details: Hugging Face Bert-Large, FP32, 16x P4d.24xl, 16x Trn1.32xl, On-Demand EC2 pricing

Pinterest and sustainability

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

Hello



I am David Chaiken, the Chief Architect at Pinterest, where I lead tech development & operations – mostly on AWS

I am passionate about environmental sustainability and currently pinning my next hiking trips

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

Snapshot: Pinterest's sustainability journey

2019

A group of passionate employees initiate conversation with AWS about carbon footprint

2020

Receive first carbon emissions report from AWS

2021

Hire Global Sustainability Lead

Launch PinPlanet, our employee climate community

Begin collecting data to calculate emissions from Pinterest's real estate portfolio

2022

Begin first GHG inventory

AWS launches their customer carbon footprint tool, for which we were beta test collaborators

Roll out our **climate misinformation policy**

Commit to purchase 100% renewable electricity for all of our offices around the world by 2023

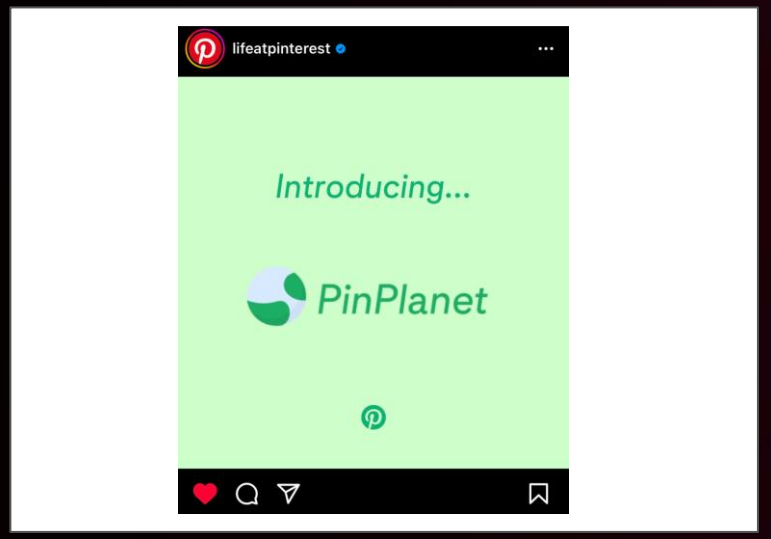
aws

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

Spotlights of Pinterest's sustainability actions



Pinterest expands its community and advertising guidelines to prohibit all content and ads with climate misinformation



Launched PinPlanet, our planet- and climate-focused employee community



Expanded on our existing efforts to save energy in our offices and reduce our global greenhouse gas emission by committing to purchase 100% renewable electricity for our offices around the world by 2023

