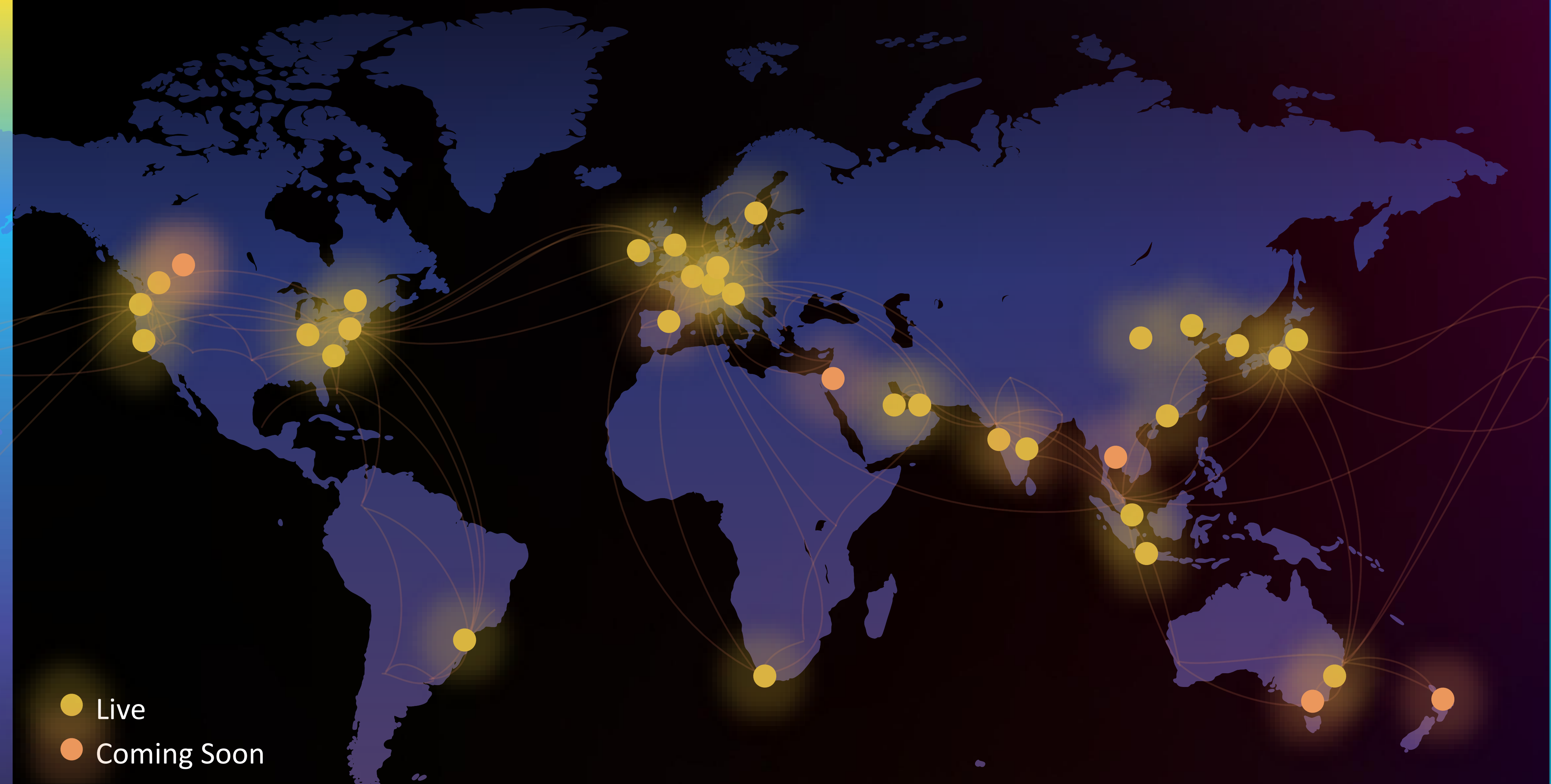
利用 Amazon EC2 最佳化雲端工作負載

Kevin Su

Sr. Solutions Architect
Amazon Web Services

AWS Regions

The most secure, extensive, and reliable Global Cloud Infrastructure



31

Live Regions

4

Coming Soon

99

Availability Zones

AWS Local Zones

RUN LATENCY-SENSITIVE APPLICATIONS AT THE EDGE USING AWS INFRASTRUCTURE AND SERVICES

LOW LATENCY

Extends AWS infrastructure services, APIs, and tools to where customers need it to support low-latency applications

FULLY MANAGED

Fully owned, managed, and supported by AWS

CITIES

New type of AWS infrastructure that places AWS compute, storage, networking, and select AWS services closer to where your end users are located

AWS Outposts

AWS INFRASTRUCTURE AND SERVICES IN YOUR ON-PREMISES LOCATION

AWS DESIGNED

Same AWS-designed infrastructure
as in AWS data centers
(built on AWS Nitro System)

FULLY MANAGED

Fully managed, monitored,
and operated by AWS as if in
AWS Regions

AWS API

Single pane of management in
the cloud providing the same
APIs and tools as
in AWS Regions

Broadest and deepest platform choice

CATEGORIES

General purpose

Burstable

Compute intensive

Memory intensive

Storage (High I/O)

Dense storage

GPU compute

Graphics intensive

+

CAPABILITIES

Choice of processor

Fast processors
(up to 4.5 GHz)

High memory footprint
(up to 24 TiB)

Instance storage
(HDD and NVMe)

Accelerated computing
(GPUs, ASICs, Video, FPGAs)

Networking
(up to 800 Gbps)

Bare metal

Size
(Nano to 112xlarge)

