CMP203

AWS re:INVENT

EC2 Foundations

Raj Pai, Director
EC2 Product Management

November 30, 2017

EC2 Foundations

Resources

Instances
Storage
Networking

Availability

Regions and AZs
Placement Groups
Load Balancing
Auto Scaling

Management

Deployment
Monitoring
Administration

Purchase Options

On Demand
Reserved
Spot

EC2 Foundations

Resources

Instances
Storage
Networking

Availability

Regions and AZs
Placement Groups
Load Balancing
Auto Scaling

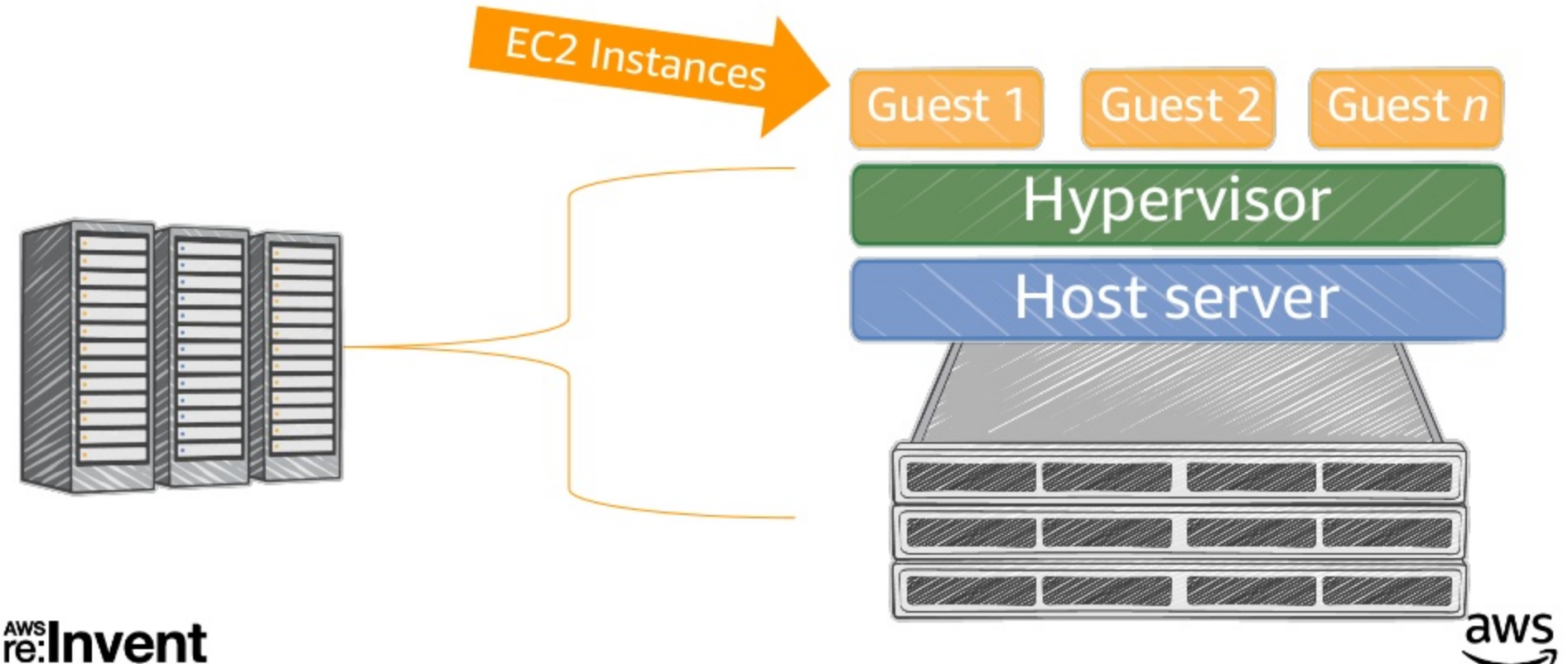
Management

Deployment
Monitoring
Administration

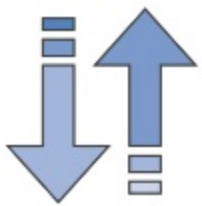
Purchase Options

On Demand
Reserved
Spot

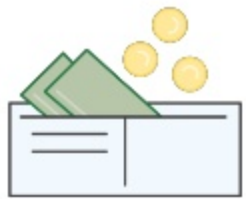
Amazon Elastic Compute Cloud (EC2): Virtual servers in the cloud



Amazon EC2 11+ years ago...



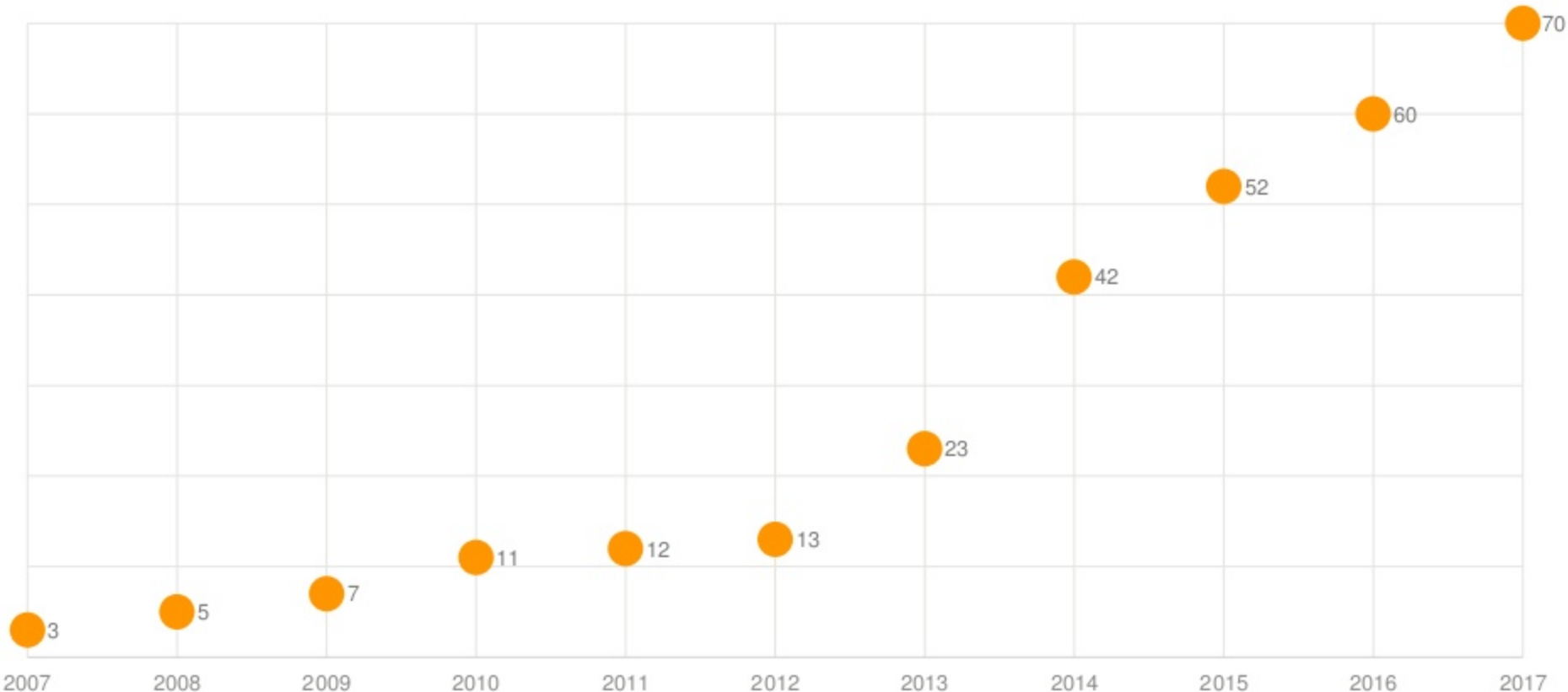
Scale up or down quickly, as needed



Pay for what you use

“One size fits all”

EC2 instance growth since then



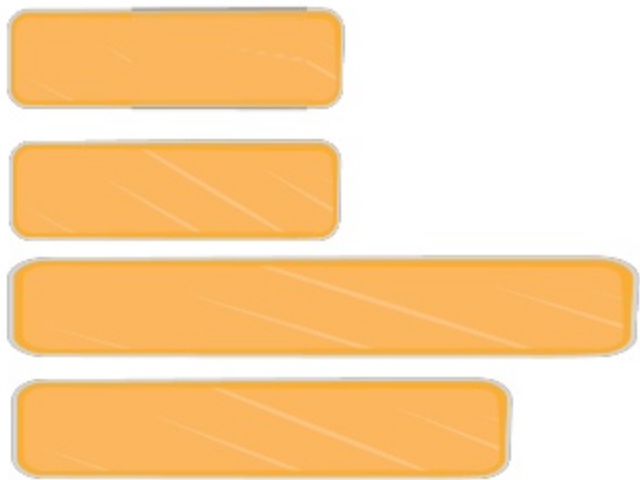
EC2 instance characteristics

CPU

Memory

Storage

Network Perf



Instance generation

i3.xlarge

Instance family

Instance size

Instance type

Amazon Machine Images (AMIs)

