aws health

AWS for Life Sciences

Accelerating HCLS Research and Discovery

Potential Impact of Next Generation Chipsets

Steve Litster Ph. D.

Senior Manager HCLS
Compute Specialists

Zheng Yang Ph. D.

WW Head, Strategy and Solutions, AI/ML, HCLS

Quote

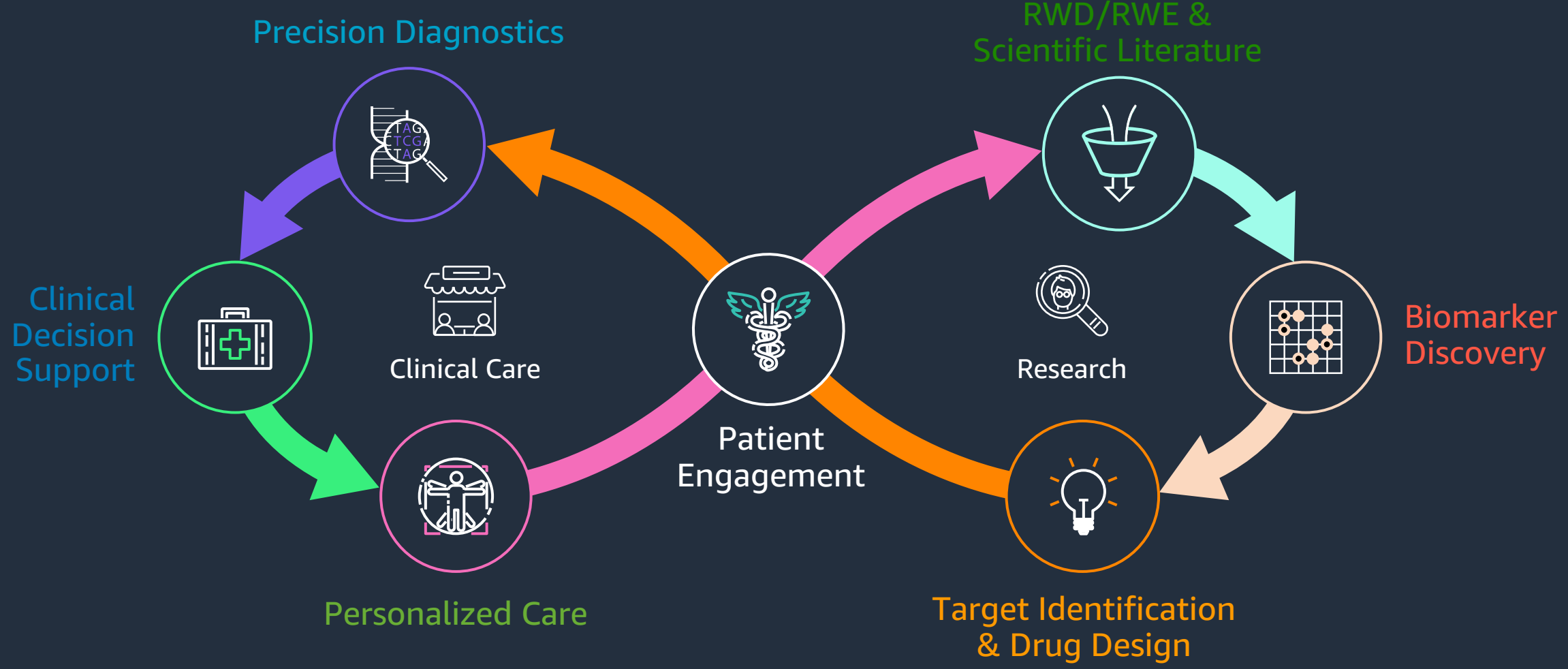
Access to supercomputers

“The science is well ahead of our ability to implement it. It's quite clear that if we could run our models at a higher resolution we could do a much better job-tomorrow-in terms of our seasonal and decadal predictions. It's so frustrating. We keep saying we need four times the computing power. We're talking just 10 or 20 million a year-dollars or pounds-which is tiny compared to the damage done by disasters. Yet it's a difficult argument to win.”

Julia Slingo OBE
Chief Scientist at the Met Office

The precision health continuum

APPLICATION & FLOW OF DATA, INFORMATION & KNOWLEDGE TO IMPROVE INDIVIDUAL AND POPULATION OUTCOMES



aws health

AWS “HPC” R&D Life Science Stack

Orchestration/Automation

ParallelCluster, Batch, CloudFormation, EKS, ECS

Visualization

NICE DCV, AppStream,

Scientific Applications

Genomics

(NextFlow, RNA-Seq, WGS, Cromwell)

Imaging

(Relion, CryoSparc, Digital Pathology)

Modeling and Simulation

(CFD, CSP, NONMEM)

Computational Chemistry

(OpenEye, MD , Virtual Screening)

Informatics Data Science

(R, SAS, Pytorch, Jupyter Notebooks.)

Compute

CPU, GPU, FPGA

Storage

S3, EFS, EBS, FSx for Lustre

Networking

EFA, ENA, VPC

Data and Security

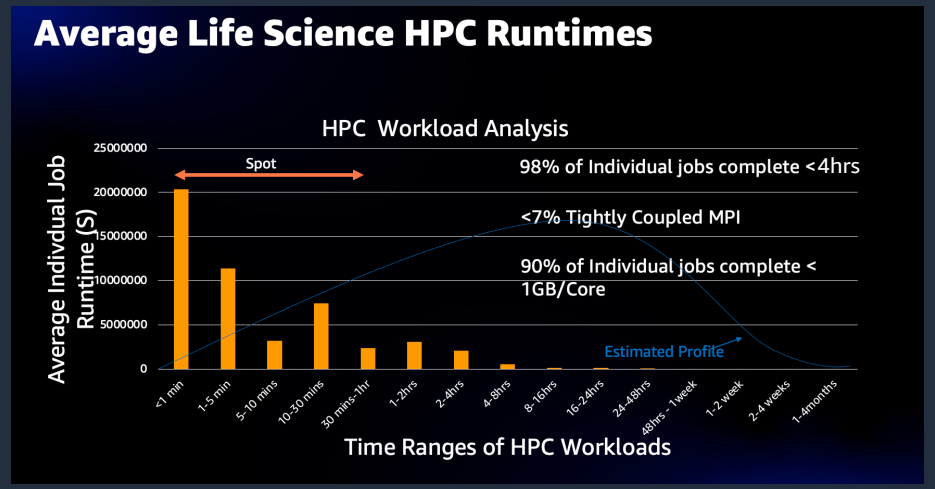
aws health

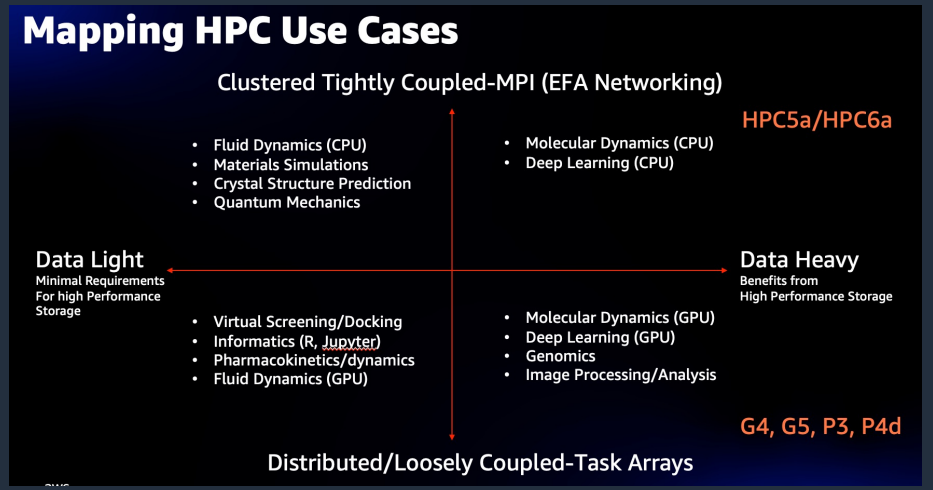
Co-develop a “Fit For Purpose” Approach

- Assessment of Workload
  - Review Data Residency, Compliance, Scientific Impact, Costs etc. requirements

- Apply Fit for Purpose Model
  - Reuse & modify proven workloads
  - Align Instance Types
  - Optimized Storage

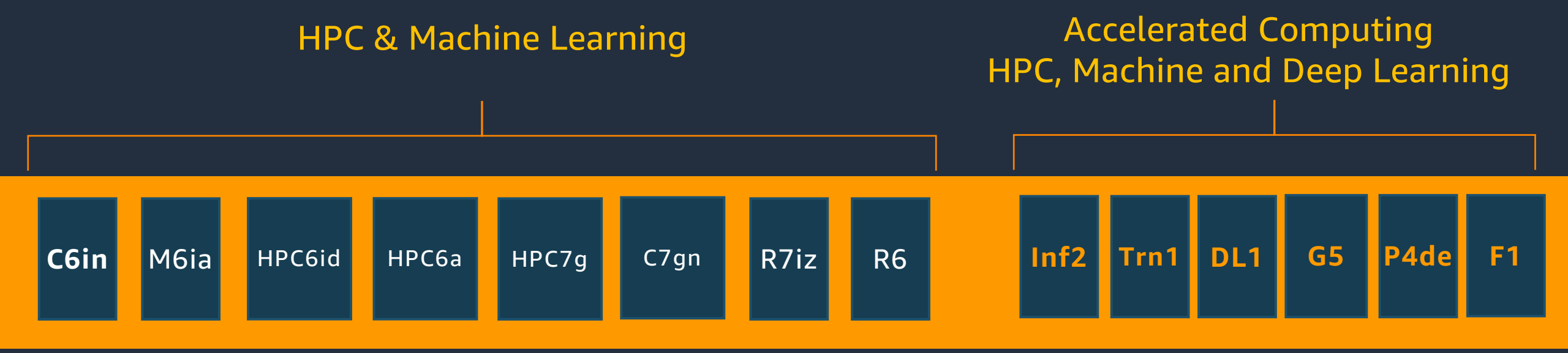
Scientific Area	Average per job [min]
Molecular Dynamics	522
Structural Biol	100
Virtual Screening	49
Modeling and Simulation	37
Electron Microcopy	36
Genomics	33
Imaging	21
Informatics	4





Broadest and Deepest Computing Choice (500+ Instances)

CPU, GPU & Custom EC2 Instances for HPC and ML



Cascade Lake CPU
Skylake CPU

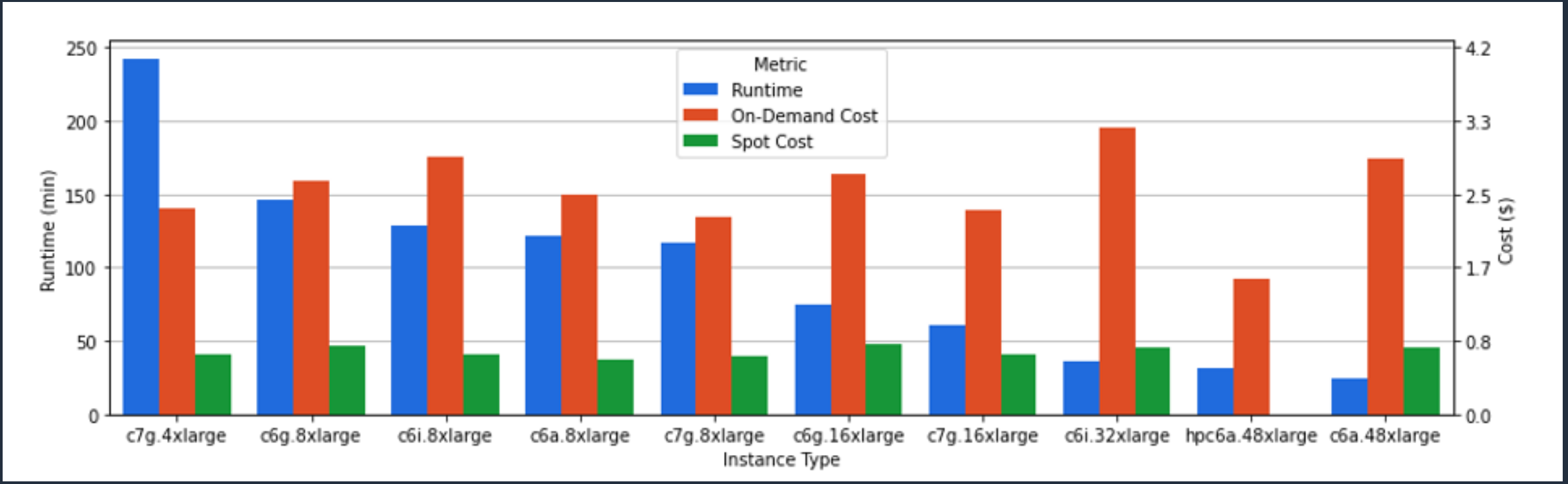
EPYC CPU

Graviton CPU
Inferentia Chip

A100, V100, T4 GPUs

Benchmarking Sentieon DNAseq pipeline

FASTQ to VCF processing with the Illumina 30X NovaSeq dataset



https://aws.amazon.com/blogs/hpc/cost-effective-and-accurate-genomics-analysis-with-sentieon-on-aws/

AWS Graviton2/3-based Amazon EC2 instances

UP TO 40% BETTER PRICE-PERFORMANCE OVER COMPARABLE X86-BASED INSTANCES

M6g, M6gd

General purpose workloads

T4g

Burstable general purpose workloads

C6g, C7gn, HPC7g NEW!

