
aws
INNOVATE
ONLINE CONFERENCE

aws INNOVATE
ONLINE CONFERENCE

分会场三：计算

Amazon Elastic Compute Cloud (Amazon EC2) 高效能计算平台的新发展

冯鹃，AWS 解决方案架构师

今日议程

1、Amazon EC2 家族概况

2、Amazon EC2 新成员和 Nitro 架构

3、 High Performance Computing (HPC) 领域的新进展

4、DEMO – 利用 ParallelCluster 快速创建 HPC 集群

Amazon EC2 家族概況

aws INNOVATE
ONLINE CONFERENCE

Amazon EC2，兼具广度和深度的计算平台

分类

通用

突发

计算密集

内存密集

存储（高I/O）

高密度存储

GPU计算

图形密集

能力

不同种类的处理器
(AWS, Intel, AMD)

高速处理器
(up to 4.0 GHz)

高内存配置
(up to 12 TiB)

实例存储
(HDD and NVMe)

加速计算
(GPUs and FPGA)

高速网络
(up to 100 Gbps)

裸金属

各种规格
(Nano to 32xlarge)

选项

亚马逊 Elastic Block Store

Elastic Graphics

Elastic Inference

175

种实例类型

涵盖几乎所有种类的工作负载和业务需求

种类丰富的计算实例

通用型