Amazon maintained

Set of Linux and Windows images

Kept up-to-date by Amazon in each region

Community maintained

Images published by other AWS users

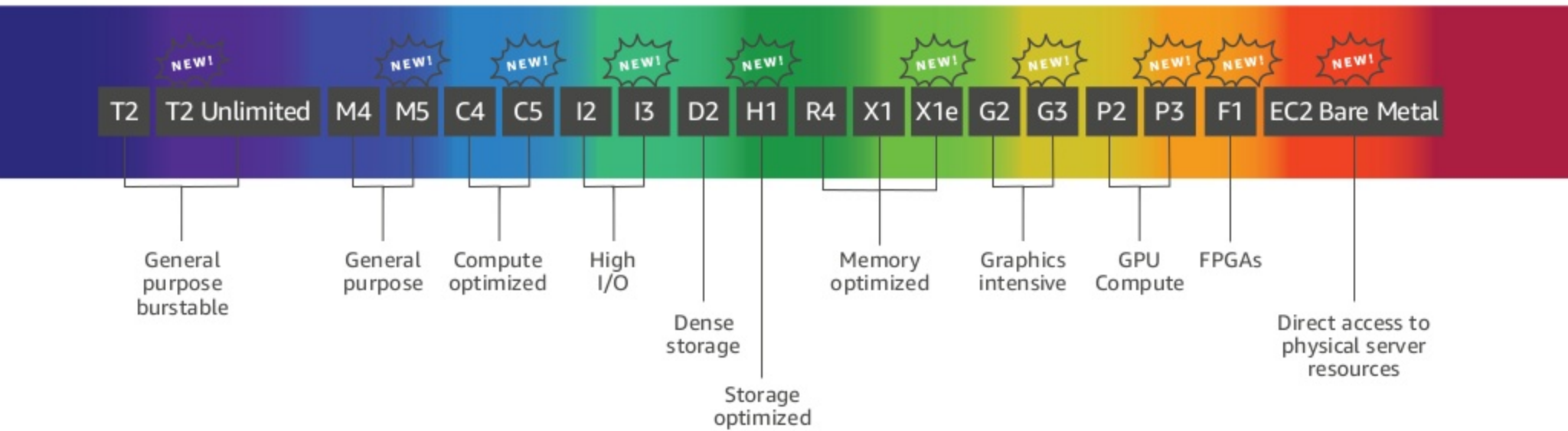
Managed and maintained by Marketplace partners

Your machine images

AMIs you have created from EC2 instances

Can be kept private or shared with other accounts

EC2 instances



General Purpose instance workloads

Web/app servers

Enterprise apps

Gaming servers

Caching fleets

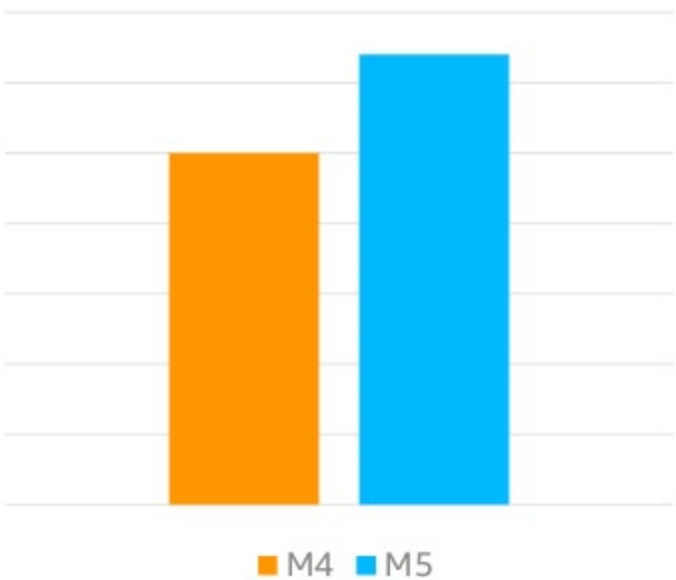
Analytics applications

Dev/test environments

NEW

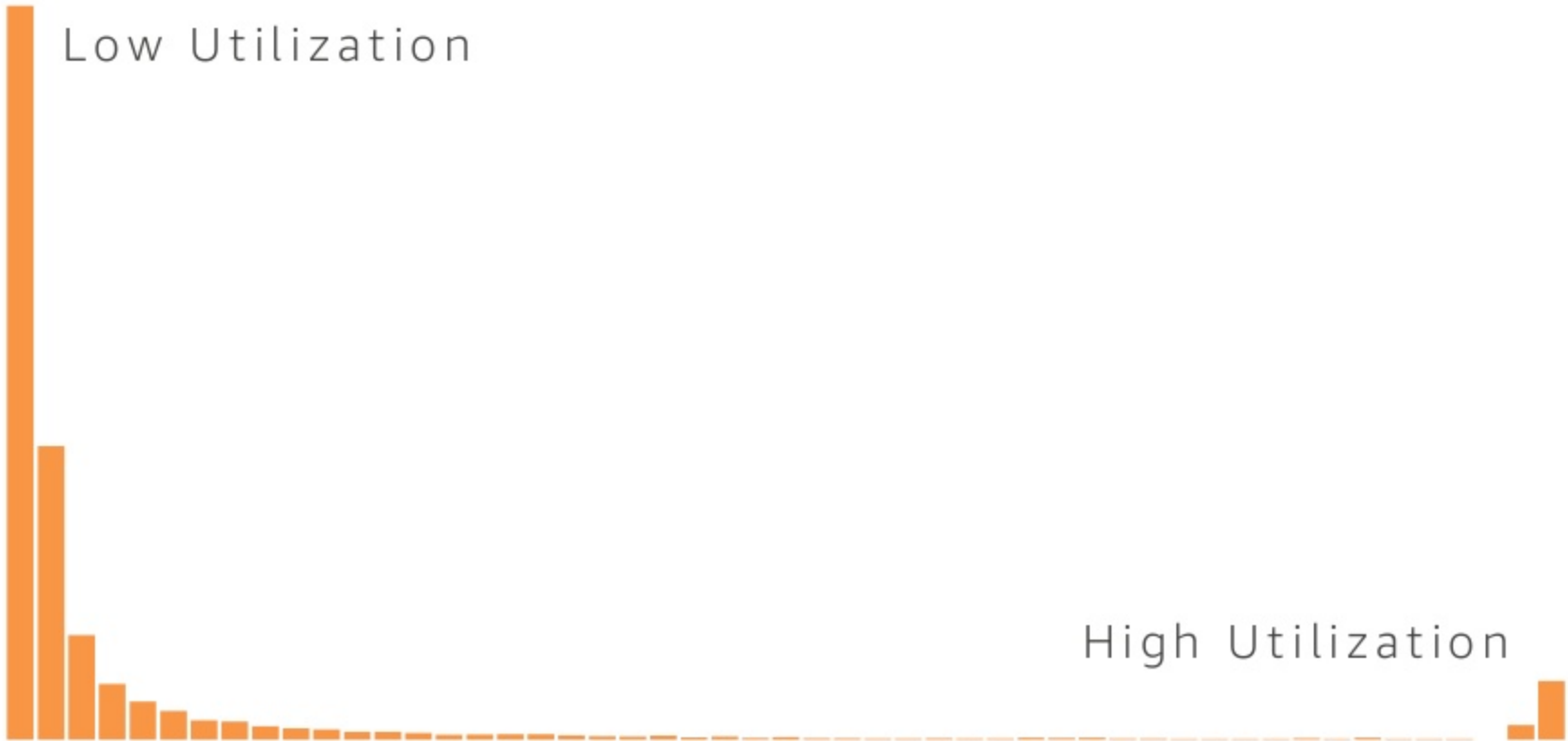
M5: Next-Generation General Purpose instance

14% price/performance improvement With M5



- Powered by 2.5 GHz Intel Xeon Scalable Processors **(Skylake)**
- New larger instance size—m5.24xlarge with **96 vCPUs** and **384 GiB of memory** (4:1 Memory:vCPU ratio)
- Improved network and EBS performance on smaller instance sizes
- Support for Intel **AVX-512** offering up to twice the performance for vector and floating point workloads

Opportunity: Most instances aren't very busy

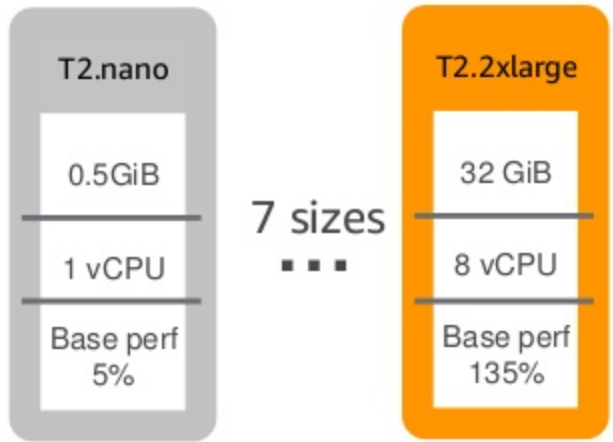


NEW

T2: General Purpose Burstable instances

T2 Burstable Performance instances provide a generous **baseline level of CPU** performance with the ability to **burst above the baseline**

Lowest cost EC2 instance at $0.0058 per hour, and available on AWS Free Tier



With **T2 Unlimited,** burst whenever you want for as long as you want

Just $0.05 per vCPU-hour over baseline, averaged over 24 hours

AWS re:Invent

aws

R4: Memory Optimized instances

- **8:1 GiB to vCPU ratio**
- **Memory-optimized** instances with Intel Xeon (Broadwell) processors
- Up to **25 Gbps NW bandwidth**

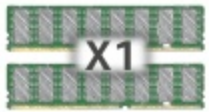
In-memory caches

In-memory databases

In-memory analytics

NEW

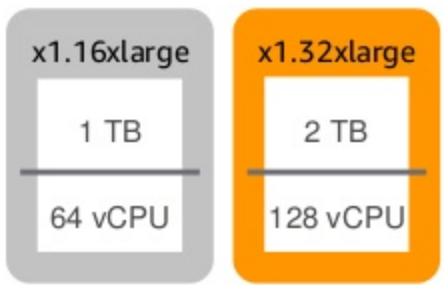
X1 and X1e—Large-Scale Memory-Optimized

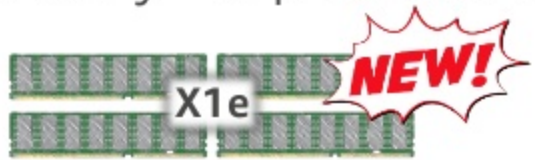


For large in-memory workloads

16:1 GiB to vCPU ratio

In-memory databases (e.g., SAP HANA), big data processing engines (Apache Spark, Presto), in-memory analytics