Compute-intensive workloads

R6g, R6gd, X2gd

Memory-intensive workloads

Im4gn, Is4gen

Storage-intensive workloads

G5g

GPU-based graphics and machine learning workloads

AVAILABLE ACROSS 23 AWS REGIONS GLOBALLY*

* Not all instances available in all regions

aws health

AWS Graviton ease of adoption

As a rule, the more current your software stack the better

Difficulty	Workload	Actions
Virtually no effort	RDS, Aurora, ElastiCache, OpenSearch, MemoryDB, & Neptune	Upgrade to latest and enjoy
Super easy	Amazon EMR –"big data" workloads	Typically, just works
Pretty easy	AWS Lambda	Typically, just works with Lambda managed runtimes or base images. Watch: JNI or Python-native modules
Quite easy	Linux – Interpreted and JIT'd languages (e.g., Java, PHP, Node.js)	Select Arm64 AMI and Install Bonus if containerized Watch: JNI or Python-native modules
More involved	Linux – Compiled languages (e.g., C/C++, Python, Go)	Select Arm64 AMI and compile Watch: port any intrinsics or assembly
Some work, high reward	Microsoft Windows – .NET	Migrate to Linux + .NET core on Arm64

Purpose built compute for ALL HPC Workloads

Compute Intensive

Hpc6a

AMD Milan 96 Cores
384GB RAM
100Gbit EFA

Compute and Network Intensive

Hpc7g

NEW

Graviton 3E, Arm 64 cores
128GB RAM
200Gbit EFA

preview

Graviton3E Processor

Up to 35% faster vector instruction performance vs Graviton 3

Data and Memory Intensive

Hpc6id

NEW

Intel Ice Lake 64 cores
1TB RAM
200Gbit EFA
15.2 TB NVME

Mem Optimized High Freq

R7iz

NEW

Intel SPR 64 cores
3.9 GHz
1 TB DDR5 RAM

preview

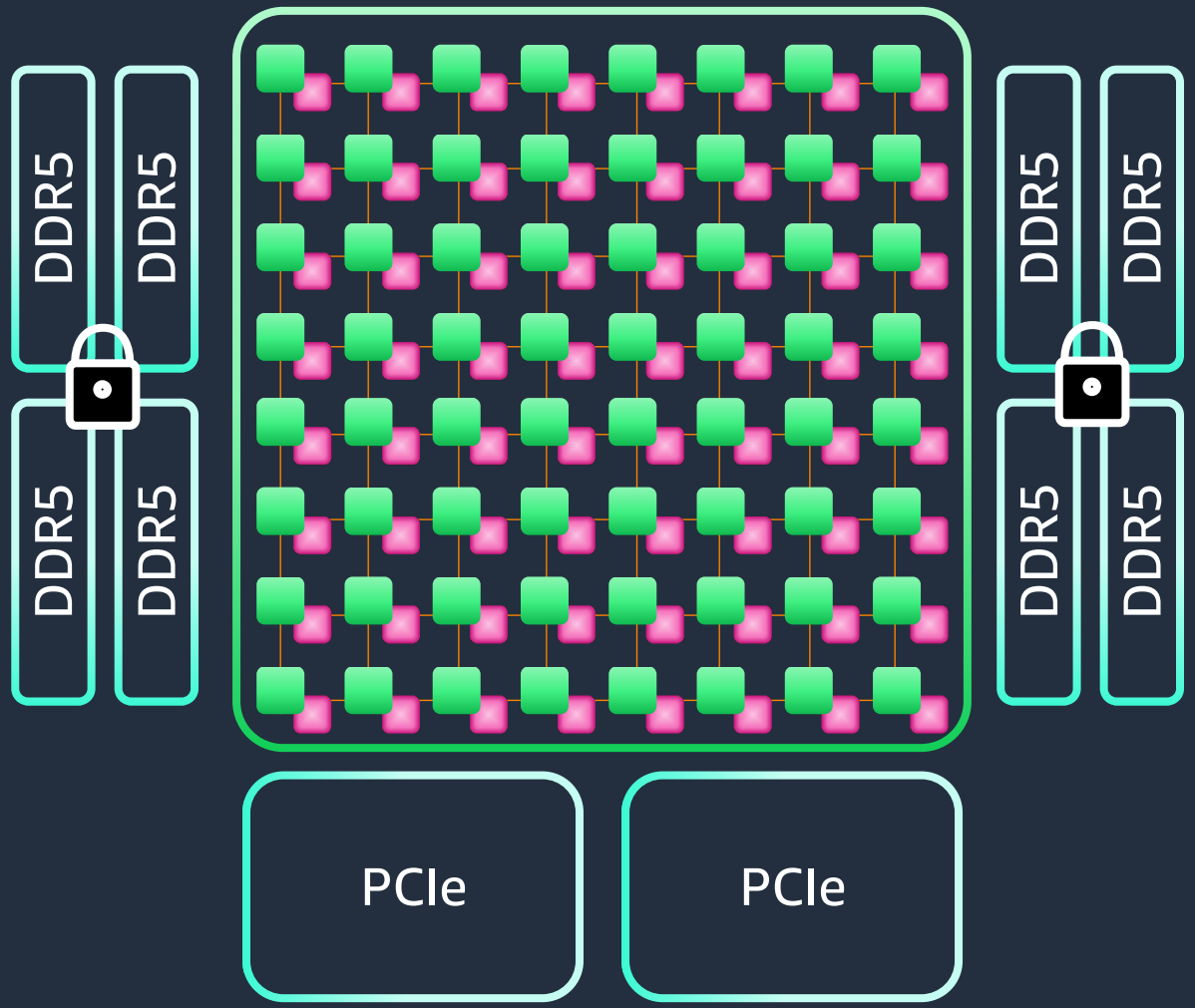
GPU

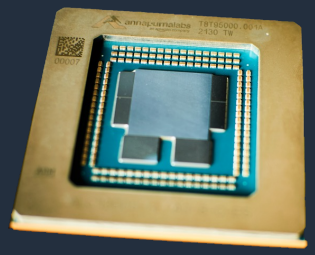
P4de

8X A100
80GB GPU Mem
1TB Mem
400 Gbit EFA
8 TB NVME

aws health

Graviton3 – Interconnect & system





Specs

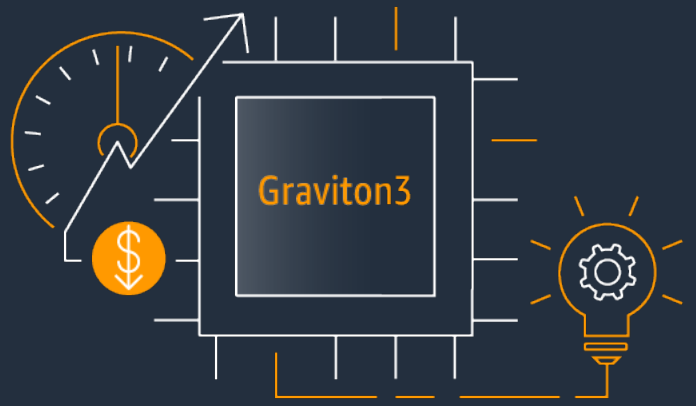
- Single Socket
- Clock Speed: 2.6Ghz
- Memory Speed: DDR5-4800
- Memory Bandwidth ~300GB/s
- Bisectional Bandwidth 2TB/s
- Scalable Vector Extension SVE (HPC)
- BFLOAT16 (SVE & Neon) 2.5x : ML, NN

aws health

https://www.youtube.com/watch?v=A3PXJtWcczI

AWS Graviton3 and Amazon EC2 C7g instances

Enabling the best price performance for workloads in Amazon EC2



Up to 25% better performance compared to Graviton2

Up to 2x higher floating-point performance, up to 2x faster cryptographic workload performance, and up to 3x better machine learning performance compared to Graviton2

First in the cloud to feature DDR5 memory

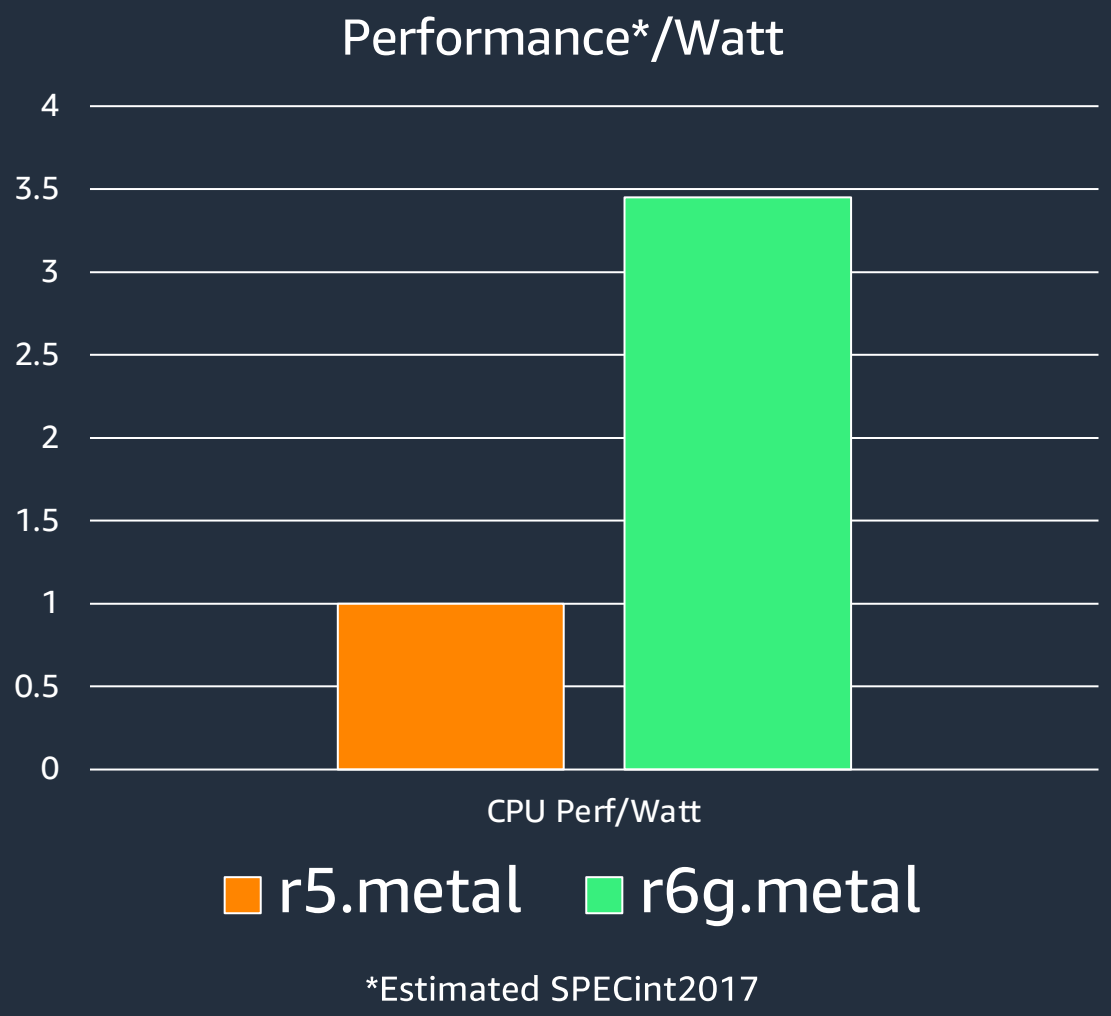
C7g instances will provide the best price performance for compute-intensive workloads in Amazon EC2