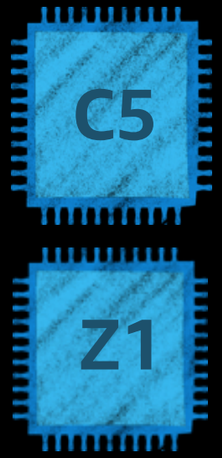
T3

M5

A1

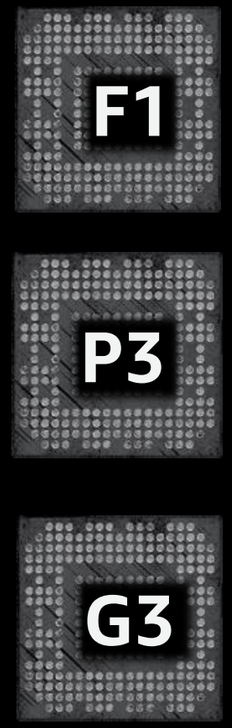
Bare Metal

计算优化型

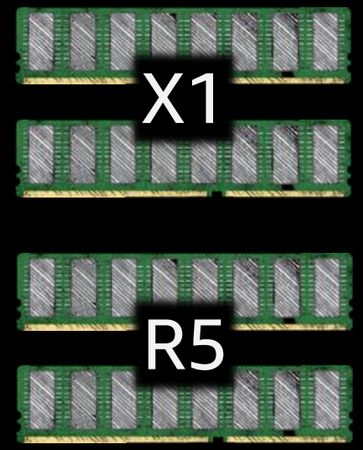


存储和 I/O 优化型

配置
GPU & FPGA



内存优化型



Amazon EC2 实例的命名规则

CPU

内存

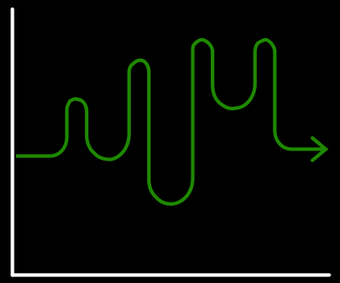
存储

网络性能

Amazon EC2 购买选项

按需实例

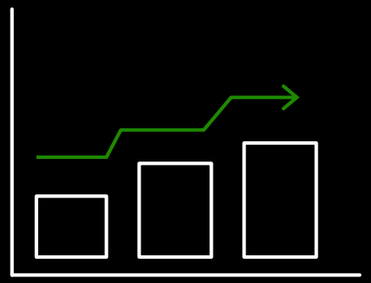
对于计算能力**按秒**进行计费，无需长期使用承诺



适合高峰负载，或者用量不明确的业务需求

预留实例

承诺1年或3年的使用周期以获得相较于按需实例来说**显著的价格优惠**



适合比较稳定的和有承诺的业务用量

Spot 实例

利用空闲的 Amazon EC2 计算能力，比按需实例能节省**最多 90%** 的成本

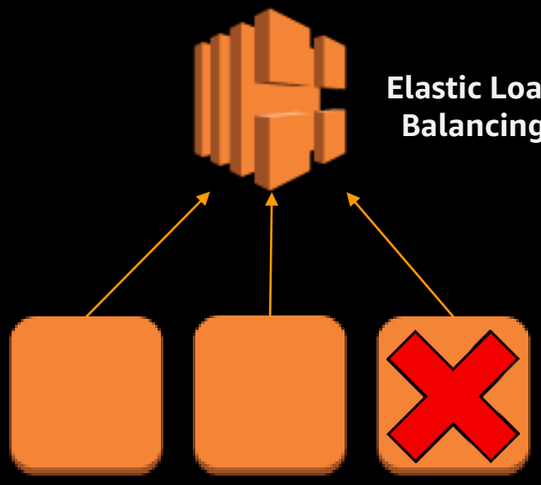
适合能容错的，灵活的和无状态的业务负载

为优化 EC2 使用体验，可将此三种选项结合起来使用！

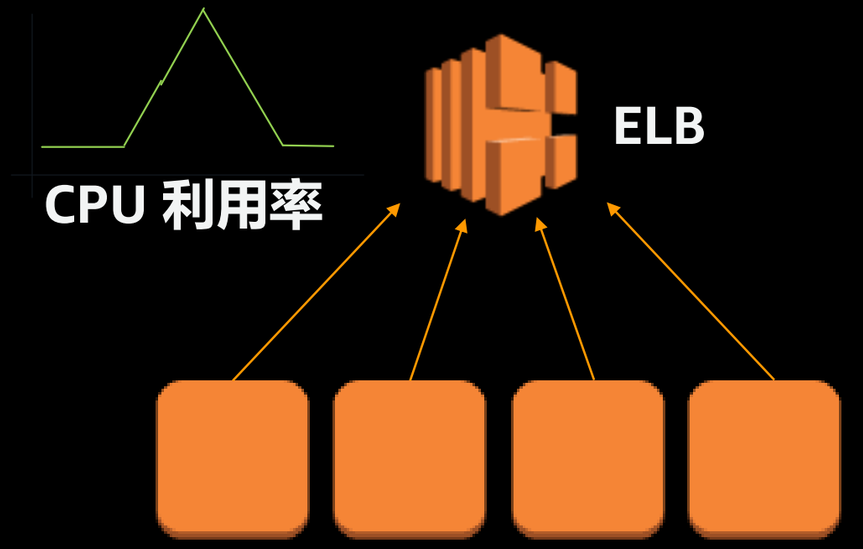
Amazon EC2 Auto Scaling 自动伸缩

自动替换不健康的实例

动态伸缩