+

OPTIONS

Amazon EBS

Amazon Elastic Inference

=

MORE THAN

600

INSTANCE TYPES

for virtually every workload and business need

Greatest variety and availability to meet your global workload needs

350+ Intel instances

16 years of partnership

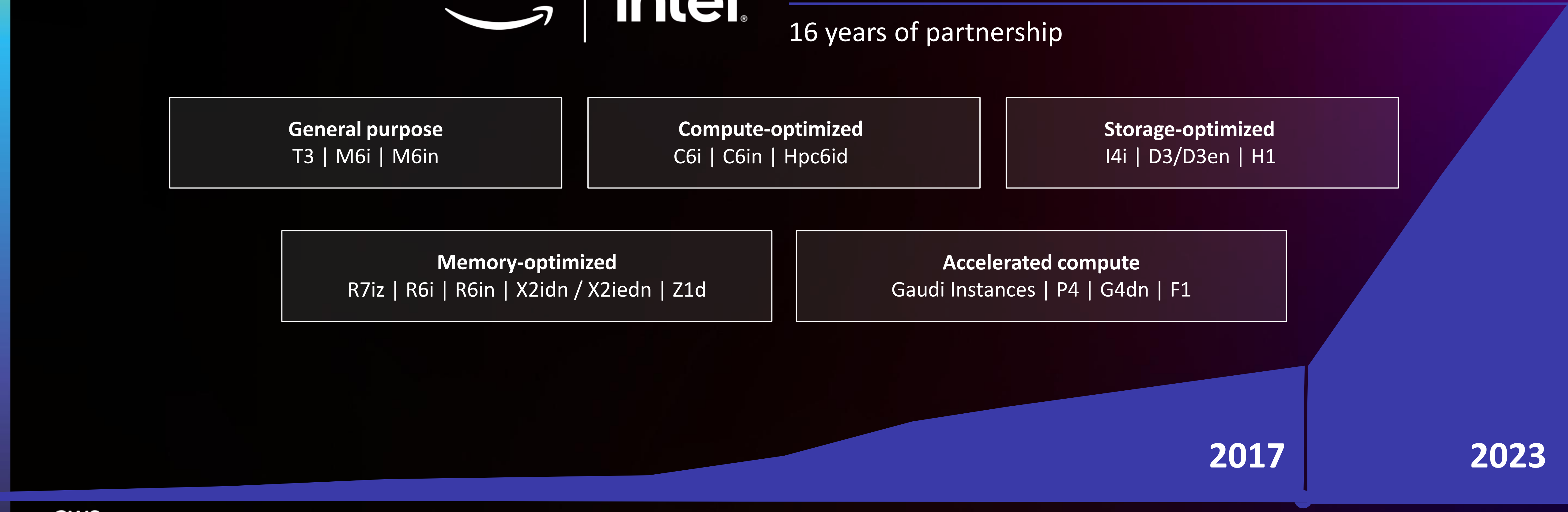
General purpose
T3 | M6i | M6in

Compute-optimized
C6i | C6in | Hpc6id

Storage-optimized
I4i | D3/D3en | H1

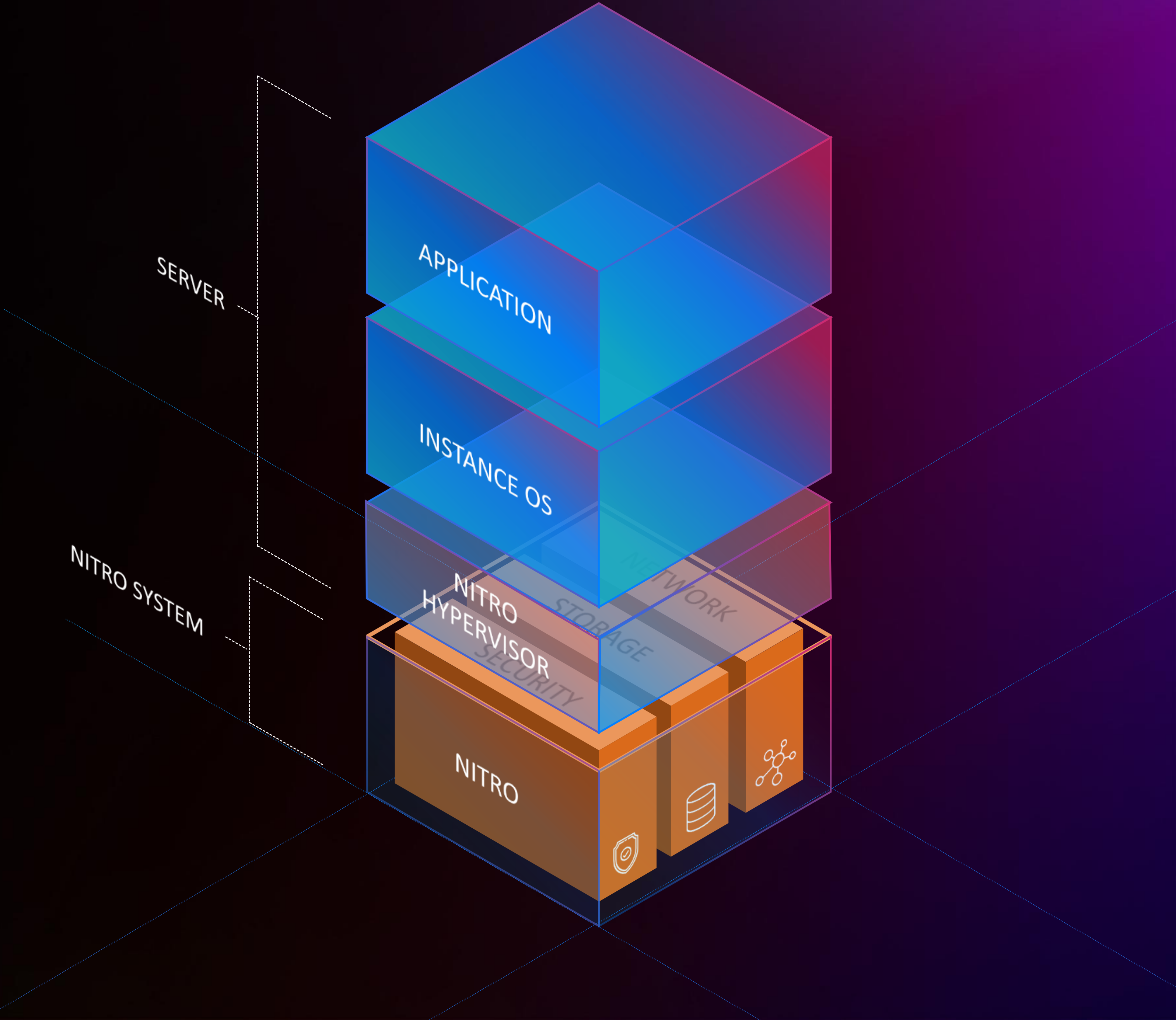
Memory-optimized
R7iz | R6i | R6in | X2idn / X2iedn | Z1d

Accelerated compute
Gaudi Instances | P4 | G4dn | F1



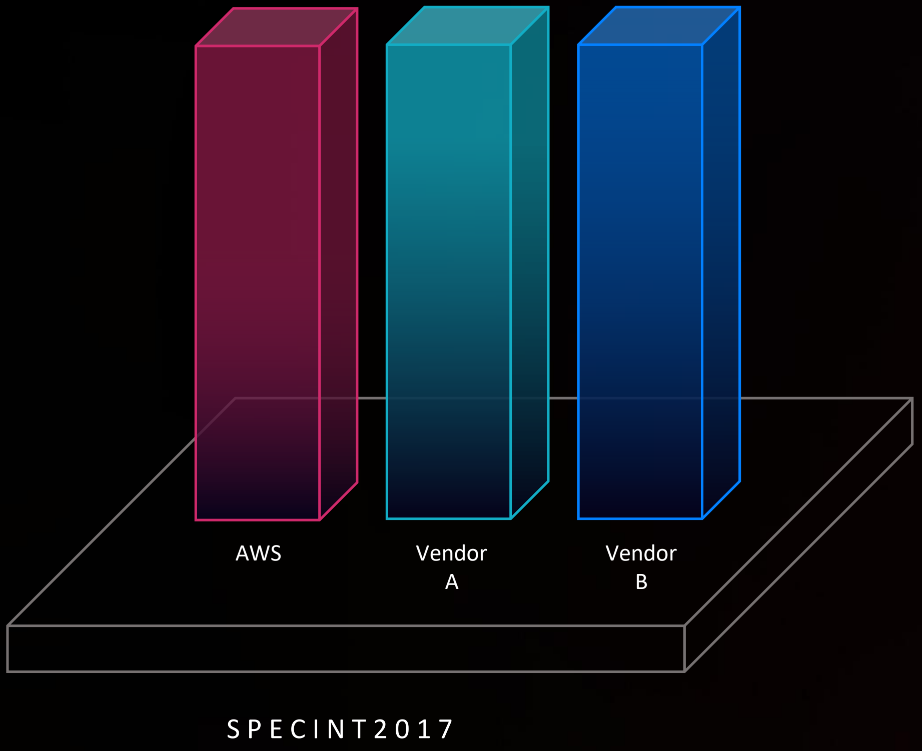
The AWS Nitro System architecture

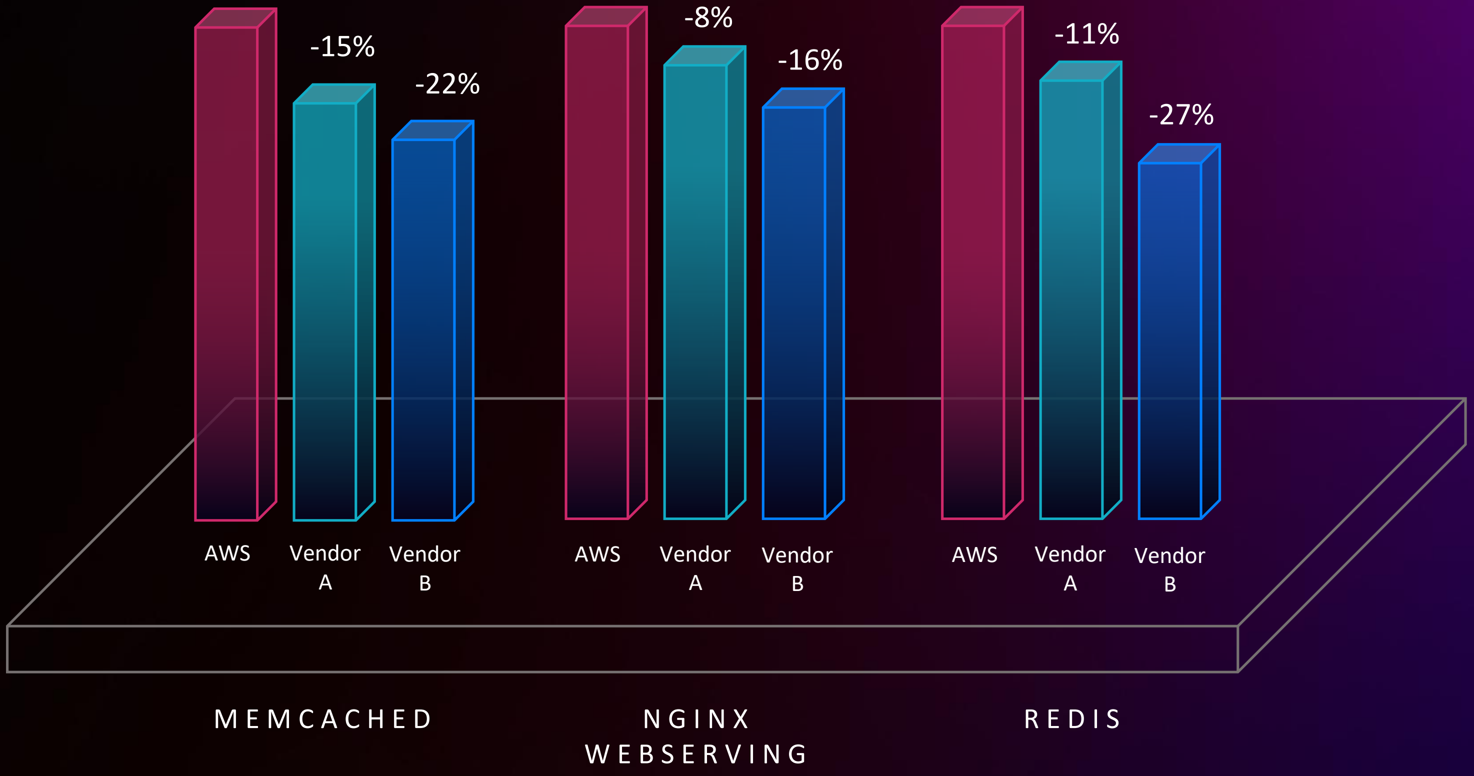
Offering strong security, performance, and innovation in the cloud