60% more energy efficient over comparable EC2 instances

https://aws.amazon.com/blogs/aws/join-the-preview-amazon-ec2-c7g-instances-powered-by-new-aws-graviton3-processors/

aws health

Sustainability: AWS Graviton Processors



- AWS Graviton2: Processor Power Efficiency up to 3.5x better performance/watt*
  - Lower power
    - Higher density
    - Lower costs
    - Lower carbon footprint

- New – Sustainability Pillar for AWS Well-Architected Framework and Customer Carbon Footprint Tool (CFFT)

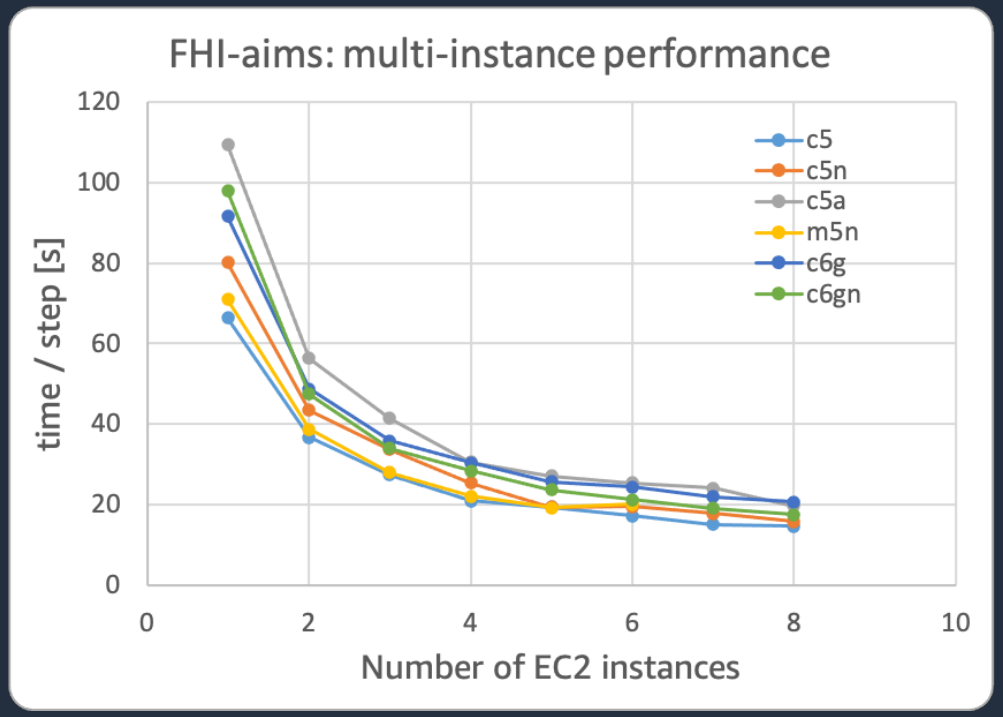
https://aws.amazon.com/blogs/aws/sustainability-pillar-well-architected-framework/
https://aws.amazon.com/blogs/aws/new-customer-carbon-footprint-tool/

aws health

Max Planck Institute: FHI-aims v21.02

Use Case: Cost Optimization and Scaling

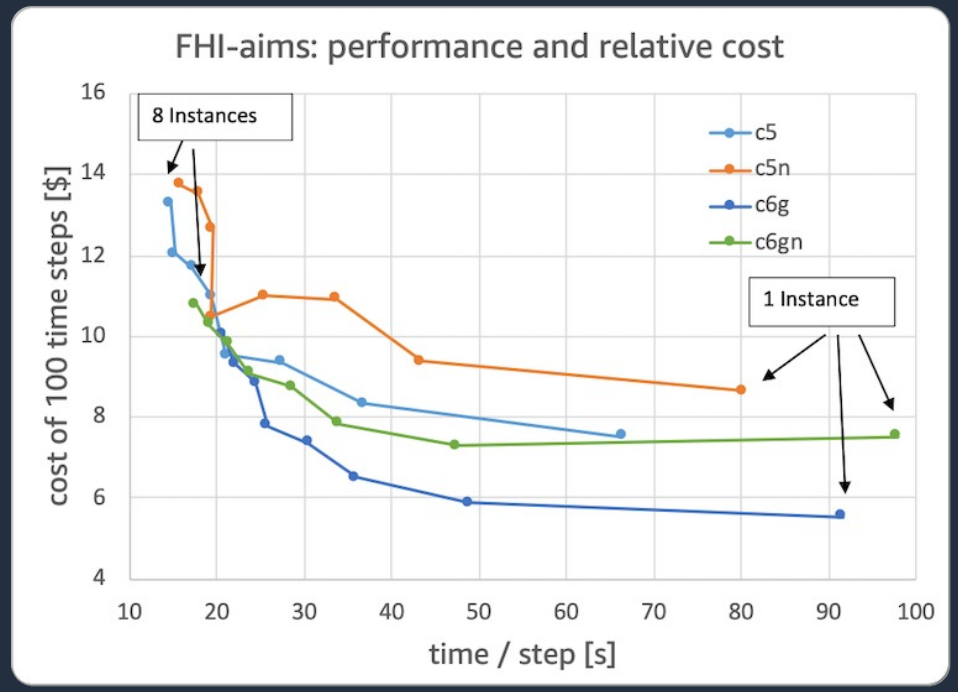
- Carbon Monoxide molecule reacting over a graphene monolayer
- All-electron full-potential numerical atomic orbital basis set code
- 20 atoms, where 18 of them forming part of a periodic structure and two of them are in a gas form
- Graphene consists of 18 carbon atoms and is sampled with a 5x5x1 k-point grid



Lower is better

c6g 25% less cost c5

c6g 33% less cost c5n



Lower is better

https://aws.amazon.com/blogs/hpc/quantum-chemistry-calculation-on-aws/

aws health

Amazon EC2 G5g instances powered by AWS Graviton2

The first Graviton-based instances to feature GPU acceleration

Get significantly lower cost-per-inference for machine learning inference over x86-based GPU instances

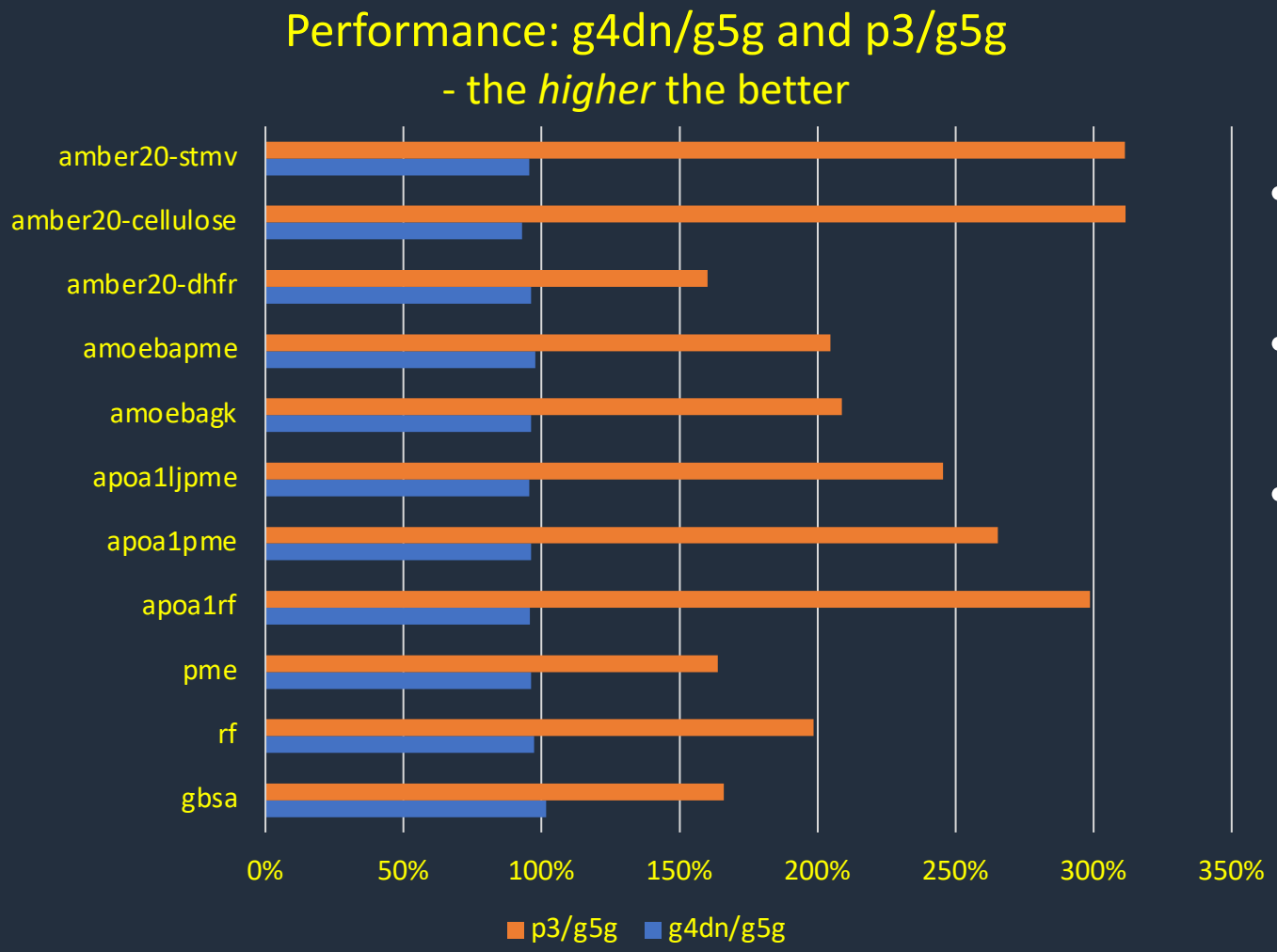
Powered by AWS Graviton2 and featuring NVIDIA T4G Tensor Core GPUs

NVIDIA NGC-ARM Containers

- Quantum Expresso
- Relion
- Autodoc-GPU
- ……

aws health

AWS Graviton Processors + NVIDIA GPU for MD Simulations - OpenMM Performance

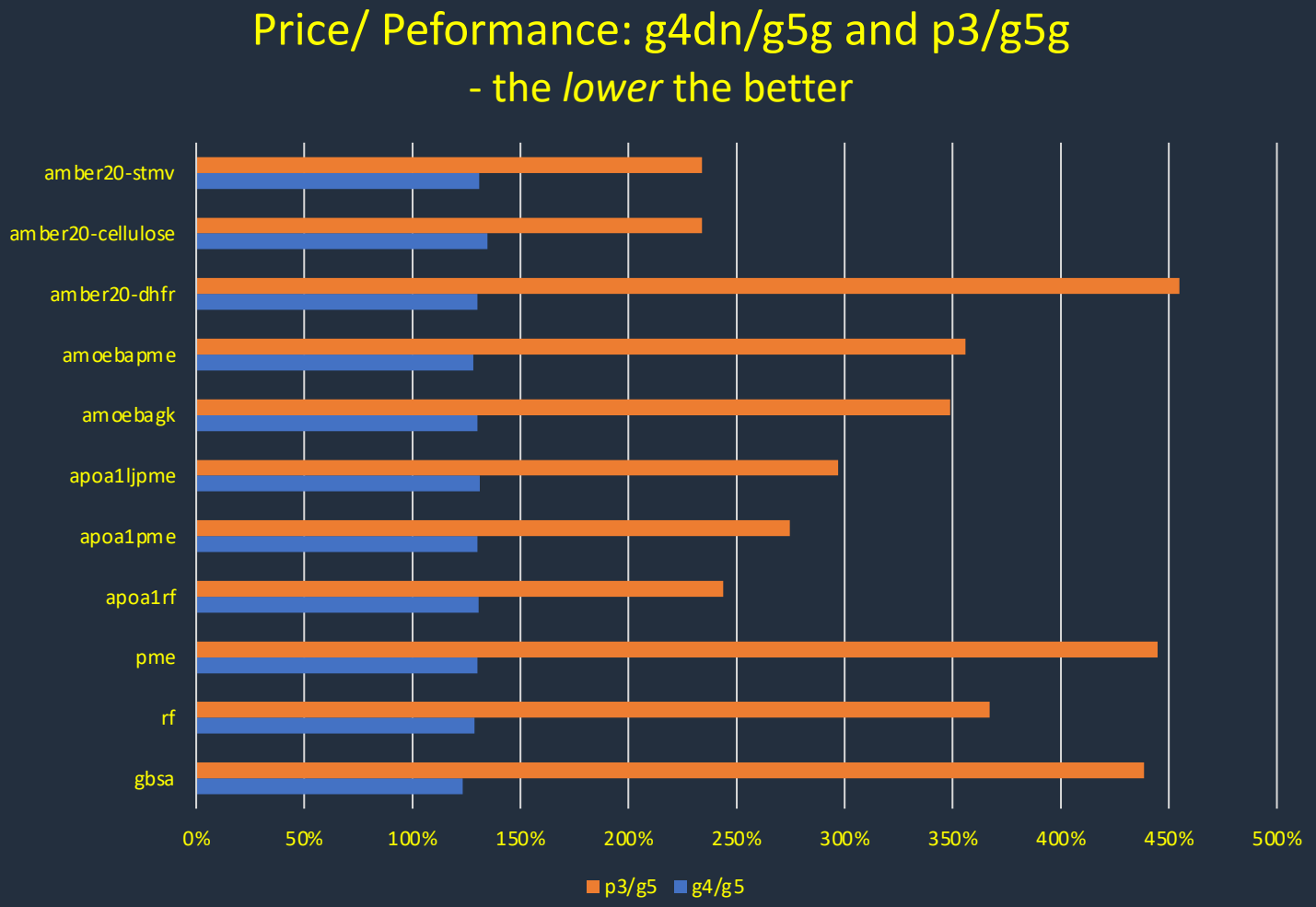


- Most MD simulation programs have been optimized for GPUs
- OpenMM benchmark on p3.2xlarge, g4dn.xlarge and g5g.xlarge
- p3.2xlarge with V100 is faster than g4dn.xlarge and g5g.xlarge with T4 (up to 3X), but at much higher cost (7X)

