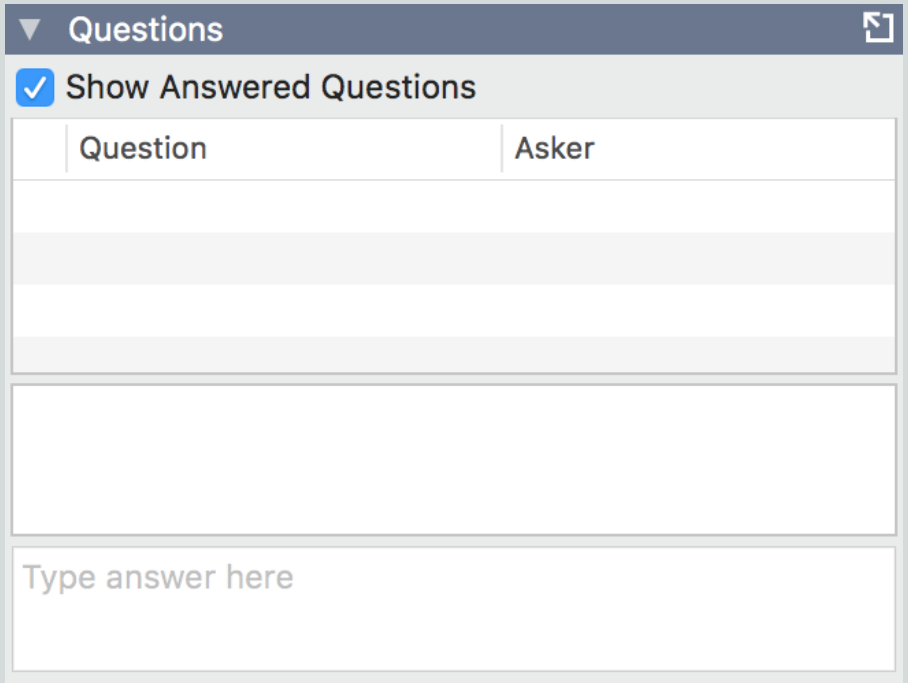
AWS Builders Program 300 – AWS Graviton2 기반 EC2 어머 이건 써야해

박선준 June Park – Solutions Architect, AWS

강연 중 질문하는 방법

AWS Builders



Go to Webinar “Questions” 창에 자신이 질문한 내역이 표시됩니다. 기본적으로 모든 질문은 공개로 답변 됩니다만 본인만 답변을 받고 싶으면 (비공개)라고 하고 질문해 주시면 됩니다.

고지 사항(Disclaimer)

본 컨텐츠는 고객의 편의를 위해 AWS 서비스 설명을 위해 온라인 세미나용으로 별도로 제작, 제공된 것입니다. 만약 AWS 사이트와 컨텐츠 상에서 차이나 불일치가 있을 경우, AWS 사이트(aws.amazon.com)가 우선합니다. 또한 AWS 사이트 상에서 한글 번역문과 영어 원문에 차이나 불일치가 있을 경우(번역의 지체로 인한 경우 등 포함), 영어 원문이 우선합니다.

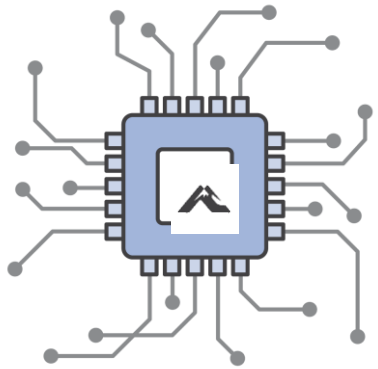
들어가기에 앞서..

- AWS Graviton2 프로세서란?
- Graviton2 프로세서는 어디에?
- Graviton2 프로세서의 성능 지표
- Graviton2와 Software Stack
- 약간의 데모

AWS Graviton2 프로세서?

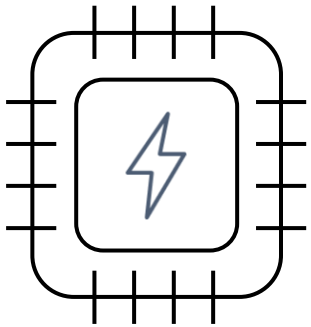
AWS 의 혁신 – 커스텀 환경

Graviton2



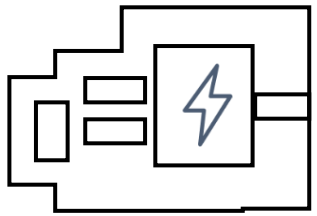
- 7nm 커스텀 실리콘
- 고성능 업무에 최적

Nitro Security Chip



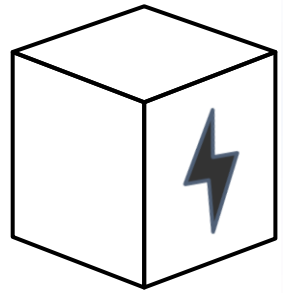
- 마더보드에 통합된 보안 칩
- 클라우드 보안에 맞춤 설계

Nitro Card



- 고성능, 가속화 및 안전
- Nitro Card for VPC
- Nitro Card for EBS
- Nitro Card for Instance Storage
- Nitro Security Chip