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

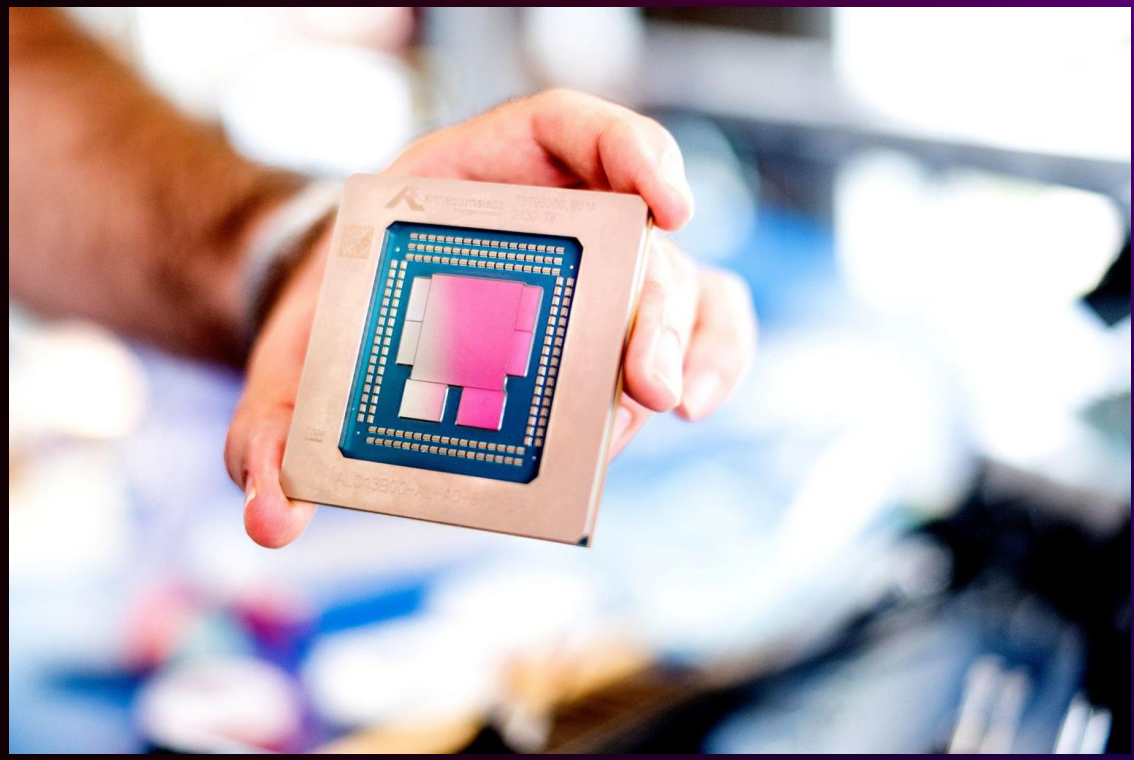
Pinterest and AWS silicon

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

AWS silicon: Performance and sustainability

Graviton

- Multi-year project at Pinterest based on TCO/ROI analysis
- Achieves performance, cost, and sustainability goals
- Memcached: first production workload, 40% performance gain; likely cost savings from this use case: $1M/year



AWS Inferentia and AWS Trainium

- Positive results from early evaluation; working with AWS on optimization and deployment
- Inferentia 2x throughput of GPU on two different transformer models

aws

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

The big picture: Aligning incentives

Frequent AMA question from Pinterest engineers:

> *“What can I do personally to help reduce Pinterest’s carbon footprint?”*

The win-win/easiest and most effective answer is:

> *“Cost efficiency is well-aligned with our carbon footprint, so work with your spend captain on the projects that can save the most money.”*

As AI technology advances, our machine learning models will increase in size; working with AWS to use AWS Inferentia and AWS Trainium will enable us to prioritize power efficiency as well as cost

As our business grows, we need to also prioritize power efficiency

aws

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

Actionable metrics and related timeframes



When driving a large engineering organization, need to think about what metrics are available, and when…

Resource utilization feedback can be immediate (daily, at Pinterest's scale)

Financial metrics true up with AWS billing on a monthly basis

The AWS customer carbon footprint tool allows us to track the reduction in our carbon footprint, driven both by our optimizations and Amazon's progress toward 100% renewable energy

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

Our work is ongoing...



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

Sustainability breakout sessions

SUS204: Sustainability in AWS global infrastructure

SUS205: Architecting sustainably and reducing your AWS carbon footprint

SUS208: Utilizing sustainability data at scale

SUS210: Modeling climate change impacts and risks at scale

SUS211: Water stewardship and renewable energy at scale

SUS212: Accelerating decarbonization and sustainability transformation

SUS301: Sustainable machine learning for protecting natural resources

SUS303: Delivering sustainable, high-performing architectures

SUS312: How innovators are driving more sustainable manufacturing

STP213: Scaling global carbon footprint management

aws

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

Sustainability workshops and chalk talks

SUS201: Detecting deforestation with geospatial images and Amazon SageMaker

SUS202: Using Amazon SageMaker, AutoGluon, and ASDI to predict air quality

SUS203: Carbon emissions reporting using modern data architecture on AWS

SUS207: Patterns for obtaining insights from sustainability data

SUS302: Optimizing architectures for sustainability

aws

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

Thank you!



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

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