EC2 instance starting pricing at US East-1: g5g.xlarge $0.42/hr; g4dn.xlarge $0.526/hr, p3.2xlarge $3.06/hr

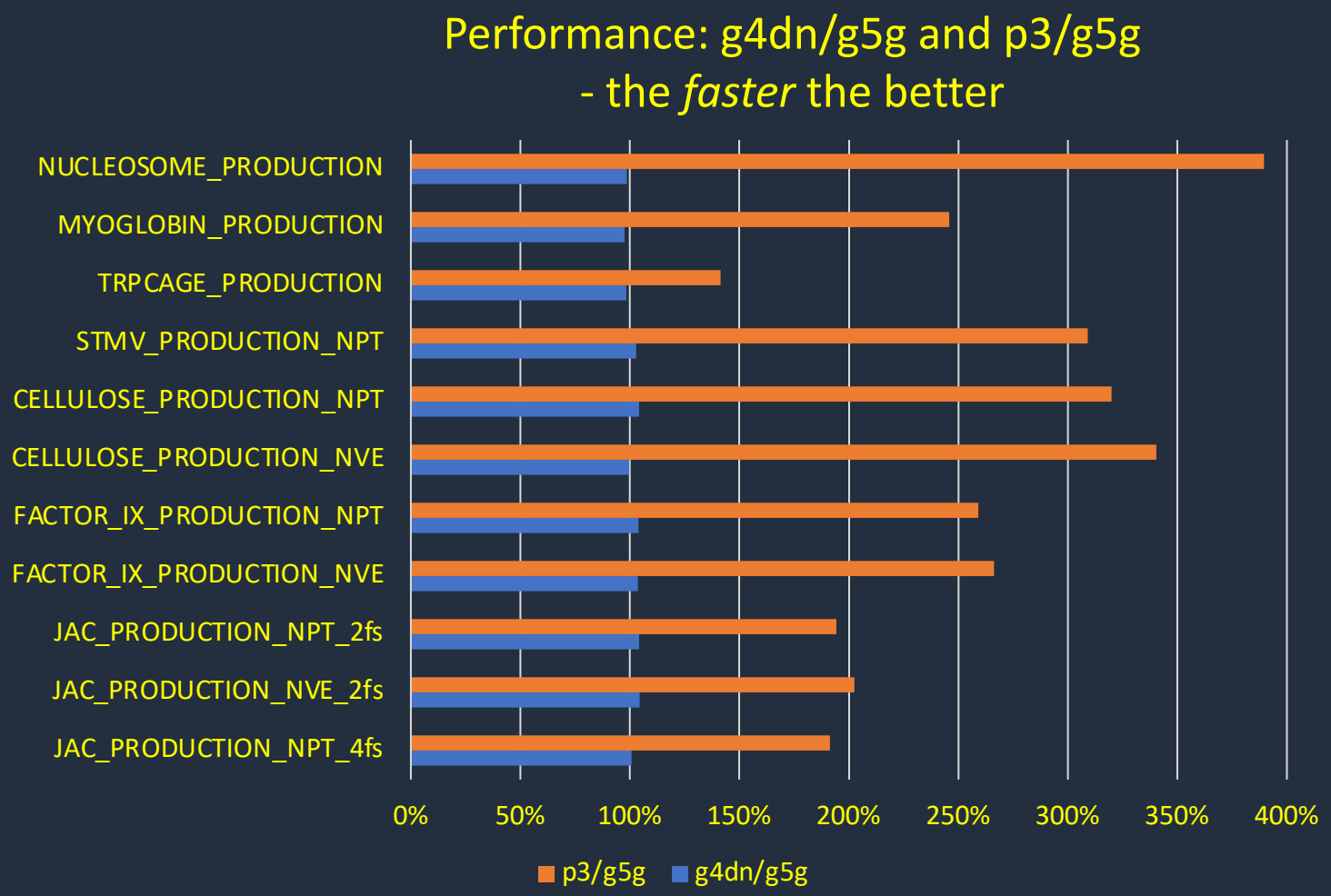
aws health

AWS Graviton Processors + NVIDIA GPU - OpenMM Price/Performance



- T4 GPUs have the best price / performance, at the sacrifice of performance
- AWS Graviton2 + T4G (g5g.xlarge) provides same performance as x84 + T4 (g4dn.xlarge) and *20+%* better price/performance

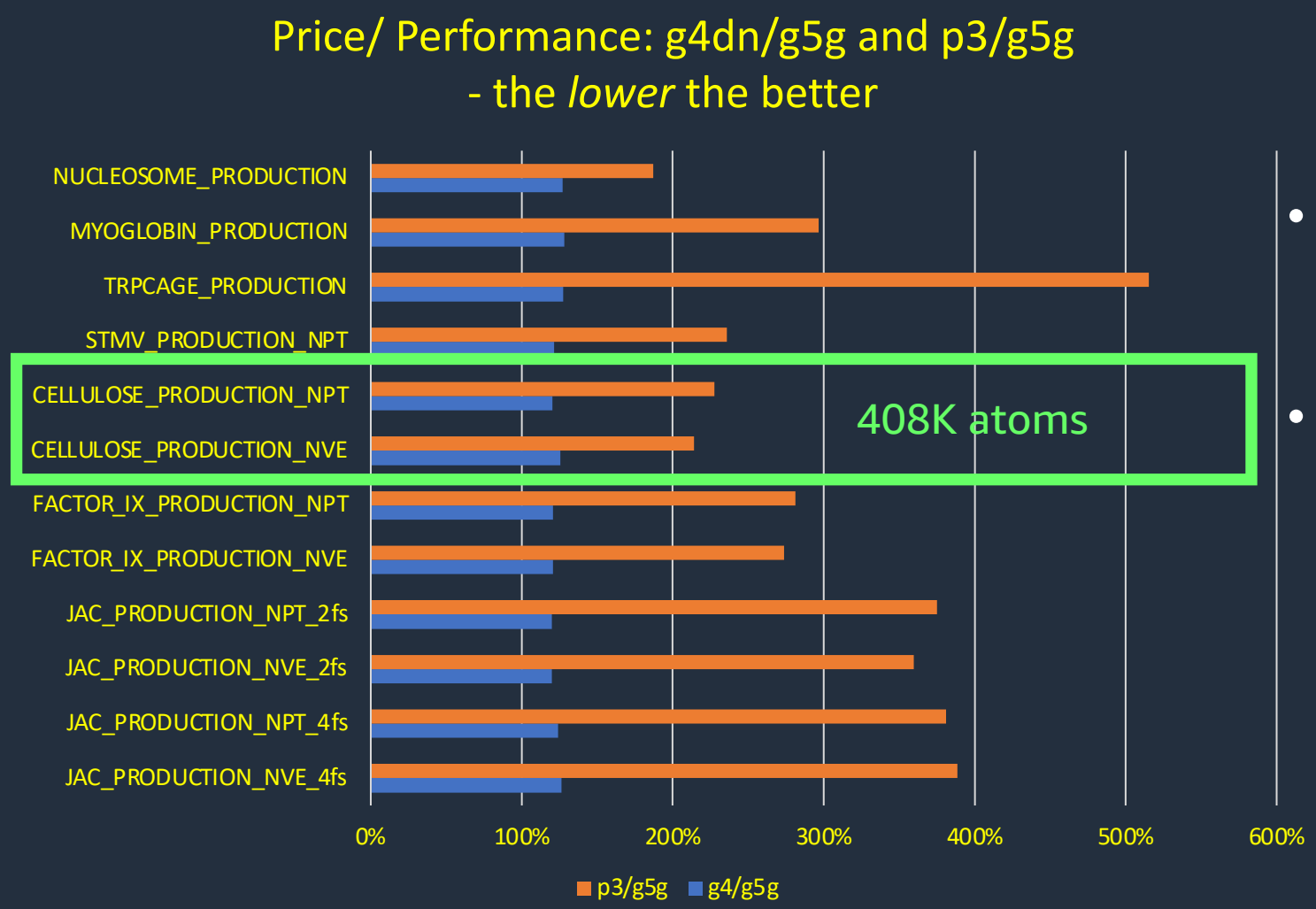
AWS Graviton Processors + NVIDIA GPU - AMBER 20/21 Performance



- p3.2xlarge with V100 is faster than g4dn.xlarge and g5g.xlarge with T4 (up to 4X), but at much higher cost (7X)

aws health

AWS Graviton Processors + NVIDIA GPU - AMBERPrice/Performance



- Again T4 GPUs have the best price / performance, at the sacrifice of performance
- AWS Graviton2 + T4G (g5g.xlarge) provides same performance as x84 + T4 (g4dn.xlarge) and *20+%* better price/performance, way better than V100.

aws health

aws health

Thank you!

Steve Litster
slitster@amazon.com

Zheng Yang
hclsyang@amazon.com

AWS Graviton getting started guide on Github

https://github.com/aws/aws-graviton-getting-started

- This guide has been assembled by our Graviton team and is designed to help customers transition and optimize their applications.
- It covers various languages and libraries, and includes tips and tricks for each.
- In general, using latest versions of operating systems, compilers, and language runtimes will provide access to latest Arm64 improvements and optimizations.