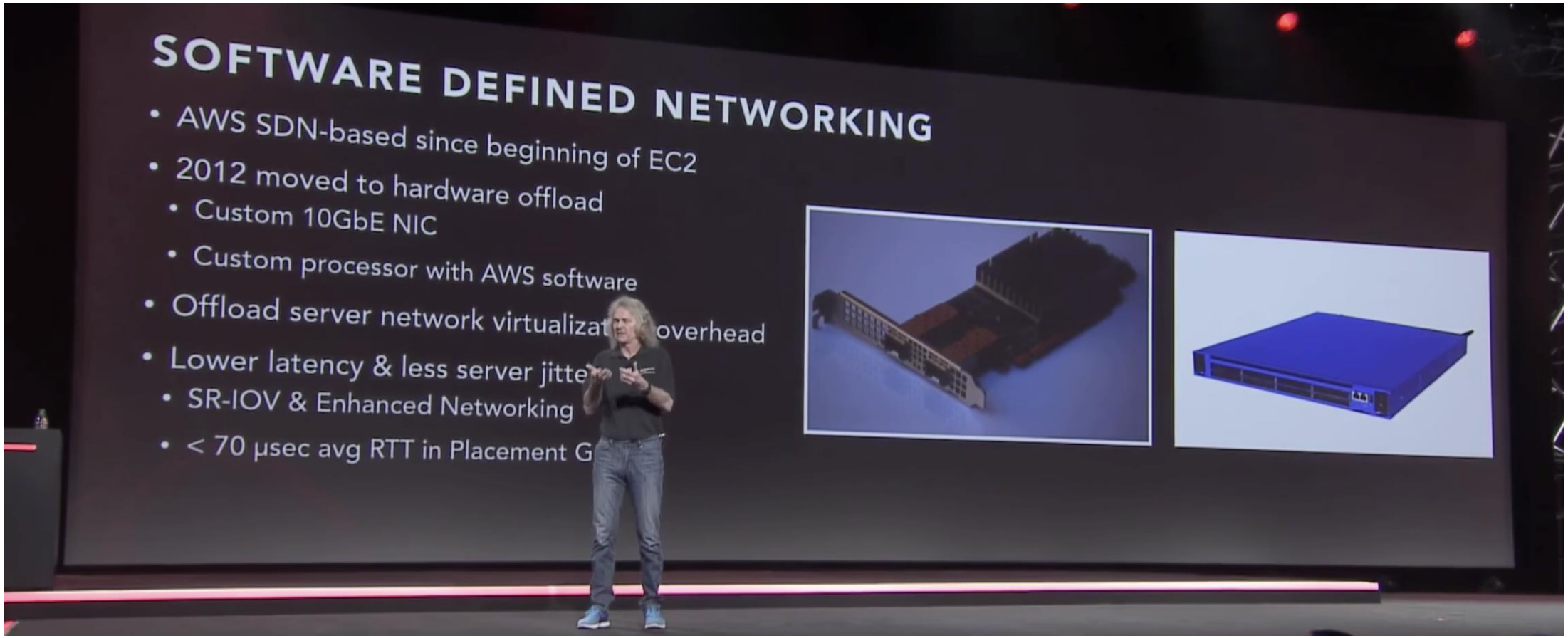
Nitro Hypervisor



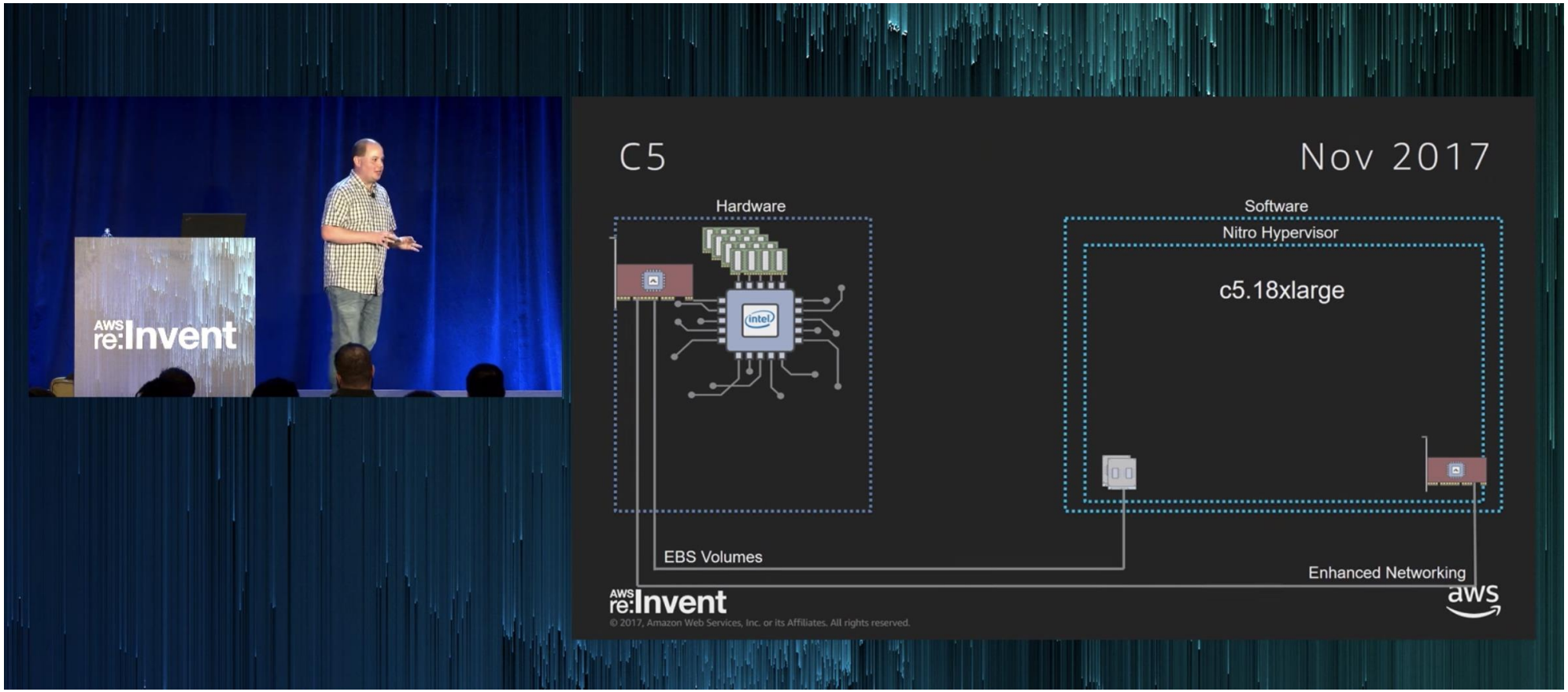
- 초 경량 hypervisor
- Memory and CPU 할당 관리
- Bare-metal 과도 같은 성능

Annapurnalabs – an Amazon Company since 2015

re:Invent 2016 - Tuesday Night Live with James Hamilton



re:Invent 2017 – Nitro Hypervisor



re:Invent 2018 – A1 Instances / Graviton

New – EC2 Instances (A1) Powered by Arm-Based AWS Graviton Processors

by Jeff Barr | on 26 NOV 2018 | in Amazon EC2, AWS Re:Invent, Launch, News, Top Posts* | Permalink | Share

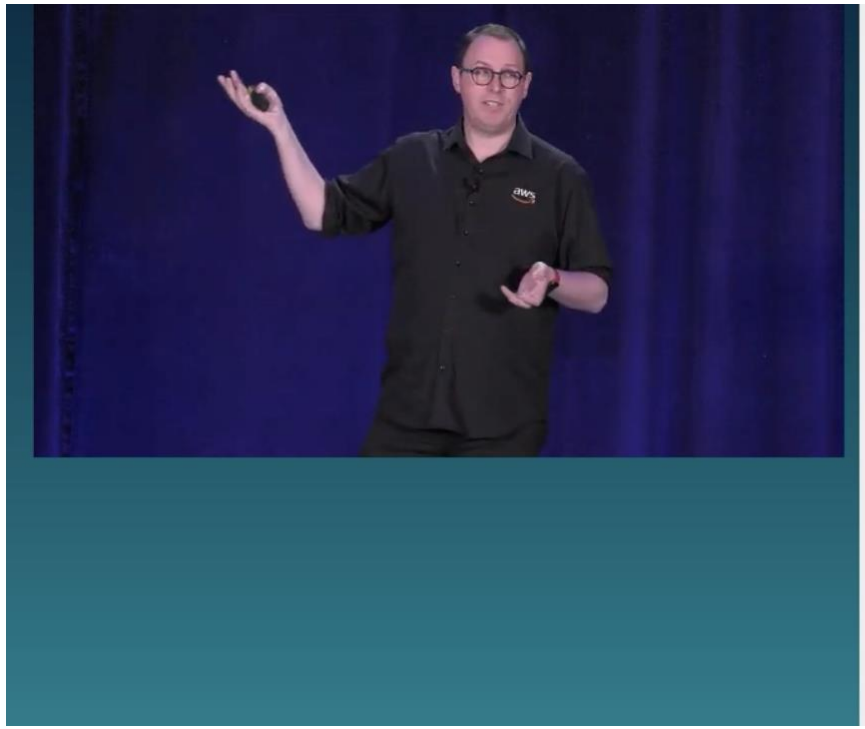
AWS Graviton Processors

Today we are launching EC2 instances powered by Arm-based AWS Graviton Processors. Built around Arm cores and making extensive use of custom-built silicon, the A1 instances are optimized for performance and cost. They are a great fit for scale-out workloads where you can share the load across a group of smaller instances. This includes containerized microservices, web servers, development environments, and caching fleets.

Jeff Barr

Jeff Barr is Chief Evangelist for AWS. He started this blog in 2004 and has been writing posts just about non-stop ever since.

re:Inforce 2019 – EC2 Networking 100Gb/s encryption



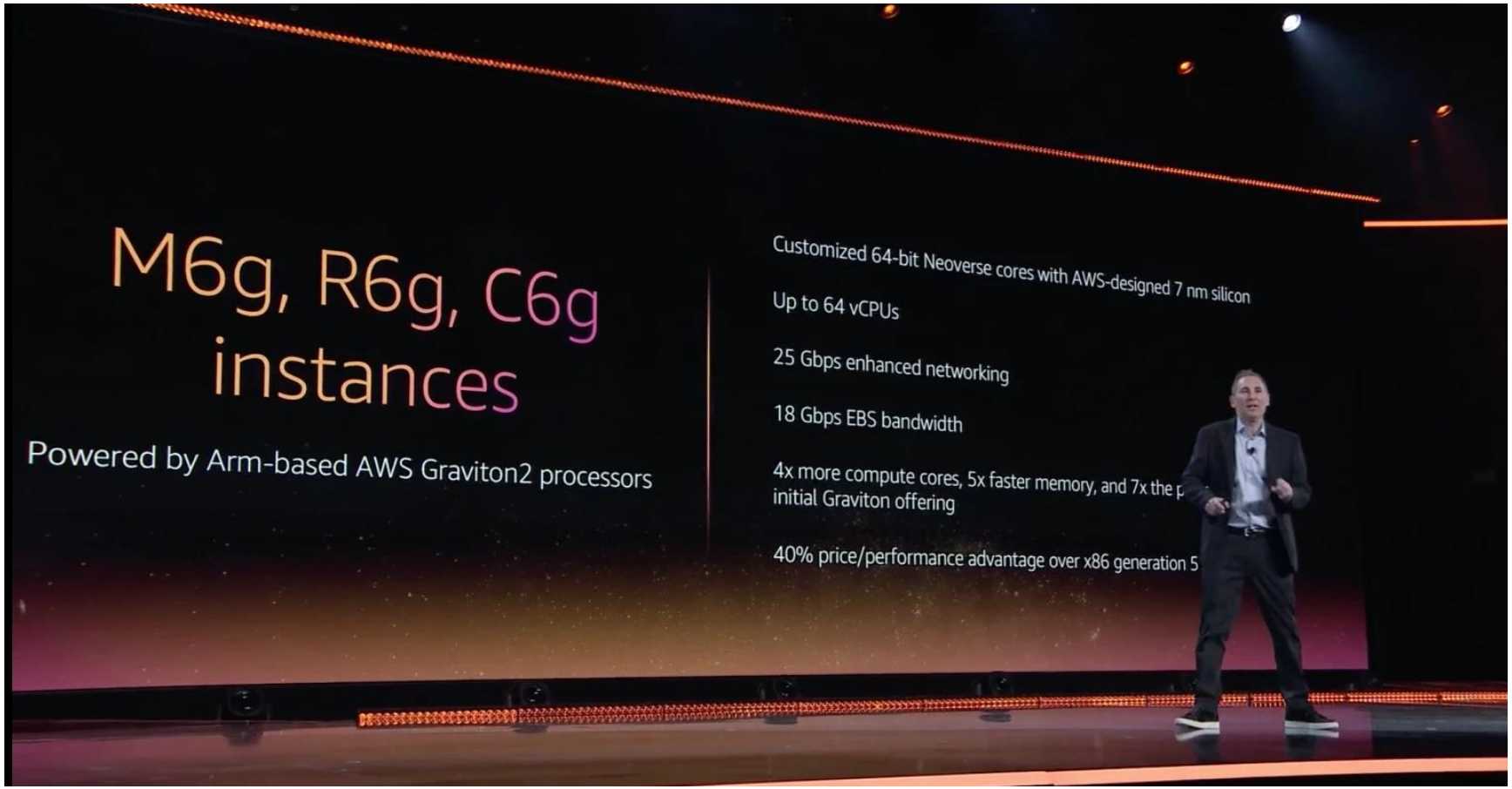
VPC encryption

Data Encryption

AWS provides secure and private connectivity between EC2 instances. In addition, we automatically encrypt in-transit traffic between C5n, I3en, and P3dn instances in the same VPC or in peered VPCs, using AEAD algorithms with 256-bit encryption. This encryption feature uses the offload capabilities of the underlying hardware, and there is no impact on network performance.