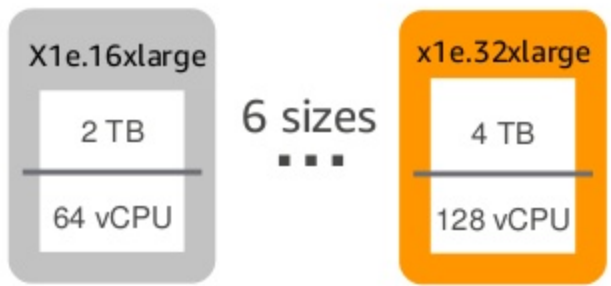




For memory-intensive workloads and very large in-memory workloads

32:1 GiB to vCPU ratio

High-performance databases, Large in-memory databases (e.g. SAP HANA), and DB workloads with vCPU based licensing (Oracle, SAP)



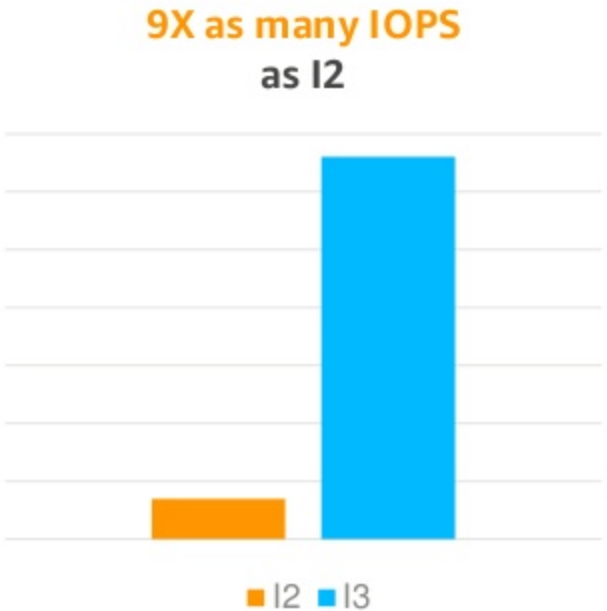
Roadmap through 2018: up to 16TB memory instances!

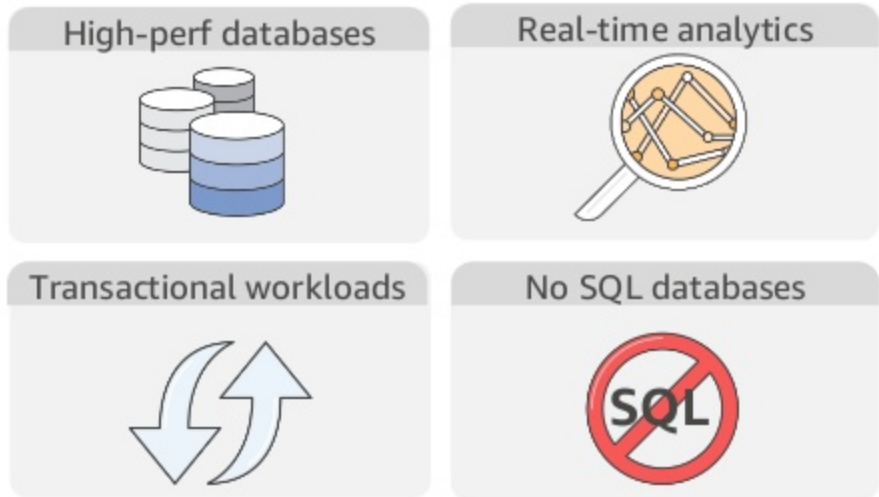
AWS re:Invent

aws

NEW

I3: I/O optimized instances





- Intel Xeon E5 v4 (Broadwell) processors, with up to 15.2 TB of locally attached NVMe SSD storage, 64 vCPUs, and 488 GiB memory
- Lowest cost per IOPS ($/IOPS)
- Offers very high Random I/O (up to 3.3 million IOPS) and disk throughput (up to 16 GB/s)
- Up to 25 Gbps NW bandwidth

AWS re:Invent

aws

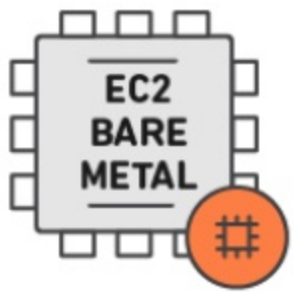
NEW

EC2 Bare Metal

EC2 Bare Metal

Run bare metal workloads on EC2
with all the elasticity, security, scale,
and services of AWS

i3.metal
36 hyperthreaded cores
15.2 TB SSD-based NVMe storage
512 GiB RAM



Designed for workloads that are not virtualized, require specific types of hypervisors, or have licensing models that restrict virtualization

Powers the VMware Cloud on AWS

NEW

Dense Storage workloads—D2 and H1

Data warehousing

HDFS

Log processing

d2.8xlarge

244 GiB

36 vCPU

48 TB HDD

- **Lowest cost per storage** ($/GB)
- Supports **high sequential disk throughput**

h1.16xlarge NEW!

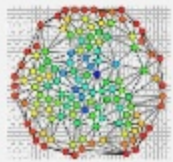
256 GiB

64 vCPU

16 TB HDD

- **More vCPUs and memory** per terabyte of disk
- **Lower costs** for big data uses cases

Big data



Kafka

MapReduce

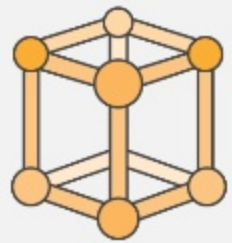
aws

Compute-intensive workloads

Batch processing

Distributed analytics

High-perf computing (HPC)



Ad serving

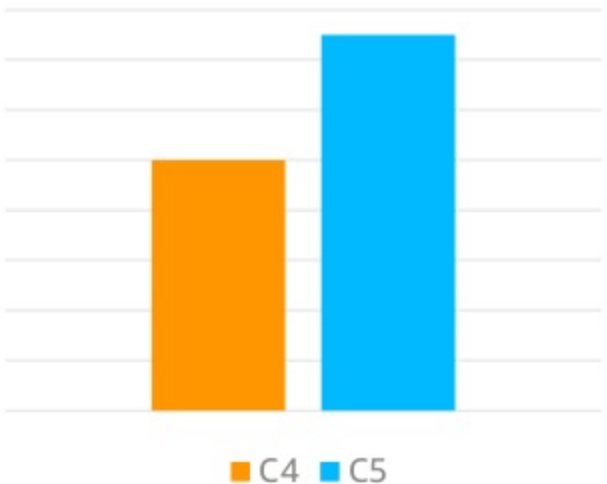
Multiplayer gaming

Video encoding

NEW

C5: Compute-optimized instances based on Intel Skylake

25% price/performance improvement over C4



- **Based on 3.0 GHz Intel Xeon Scalable Processors (Skylake)**
- **Up to 72 vCPUs and 144 GiB of memory (2:1 Memory:vCPU ratio)**
- **25 Gbps NW bandwidth**
- **Support for Intel AVX-512**

NETFLIX

"We saw significant performance improvement on Amazon EC2 C5, with up to a 140% performance improvement in industry standard CPU benchmarks over C4."

GRAIL

"We are eager to migrate onto the AVX-512 enabled c5.18xlarge instance size… . We expect to decrease the processing time of some of our key workloads by more than 30%."

AWS re:Invent

Accelerated computing on AWS

Parallelism increases throughout

CPU: High speed, highly flexible

GPU/FPGA: High throughput, high efficiency

GPUs and FPGAs can provide **massive parallelism** and **higher efficiency** than CPUs for many categories of applications

NEW

High-performance graphics with G3

Ideal for workloads needing massive parallel processing power

Visualizations

Cloud workstation

3D rendering

Video encoding

Virtual reality

4 GPUs, 64 vCPUs, 488 GiB of host memory, and 20 Gbps of network bandwidth

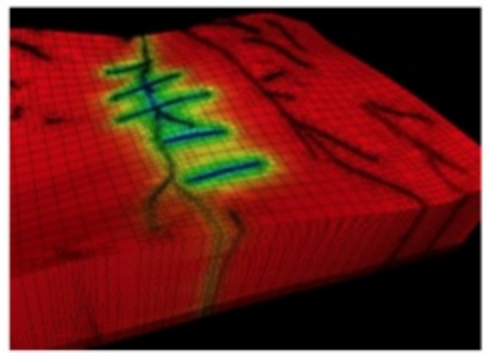
Tesla M60 GPU offers 8 GB of GPU memory, 2048 parallel processing cores and a hardware encoder
10 H.265 (HEVC) 1080p30 streams
18 H.264 1080p30 streams

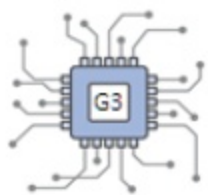
Seismic exploration and analytics for oil and gas

"The exploration and production models are increasingly complex with very large datasets, 3D and dynamic algorithms, security, and global reach… . Amazon EC2 G3 instances enable Landmark to deliver value to our clients in ways that were not possible before."