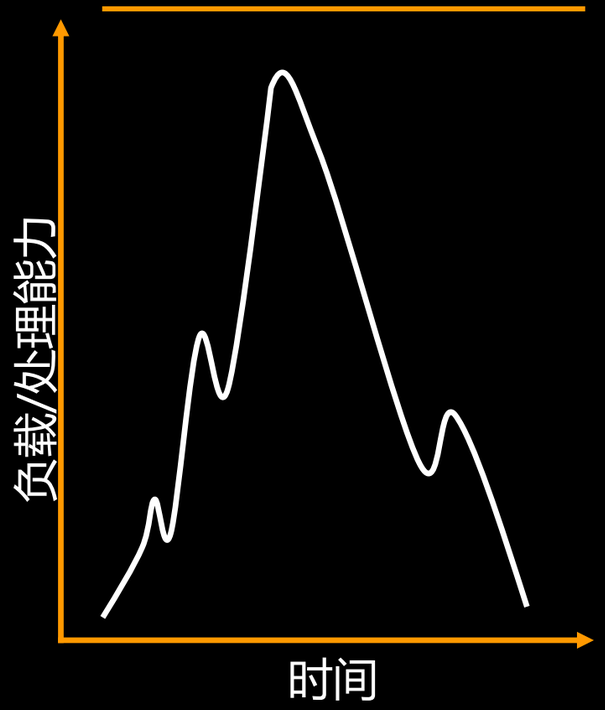


Auto Scaling group 自动伸缩组

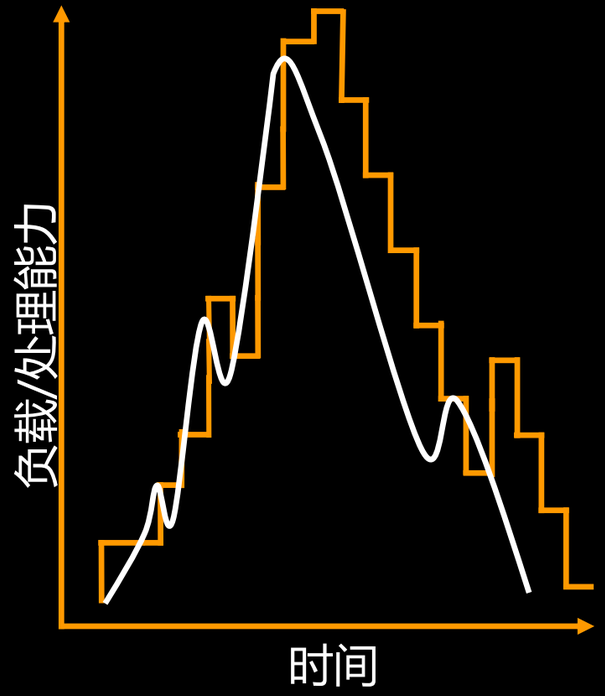


Auto Scaling group 自动伸缩组

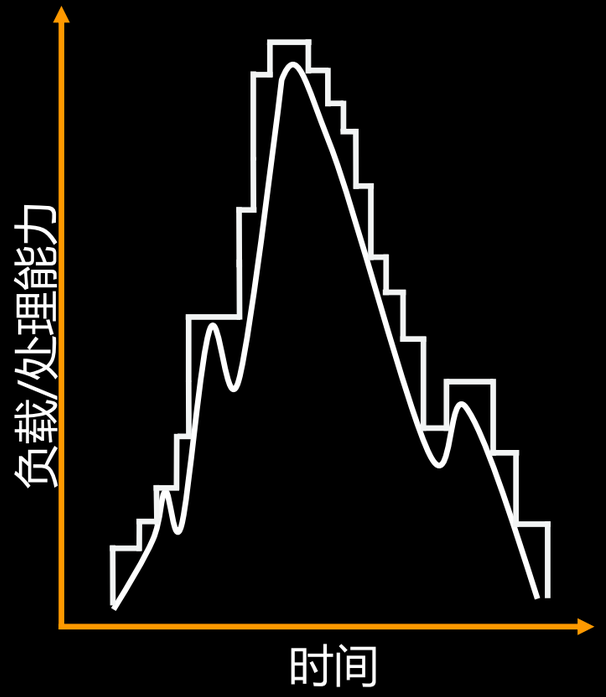
使用预测性伸缩来提前调整计算能力



一般数据中心的容量规划



云上目标跟踪方式的容量规划



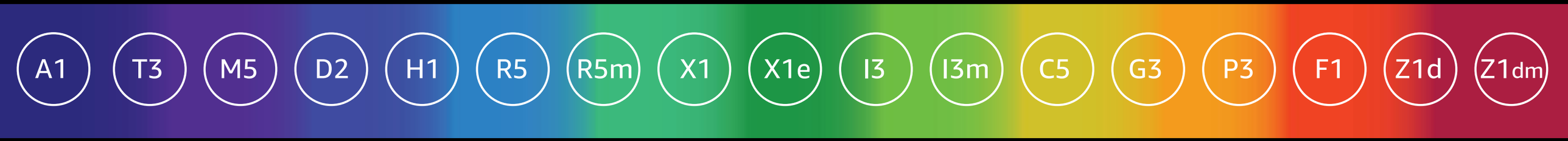
云上预测性伸缩方式的容量规划

✓ **在业务高峰来到前，提前将需要较长启动时间的实例准备好**
✓ **防止由于偶然的低负载导致集群过早的收缩**
✓ **自动化的实现随着使用模式的变化来调整伸缩策略**

Amazon EC2 新成员和 Nitro 架构

aws INNOVATE
ONLINE CONFERENCE

持续创新 – 实例类型



基于 ARM 的应用 | 突发型 | 通用型 | 高密度存储型 | 大数据优化型 | 内存优化型 | 内存计算型 | 内存密集型 | 高 I/O | 裸金属高 I/O 型 | 计算密集型 | 图形密集型 | 通用 GPU 型 | FPGA型 | 计算和内存密集型

Amazon EC2 弹性 GPU

- 用于加强 Amazon EC2 实例的图形加速能力

Amazon EC2 队列

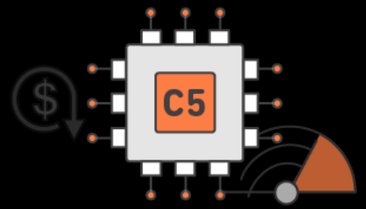
- 规划简单
- 大规模
- 灵活的容量分配

NEW!