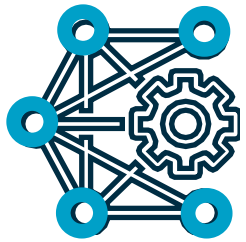
- Implemented in AWS hardware, by Annupurna Labs, as part of Nitro
- We encrypt your data AND our network virtualization protocol
- Encryption is applied within and between availability zones
- Forward-secrecy for between hours and one day

re:Invent 2019 – Graviton2 / [CMR]6g

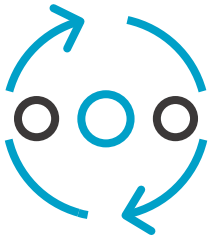


AWS 의 혁신 – AWS Graviton 프로세서

64-bit Arm Neoverse 코어 기반의 커스텀 AWS 실리콘

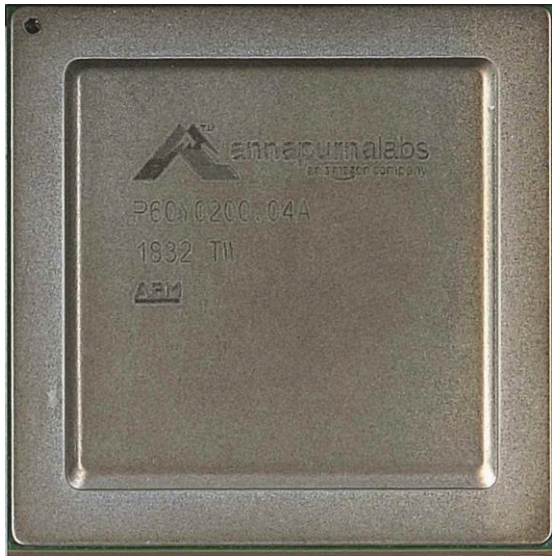


클라우드 환경 및 업무에 최적화

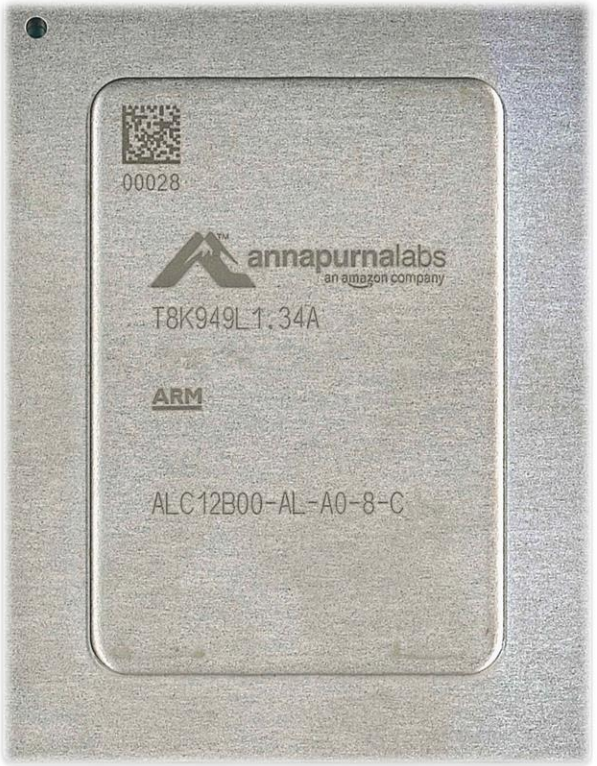


고객지향적인 AWS 의 사상에 따라 빠르게 혁신하고, 빌드하며 변화

AWS 의 혁신 – AWS Graviton 프로세서



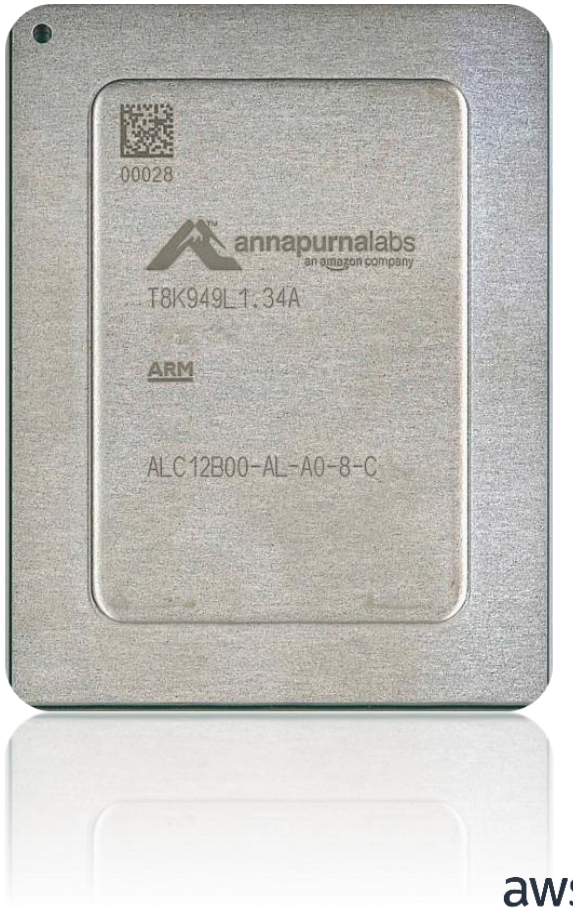
AWS Graviton 1세대 프로세서
16nm 실리콘 공정
~50억개의 트랜지스터로 구성
첫 ARM 기반 AWS 프로세서



AWS Graviton2 processor
4 배 많은 vCPU
7 배 높은 CPU 성능
개별 vCPU 마다 2배까지 높은 성능
7nm 실리콘 공정
~300 억개의 트랜지스터

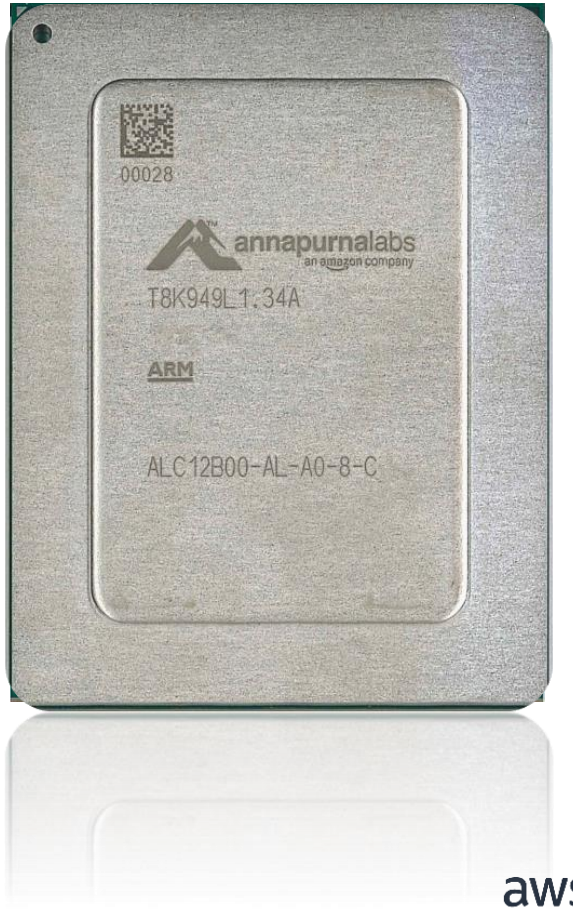
AWS Graviton2 프로세서 – 그 안에는..

- Arm® Neoverse™ N1 cores
- Arm v8.2 compliant
- Arm 과의 협업을 통해 생성된 N1
  - Large 64KB L1 caches and 1MB L2 cache/vCPU
  - Coherent Instruction cache
  - 인터럽트, 가상화 및 context switching 에 낮은 오버헤드
  - 4-wide front-end, with 8-wide dispatch/issue
  - Dual-SIMD units
  - 기계 학습 인퍼런스 업무에 가속화 제공: int8, fp16
- 모든 vCPU 는 물리코어로 제공
  - 하이퍼스레딩 (SMT) 이 아닌 개념



AWS Graviton2 프로세서 – 그 안에는..

- 64 cores connected together with a mesh
- ~2TB/s bisection bandwidth
- 32MB LLC private caches over 100MB of user-accessible caches
- No NUMA concerns
  - Every core sees the same path to memory and to other cores
- 64 lanes of PCIe gen4
- 8x DDR4-3200 channels → over 200GB/s
  - Always AES-256 encrypted DRAM with ephemeral key
  - Uniform memory latency from all CPU cores
- 1Tbit/s of compression accelerators
  - 2xlarge and larger instances will have a compression device



Graviton2 프로세서는 어디에?

차세대 코어 기반 EC2 인스턴스

M6g

범용성 업무

1:4
vCPU:memory

R6g

메모리 최적화

1:8
vCPU:memory

C6g

컴퓨팅 업무 최적화

1:2
vCPU:memory

T4g

범용 - 버스팅

Free trial 제공

C6gn

컴퓨팅+고성능 네트워크

100 Gbps Networking
38 Gbps EBS

이외에도 :

NVMe 로컬 인스턴스 스토리지 기반의 M6gd, R6gd, C6gd 타입

M, C, and R 타입에 .metal 인스턴스

AWS Graviton2 프로세서 기반의 8개 신규 EC2 인스턴스

Graviton2 프로세서 : ElastiCache

ElastiCache는 45% 의 성능대비 가격 개선을 앞선 세대 인스턴스 대비 제공하며 Graviton2 인스턴스가 기본적으로 사용됩니다

Amazon ElastiCache now supports M6g and R6g Graviton2-based instances

Posted On: Oct 8, 2020

Amazon ElastiCache is announcing the launch of ElastiCache for Redis and Memcached on Graviton2 M6g and R6g instance families. Customers choose Amazon ElastiCache for workloads that require ultra-low latency and high throughput, and can now enjoy up to a 45% price/performance improvement over previous generation instances. Graviton2 instances are now the default choice for ElastiCache customers.