Nitro performance for real-world workloads

Amazon EC2 instances can deliver over 15% higher throughput performance





The Nitro System

Nitro Cards

VPC Networking

Amazon Elastic Block Store (Amazon EBS)

Instance Storage

System Controller

Nitro Security Chip

Integrated into motherboard

Protects hardware resources

Hardware Root of Trust

Nitro Hypervisor

Lightweight hypervisor

Memory and CPU allocation

Bare Metal-like performance

Innovating with Intel

16 YEARS OF COLLABORATION AND INNOVATION WITH AWS

Collaboration

Deep engineering collaboration across AWS portfolio

Extensive integration

Over 350 Amazon EC2 instances are powered by Intel processors

Fastest

Fastest processor in the cloud and widest selection of Ice Lake instances

Recent intel-based instances

I4I
STORAGE-OPTIMIZED

X2IDN
MEMORY-OPTIMIZED

X2IEDN
MEMORY-OPTIMIZED

X2IEZN
HIGH-FREQUENCY

NEW
M6ID
GENERAL PURPOSE

NEW
C6ID
COMPUTE-OPTIMIZED

NEW
R6ID
MEMORY-OPTIMIZED

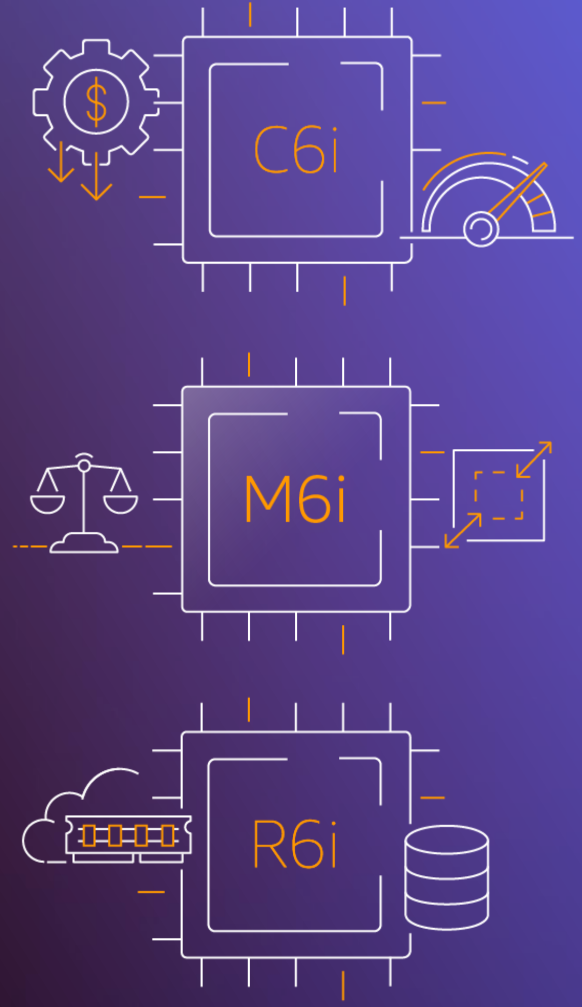
Amazon EC2 C6id, M6id, and R6id instances

AVAILABLE NOW

NEW!

EC2 instances powered by 3rd gen Intel Xeon Scalable processor and NVMe attached storage

- Equipped with up to 7.6 TB of local NVMe-based SSD block-level storage for workloads that needs access to high-speed, low-latency storage
- Deliver up to 15% better price performance compared to previous gen C5d, M5d, and R5d instances
- Up to 2x faster networking and 20% higher memory bandwidth
- Support for Total Memory Encryption (TME)
- Ideal for core computing workloads that need access to high-speed, low latency storage.



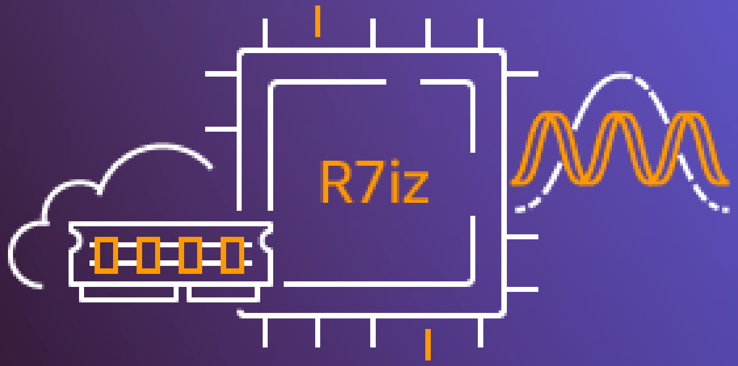
Amazon EC7 R7iz instances

PREVIEW

NEW!

High-frequency memory-optimized instances powered by 4th generation Intel Xeon Scalable processor

- Up to 128 vCPU, up to 1 TiB of memory to provide up to 2.6x more vCPU and memory compared to comparable high frequency instances
- Up to 20% higher performance when compared to comparable high frequency instances
- First x86-based EC2 instance to use DDR5 memory and deliver up to 2.4x higher mem bandwidth over comparable high frequency instances
- Designed for workloads such as front-end Electronic Design Automation (EDA), relational database workloads with high per-core licensing fees, and financial, actuarial, and data analytics simulation workloads



High performance computing (HPC)

Intel-based EC2 instances power the most computationally demanding applications in a cost-effective way at scale. Intel and AWS offer a comprehensive set of compute, networking, storage, and visualization technologies to give customers an ideal environment for HPC workloads. Coupled with an extensive partner ecosystem, customers are empowered to innovate more freely.

Workload	Instance family	Instance family	Best use cases	Notable features
HPC	Compute networking performance	C6in	• Ideal choice for HPC workloads, data lakes • Network appliances that can take advantage of improved network throughput and packet rate performance	• Up to 200 Gbps network bandwidth • 80 Gbps of EBS bandwidth • EFA support on the 32xlarge and metal sizes
	Compute performance	C6i	• Optimized for compute-intensive workloads • Deliver cost-effective high performance at a low price per compute ratio	• AVX-512 • 4GB/core memory • Intel Total Memory Encryption (TME)
	Fastest compute	R7iz z1d M5zn	• R7iz and z1d targets both memory- and compute-intensive apps • R7iz and z1d is ideal for EDA, gaming, and certain relational database workloads with high per-core licensing costs	• High single-thread performance with sustained all core frequency up to 4.5 GHz • z1d = 16 GiB/vCPU memory • z1d has up to 1.8 TB of instance storage • M5zn – up to 100 Gbps network bandwidth
	Balanced networking	M6i (+M6in, M6idn)	• General purpose instance that provides a balance of compute, memory, and network resources • Good for many applications including web, application and gaming servers, and small to mid-size databases	• 8 GB/core memory • Up to 200 Gbps network bandwidth (M6in) • Up to 7.6 TB of instance storage

NEW

Amazon EC2 Hpc6id instances

Best price performance for memory and data-intensive HPC workloads in Amazon EC2

200G networking with EFA

2x higher Elastic Fabric Adaptor performance over current generation HPC instances for increased application performance