Elastic Inference

- 可减少多达 75% 的深度学习推理工作成本
- 与 Amazon EC2 Auto Scaling 配合，根据需要扩展或收缩推理加速能力

Amazon EC2 C5: 基于 Intel Skylake 的计算优化型实例



与 C4 实例相比性价比

最大可提高 49%

采用定制的 3.0 GHz Intel Xeon 扩展型处理器 (Skylake)

多达 72个 vCPUs 及 144 GiB 内存 (内存:vCPU配比为2:1)

25 Gbps 网络带宽

支持 Intel AVX-512

C5d 实例支持本地基于 NVMe 的 SSD 存储

批量作业处理
Batch Processing

分布式分析
Distributed
Analytics

高性能计算
HPC

机器/深度学习推理
Machine/Deep
Learning Inference

广告投放
AD Serving

高度可扩展的多人游戏Highly Scalable Multiplayer Gaming

视频编码
Video Encoding

Amazon EC2 R5: 内存优化型实例

与 R4 实例相比，R5 实例每 GiB 价格降低多达 50%

r5.large	6种大小	r5.24xlarge
16 GiB	● ● ●	768 GiB
2 vCPU		96 vCPU

内存优化型实例提供 8:1 的内存: vCPU 配比

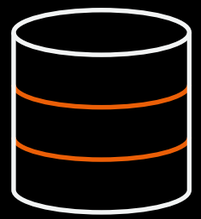
采用2.5 GHz Intel Xeon 扩展型处理器 (Skylake)

多达 25 Gbps 网络带宽

R5d实例包含最大 3.6 TB的本地 NVMe SSD存储

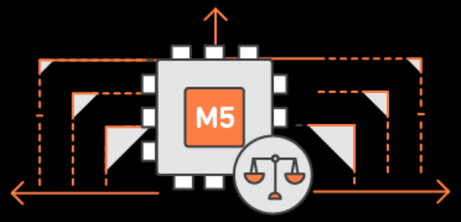
内存缓存

高性能数据库



大数据分析

Amazon EC2 M5: 通用型实例



与 M4 实例相比性价比
最大可提高 47%

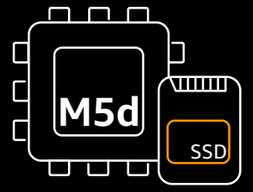
平衡的计算，内存和网络资源配置

采用 2.5 GHz Intel Xeon 扩展型处理器 **(Skylake)**

M5系列中最大的实例 m5.24xlarge 配置了 **96个 vCPUs 及 384 GiB 内存**

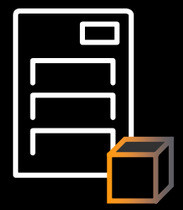
较小的实例也拥有经过提升的网络和EBS性能

支持Intel **AVX-512** ，对于向量和浮点计算负载可以提供最多2倍的性能

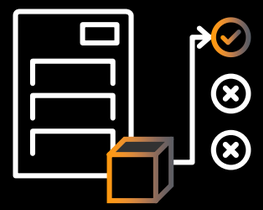


M5d: 可提供高性能的
本地 NVMe SSD 存储

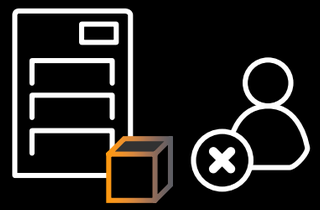
裸金属实例



非虚拟化



具体的管理程序



限制性的，客户许可控制

适用于特定工作

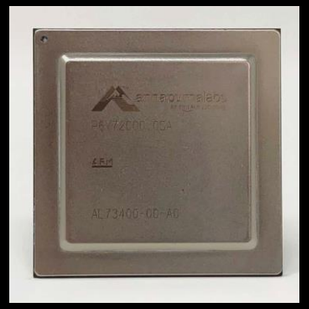
- i3.metal
- u-6tb1.metal
- u9tb1.metal
- u-12tb1.metal
- r5.metal
- r5d.metal
- z1d.metal
- z1d.metal

即将发布！

广泛的可用实例类型

A1: Amazon EC2 中的首款 ARM 实例

NEW!