Source: https://aws.amazon.com/about-aws/whats-new/2020/10/amazon-elasticache-now-supports-m6g-and-r6g-graviton2-based-instances/

Graviton2 프로세서 : RDS

Graviton2 기반 인스턴스를 통해 35% 까지 개선된 성능과 52% 개선된 성능/가격을 RDS 오픈소스 데이터베이스 엔진에서 제공합니다.

Achieve up to 52% better price/performance with Amazon RDS using new Graviton2 instances

Posted On: Oct 15, 2020

AWS Graviton2-based database instances are now generally available for Amazon Relational Database Service (RDS). Graviton2 instances provide up to 35% performance improvement and up to 52% price/performance improvement for RDS open source databases depending on database engine, version, and workload. You can launch these database instances when using Amazon RDS for MySQL, Amazon RDS for PostgreSQL, and Amazon RDS for MariaDB. Support for Amazon Aurora is coming soon.

AWS Graviton2 processors are custom built by Amazon Web Services using 64-bit Arm Neoverse cores and deliver several performance optimizations over first-generation AWS Graviton processors. This includes 7x the performance, 4x the number of compute cores, 2x larger private caches per core, 5x faster memory, and 2x faster floating-point performance per core. Additionally, the AWS Graviton2 processors feature always-on fully encrypted DDR4 memory and 50% faster per core encryption performance.

Source: https://aws.amazon.com/about-aws/whats-new/2020/10/achieve-up-to-52-percent-better-price-performance-with-amazon-rds-using-new-graviton2-instances/

Graviton2 프로세서 : EMR

Amazon EMR 는 35% 까지 낮은 가격 그리고 15% 까지 개선된 성능을 이전 아키텍처 기반의 인스턴스 대비 제공합니다

Amazon EMR now provides up to 30% lower cost and up to 15% improved performance for Spark workloads on Graviton2-based instances

Posted On: Dec 4, 2020

Amazon EMR now supports Amazon EC2 M6g, C6g and R6g instances with EMR Versions 6.1.0, 5.31.0 and later. These instances are powered by AWS Graviton2 processors that are custom designed by AWS utilizing 64-bit ArmNeoverse cores to deliver the best price performance for cloud workloads running in Amazon EC2. Please read our blog for more information.

On Graviton2 instances, Amazon EMR runtime for Apache Spark provides an additional cost savings of up to 30%,and improved performance of up to 15% relative to equivalent previous generation instances. Additionally, TPC-DS3 TB benchmark queries run up to 32 times faster using Amazon EMR runtime for Apache Spark.

https://aws.amazon.com/about-aws/whats-new/2020/12/amazon-emr-now-provides-up-to-30-lower-cost-and-up-to-15-improved-performance/

운영체제

Commercial

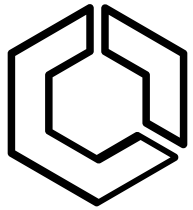
- Amazon Linux 2
- Red Hat Enterprise Linux 8.2+
- 18.04 LTS, 20.04 LTS
- SLES 15 SP2

Community

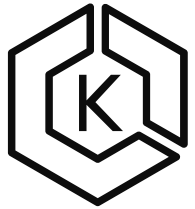
- CentOS
- Debian
- fedora
- FreeBSD
- NetBSD

컨테이너

Runtimes



Amazon ECS

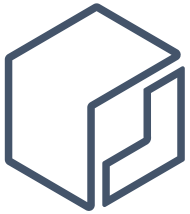


Amazon EKS

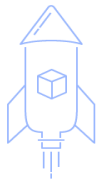
Docker

Kubernetes

Multi-Arch Registries



Amazon ECR

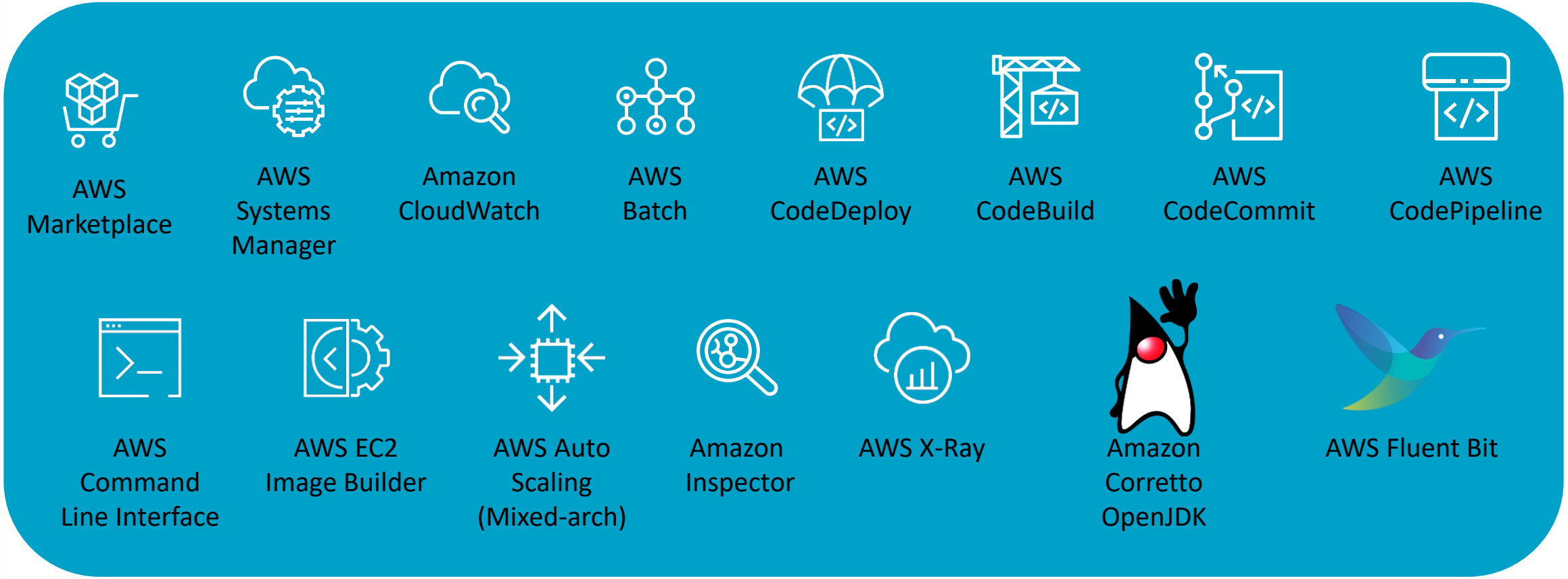


Bottlerocket (Container OS)
(github.com/bottlerocket-os)

Firecracker (MicroVMs)
(github.com/firecracker-microvm)