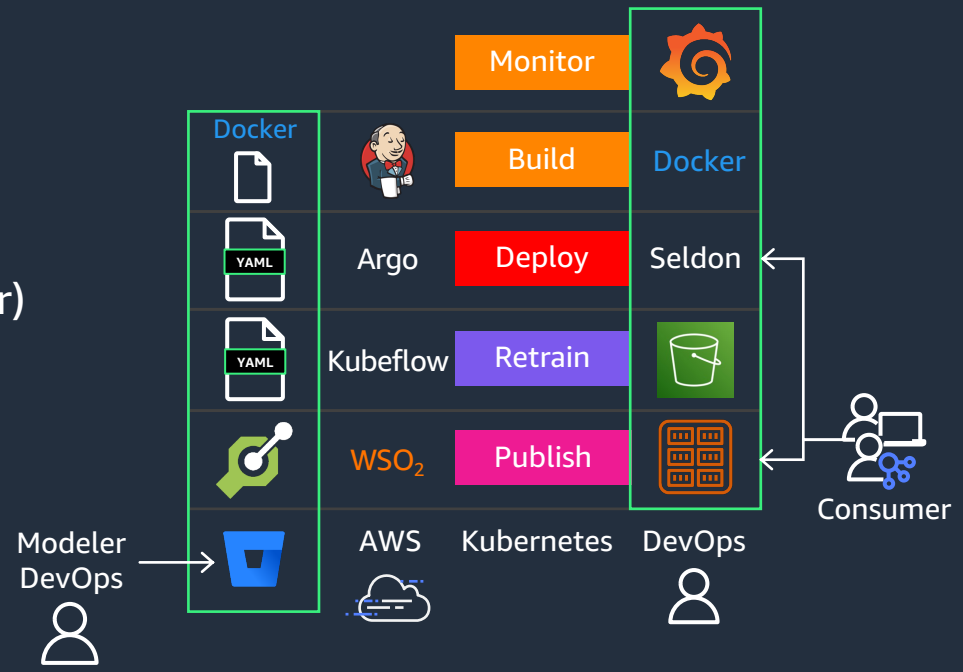
Predictive insights platform

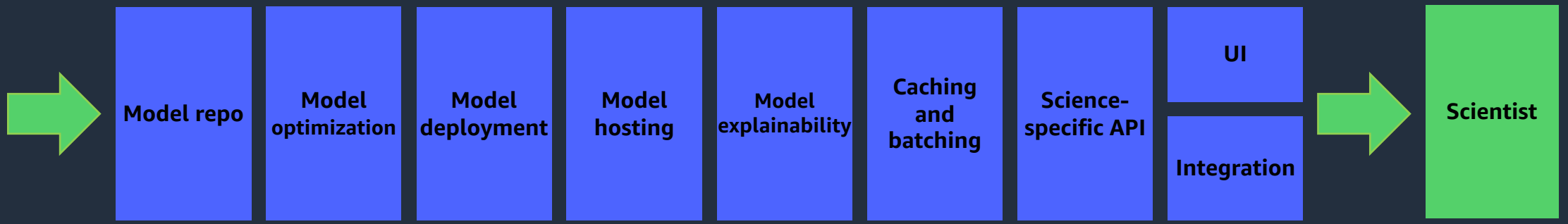
SCOPE

- Scalable predictive platform for all drug discovery
- Load models in less than a day

MODEL SOURCE

- SCP – AstraZeneca HPC
- GPU workstations
- AI Bench (Amazon SageMaker)
- MELLODDY consortium