为横向扩展型应用进行
成本和性能的优化

a1.medium	•••	a1.4xlarge
2 GiB	*5 种实例大小*	32 GiB
1 vCPU		16 vCPU

最高可达 45% 的成本节省

AWS Graviton 处理器配备
基于 ARM 的内核和定制硅片

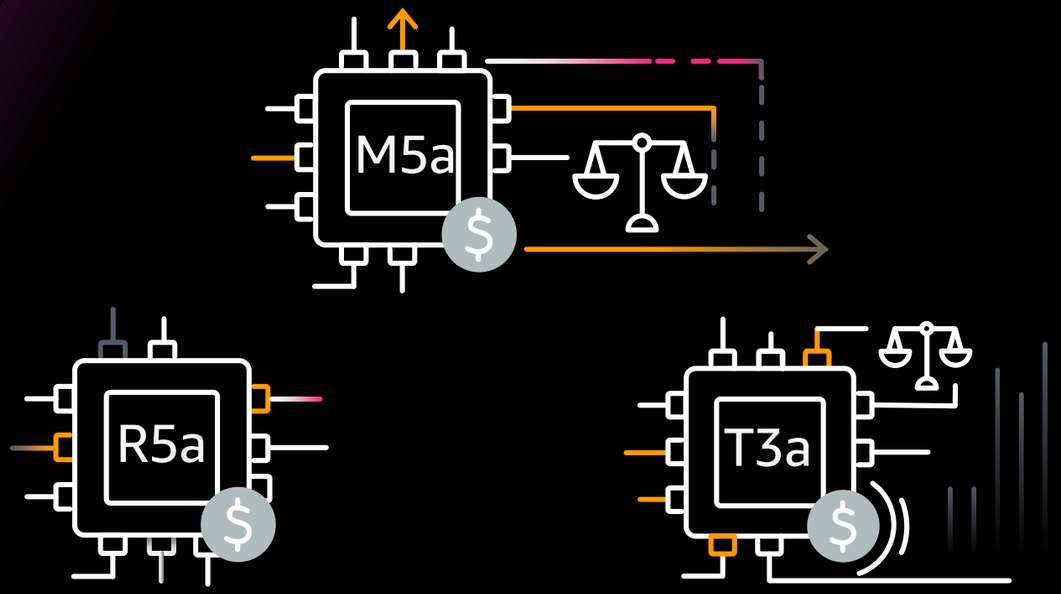
为您的工作负载
提供灵活的选择

更低的成本

采用 AWS Nitro 系统以提供
最大化的资源效率

基于 AMD 的实例

采用了 AMD EPYC 处理器的 R5a, M5a 和 T3a 实例



在AWS通用和内存优化实例上选择处理器

基于AMD的实例价格低10%

大多数应用程序可以在基于AMD的变体上运行，几乎不需要修改

提供更广泛的处理器和架构选择

Intel® Xeon® Scalable (Skylake) 处理器

NVIDIA V100 Tensor Core GPUs

AMD EPYC 处理器

AWS Graviton 处理器

为您的应用和业务负载选择最适合的计算资源

Amazon EC2 宣布已面向 AWS 中国（北京）区域（由光环新网运营）和 AWS 中国（宁夏）区域（由西云数据运营）提供 C5、C5d、R5 和 R5d

发布于: Feb 12, 2019

Amazon EC2 已面向 AWS 中国（北京）区域（由光环新网运营）和 AWS 中国（宁夏）区域（由西云数据运营）提供下一代计算优化 C5 和 内存优化 R5 实例。C5 和 R5 实例提供最新一代 Intel® Xeon Platinum 处理器（之前代号为 Skylake），并且采用了 Nitro 系统，结合了专用硬件和重量更轻的虚拟机器监视器，旨在实现与裸机服务器别无二致的优秀性能。

C5 实例提供了 EC2 产品系列中最佳的价格/计算性能比，并且与 C4 实例相比，价格/性能比提高了 49%。C5 实例非常适合运行计算密集型工作负载，例如批处理、分布式分析、高性能计算 (HPC)、机器/深度学习推理、广告投放、高度可扩展的多人游戏和视频编码。

与 R4 实例相比，R5 实例为每个 vCPU 提供额外 5% 的内存，且每 GiB 价格低 50%。R5 实例非常适用于高性能数据库、分布式内存缓存、内存数据库和大数据分析等应用程序。