aws

AWS 툴과 서비스



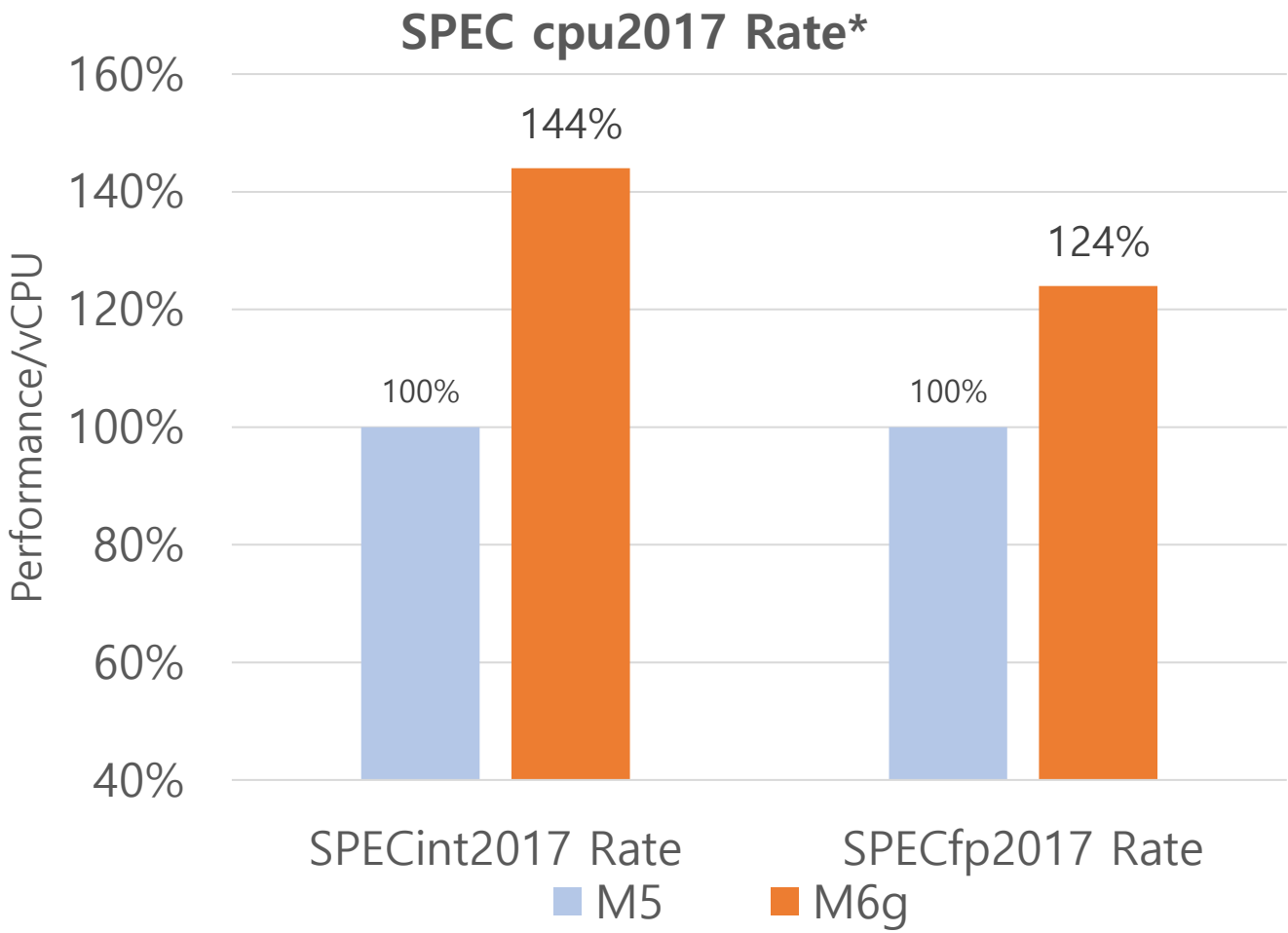
AWS Graviton 에코시스템

Databases	Cassandra, KeyDB, MariaDB, Memcached, MongoDB, MySQL, PostreSQL, Redis, ScyllaDB
Configure & Monitor	Chef, Datadog, Dynatrace, Honeycomb, New Relic, Splunk, Terraform
Secure	CrowdStrike, Qualys, Rapid7, Snyk, Tenable
Build/Test	CircleCI, GitHub Actions, GitLab, Jenkins, Travis CI

Graviton2 프로세서의 성능

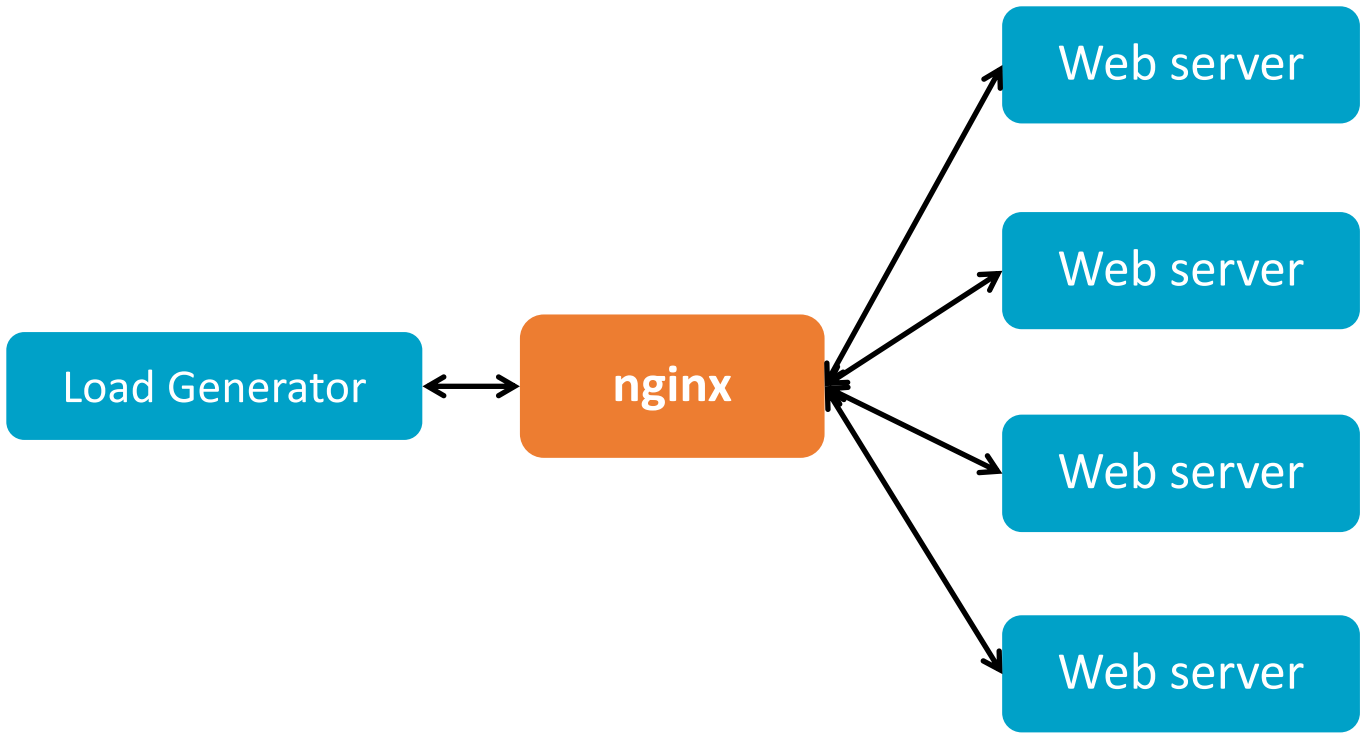
SPEC cpu2017

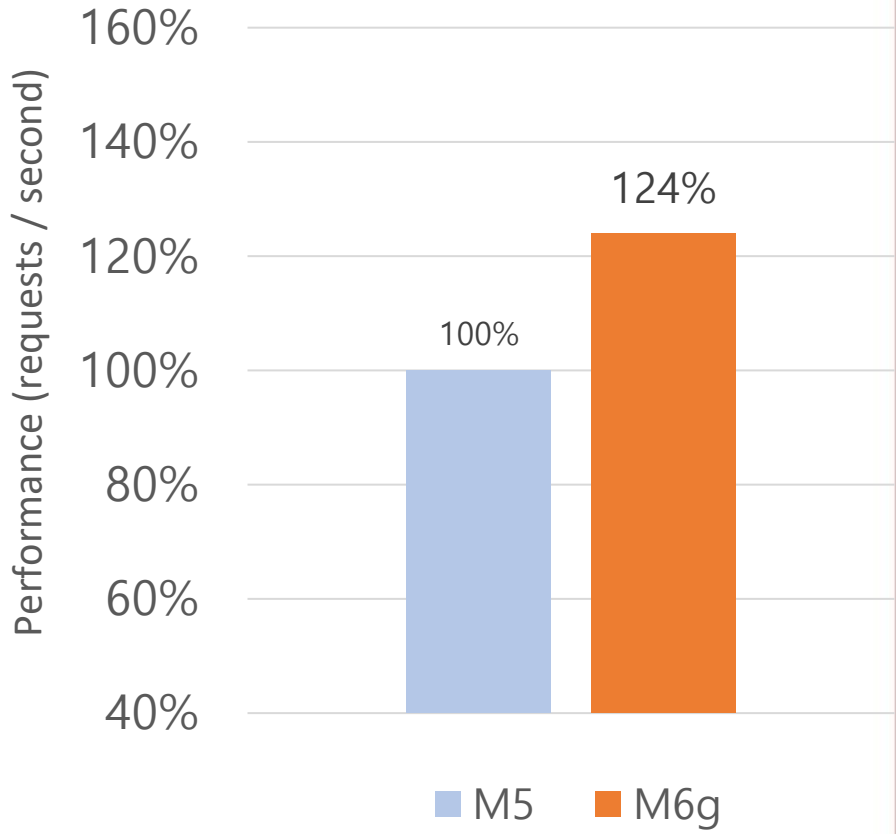
- CPU 특화된 업계 표준 벤치마크 지표
- 동시에 모든 vCPU 에서 작동
- vCPU 의 성능 비교



** All SPEC scores estimates, compiled with gcc v9 -O3 -march=native, run on largest single-socket size for each instance type tested.*