- Chandra Yeleshwarapu,
Global Head of Services and Cloud
Landmark, Halliburton

HALLIBURTON





aws re:Invent

aws

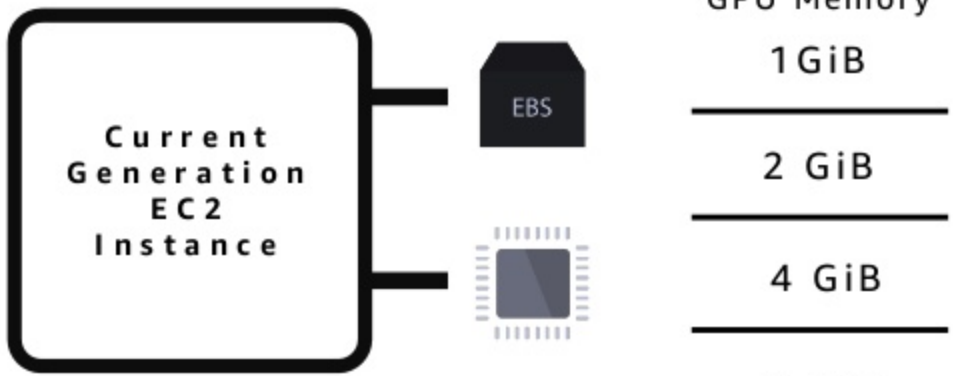
NEW

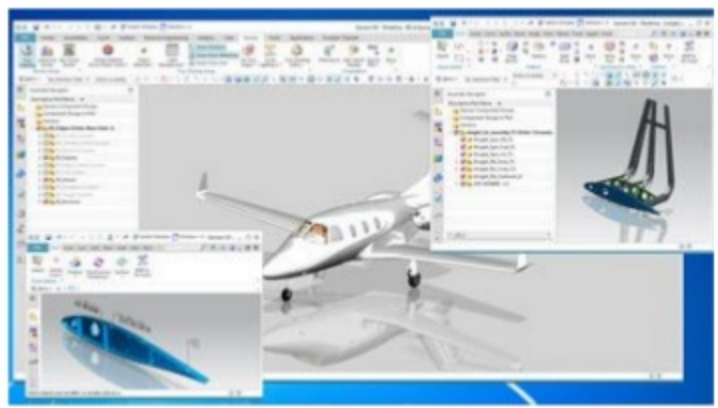
Graphics Acceleration: Elastic GPUs

Allows customers to add **low-cost graphics acceleration** to Amazon EC2 instances over the network

Come in a wide range of sizes; you can **attach GPUs to a wide range of EC2 instances** to achieve optimal performance

OpenGL compliant, giving you the confidence to run any graphics-intensive application





Use Cases for GPU Compute

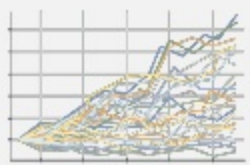
Machine learning/AI

Natural language processing

Image and video recognition

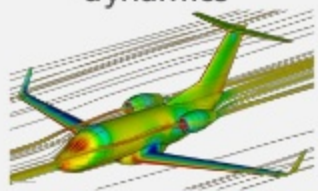
Autonomous vehicle systems

Recommendation systems



High-performance computing

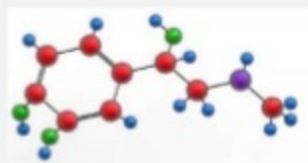
Computational fluid dynamics



Financial and data analytics

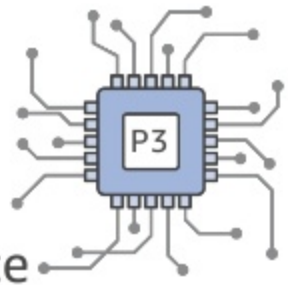
Weather simulation

Computational chemistry



NEW

Next Generation of GPU Compute Instances—P3 Instances



- Industry's most powerful GPU-based platform
- Based on NVIDIA's latest GPU Tesla V100
- 1 PetaFLOP of computational performance in a single instance
- Provides up to 14X performance improvement over P2 for machine learning use cases
- Up to 2.6X performance improvement over P2 for HPC use cases

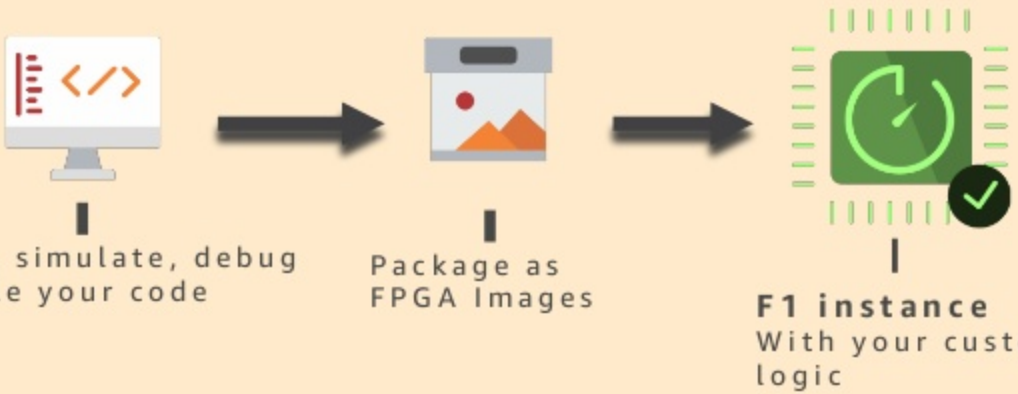
Instance Size	GPUs	Accelerator (V100)	GPU Peer to Peer	GPU Memory (GB)	vCPUs	Memory (GB)	Network Bandwidth	EBS Bandwidth
P3.2xlarge	1	1	No	16	8	61	Up to 10Gbps	1.7Gbps
P3.8xlarge	4	4	NVLink	64	32	244	10Gbps	7Gbps
P3.16xlarge	8	8	NVLink	128	64	488	25Gbps	14Gbps

NEW

F1 instances—First cloud instance with FPGA

Speed up applications over 30x

- Financial computing
- Genomics sequencing
- Engineering simulations
- Image and video processing
- Big data and ML
- Security, compression



Mipsology

NATIONAL INSTRUMENTS

edico genome

Falcon COMPUTING

TERADEEP

MAXELER Technologies
MAXIMUM PERFORMANCE COMPUTING

RYFT
ACTIONABLE INTELLIGENCE FROM COMPLEX DATA

Reconfigure.io

Atomic Rules

Titan IC

NGCODEC

aws re:Invent

aws

World's Record Genomics Processing on F1

edico genome

DRAGEN PIPELINES APPLICATIONS NEWS COMPANY CONTACT US

Oct. 19
2017

Children's Hospital of Philadelphia And Edico Genome Achieve Fastest-Ever Analysis Of 1,000 Genomes

GUINNESS WORLD RECORDS™
RECORD HOLDER

ORLANDO, Fla., Oct. 19, 2017 — The Children's Hospital of Philadelphia (CHOP) and Edico Genome today set a new scientific world standard in rapidly processing whole human genomes into data files useable for researchers aiming to bring precision medicine into mainstream clinical practice. Utilizing Edico Genome's DRAGEN™ Genome Pipeline, deployed on 1,000 Amazon EC2 F1 instances on the Amazon Web Services (AWS) Cloud, 1,000 pediatric genomes were processed in two hours and twenty-five minutes.

Back

EC2 Foundations

Resources

Instances

Storage

Networking

Availability

Regions and AZs

Placement Groups

Load Balancing

Auto Scaling

Management

Deployment

Monitoring

Administration

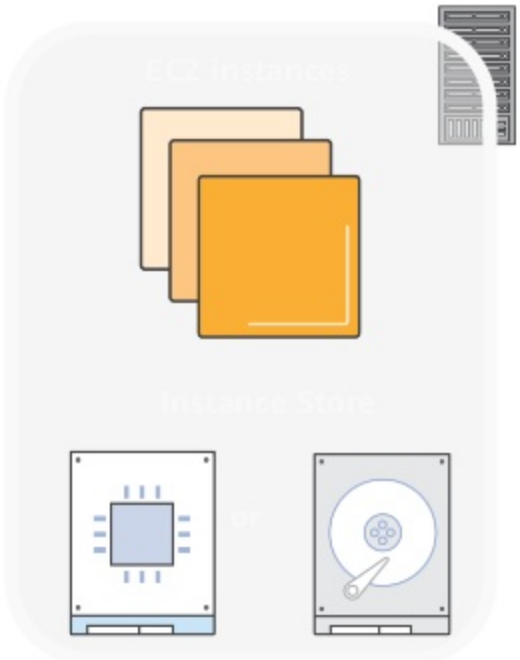
Purchase Options

On Demand

Reserved

Spot

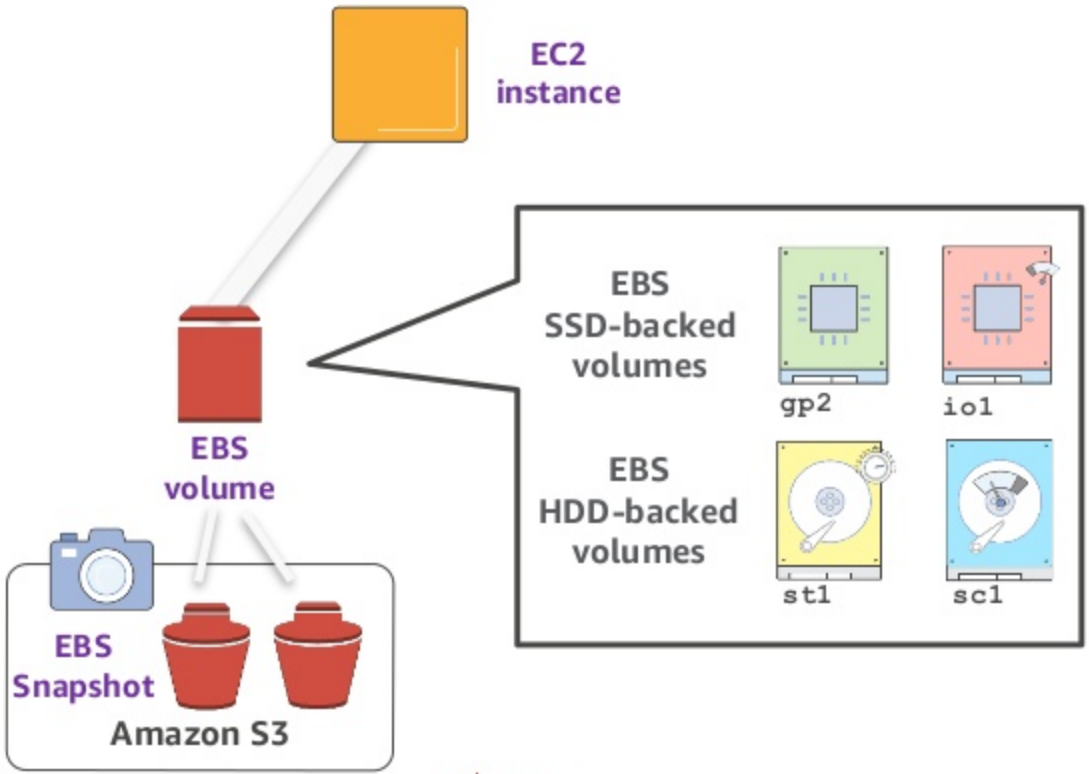
Amazon EC2 instance store



- Local to instance
- Non-persistent data store
- Data not replicated (by default)
- No snapshot support
- SSD or HDD