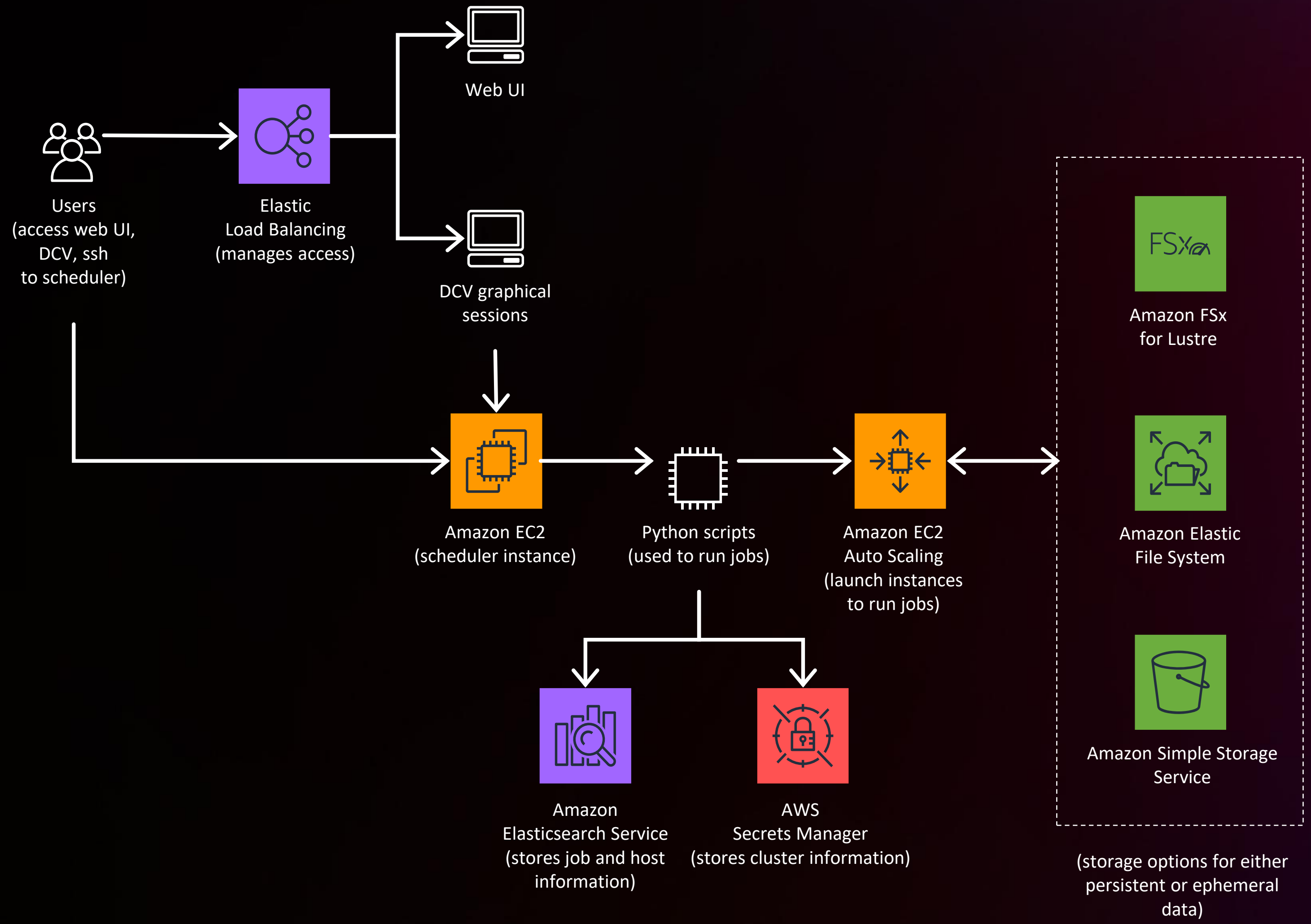
Price performance benefits

Up to 2.2x better price-performance for data-intensive HPC workloads such as Finite Element Analysis (FEA) over comparable x86-based instances

Optimized for for data intensive HPC workloads

1TB of instance memory and 15.2 TB of NVMe storage to accelerate seismic, energy, and FEA workloads

Scale-Out Computing on AWS



Users (access web UI, DCV, ssh to scheduler)

Elastic Load Balancing (manages access)

Web UI

DCV graphical sessions

Amazon EC2 (scheduler instance)

Python scripts (used to run jobs)

Amazon EC2 Auto Scaling (launch instances to run jobs)

Amazon FSx for Lustre

Amazon Elastic File System

Amazon Simple Storage Service

(storage options for either persistent or ephemeral data)

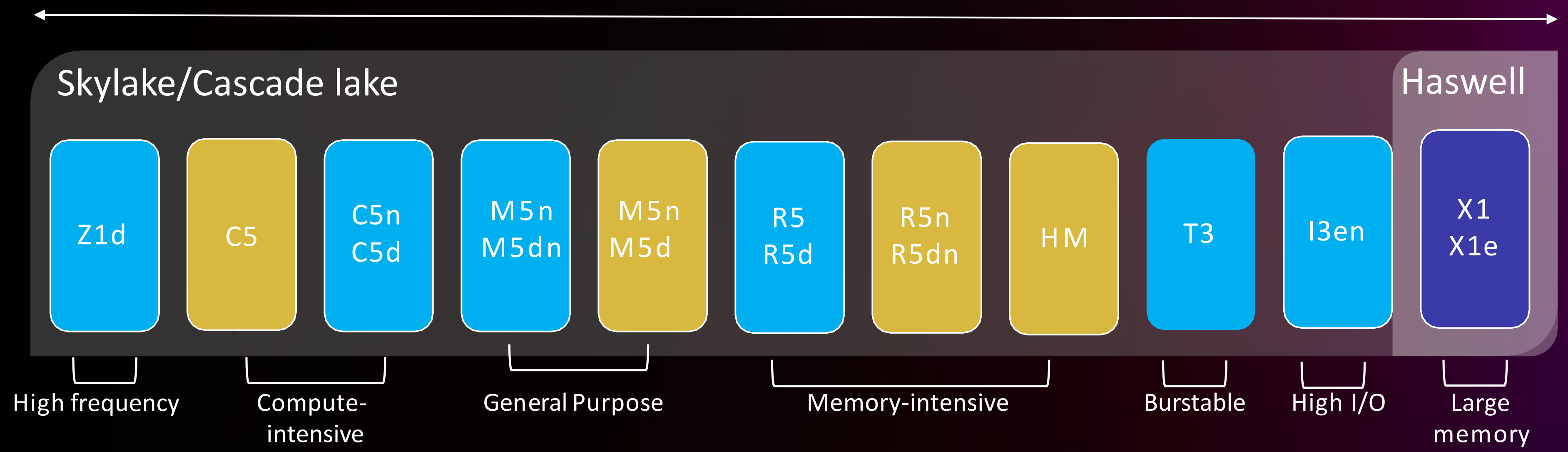
Amazon Elasticsearch Service (stores job and host information)

AWS Secrets Manager (stores cluster information)

Intel-based Amazon EC2 instances for ML

Second-generation Intel Xeon Scalable processor

DL boost for inference, single-node training



Skylake/Cascade lake

Haswell

Z1d

C5

C5n
C5d

M5n
M5dn

M5n
M5d

R5
R5d

R5n
R5dn

HM

T3

I3en

X1
X1e

High frequency

Compute-intensive

General Purpose

Memory-intensive

Burstable

High I/O

Large memory

M5, C5 instances are suitable for all computer vision, ML, and DL inference workloads

R5 instances are for memory-intensive workloads that use 3D-CNN/BERT- large/T5 topologies with memory requirement more than 192 GB

T3 instances are better suited for ML applications and low-compute DL inference applications

C5n instances are suitable for distributed deep learning training due to the high NW performance required for inter-node communication

Bare metal instances are preferred for large topologies such as HVM-based instances, which add ~10% performance overhead

For more info, visit https://aws.amazon.com/ec2/instance-types/

Habana Gaudi-based instances – DL1

ML TRAINING POWERED BY NEW HABANA GAUDI PROCESSORS FROM INTEL

habana®
An Intel Company

New Amazon EC2 instances built specifically for ML training and powered by up to 8 new Habana Gaudi processors from Intel

Will deliver up to 40% lower cost to train deep learning models over GPU-based instances

Will allow customers to iterate and train models more frequently

Benefit from full stack of Amazon EC2 services – DL AMIs, DLC for containerized applications, ultimately Amazon SageMaker

Developers can implement Gaudi-based instances via Amazon ECS and Amazon EKS for containerized applications

Will support common frameworks like TensorFlow and PyTorch

Wide range of ML workloads for applications including NLP, image classification, object detection, and recommendation systems

For efficient scaling across multiple Gaudi-based Amazon EC2 instances, support for AWS Elastic Fabric Adapter

Resource optimization

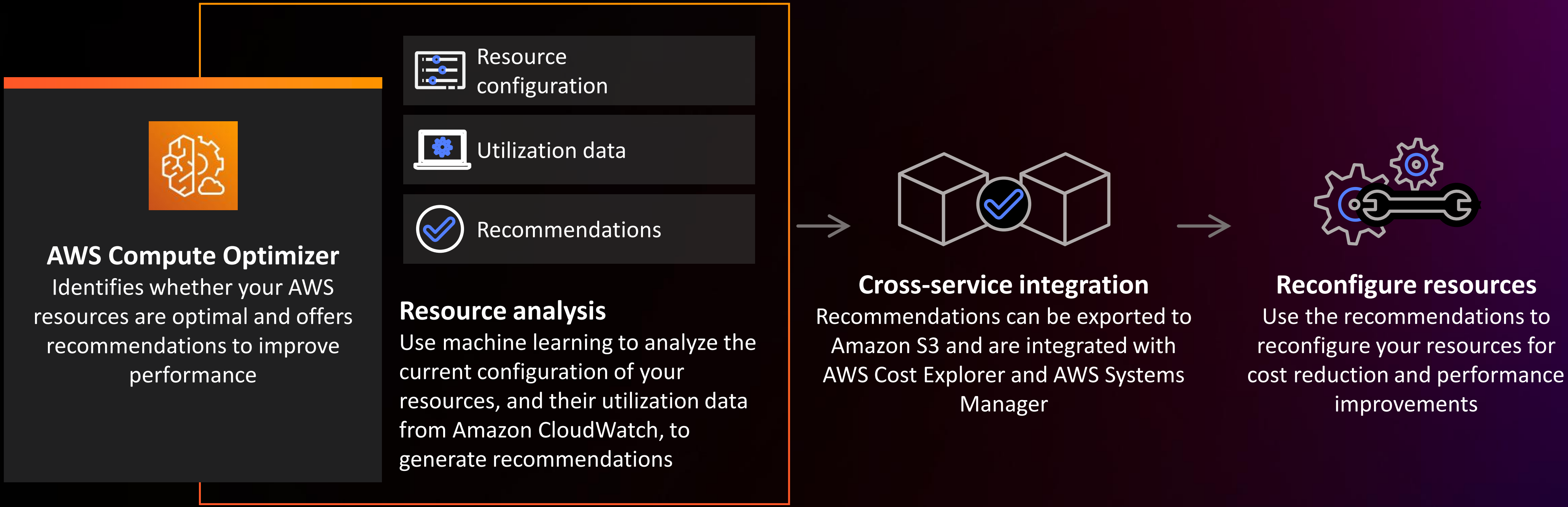
Cost

Maximize value you derive from your spend

Performance

Ensure your provisioned capacity meets workload requirements

AWS Compute Optimizer



AWS Compute Optimizer
Identifies whether your AWS resources are optimal and offers recommendations to improve performance