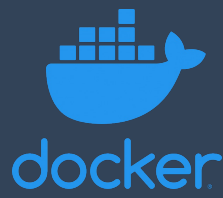


Broad Graviton Support for Containers

Orchestrators


Amazon ECS

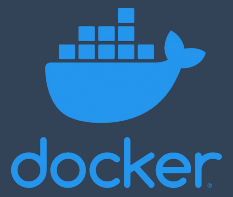

Amazon EKS


Docker Swarm


Kubernetes

Image registries


Amazon ECR


Docker Hub

Container-optimized Linux distros


Bottlerocket

Serverless


AWS Fargate

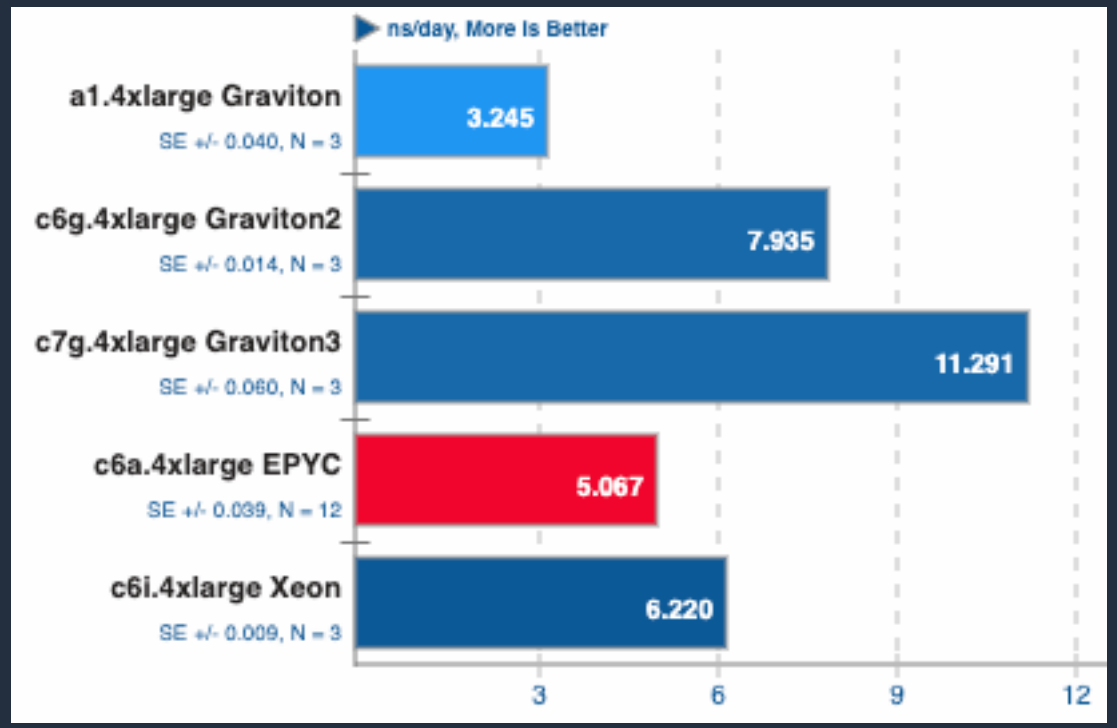

AWS Lambda

aws health

LAMMPS and GROMACS Benchmarks C7g.4xlarge

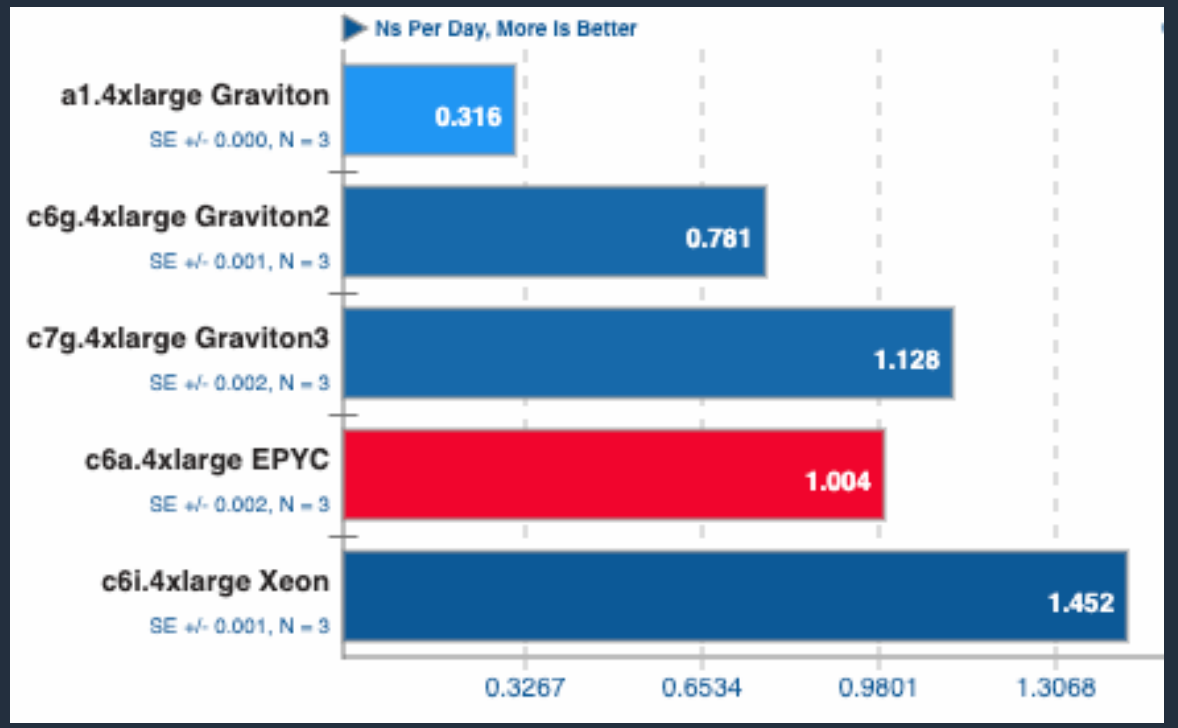
C7g.4xlarge: 16 vCPU, 32GB RAM

LAMMPS



Rhodopsin

GROMACS

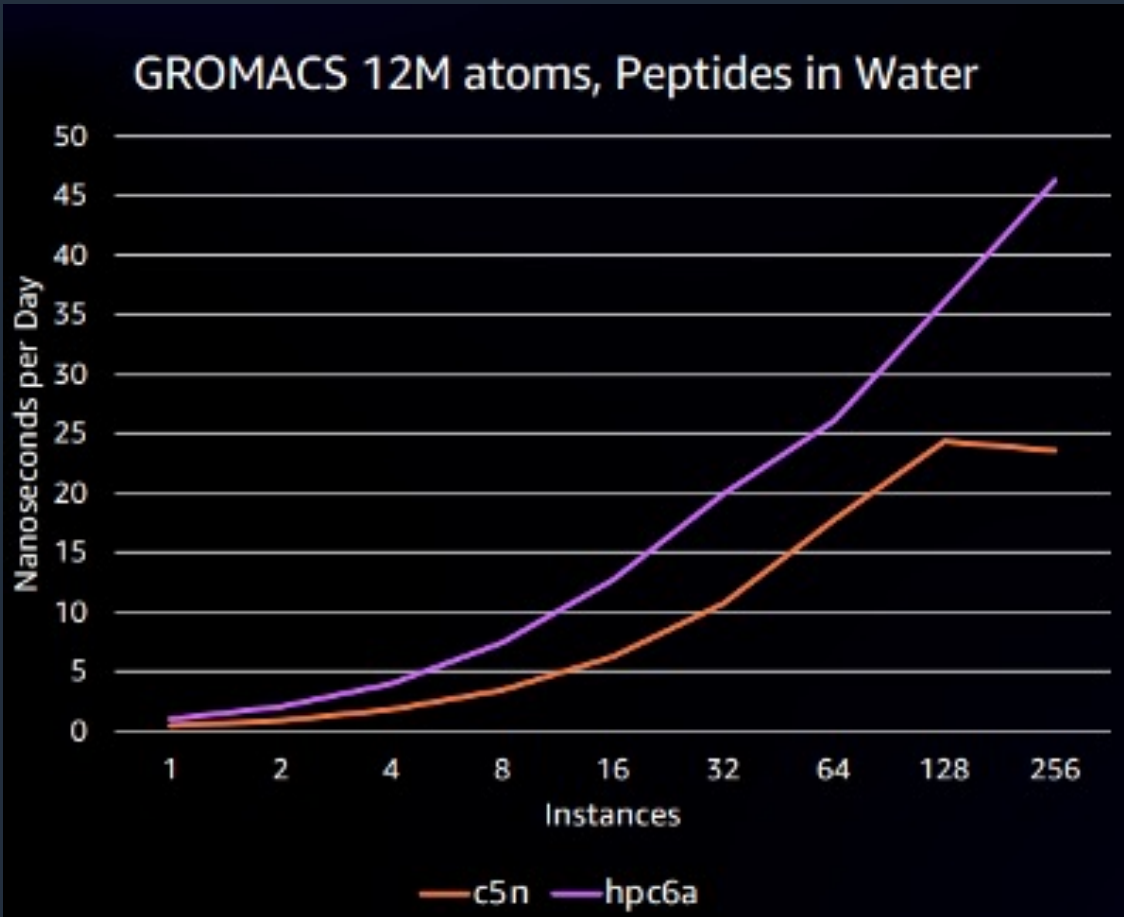


Water_DMX50

aws health

https://openbenchmarking.org/

GROMACS: Benchmarking the HPC6a Instance

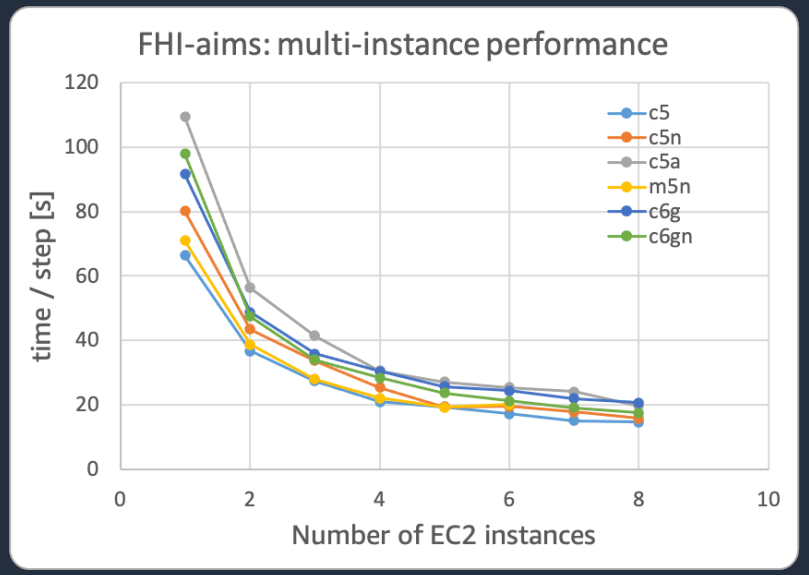


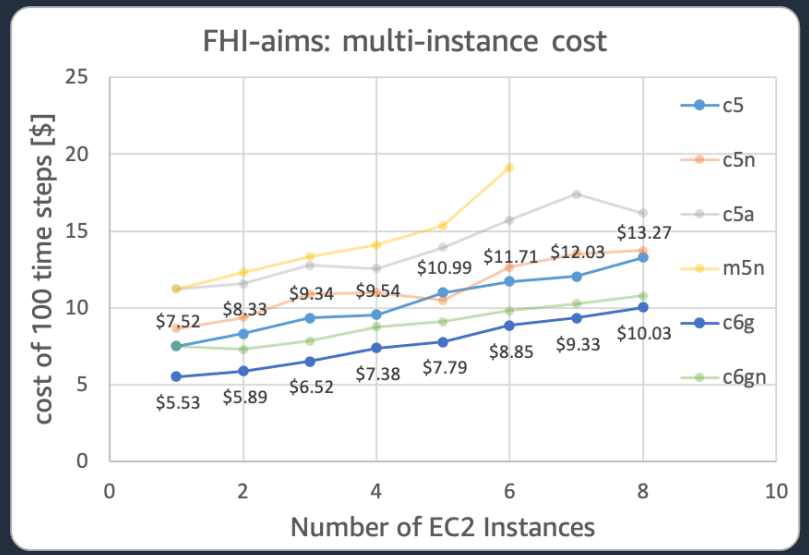
Hpc6a vs Price Performance Improvement

	16 Instances	64 Instances	128 Instances	256 Instances
Hpc6a	64%	50%	50%	62%

FHI-aims v21.02

	x86 architecture	ARM Graviton2
Operative System	Amazon Linux 2	Amazon Linux 2
Compiler	Intel OneAPI Compilers 2021.2	GNU 10.2.0
Numerical library	Intel OneAPI MKL 2021.2	ARM Perf 21, Scalapack 2.1
MPI library	Intel OneAPI MPI 2021.2	OpenMPI 4.1.1





https://aws.amazon.com/blogs/hpc/quantum-chemistry-calculation-on-aws/

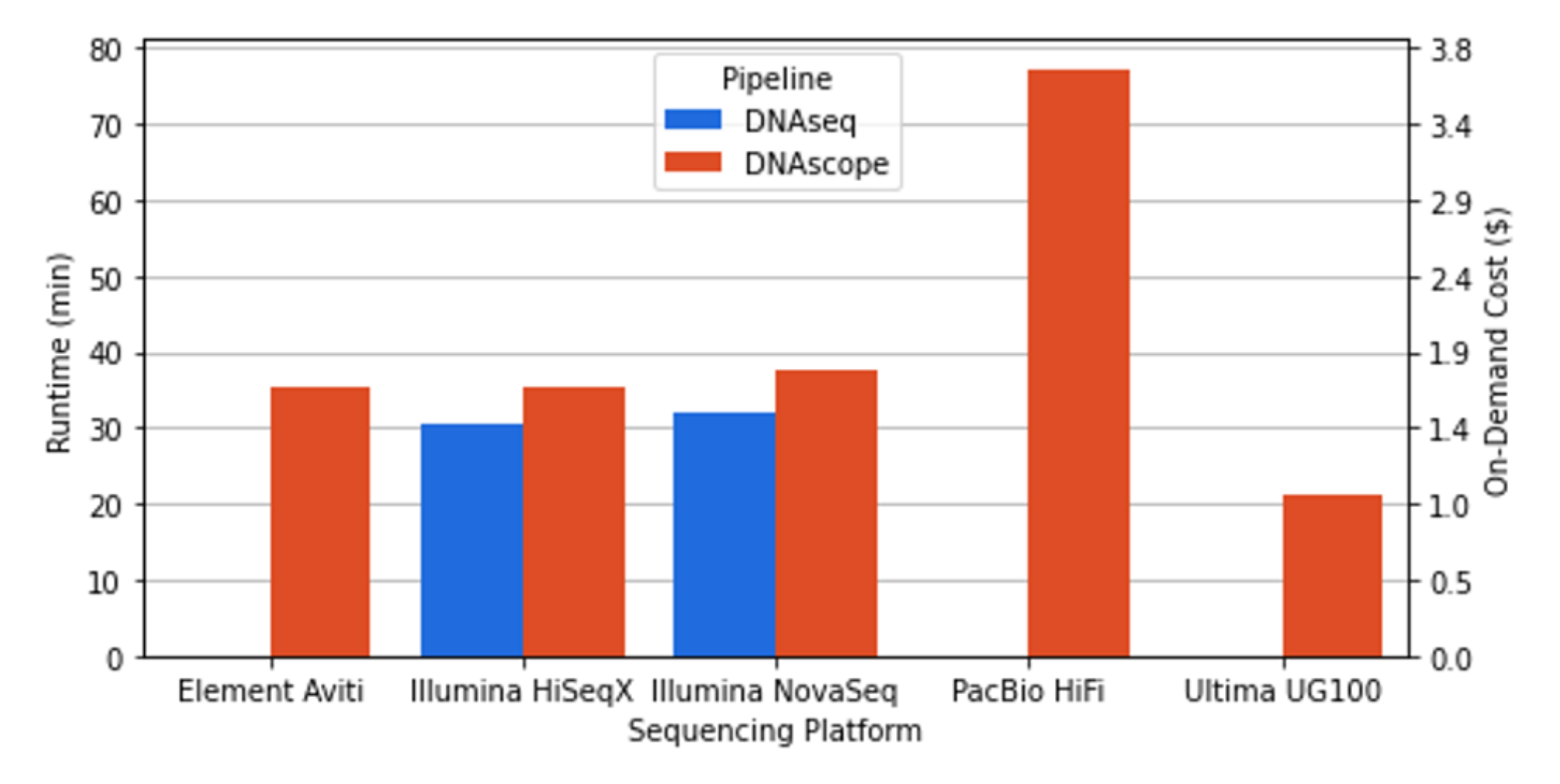
aws health

AWS managed services supporting Graviton2

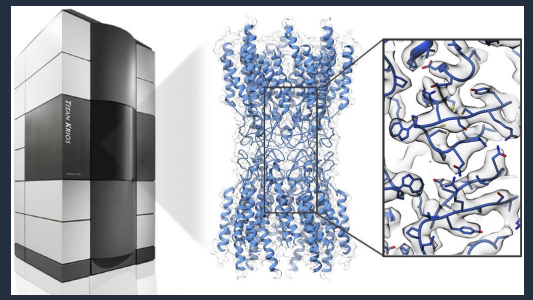
EXTENDING THE GRAVITON2 PRICE PERFORMANCE TO MANAGED SERVICES

- RDS: Graviton2 instances provide up to 35% performance improvement and up to 52% price/performance improvement for open source databases depending on database engine, version, and workload.
- Aurora: Graviton2 instances provide up to 20% performance improvement and up to 35% price/performance improvement for Aurora depending on database size.
- EMR: Amazon EMR provides up to 35% lower cost and up to 15% improved performance for Spark workloads on Graviton2-based instances versus previous generation instances
- Elasticache: Up to a 45% price/performance improvement over previous generation instances. Graviton2 instances are now the default choice.
- OpenSearch: Up to 38% improvement in indexing throughput, 50% reduction in indexing latency, and 30% improvement in query performance when compared to the corresponding x86-based instances from the current generation (M5, C5, R5)
- Lambda: AWS Lambda Functions Powered by AWS Graviton2 Processor – Run Your Functions on Arm and Get Up to 34% Better Price Performance
- DocumentDB: Achieve up to 30% better performance with Amazon DocumentDB (with MongoDB compatibility) using new Graviton2 instances

Runtime and On-Demand compute cost of the Sentieon DNAseq and DNAscope pipelines on a hpc6a.48xlarge



Cryo-Electron Microscopy | Target Identification



Cryo-EM & Protein Folding

- Cryo-Electron Microscopy is is a biophysical technique that can be used to determine the 3D structures of biological macromolecules and assemblies.

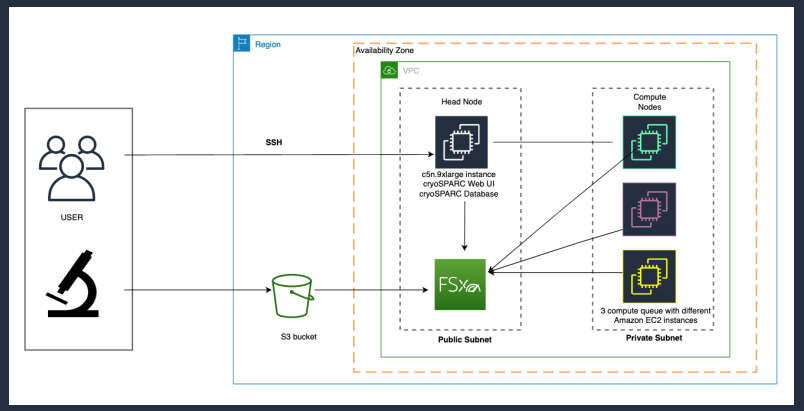
	GPUs	CGRP	GLP1R	Spike
Patch motion (M)	8	1 h 40 min	2 h 29 min	-
Patch CTF (M)	8	35 min	38 min	13 min
Blob picker	1	33 min	30 min	25 min
Particle extraction	8	15 min	12 min	3 min
2D classification	4	5 h 1 min	4 h 17 min	1 h 28 min
Heterogenous refinement	1	12 h 14 min (5 rounds)	10 h 38 min (4 rounds)	20 min
Ab Initio Reconstruction	1	6 h 34 min	13 h 26 min	-
Particle re-extraction	8	13 min	11 min	5 min
Non-uniform refinement	1	8 h 57 min	5 h 39 min	3 h 7 min
Total runtime		36 h 2 min	38 h	5 h 41 min

Key Industry Trends

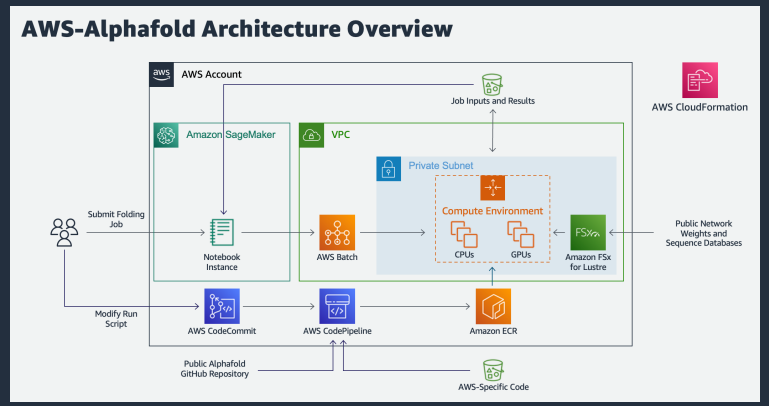
- Cryo-EM is becoming the #1 technique for structure based target discovery & validation
- Long Term/High Cost customer investment –needs to run 24x7
- Single microscope can generate 1PB/Year
- Compute and storage remain significant bottlenecks