同时也向这些区域提供带有本地实例存储（C5d 和 R5d）的 C5 和 R5 实例。

所有 EC2 定价选项（包括按需、预留和竞价实例）均支持 C5 和 R5 实例（按需定价可提供高达 90% 的折扣）

原有的 Amazon EC2 主机

服务器

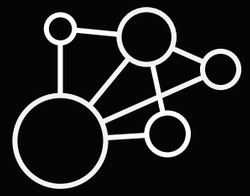
配备 Nitro 系统的 Amazon EC2 主机

服务器

客户实例

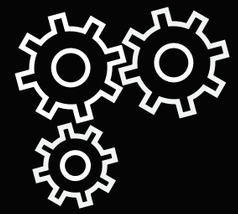
Nitro Hypervisor

NITRO系统



网络

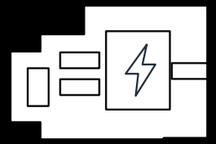
存储



管理，安全和监控

创新的 AWS Nitro 系统

Nitro卡

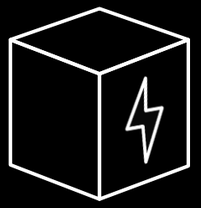


本地 NVMe 存储
EBS 存储
网络，监控和安全

Nitro 安全芯片

与服务器主板集成
保护硬件资源

Nitro Hypervisor



轻量级的 hypervisor
内存和 CPU 分配
近似裸金属的性能

为亚马逊 EC2 实例的快速设计和交付而设计的构件块

提供性能上与裸金属几无差别的虚拟化实例

在亚马逊 EC2 上运行裸金属业务负载，且仍具备 AWS 所有的弹性，安全性和扩展等特性

HPC 领域的新进展

aws INNOVATE
ONLINE CONFERENCE