NEW

Amazon Elastic Block Store (EBS)



- Block storage as a service
- Create, attach volumes through an API
- Service accessed over the network
- Select storage and compute based on your workload
- Volumes persist independent of EC2
- Detach and attach between instances
- Choice of magnetic and SSD-based volume types
- Supports Snapshots: Point-in-time backup of modified volume blocks

NEW! *Elastic Volumes let you increase volume size or change volume type*

AWS re:Invent

aws

EC2 Foundations

Resources

Instances
Storage
Networking

Availability

Regions and AZs
Placement Groups
Load Balancing
Auto Scaling

Management

Deployment
Monitoring
Administration

Purchase Options

On Demand
Reserved
Spot

Amazon Virtual Private Cloud (VPC)

Virtual Private Cloud

Provision a logically isolated cloud where you can launch AWS resources into a virtual network

Security Groups & ACLs

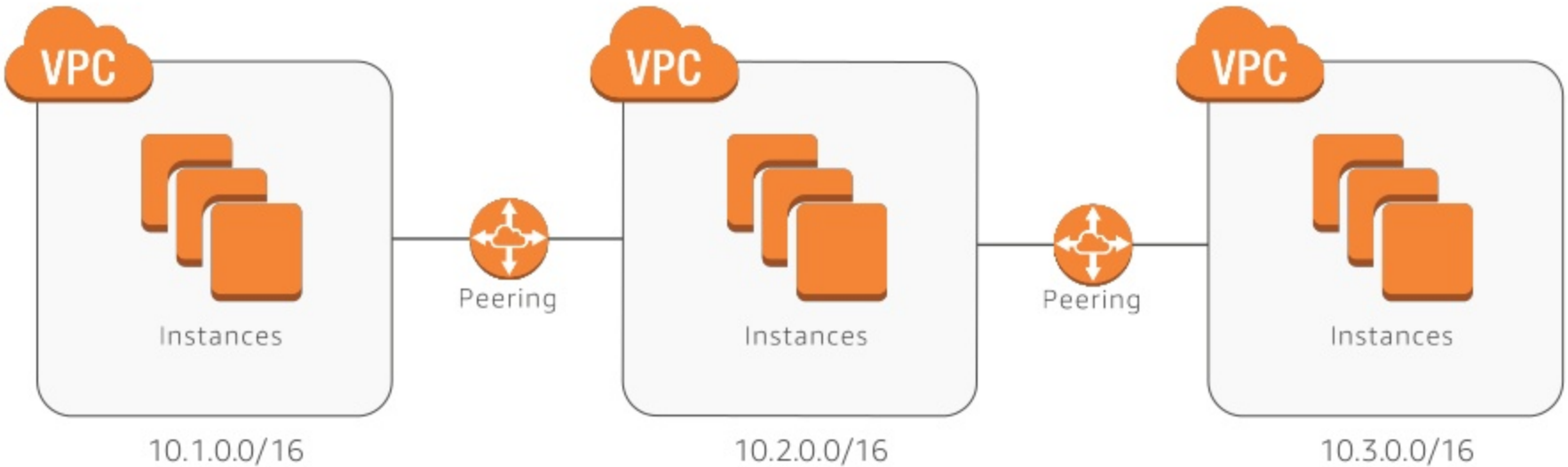
NAT Gateway

Flow Logs

NEW

VPC Peering

A networking connection between two VPCs



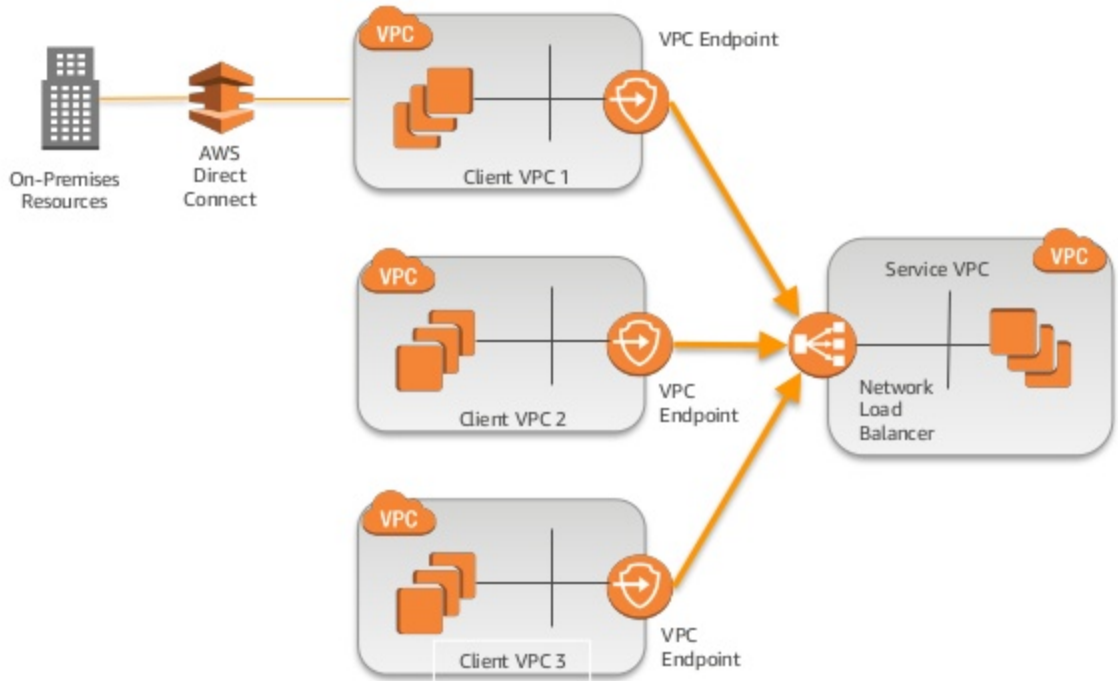
Now available: Inter-region Peering!

NEW

AWS PrivateLink

Share services privately
between VPCs and
on-premises networks

Secure, Scalable, Reliable



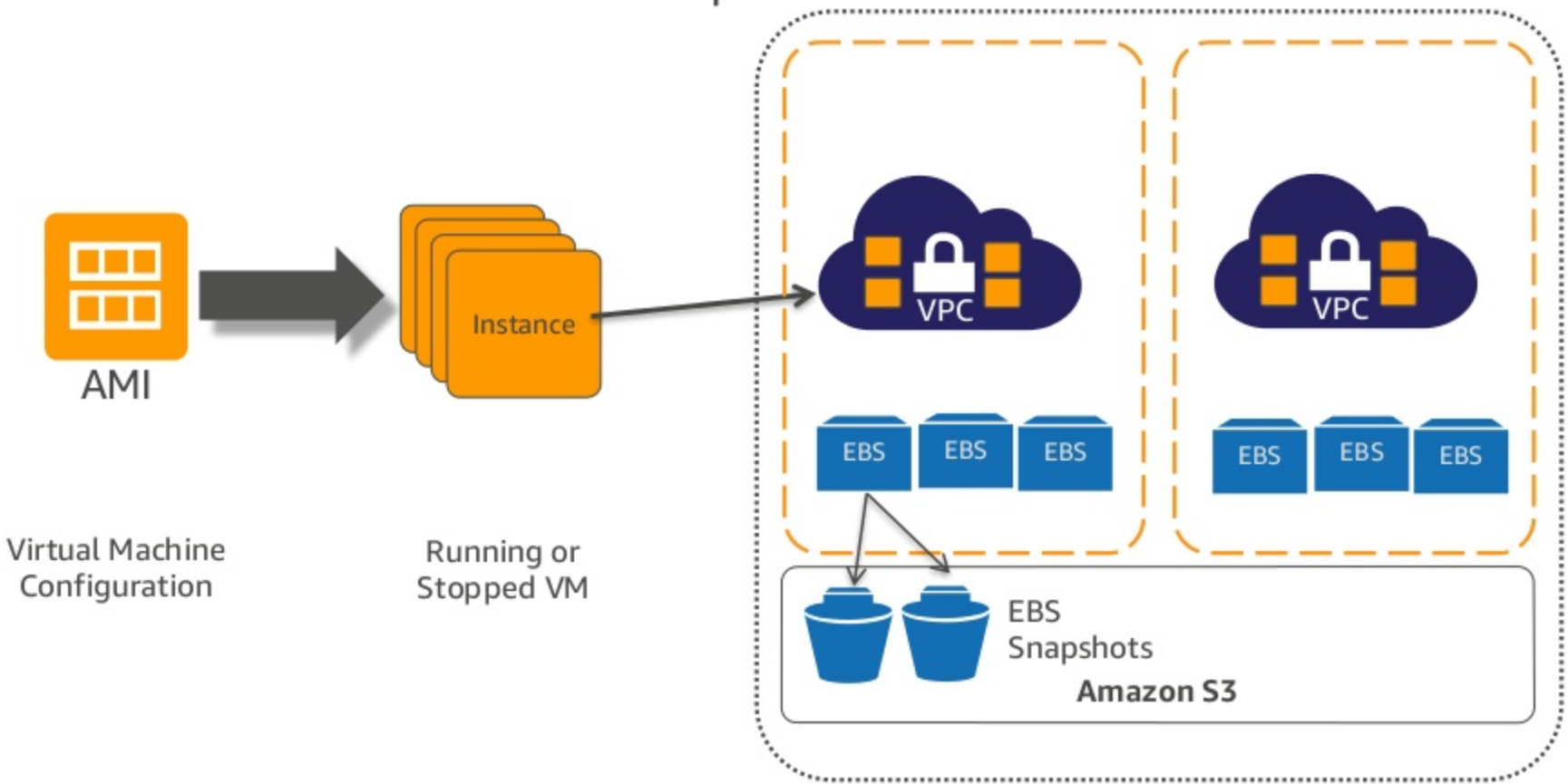
Customers and Partners

ca technologies · Vanguard · twilio · TIBCO · snowflake · SIGOPT · salesforce heroku · APPDYNAMICS · GE · Expedia · dynatrace · CISCO Stealthwatch Cloud · AUTODESK. · aqua · Alfresco