Solutions

Cryo-EM Architecture



Protein Folding-Alphafold

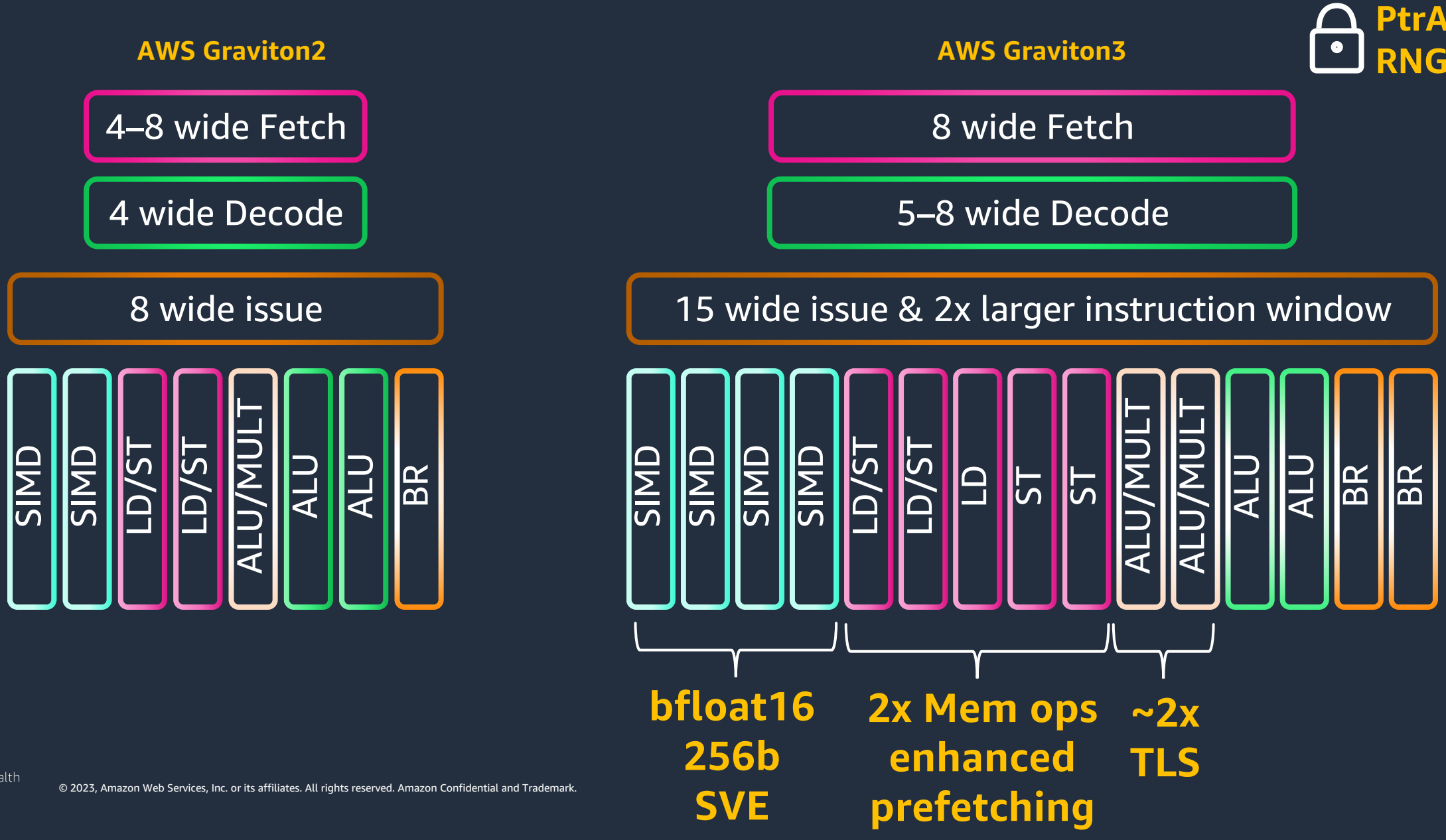


ThermoFisher SCIENTIFIC

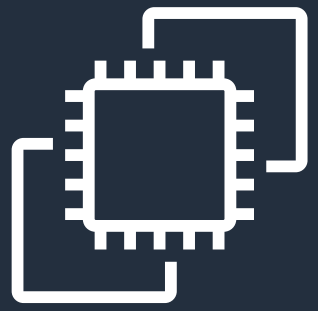
STRUCTURA BIOTECHNOLOGY

aws health

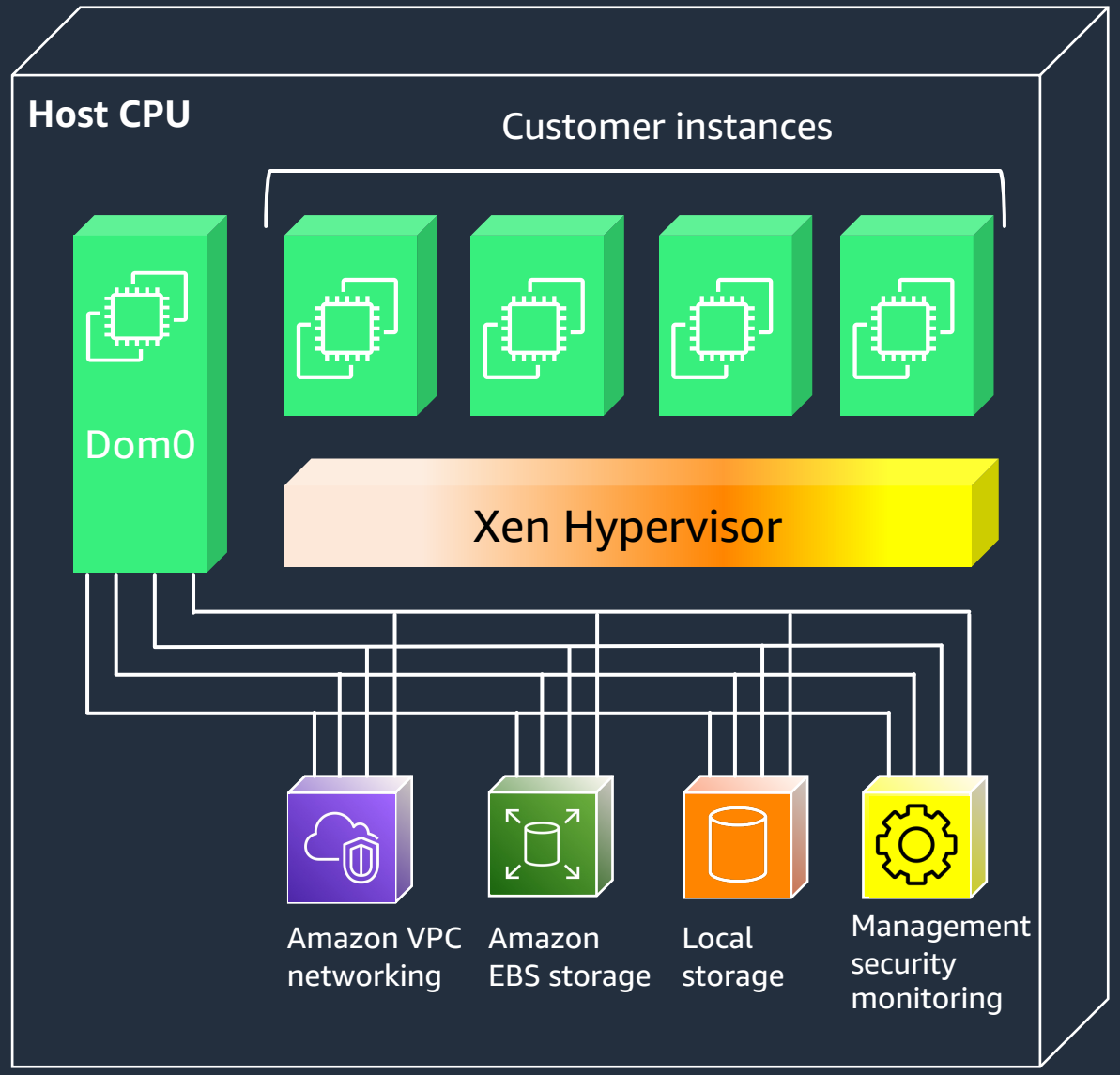
Graviton3 CPU enhancements



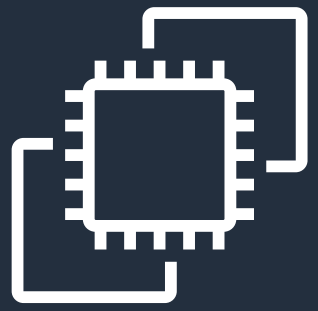
Before Nitro . . .