为什么要在 AWS 上运行 HPC（High Performance Computing）

提供在传统数据中心里无法获得的
近似于无限的 可扩展性和灵活性

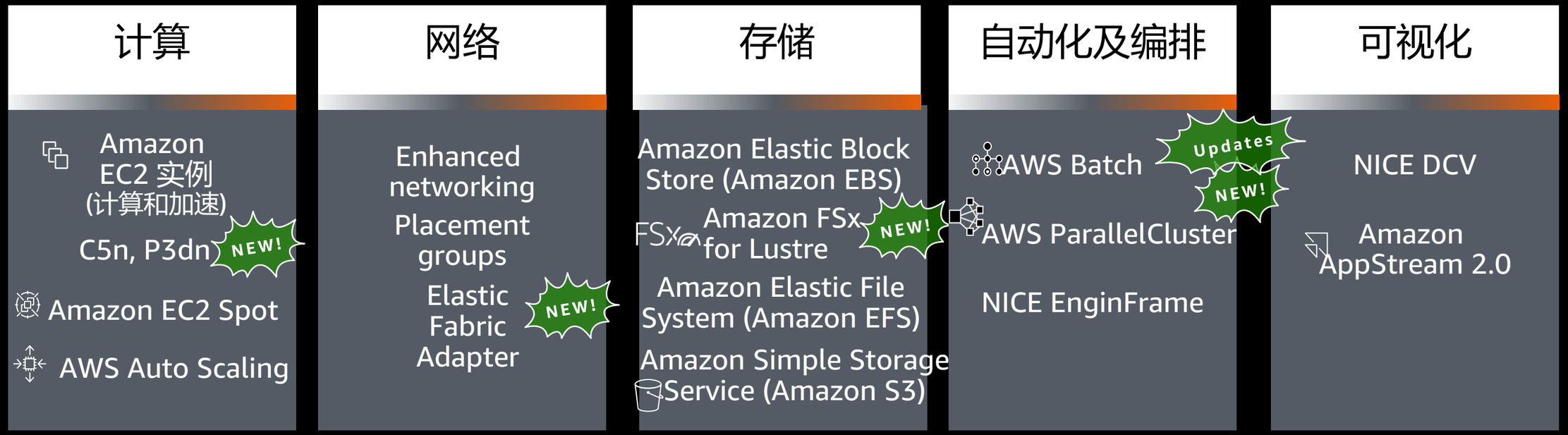
更好的 ROI

全球范围内安全的访问集群，
以此**广泛的提升协作**

弹性的配置选择和快速可重复的
资源配置方案，并确保**成本的优化**

更快的获得
业务成效

AWS High Performance Compute (HPC) 解决方案组件



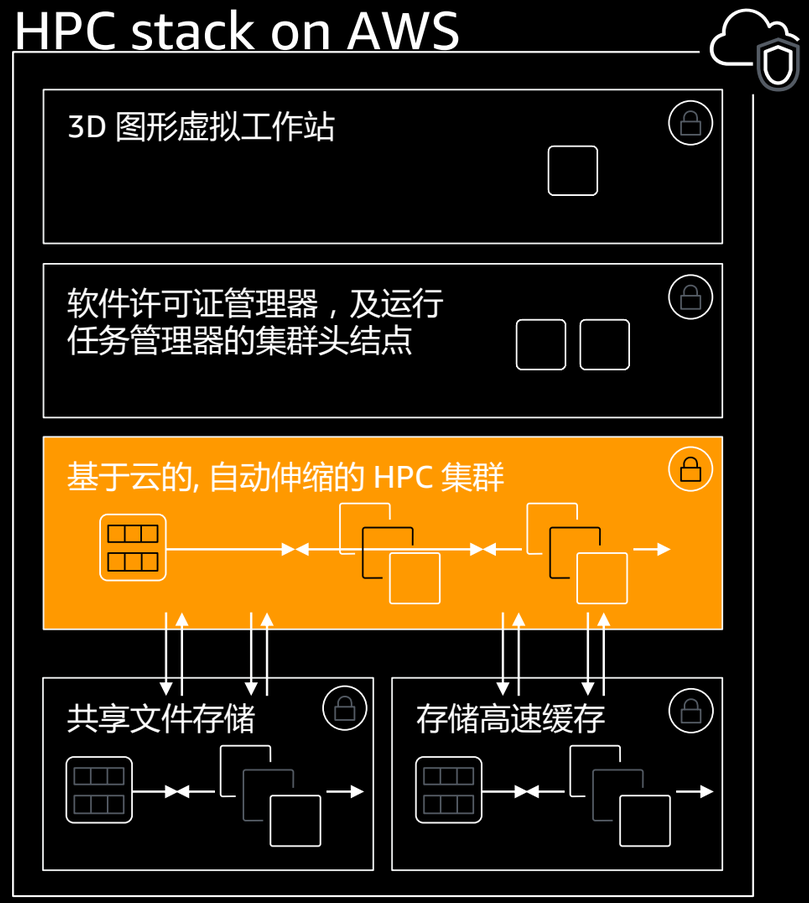
HPC 基础架构的创新

高时钟主频计算实例：Z1d

Z1d 专为内存密集以及计算密集的应用进行优化

- 定制的 Intel Xeon 扩展型处理器
- 最高可提供持续的 4 GHz 时钟主频, 全加速性能
- 最大可提供 385GiB DDR4 内存
- 增强型网络，最大可提供 25 GB 吞吐

特点　采用Intel Xeon 扩展型 (Skylake) 高主频处理器



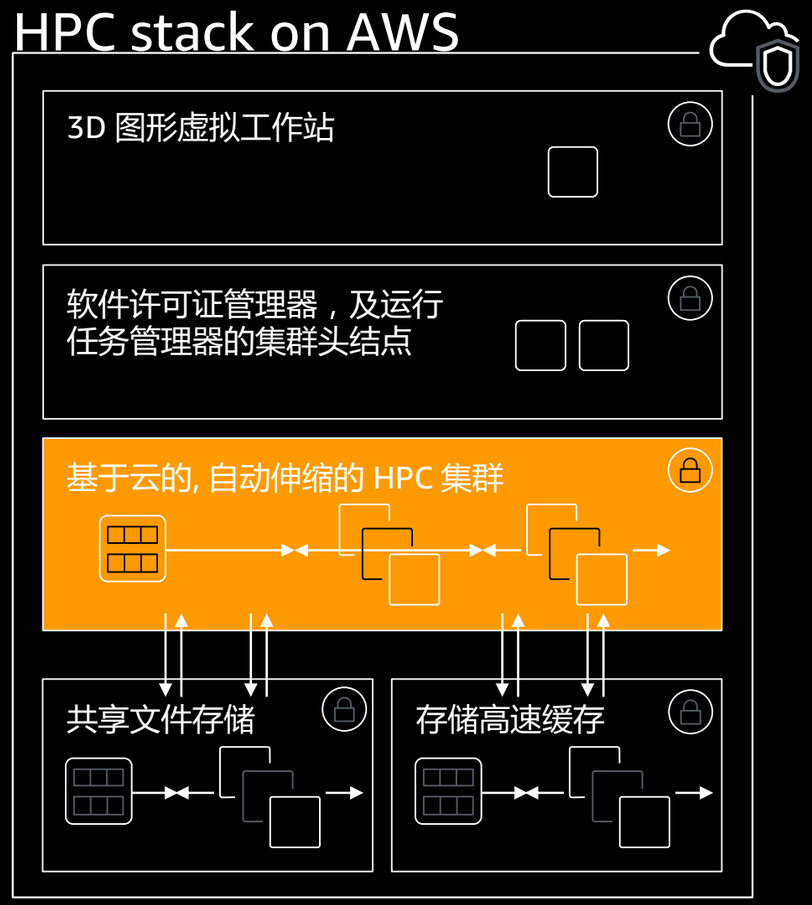
HPC 基础架构的创新

高带宽计算实例：C5n

高度可伸缩的性能

- C5n 可提供高达 100 Gbps 的网络带宽
- 极大的提升最大带宽，PPS，和 packet 处理能力
- 特定设计的 Nitro 网络卡
- 专为网络密集型负载设计，适用场景包括分布式集群和数据库负载，HPC，实时通信和视频流