AWS re:Invent

aws

EC2 Resources Recap



EC2 Foundations

Resources

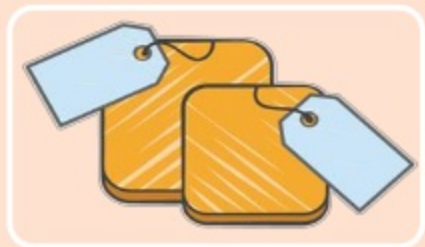
Instances
Storage
Networking

Availability

Regions and AZs
Placement Groups
Load Balancing
Auto Scaling

Management

Deployment
Monitoring
Administration



Purchase Options

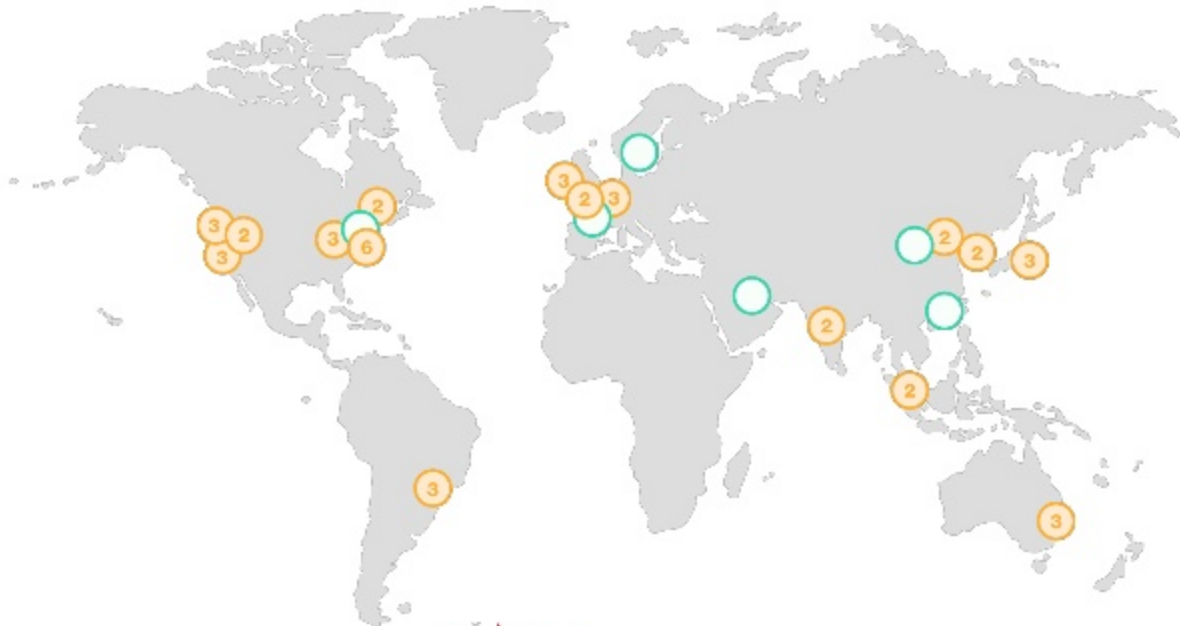
On Demand
Reserved
Spot

NEW

AWS global infrastructure

16 regions

A region is a physical location in the world where we have multiple Availability Zones



44 Availability Zones

Distinct locations that are engineered to be insulated from failures in other Availability Zones

SLA of 99.99% availability NEW!

*The AWS Cloud has announced plans to expand with 17 new Availability Zones in six new geographic Regions: Bahrain, China, France, Hong Kong, Sweden, and a second AWS GovCloud Region in the U.S.

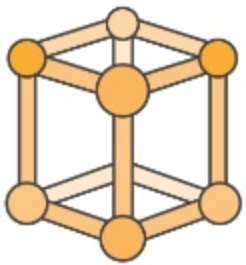
AWS re:Invent

aws

NEW

Placement Groups

Placement Groups enable you to influence our selection of capacity for member instances, optimizing the experience for a workload



CLUSTER

EC2 places instances closely together in order to optimize the performance of inter-instance network traffic

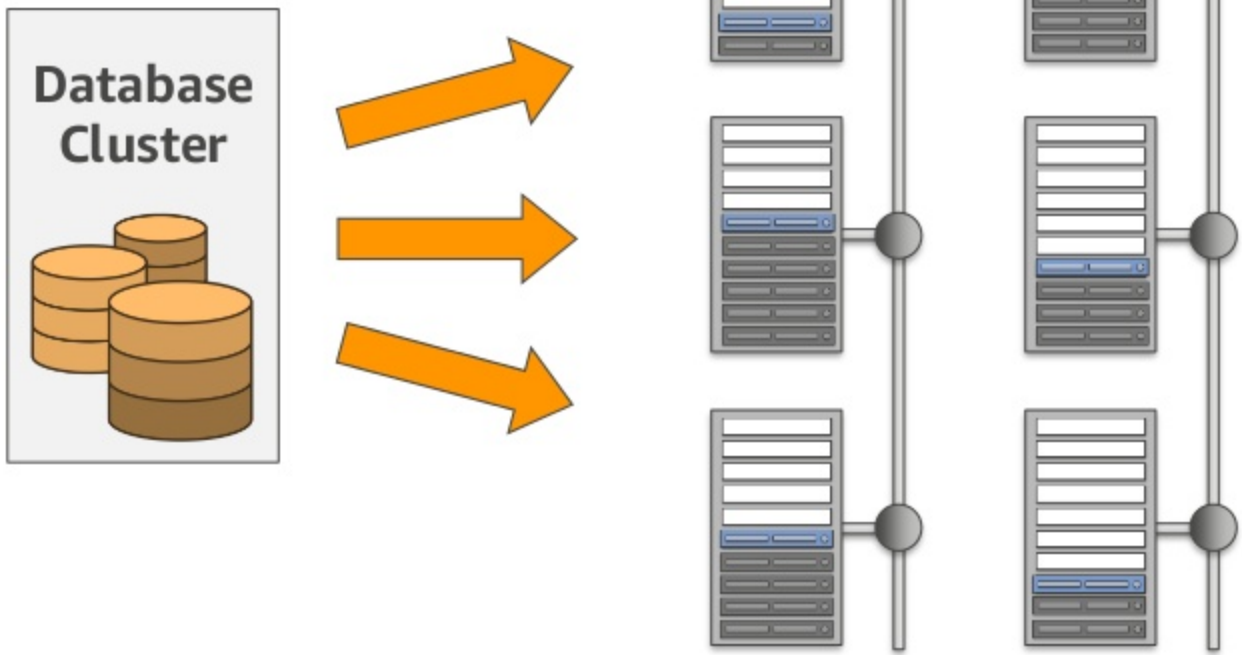
SPREAD

EC2 places instances on distinct hardware in order to help reduce correlated failures

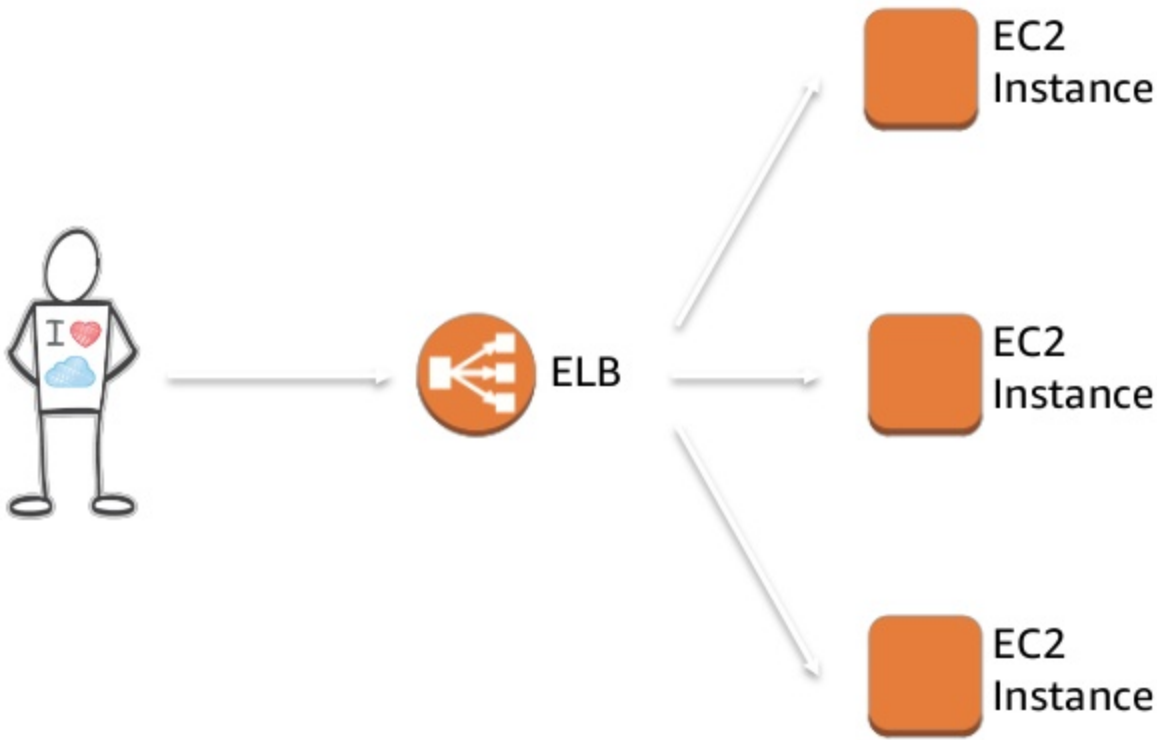
NEW

Spread Placement Groups

When deploying a No SQL database cluster in EC2, Spread Placement will ensure the instances in your cluster are on distinct hardware, helping to insulate a single hardware failure to a single node