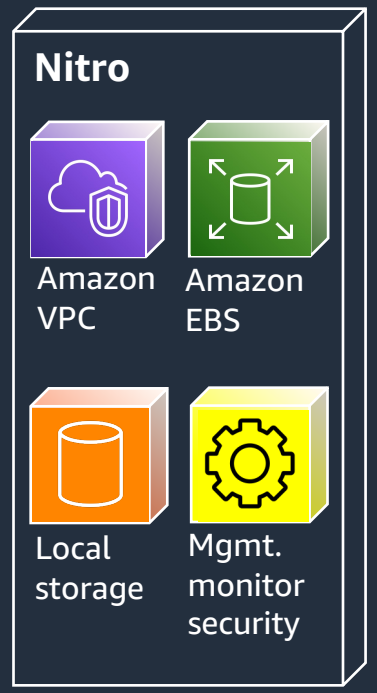
AWS Nitro System

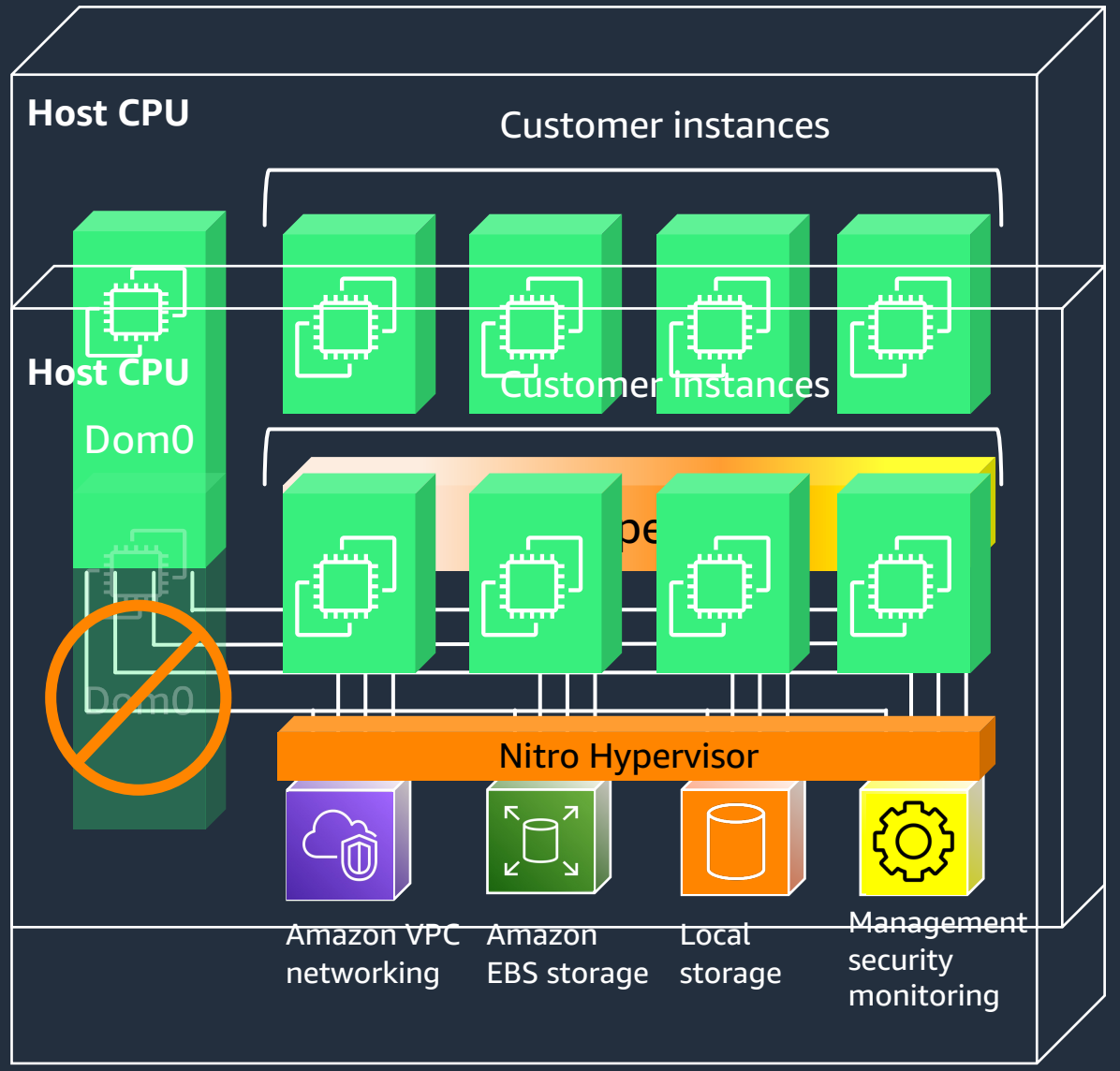


AWS Nitro System

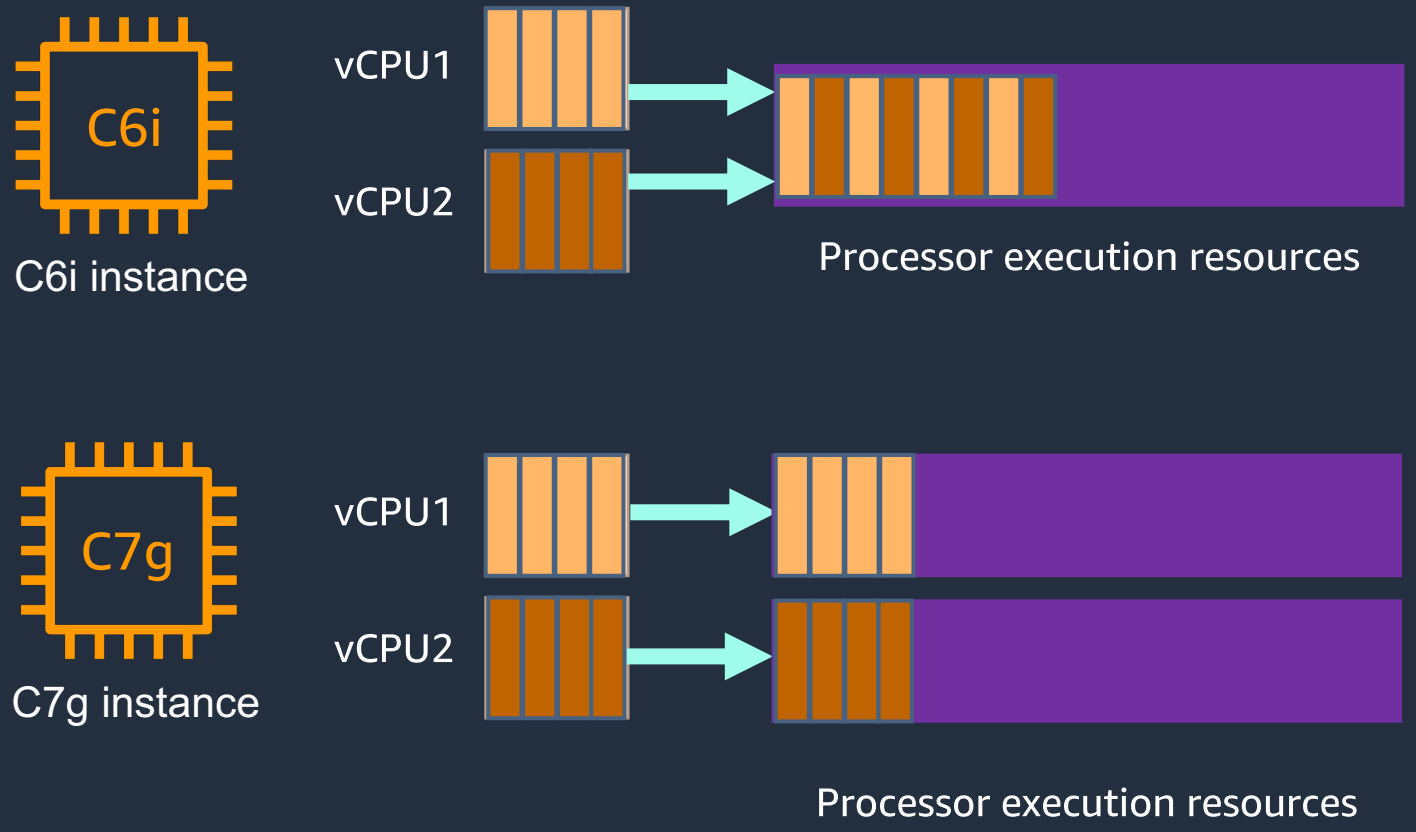


AWS Nitro System





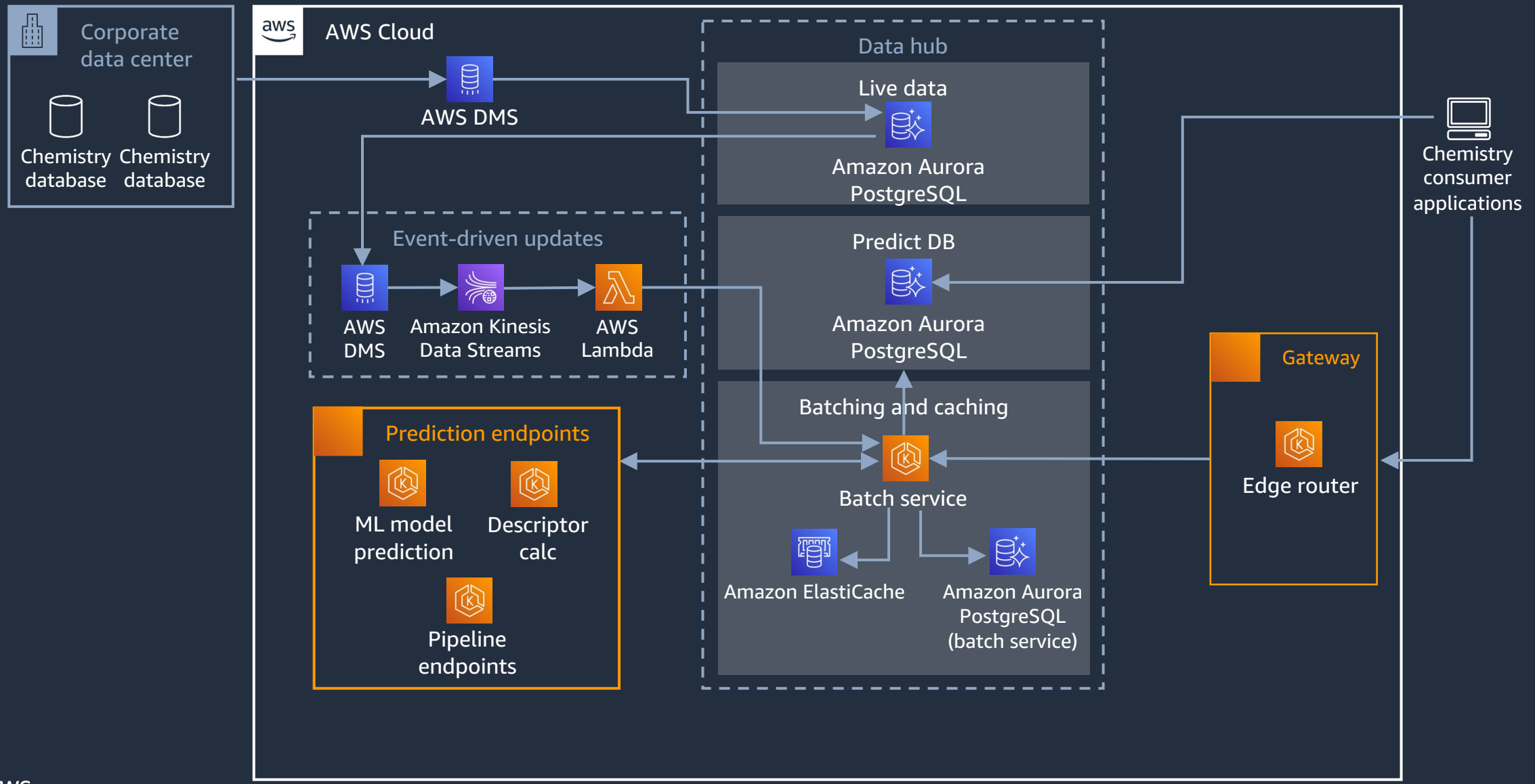
Graviton3 – vCPU



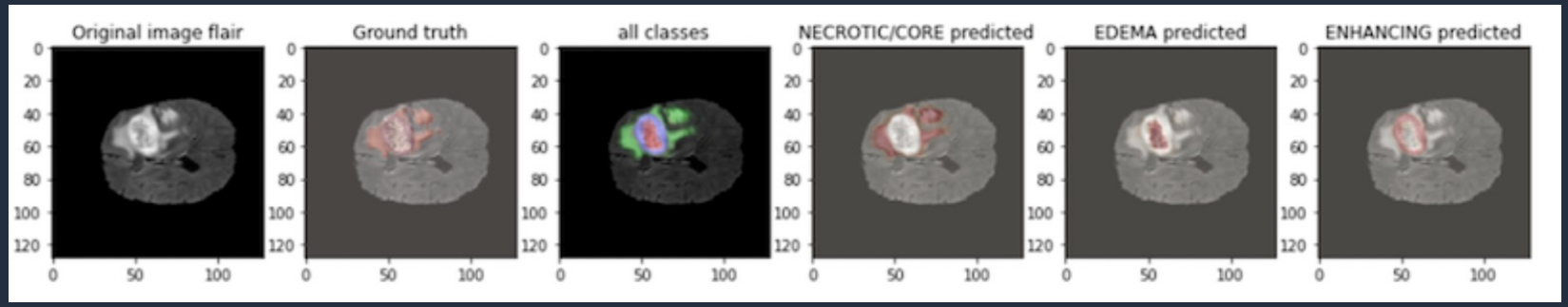
Every vCPU is a physical core

No simultaneous multi threading (SMT)

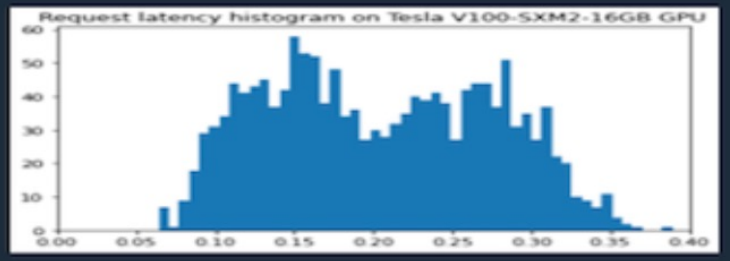
Predictive insights platform: Multi Modal Analysis

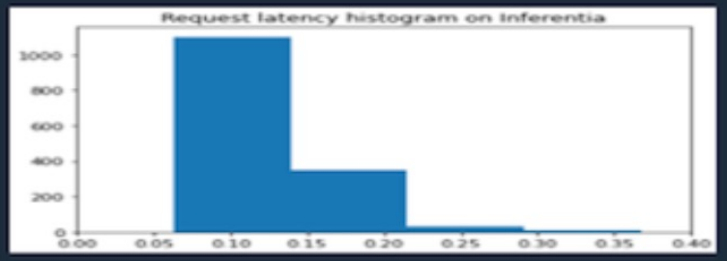


Brain tumor segmentation at scale using AWS Inferentia



AWS Inferentia Reduced Latency by 43% and Increased Throughput by 140% over V100





Tesla V100-SXM2-16GB GPU

- 95% of requests take less than 323ms
- Rough request throughput/second is 23.63

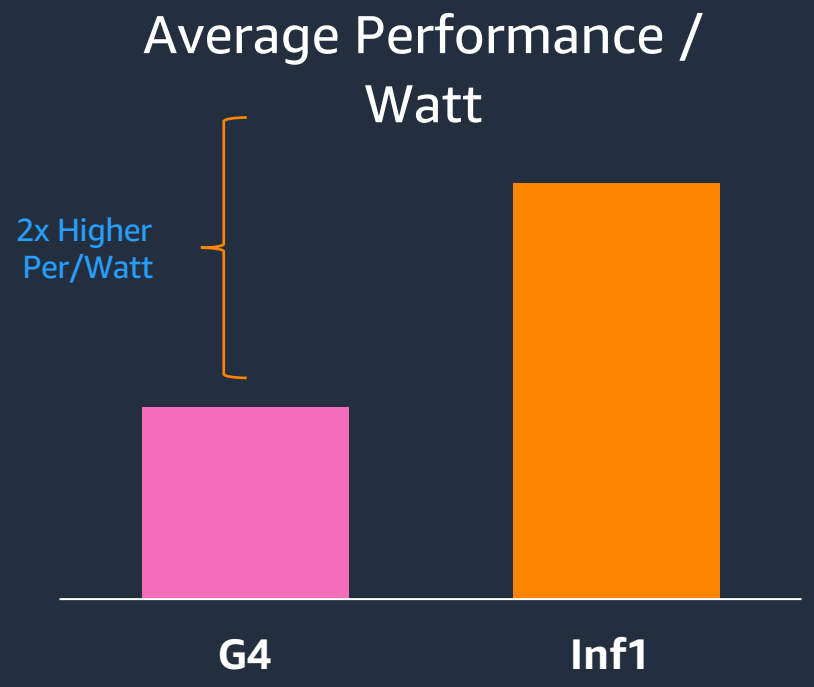
Inferentia

- 95% of requests take less than 185 ms
- Rough request throughput/second is 33

https://aws.amazon.com/blogs/machine-learning/brain-tumor-segmentation-at-scale-using-aws-inferentia/

aws health

High Efficiency Enables Sustainable ML Inference at Scale



- Inf1 reduces carbon footprint for ML Inference when compared to higher power GPUs.
- Over 2x higher average perf / watt over G4 instances