- Resource configuration
- Utilization data
- Recommendations

Resource analysis
Use machine learning to analyze the current configuration of your resources, and their utilization data from Amazon CloudWatch, to generate recommendations

Cross-service integration
Recommendations can be exported to Amazon S3 and are integrated with AWS Cost Explorer and AWS Systems Manager

Reconfigure resources
Use the recommendations to reconfigure your resources for cost reduction and performance improvements

Example: Amazon EC2 instances

M5.2xlarge

vCPU: 8
RAM: 32 GiB
Instance storage: EBS only
Network: Up to 10 Gbps
Estimated monthly cost: $280.32

- ~40% CPU utilization during the day
- ~10% CPU utilization during the night
- ~30% RAM usage throughput
- <1 Mbps network usage more than 99% of the time
- <2 IOPS more than 99% of the time

Option 1

M5.xlarge

vCPU: 4
RAM: 16 GiB
Instance storage: EBS only
Network: Up to 10 Gbps
Estimated monthly cost: $140.16
Savings: 50.0%
Risk: Low

Option 2

T3.xlarge

vCPU: 4
RAM: 16 GiB
Instance storage: EBS only
Network: Moderate
Estimated monthly cost: $121.47
Savings: 56.7%
Risk: Medium

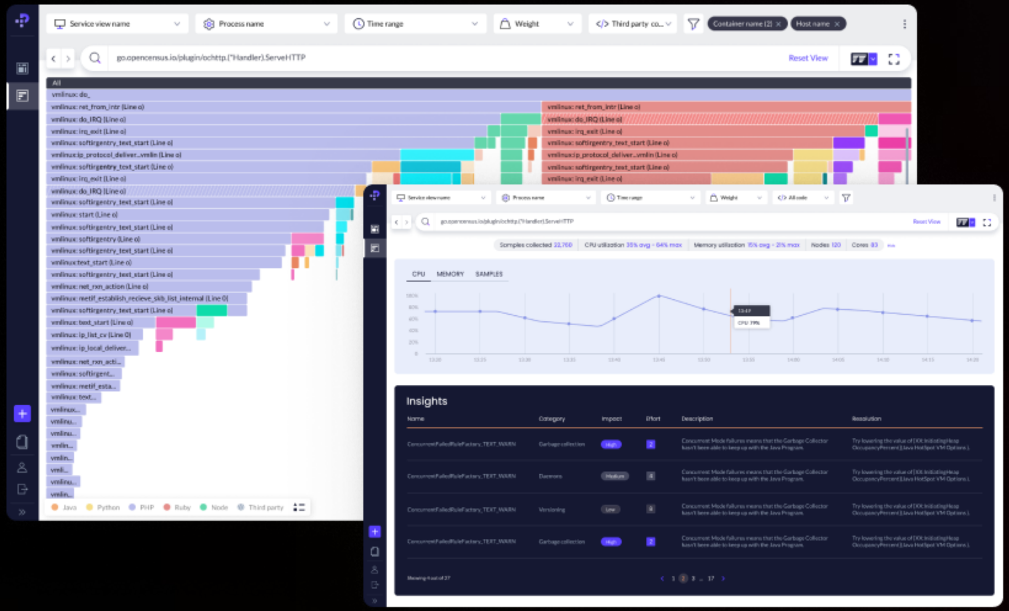
Option 3

R5.large

vCPU: 2
RAM: 16 GiB
Instance storage: EBS only
Network: Up to 10 Gbps
Estimated monthly cost: $91.98
Savings: 67.2%
Risk: Medium

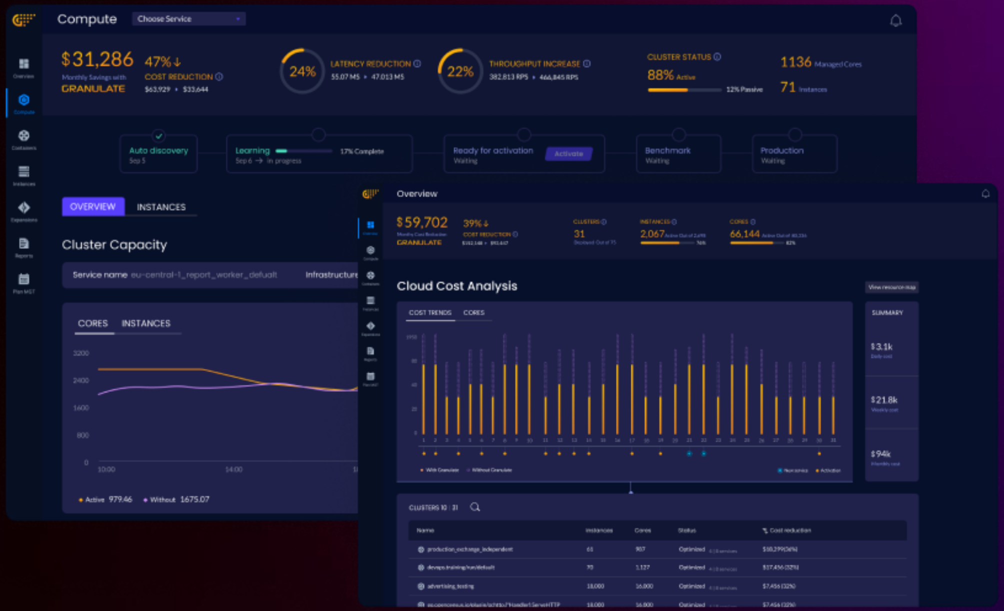
Continuous profiling and continuous optimization

gProfiler



+

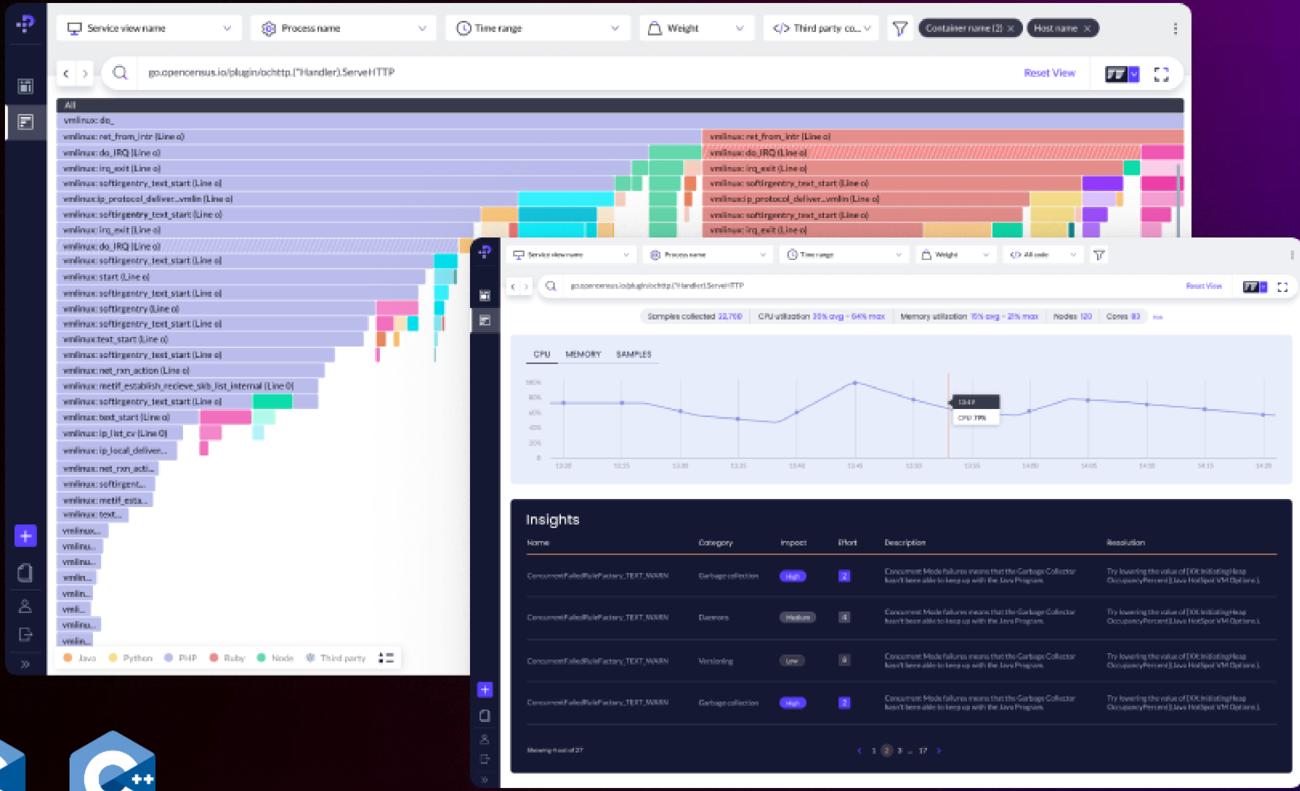
GRANULATE



gProfiler

PRODUCTION PROFILING, MADE EASY

- Low-overhead continuous production CPU profiling
- Free and open source
- SaaS and standalone deployments
- One of the first optimization-driven profilers
- Purpose-built for collaboration
- Wide runtime coverage:


Service view name
Process name
Time range
Weight
Third party co...
Container name (2)
Host name
go.opencensus.io/plugin/ochttp.(*Handler).ServeHTTP
Reset View
All
vmlinux: do_
vmlinux: net_from_intr (Line o)
vmlinux: do_IRQ (Line o)
vmlinux: irq_exit (Line o)
vmlinux: softirqentry_text_start (Line o)
vmlinux: ip_protocol_deliver...vmlin (Line o)
vmlinux: start (Line o)
vmlinux: net_rxn_action (Line o)
Java
Python
PHP
Ruby
Node
Third party
CPU
MEMORY
SAMPLES
Insights

Name	Category	Impact	Effort	Description	Resolution
ConcurrentFailedRuleFactory_TEXT_WARN	Garbage collection	High	2	Concurrent Mode failures means that the Garbage Collector hasn't been able to keep up with the Java Program.	Try lowering the value of [XX:InitiatingHeapOccupancyPercent] Java HotSpot VM Options].
ConcurrentFailedRuleFactory_TEXT_WARN	Daemons	Medium	4	Concurrent Mode failures means that the Garbage Collector hasn't been able to keep up with the Java Program.	Try lowering the value of [XX:InitiatingHeapOccupancyPercent] Java HotSpot VM Options].
ConcurrentFailedRuleFactory_TEXT_WARN	Versioning	Low	8	Concurrent Mode failures means that the Garbage Collector hasn't been able to keep up with the Java Program.	Try lowering the value of [XX:InitiatingHeapOccupancyPercent] Java HotSpot VM Options].
ConcurrentFailedRuleFactory_TEXT_WARN	Garbage collection	High	2	Concurrent Mode failures means that the Garbage Collector hasn't been able to keep up with the Java Program.	Try lowering the value of [XX:InitiatingHeapOccupancyPercent] Java HotSpot VM Options].

Showing 4 out of 27

‹ 1 2 3 ... 17 ›

Java

Optimization use case

Regex hogging CPU

gProfiler process

Background

Deployment of new service

10-machine cluster

gProfiler deployed on inception

Investigation

40% of CPU spent performing regex

Regex is not CPU-efficient

Optimization

Replace regex by finite state machine

Reduced CPU utilization from 40% to 9%

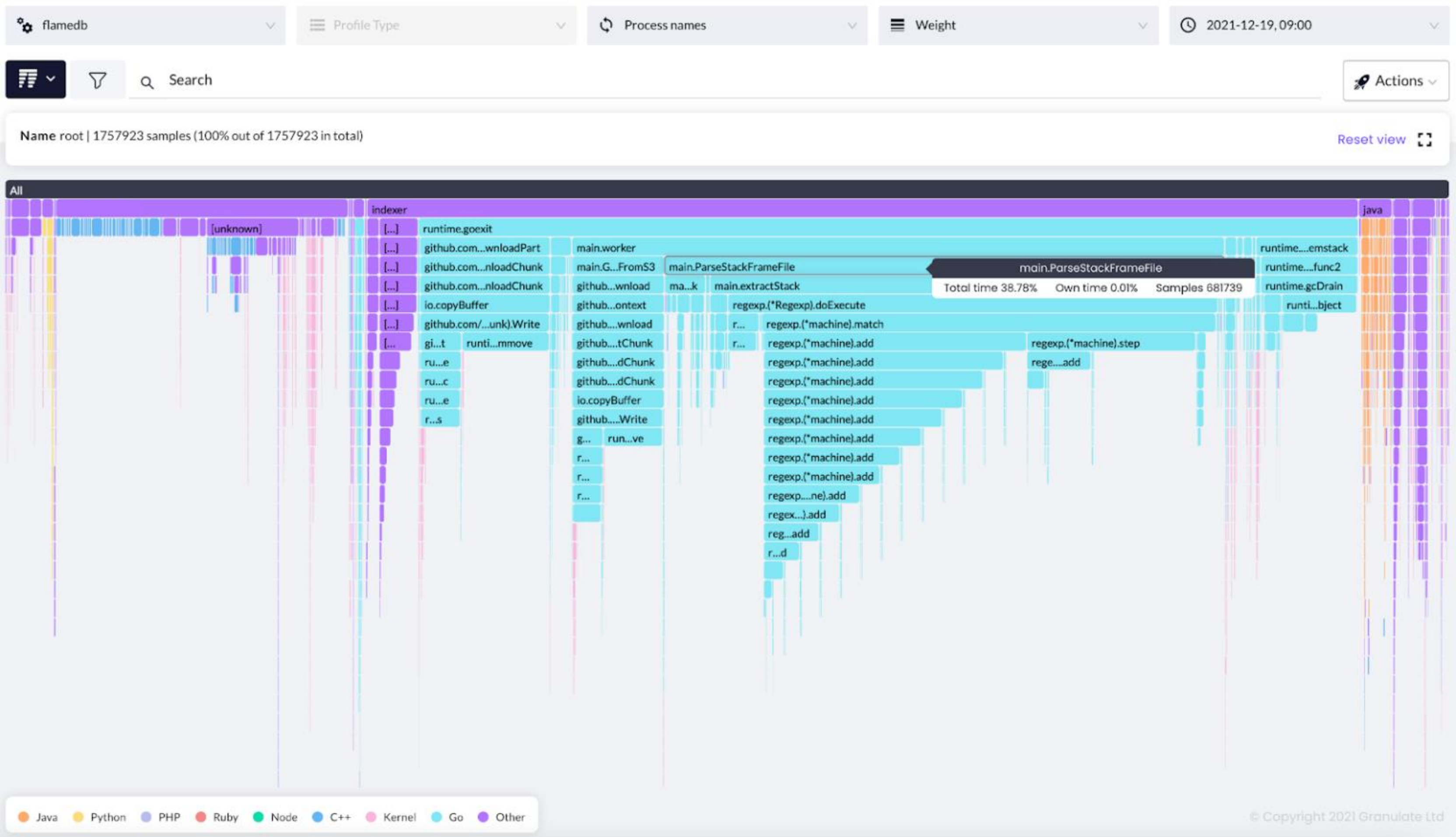
Results

CPU utilization of cluster drops from 50% to 25%

Cluster size decreased to 4 machines

Regex hogging CPU

Before optimization



flamedb | Profile Type | Process names | Weight | 2021-12-19, 09:00

Search

Actions

Name root | 1757923 samples (100% out of 1757923 in total)