Elastic Load Balancing



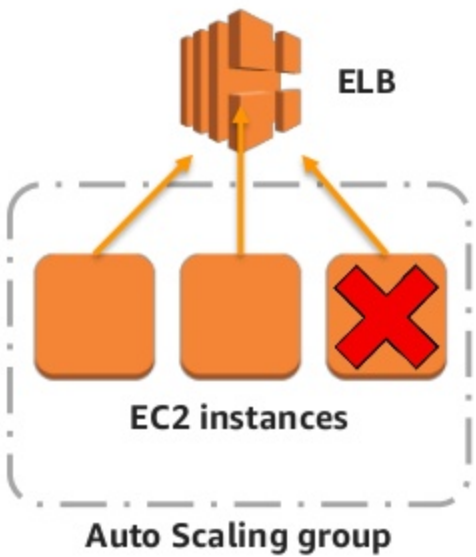
Load balancer used to route incoming requests to multiple EC2 instances, Containers, or IP addresses in your VPC

Elastic Load Balancing provides **high-availability** by utilizing multiple Availability Zones

Auto Scaling

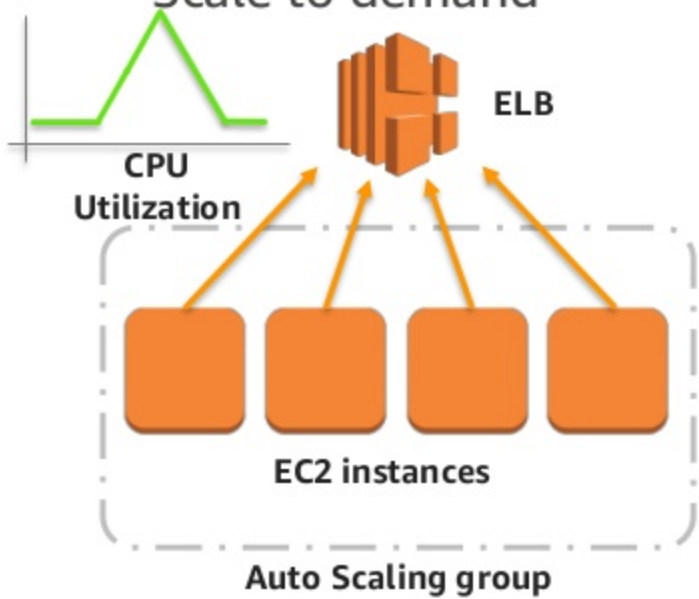
Fleet management

Replace unhealthy instances



Dynamic scaling

Scale to demand



EC2 Foundations

Resources

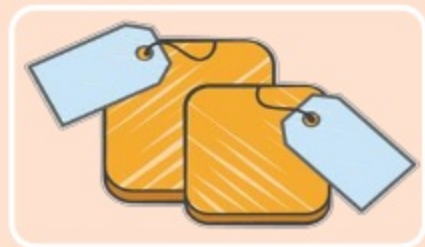
Instances
Storage
Networking

Availability

Regions and AZs
Placement Groups
Load Balancing
Auto Scaling

Management

Deployment
Monitoring
Administration

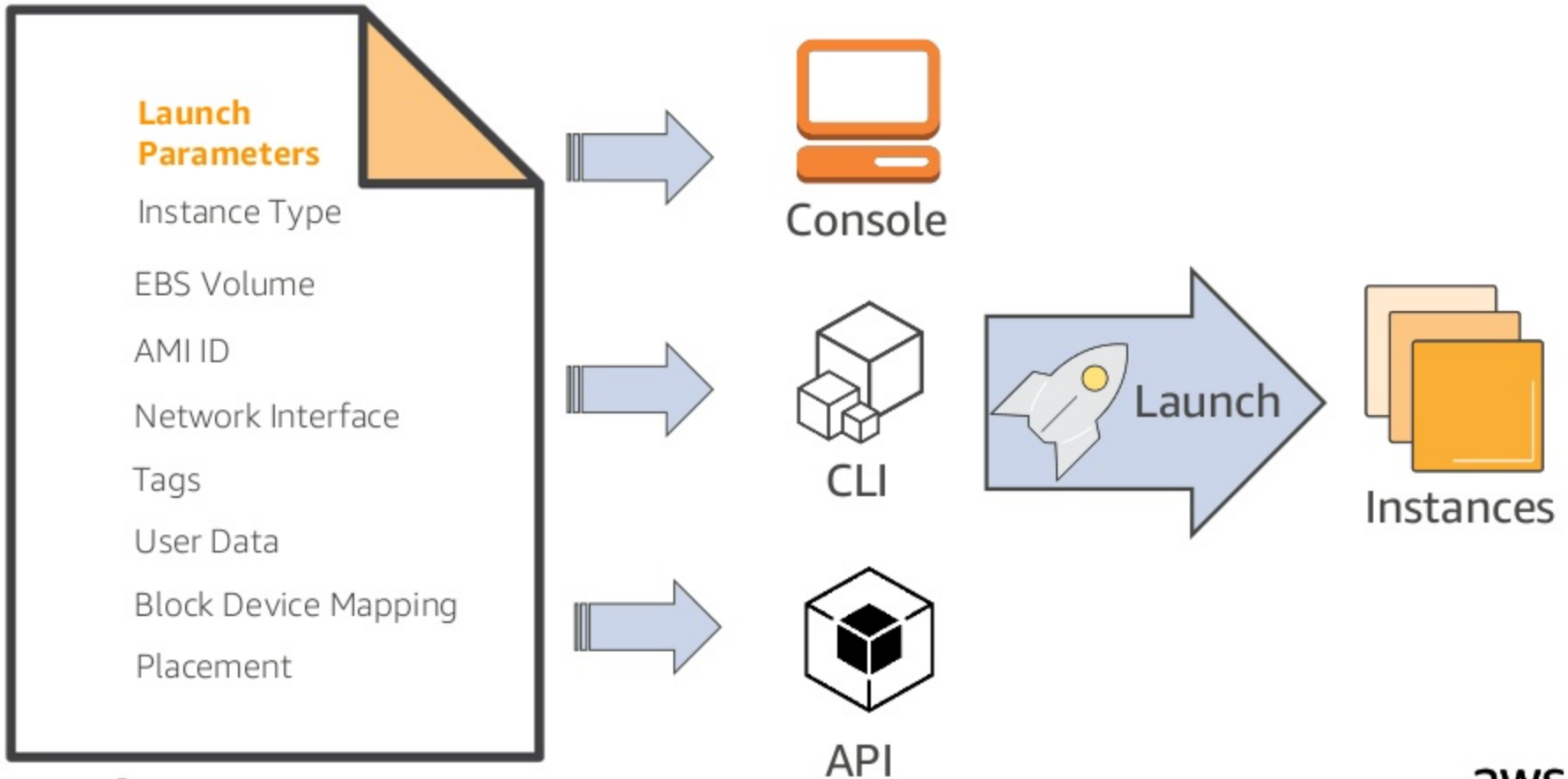


Purchase Options

On Demand
Reserved
Spot

NEW

Launching Instances with Launch Templates



NEW

Launch Templates

Templatize launch requests in order to streamline and simplify future launches in Auto Scaling, Spot Fleet, and On-Demand Instances

CONSISTENT EXPERIENCE

SIMPLE PERMISSIONS

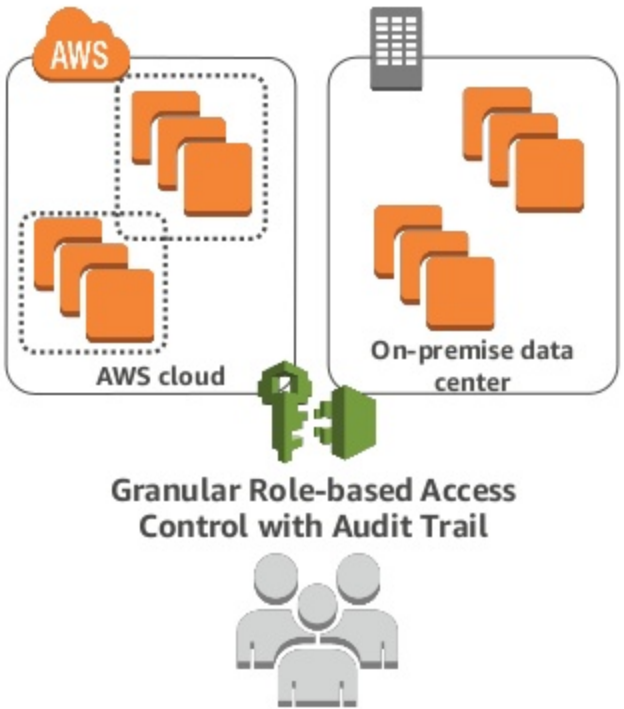
GOVERNANCE & BEST PRACTICES

INCREASE PRODUCTIVITY

Amazon CloudWatch

- Monitoring service for AWS cloud resources and the applications you run on AWS
- You can use Amazon CloudWatch to collect and track metrics, collect and monitor log files, set alarms, and automatically react to changes in your AWS resources