特点　采用Intel Xeon 扩展型 (Skylake) 处理器



HPC 基础架构的创新

适合 HPC 应用的高性能网络互联组件：EFA

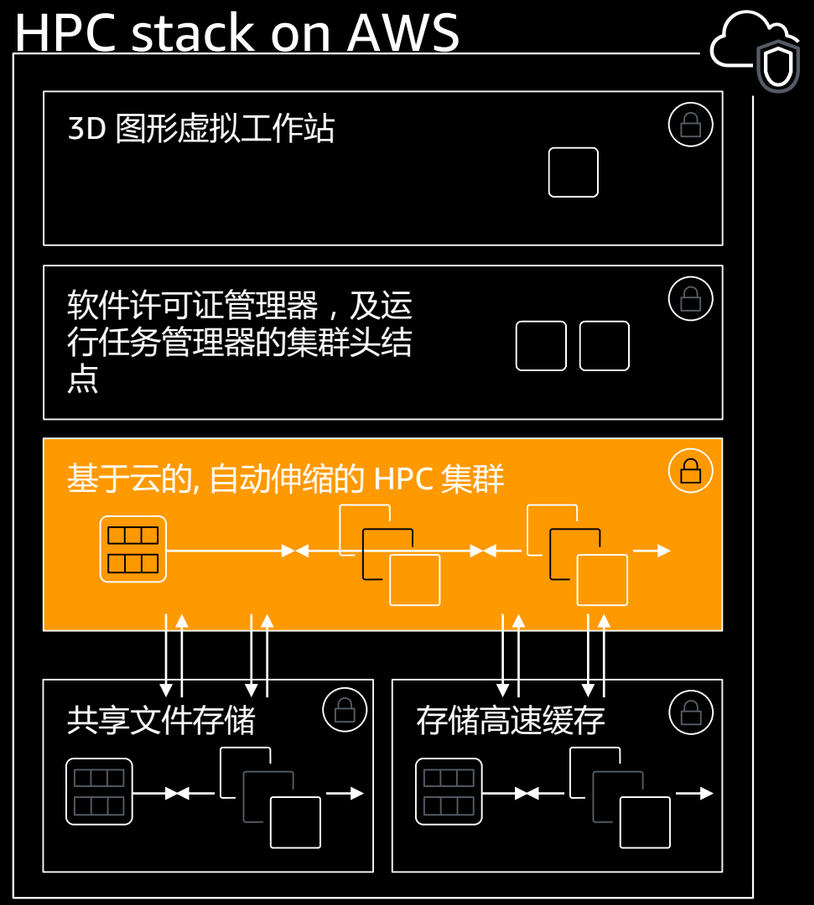
使 AWS 上基于 MPI 的应用更具可扩展性

- 提升应用性能
- 使 HPC 应用得以更好的利用 AWS 云的灵活性和弹性
- 支持基于工业标准 MPI 的 HPC 设施和应用向 AWS 迁移而无需修改代码

NEW!

EFA

Elastic Fabric Adapter



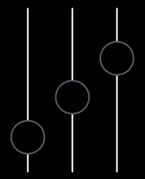
HPC 基础架构的创新

全托管的高性能共享文件系统：Amazon FSx for Lustre

高度可伸缩的性能

- 100+ GiB/s 的吞吐能力
- 数百万 IOPS
- 可持续的低时延

高性能



并行分布式
文件系统

性能参数优化

HPC 基础架构的创新

易于使用的集群管理工具：AWS ParallelCluster

创建和管理 HPC 集群

- 简化云上的HPC 部署，已与流行的 HPC schedulers 进行集成
- 基于 AWS CloudFormation, 易于针对特定应用或项目需要进行调整
- 现已于 AWS Batch 集成

DEMO – 利用 ParallelCluster 快速创建 HPC 集群

aws INNOVATE
ONLINE CONFERENCE

议程回顾

1、Amazon EC2 家族概况 – 175种，通用型，计算优化型，GPU，FPGA ……

2、Amazon EC2 新成员和 Nitro 架构 – C5，R5，A1，AMD-based，Nitro

3、HPC 领域的新进展 – Z1d，C5n，EFA，FSx for Lustre，ParallelCluster

4、DEMO – 利用 ParallelCluster 快速创建 HPC 集群

感谢参加 AWS INNOVATE 2019 在线技术大会

我们希望您在这里找到感兴趣的内容！

也请帮助我们完成**投票打分**和**反馈问卷**。

欲获取关于 AWS 的更多信息和技术内容，可以通过以下方式找到我们：


微信公众号：AWSChina


新浪微博：https://www.weibo.com/amazonaws/


领英：https://www.linkedin.com/company/aws-china/


知乎：https://www.zhihu.com/org/aws-54/activities/


视频中心：http://aws.amazon.bokecc.com/


更多线上活动：https://aws.amazon.com/cn/about-aws/events/webinar/

aws INNOVATE
ONLINE CONFERENCE