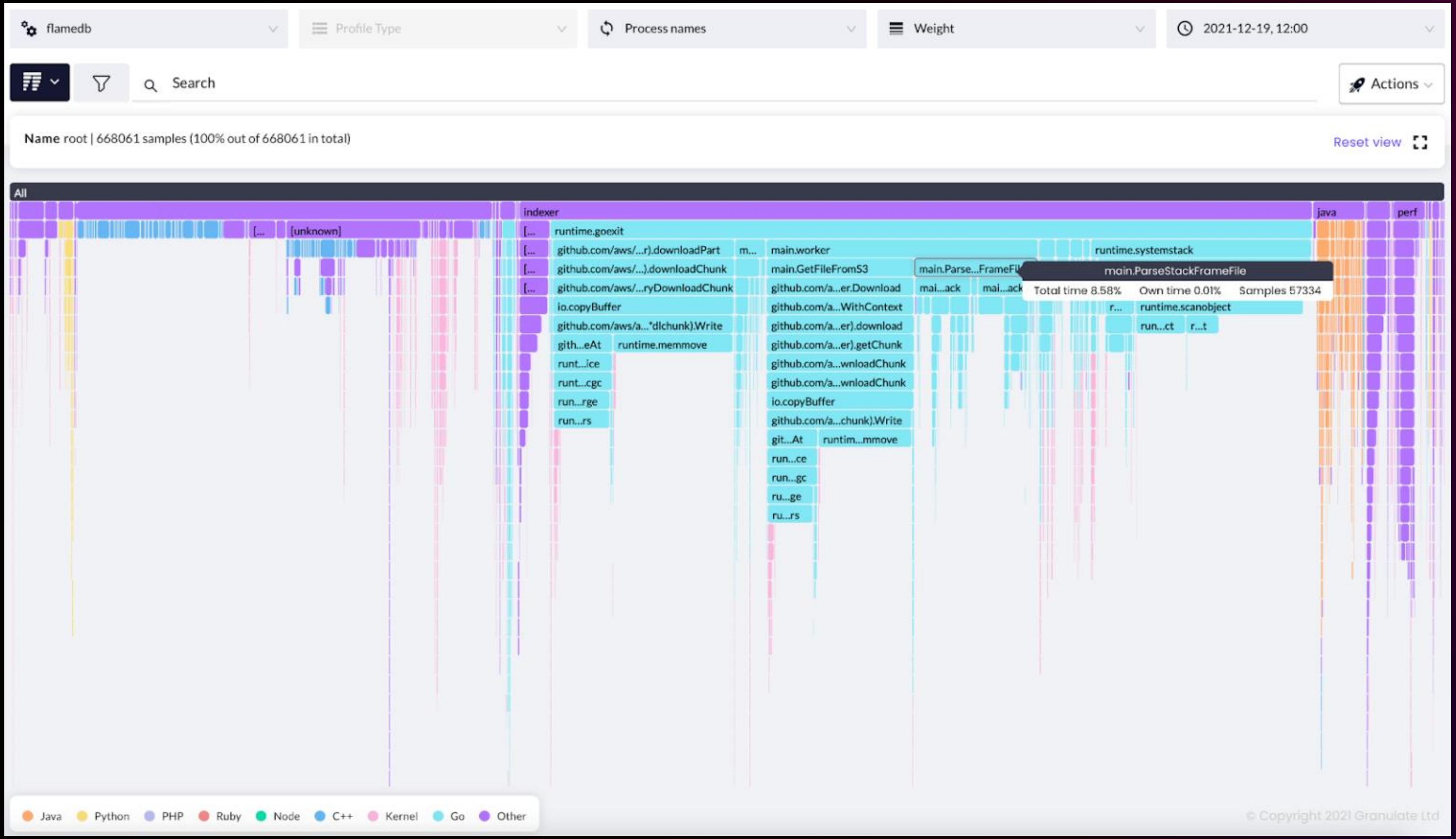
Reset view

main.ParseStackFrameFile

Total time 38.78% Own time 0.01% Samples 681739

Regex hogging CPU

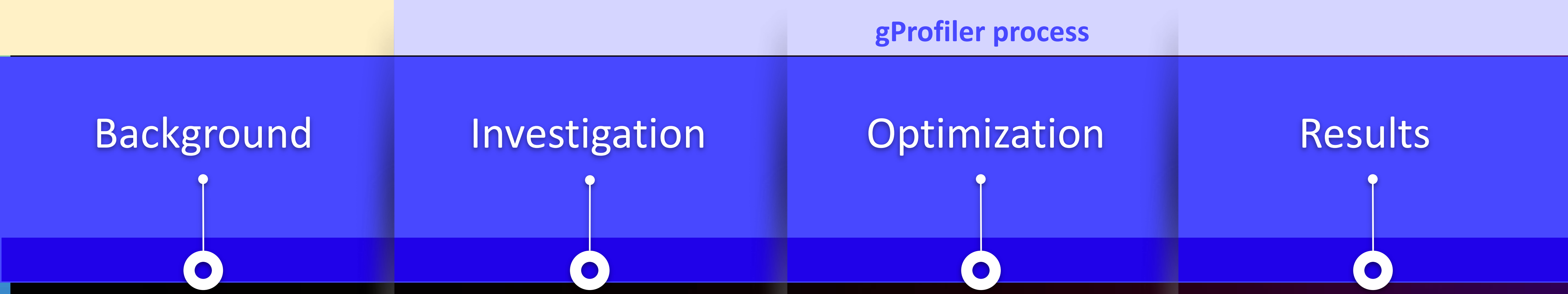
After optimization


flamedb
Profile Type
Process names
Weight
2021-12-19, 12:00
Search
Actions
Reset view
All
indexer
java
perf
[unknown]
runtime.goexit
github.com/aws/...r).downloadPart
github.com/aws/...).downloadChunk
github.com/aws/...ryDownloadChunk
io.copyBuffer
github.com/aws/a...*dlchunk).Write
gith...eAt
runtime.memmove
main.worker
main.GetFileFromS3
main.Parse...FrameFil
github.com/a...er.Download
github.com/a...WithContext
github.com/a...er).download
github.com/a...er).getChunk
github.com/a...wnloadChunk
io.copyBuffer
github.com/a...chunk).Write
runtime.systemstack
runtime.scanobject
main.ParseStackFrameFile
Java
Python
PHP
Ruby
Node
C++
Kernel
Go
Other
© Copyright 2021 Granulate Ltd

Name root | 668061 samples (100% out of 668061 in total)