EC2 Systems Manager



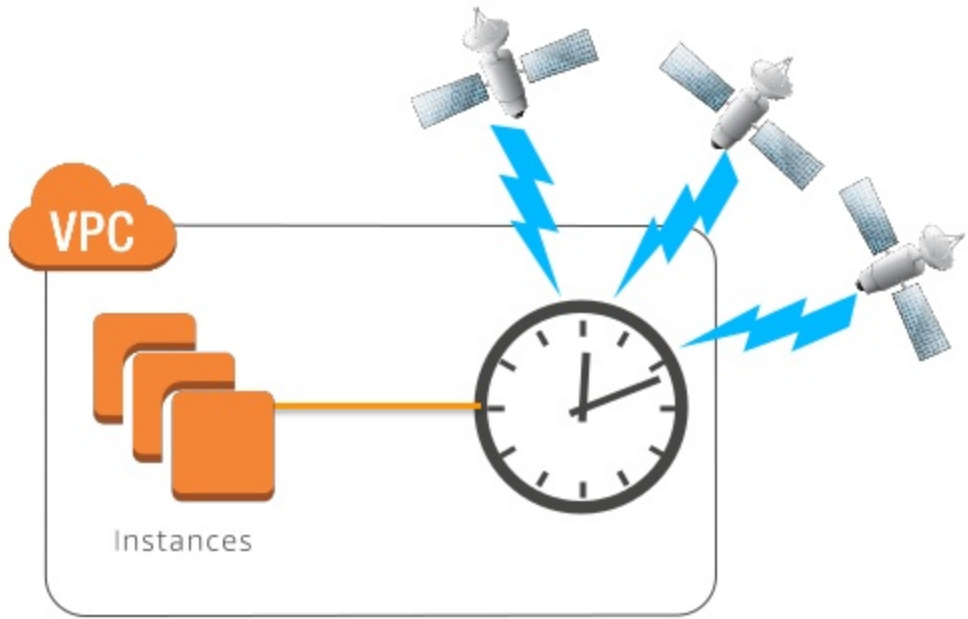
- Securely manage Windows and Linux instances, EC2, or on-premises
- Stay compliant with patching, config drift management, and software inventory
- Automate daily tasks with delegated administration and approval
- Centrally manage secrets and config items

NEW

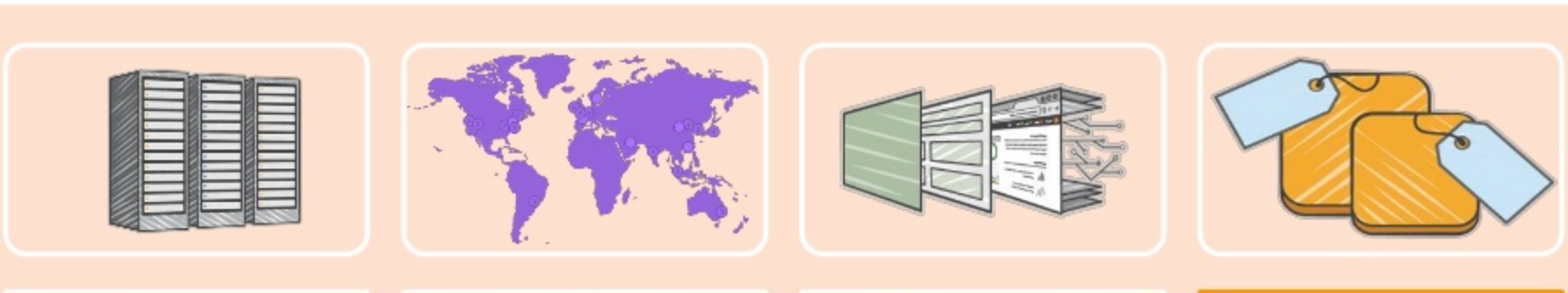
Amazon Time Sync Service

Highly reliable service with redundant array of satellite and atomic clock sources

Available globally!



EC2 Foundations



Resources

Instances
Storage
Networking

Availability

Regions and AZs
Placement Groups
Load Balancing
Auto Scaling

Management

Deployment
Monitoring
Administration

Purchase Options

On Demand
Reserved
Spot

NEW

EC2 Purchasing Options

On-Demand

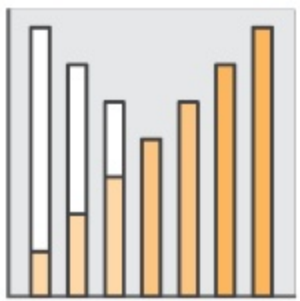
Pay for compute capacity **by the second** with no long-term commitments

Spiky workloads, to define needs

Reserved

Make a 1- or 3-year commitment and receive a **significant discount** off of On-Demand prices

Committed, steady-state usage



Spot

Spare EC2 capacity at a **savings of up to 90%** off of On-Demand prices

Fault-tolerant, dev/test, time-flexible, stateless workloads

NEW! *Per Second Billing for EC2 Linux instances & EBS volumes* NEW!

AWS re:Invent

aws

NEW

EC2 Reserved Pricing

Discount up to 75% off of the On-Demand price

Steady state and committed usage

1- and 3-year terms

Reserve Capacity or opt for flexibility across AZs and instance sizes

Convertible RIs
Change instance family, OS, tenancy, and payment

1-Year Convertible RIs NEW!

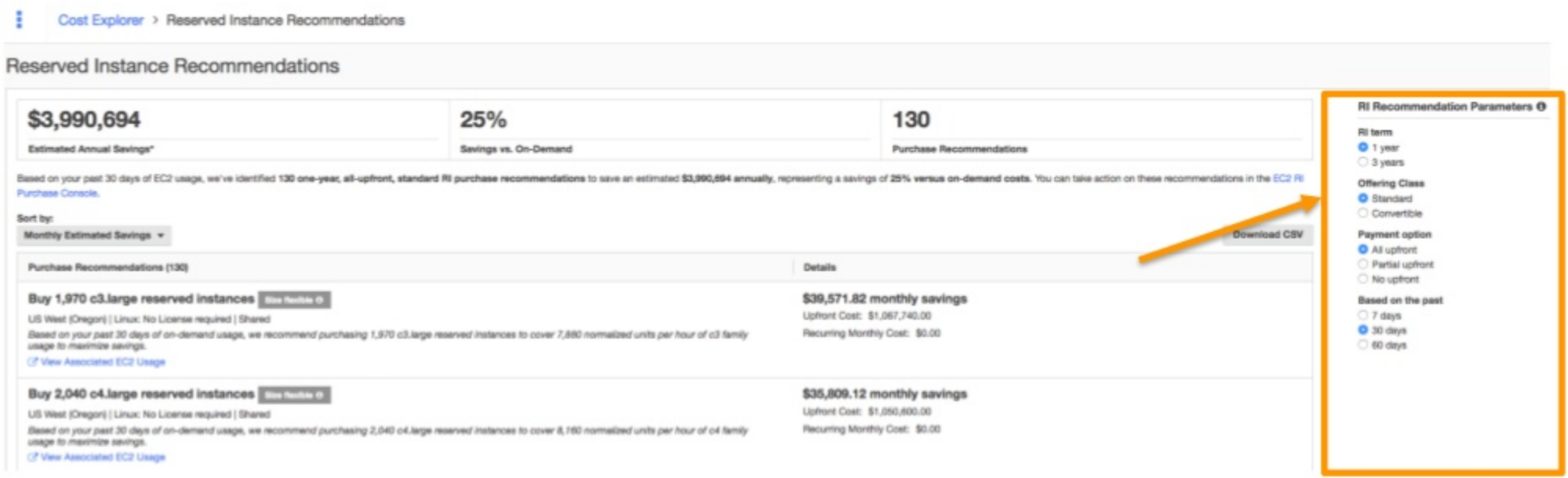
Payment flexibility with 3 upfront payment options *(all, partial, none)*

AWS re:Invent

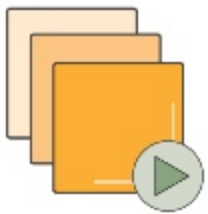
aws

NEW

Reserved Instance Recommendations

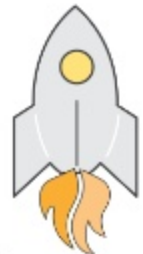


EC2 Spot Pricing



Spare EC2 Capacity that AWS can reclaim with 2-minutes notice

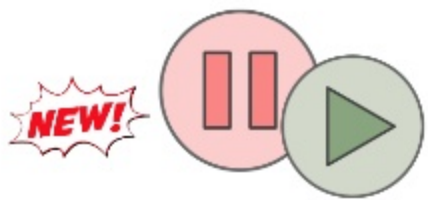
Savings up to 90% off of the On-Demand price



Turbo Boost your results with Spot Fleet

Eliminate the bid!

No need to learn new APIs



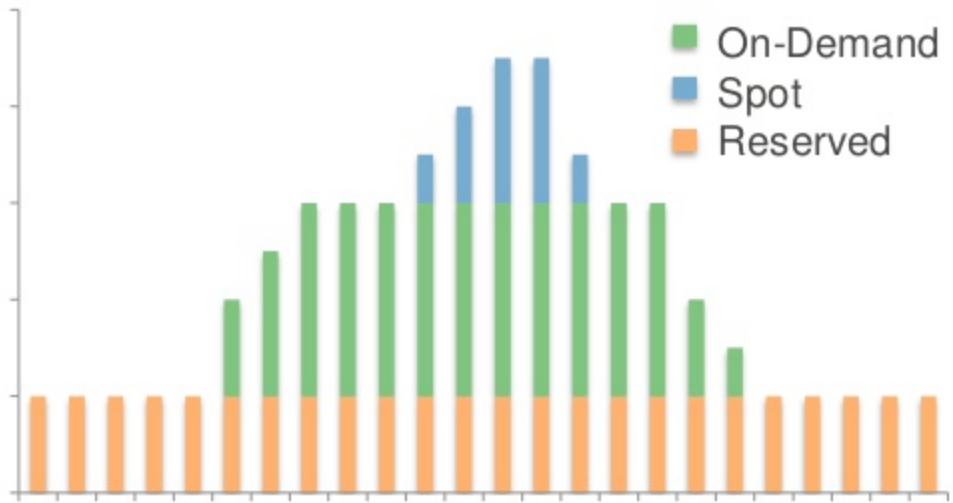
Pause and resume with Stop/Start and Hibernate

To optimize EC2, combine all 3 options

1. Use Reserved Instances for known/steady-state workloads

2. Scale using Spot, On-Demand or both

3. AWS services make this easy and efficient (e.g., Auto Scaling, Spot fleet, ECS, EMR, Thinkbox Deadline, AWS Batch)



EC2 Foundations

Resources

Instances
Storage
Networking

Availability

Regions and AZs
Placement Groups
Load Balancing
Auto Scaling

Management

Deployment
Monitoring
Administration

Purchase Options

On Demand
Reserved
Spot

AWS re:Invent

Thank you!

aws