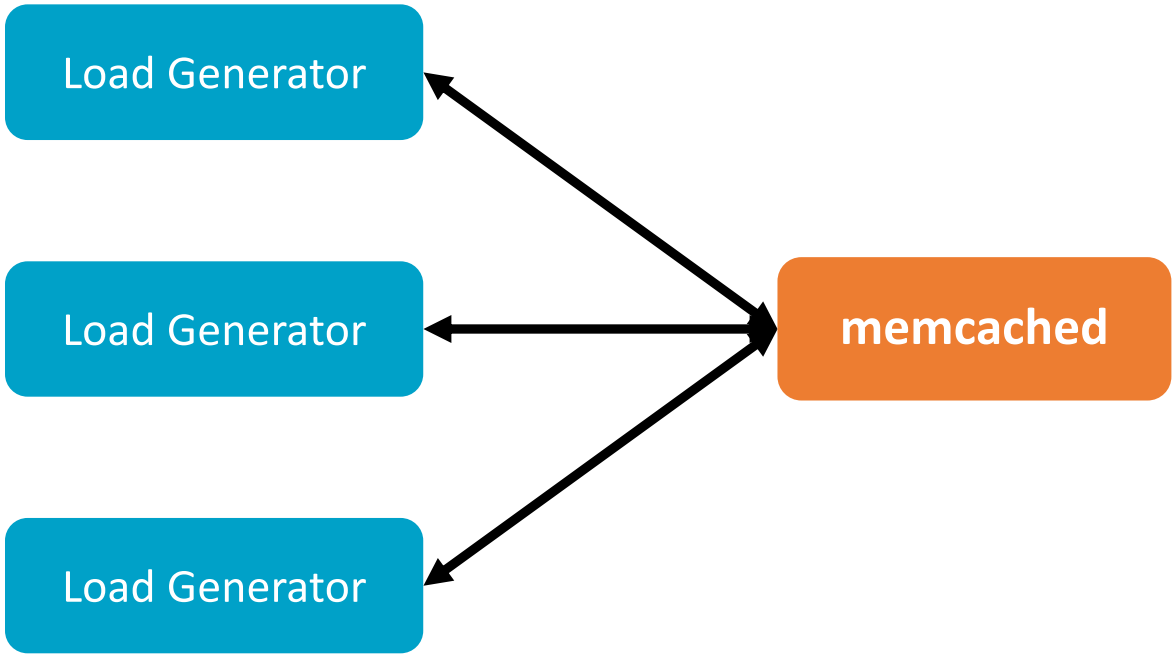
NGINX 부하 분산 성능

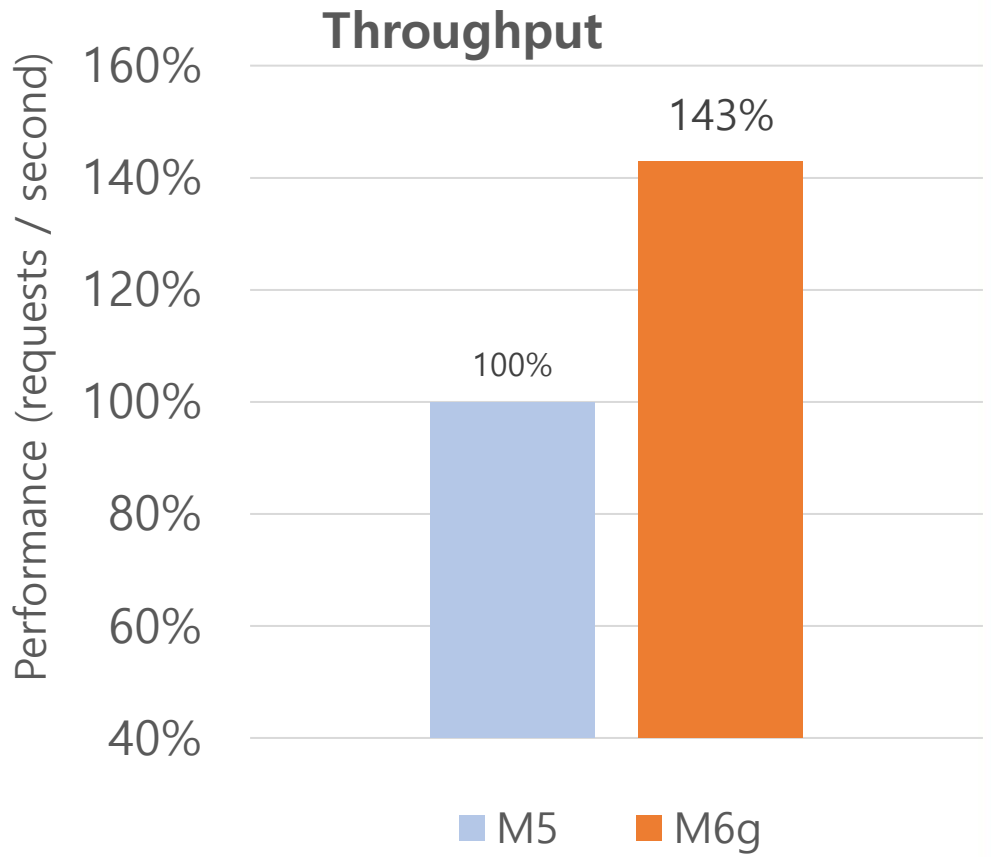




NGiNX v1.15.9, 512 clients, 128 GET/POST payloads, all HTTPS connections, AES128-GCM-SHA256, OpenSSL 1.1.1, 4 target machines, all tests run on 4xl size; load generator c5.9xl; web servers c5.4xls; All servers run in a cluster placement group.

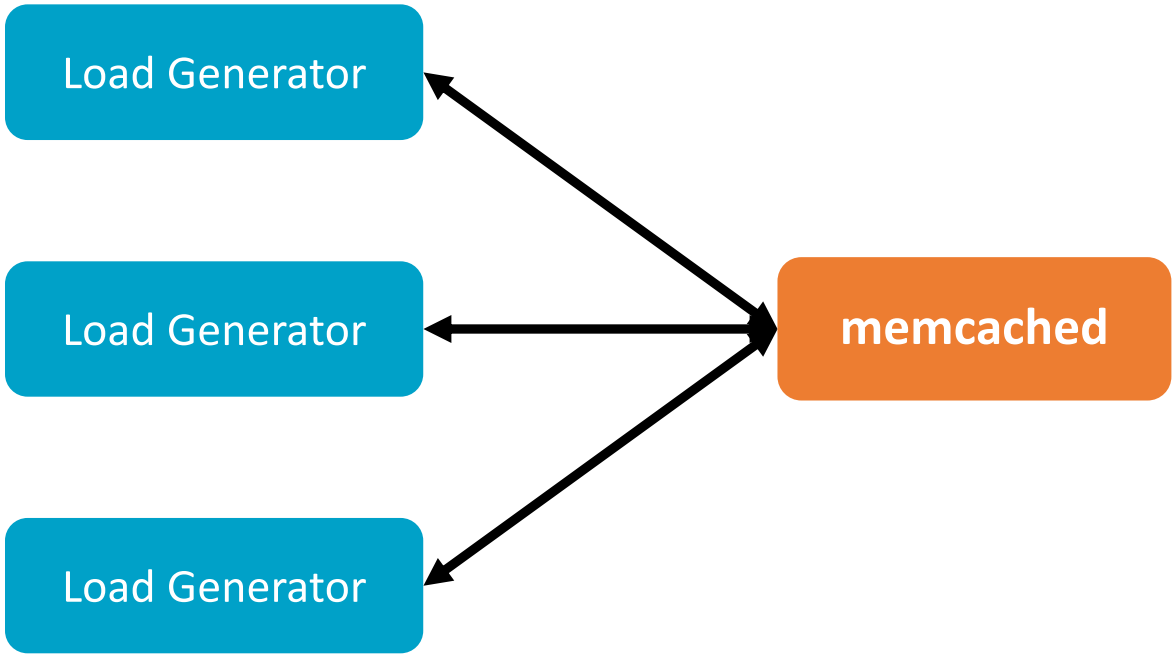
Memcached 캐싱 성능

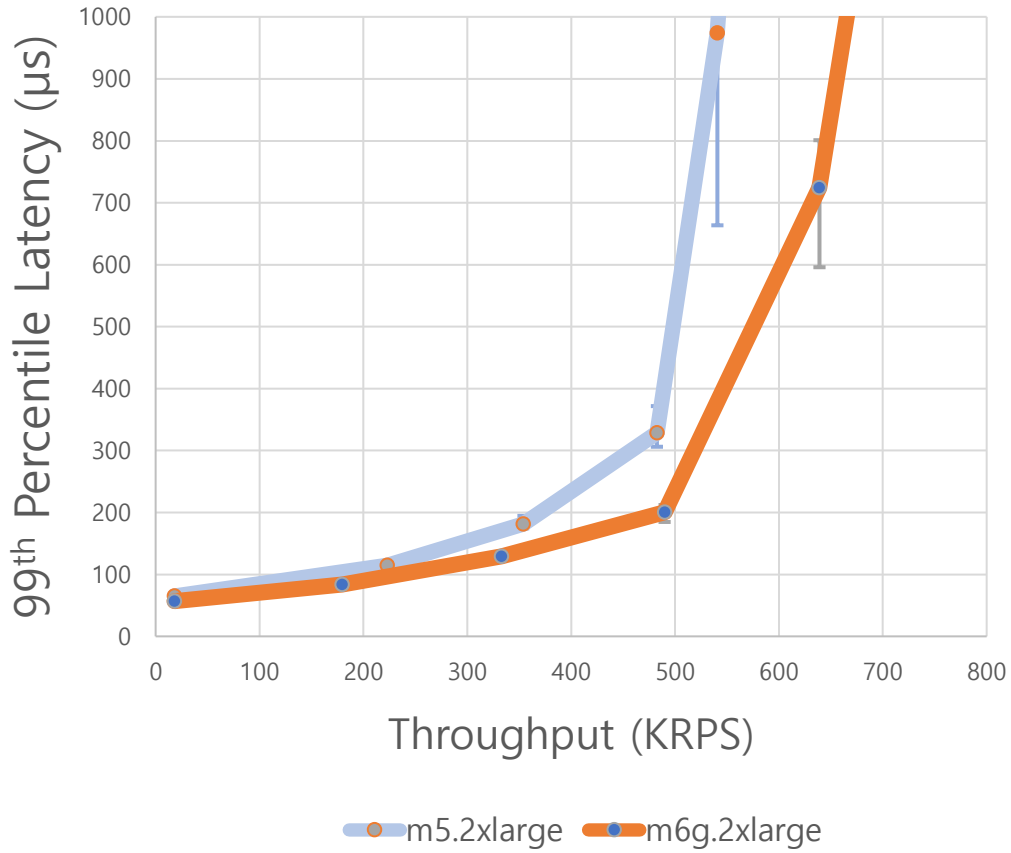




Memcached v1.5.16, 16B keys, 128B values, 7.8M KV-pairs, 576 connections for load generation from 2x c5.9xlarge instances; 16 additional connections measuring latency from 1 additional c5.9xlarge, each connection maintains 4096 outstanding requests; All servers in a cluster placement group.

Memcached 캐싱 성능





Memcached v1.5.16, 16B keys, 128B values, 7.8M KV-pairs, 576 connections for load generation from 2x c5.9xlarge instances; 16 additional connections measuring latency from 1 additional c5.9xlarge, each connection maintains 4096 outstanding requests; All servers in a cluster placement group.