main.ParseStackFrameFile
Total time 8.58% Own time 0.01% Samples 57334

Optimization use case



gProfiler process

Background

Turned to profiling to reduce cloud costs

Django (Python) application

Investigation

18.5% of CPU time spent on _get_pk_val function

Indicates unnecessary queries to the database

Optimization

Cache the returned values

Compare the values and spare unnecessary database operations

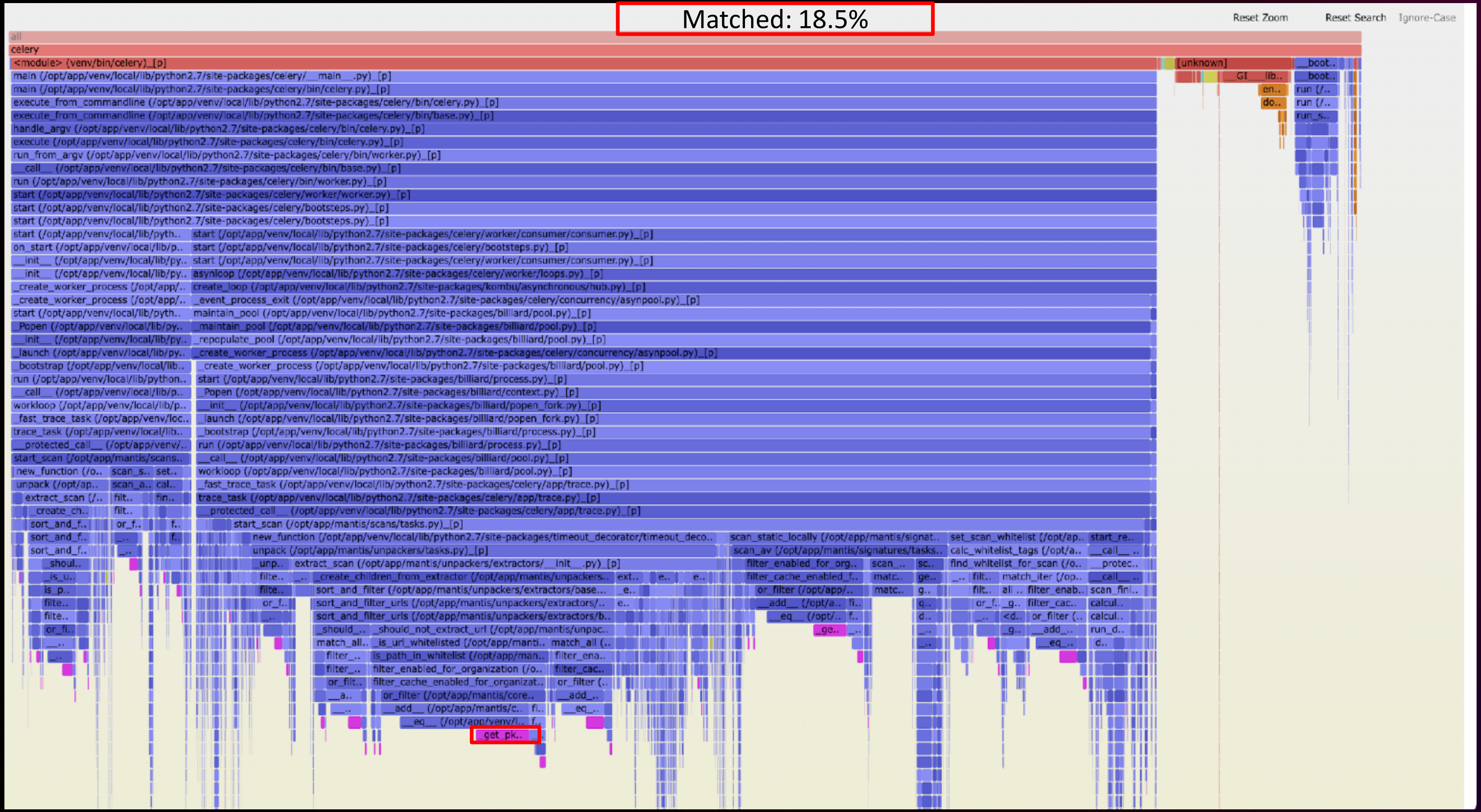
Results

3% of CPU time spent on _get_pk_val function

Cluster size decreased by 15%

Deep vs. shallow comparison

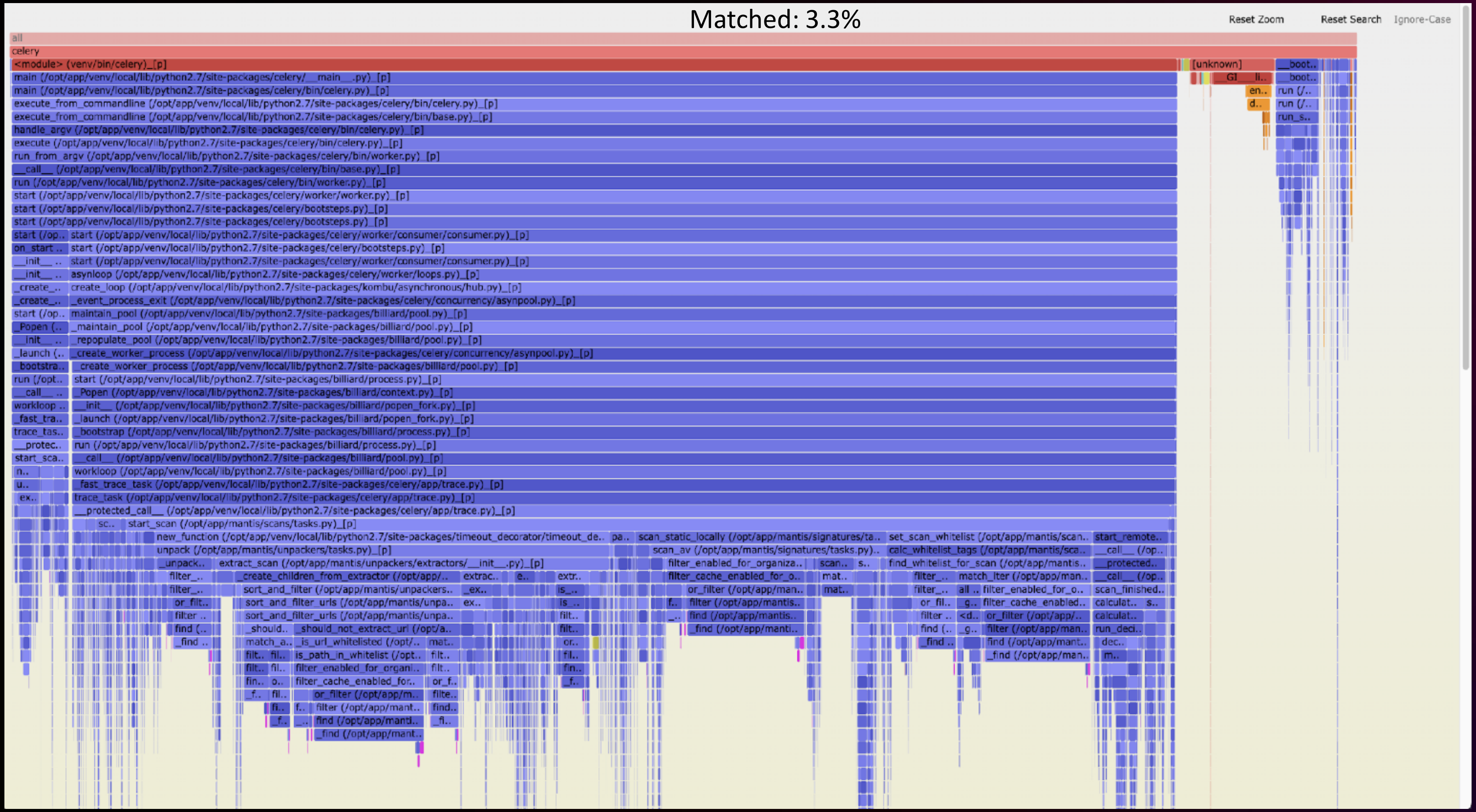
Before optimization



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Deep vs. shallow comparison

After optimization



Granulate optimization process

UP TO 2 WEEKS

1

PROFILING

Start by identifying potential for optimizations on relevant customer workloads using gProfiler

2

LEARNING

Agent learns workload data patterns and makes optimized resource management decisions in real time

3

OPTIMIZING

Immediately lower CPU utilization and latency with adjusted OS- and runtime-level resource allocation

4

COST REDUCTION

Realize lower costs by leveraging improved machine performance to reduce cluster size and compute spending

Use cases

LINUX-BASED

*80% of production workloads run on Linux

EVERY ARCHITECTURE

Cloud

Hybrid

On premises

MICROSERVICES

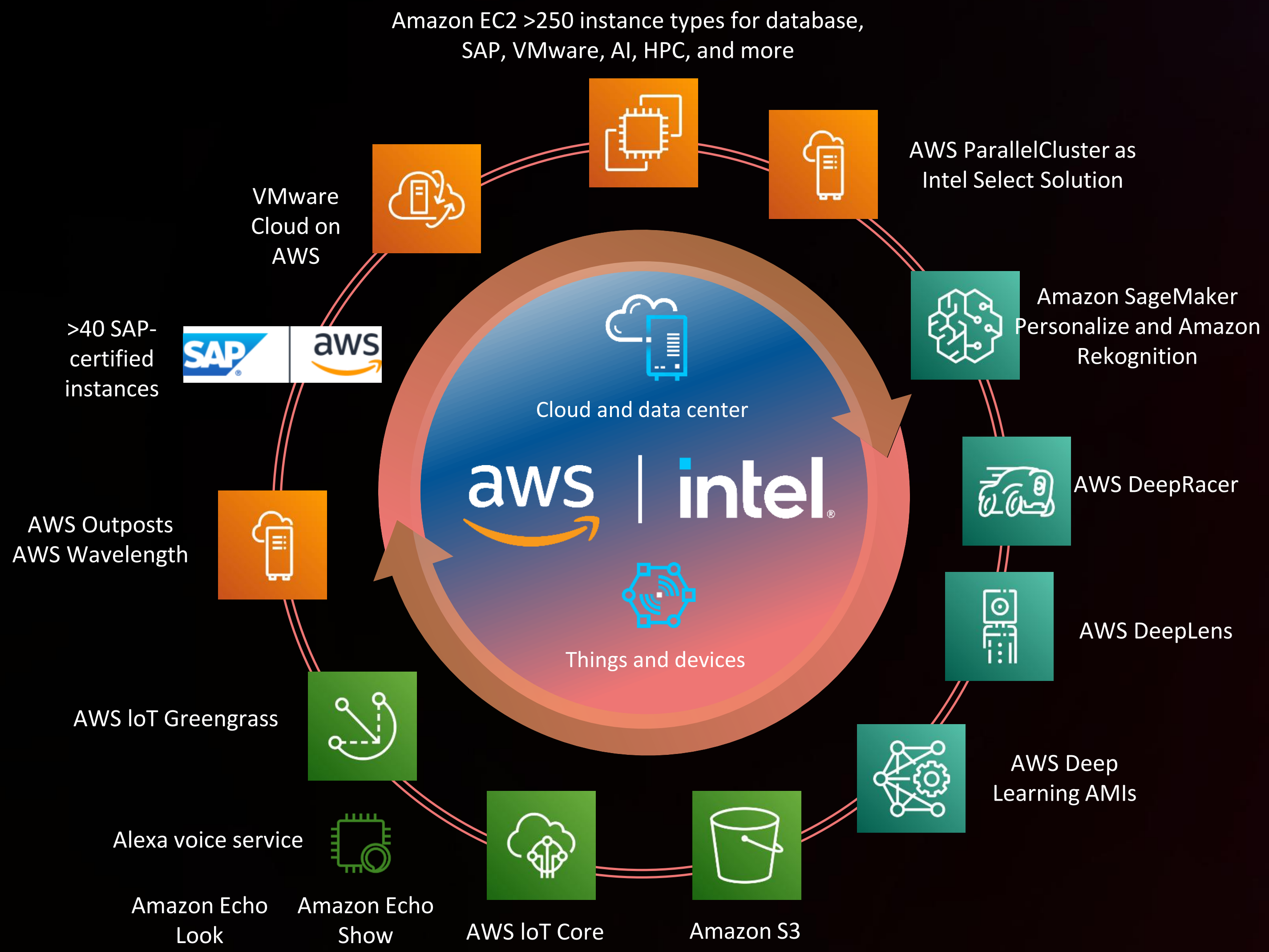
COMPUTE

BIG DATA

APACHE Spark

Elasticsearch

Takeaways



Amazon EC2 >250 instance types for database, SAP, VMware, AI, HPC, and more
AWS ParallelCluster as Intel Select Solution
Amazon SageMaker Personalize and Amazon Rekognition
AWS DeepRacer
AWS DeepLens
AWS Deep Learning AMIs
Amazon S3
AWS IoT Core
Amazon Echo Show
Amazon Echo Look
Alexa voice service
AWS IoT Greengrass
AWS Outposts AWS Wavelength
>40 SAP-certified instances
SAP
aws
VMware Cloud on AWS
Cloud and data center
aws | intel
Things and devices

- Close collaboration between Intel and AWS has resulted in excellent end user experience and customer successes
- Instance types with the best TCO on Intel can accelerate your customers' applications across a variety of workloads
- Existing solutions for deployment with many successful outcomes can deliver both high performance and cost savings
- Boost application performance and reduce infrastructure cost with continuous profiling and continuous optimization

Thank you!

kevinshs@amazon.com