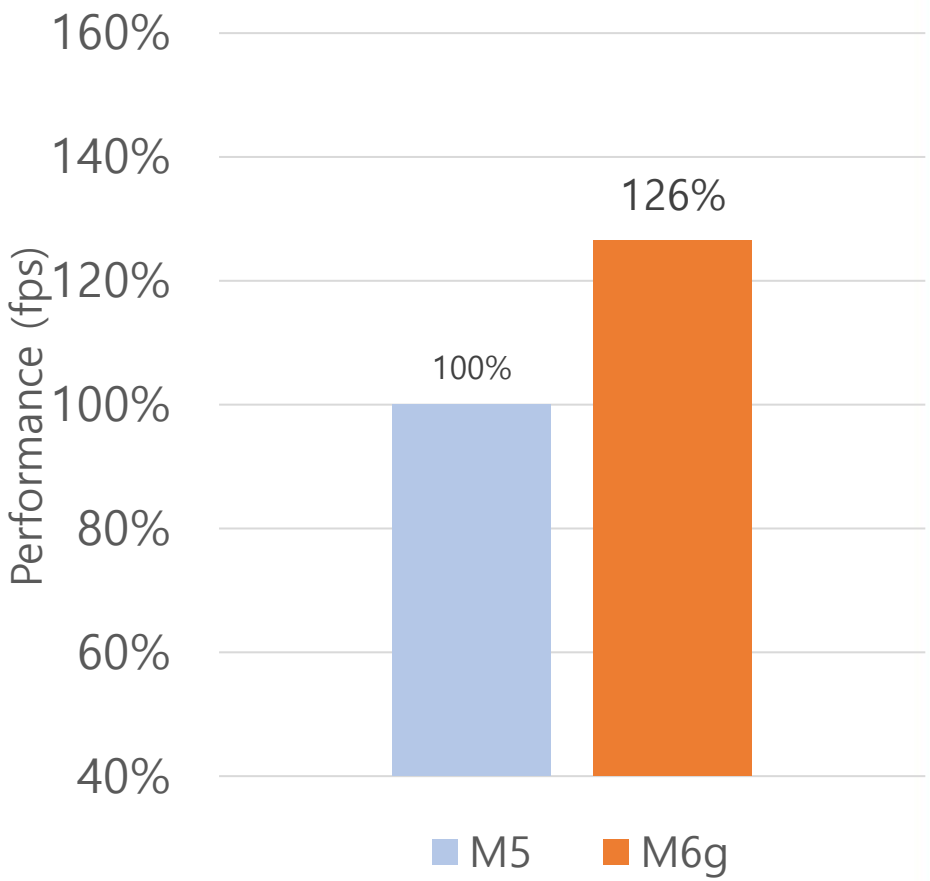
x264 비디오 인코딩

인터넷에서 60% 이상의 다운 스트림을 차지하는 비디오 컨텐츠

비디오 인코딩을 통해 대역폭과 성능을 개선

libx264 소프트웨어 기반 인코더를 통해 H.264 기반 1080p 영상 압축 테스트

초당 프레임 처리 속도 26% 이상의 성능 개선



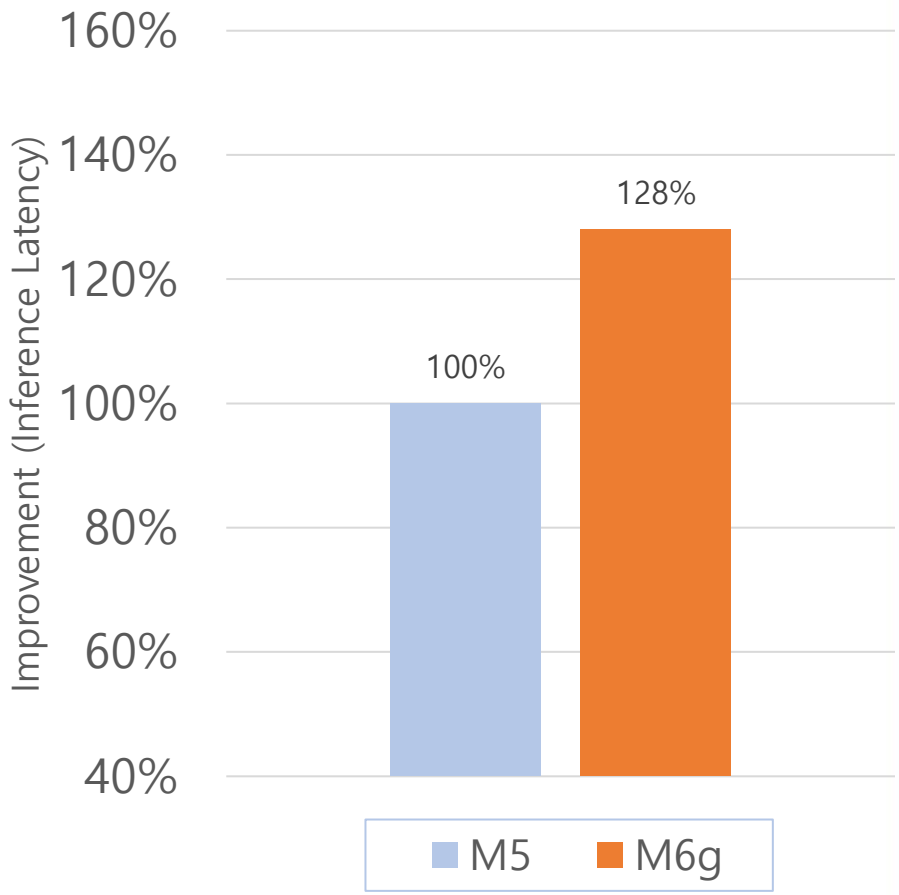
x264 (3759fcb7), 4xl instance size, medium preset, input uncompressed 1080p50, output encoded h.264 1080p50

기계 학습 - ML

BERT: 인코더
- 머신러닝의 필수 단계인 특성추출/표현
- 심층신경망을 통해 텍스트의 특성 표현

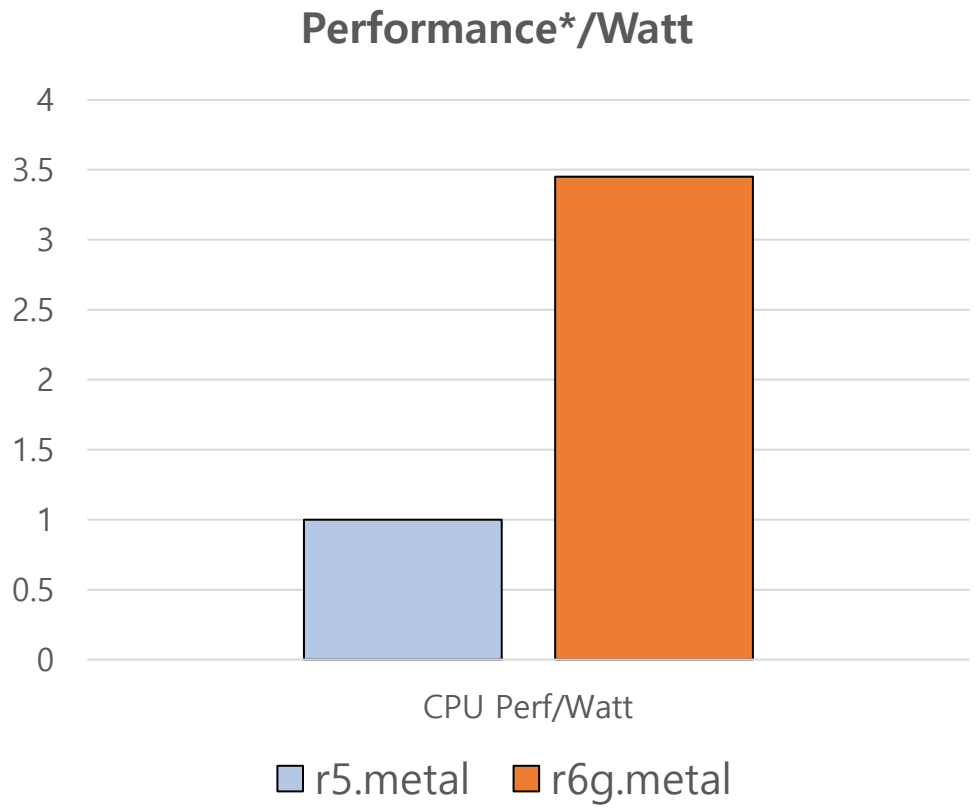
Graviton2 에 탑재된 fp16와 int8의 지원기능으로 기계학습 가속화

M6g는 이전세대의 M5 보다 CPU 기반의 추론 업무에 월등한 결과 제공
- M5 with AVX-512 is limited to FP32
- With FP16, M6g performs better



BERT classification using TVM and 64 length sequence on CPUs
Batch size of one; dedicated instances on .xlarge size

CPU 전력 효율화



*Estimated SPECint2017

낮은 전력 사용량:

→ 높은 밀도

→ 낮은 비용

→ 낮은 탄소 수치

어떻게 쓰나요?

컨테이너

- 컨테이너 이미지의 경우, **아키텍처 특화** 되어있으며 이미지가 arm64 기반이어야 함
- Best practices: **다중 아키텍처 Manifest 를 사용하는 병렬 빌드 파이프라인**
- DockerHub과 Amazon ECR은 멀티-이케텍처 Manifest list 를 지원
- **Amazon ECS**와 **Amazon EKS**는 AWS Graviton과 Graviton2 인스턴스를 이미 지원하며 x86과 ARM 클러스터의 조합으로 이용 가능
- **Bottlerocket** (컨테이너를 위한 리눅스 기반OS)역시 Graviton2/arm64 지원:

https://github.com/bottlerocket-os/bottlerocket

aws

ARM 아키텍처를 지원하는 AMI 사용방법

- AMI ID를 CloudFormation 템플릿에 하드코딩 없이 **AWS Systems Manager Parameter Store**를 사용하시면 편합니다

```
Parameters:
  LatestAmiId:
    Type: 'AWS::SSM::Parameter::Value<AWS::EC2::Image::Id>'
    Default: '/aws/service/ami-amazon-linux-latest/amzn2-ami-hvm-arm64-gp2'

Resources:
 Graviton2Instance:
    Type: 'AWS::EC2::Instance'
    Properties:
      ImageId: !Ref LatestAmiId
      InstanceType: 'c6g.medium'
```

C / C++

- 기존에 컴파일 되었거나 Low Level 코드: 최신 컴파일러로 **리컴파일 필요**
- 어셈블리, instrinsics 또는 AVX: **포팅 필요** (sse2neon 사용 가능)
- Large-System Extensions (LSE) atomic operations 은 기존 Amazon Linux 2 와 Ubuntu 20.04 를 통해 최적화된 libc 제공
- 새로운 머신러닝 코드: **ARM dot-product instructions** (점곱)이나 **float16** (half-precision floating-point)을 통해 빠르고 정확한 결과값 생성
- ARM performance 라이브러리는 Graviton2 인스턴스에서 **무료**로 사용이 가능합니다.

https://developer.arm.com/tools-and-software/server-and-hpc/downloads/arm-performance-libraries

	GCC < 9	GCC >= 9
Core specific Tuning – Amazon Linux 2	-mtune=neoverse-n1	
Core specific tuning – Other distributions	-mtune=cortex-a72	-mtune=neoverse-n1

Java (1)

- Java 애플리케이션은 bytecode 로 컴파일되어 리컴파일 필요 없음
- 일반적으로 arm64 기반 환경에서 즉시 사용 가능

- Amazon Corretto (https://aws.amazon.com/corretto/):
  - 장기 지원 제공, OpenJDK의 운영 환경이 준비된 버전
- Version 8, 11 및 15 제공
- 11버전 이상은 Graviton2 기반의 인스턴스에서 최적의 성능을 제공
- Corretto 11 – 2020년 10월 릴리즈:
  - Enabled -moutline-atomics for aarch64
  - Optimized I2L / L2I conversions on aarch64
  - Faster Math.signum(fp) on aarch64
  - Various other aarch64 improvements

Java (2)

자바 JARs 파일의 경우, 아키텍처 특화된 공유 객체를 JNI 호출을 통해 포함시킬 수 있습니다.

- 방법을 살펴보자면 :

```
$ unzip foo.jar
$ find . -name "*.so" | xargs file
```

다음, x86_64 ELF 객체
에 aarch64 대응하는 항목을 확인

- 멀티 아키텍처 JAR 빌드를 하려면 아래의 링크를 참조 바랍니다.

https://github.com/aws/aws-graviton-getting-started/blob/master/java.md

Python

- Python은 하이-레벨의 컴파일이 불필요한 인터프리티드 언어로 리컴파일이 필요 없습니다.
- arm64/Graviton2 기반 환경에서 완벽히 지원됩니다.

- 단점으로는 : pip (Python package installer)을 사용시, PPI(Python Package Index)에서 패키지를 받아오는 경우 C / C++ 연속 파일이 컴파일 안된 상태로 다운로드 되어 총 설치시간에 컴파일이 포함되어 패키지 사이즈가 큰 경우 오랜 시간이 소요 될 수 있습니다.

장시간 소요되는 설치시간을 줄이기 위해 AWS는 현재 패키지를 프리-컴파일하여 제공합니다:
  - 해결된 패키지: numpy, scipy, cffi, pywavelets, pillow, bcrypt, sqlachemy, ...
  - 해결 중: pytorch, tensorflow, scikit-image, twisted, matplotlib, ...

추가로, AWS Graviton2 의 Getting Started 가이드문서를 통해 보다 자세한 내용을 확인하세요:
https://github.com/aws/aws-graviton-getting-started/blob/master/python.md

.NET workload

.NET Core runs perfectly on Amazon EC2 Linux Instances

AWS now provides a free tool to accelerate the porting of .NET applications to .NET Core:
https://aws.amazon.com/porting-assistant-dotnet/

.NET 5 (released Nov 2020) brings significant arm64 performance improvements (**up to 17 %** on Techempower benchmarks),

Making **Graviton2-based instances** a platform of choice to run those Workloads

aws

NodeJS, Ruby

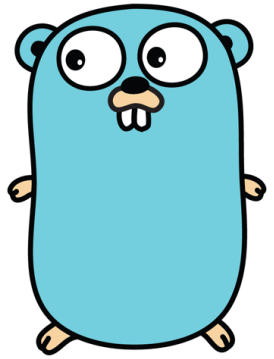
- NodeJS – Based on Chrome's V8 JavaScript engine
  - Fully supported on arm64 / linux – Binaries available for your favorite linux distribution or directly from http://nodejs.org

- Ruby – Fully supported on arm64 / linux, available from your linux distribution
  - AWS contributed optimization (up to 20% faster on string operations):

https://github.com/ruby/ruby/commit/511b55bcefc81c036294dc9a544d14bd342acd3b

Go

- Compiled language: **applications need to be recompiled**
- Cross-compiler is built-in: set GOOS and GOARCH



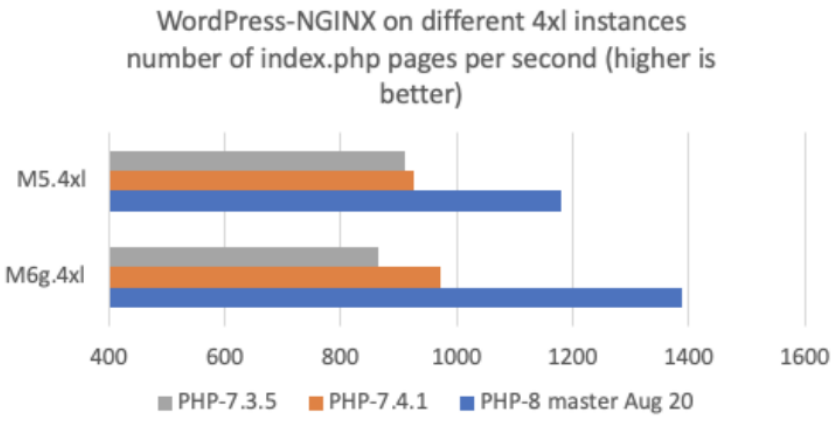
```
$ env GOOS=linux GOARCH=arm64 go build ...
```

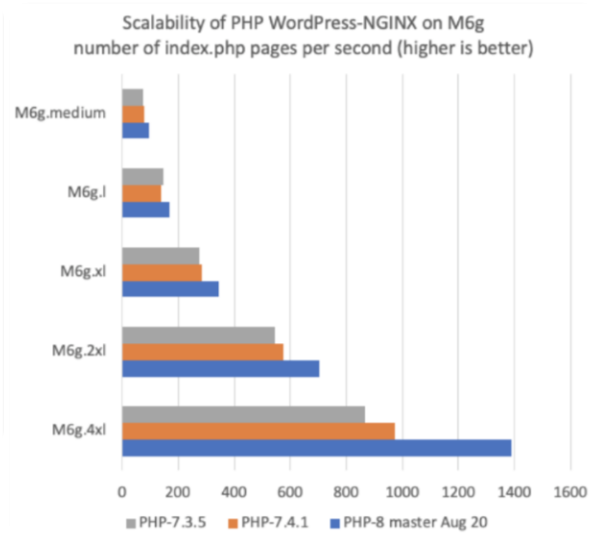
- Future developments – Go 1.16 (expected Jan 2021):
- Support for **ARMv8.1 Atomics** – better mutex
- Copy perf improvements (**memmove arm64 specific optimizations**)

aws

PHP

- PHP is interpreted, **no need to modify applications** to run on Graviton2
- AWS contributions to PHP-7.4: up to **37% better perf.** vs PHP-7.3 on Graviton2
- Future developments in **PHP-8: faster regexp engine** (PCRE2 10.34 vectorized with NEON, **up to 8x faster**), improved **toupper / tolower** (speedup **16.5x**), Opcache JIT with arm64 s upport





AWS contributions to PHP-7.4

Function	Speedup	Commits to PHP-7.4
inc/dec	1.5x	https://github.com/php/php-src/pull/4094
add/sub	1.82x	https://github.com/php/php-src/pull/4095
hash_init	1.61x	https://github.com/php/php-src/pull/4096
hash_vect	1.72x	https://github.com/php/php-src/pull/4126
crc32	29x	https://github.com/php/php-src/pull/4108
rev64	7.8x	https://github.com/php/php-src/pull/4109
base64 encode	3.5x	https://github.com/php/php-src/pull/4381
base64 decode	2.15x	https://github.com/php/php-src/pull/4381
string addslashes	2.8x	https://github.com/php/php-src/pull/4396
string stripslashes	4.9x	https://github.com/php/php-src/pull/4396

데모

june@JP:~$

june@JP:~$

정리해 보자면...

- Graviton2 기반의 인스턴스는 40% 이상의 가격대비 성능을 기존 x86 인스턴스 대비 제공합니다.
- Graviton2 를 체험해보기 위해 T4g 인스턴스에 무료 체험 기간을 제공하고 있습니다
- 보다 자세한 내용은 세부적인 내용이 제공되는 기술문서를 참고 부탁드립니다.

https://aws.amazon.com/ec2/graviton

마지막으로 다양한 Best Practice 를 제공해드리며 아래에서 찾아보실 수 있습니다.

실습 – https://graviton.june.nz

더 나은 세미나를 위해
여러분의 의견을 남겨주세요!

▶ 질문에 대한 답변 드립니다.

AWS Builders Program 300 –

감사합니다.

실습 – https://graviton.june.nz

aws