成本优化支柱

AWS 架构完善的框架

2020 年 7 月

声明

客户负责对本文档中的信息进行独立评估。本文档：(a) 仅供参考，(b) 代表 AWS 当前的产品服务和实践，如有变更，恕不另行通知，以及 (c) 不构成 AWS 及其附属公司、供应商或授权商的任何承诺或保证。AWS 产品或服务均“按原样”提供，没有任何明示或暗示的担保、声明或条件。AWS 对其客户的责任和义务由 AWS 协议决定，本文档与 AWS 和客户之间签订的任何协议无关，亦不影响任何此类协议。

目录

摘要

本白皮书重点介绍 Amazon Web Services (AWS) 架构完善的框架的成本优化支柱。文中提供了指导，可帮助客户在 AWS 环境的设计、交付和维护过程中应用最佳实践。

经过成本优化的工作负载能够充分利用所有资源，以尽可能低的价格实现成果，并满足您的功能要求。本白皮书提供了有关在组织内增强能力、设计工作负载、选择服务、配置和运营服务以及应用成本优化技术的深入指导。

引言

AWS 架构完善的框架能够帮助您理解您在 AWS 上构建工作负载时所做的决策。此框架提供了在云中设计和运行可靠、安全、高效且经济实惠的工作负载的架构最佳实践。它提供了一种统一的方法，使您能够根据最佳实践衡量架构，并确定需要改进的方面。我们相信，拥有架构完善的工作负载能够大大提高实现业务成功的可能性。

该框架基于五大支柱：

- 卓越运营
- 安全性
- 可靠性
- 性能效率
- 成本优化

本白皮书重点介绍成本优化支柱以及如何以最高效的方式使用服务和资源来构建工作负载，从而以最低的价格实现业务成果。

您将了解如何在组织内应用成本优化支柱的最佳实践。对传统的本地解决方案进行成本优化可能比较困难，因为您必须要预测未来容量和业务需求，同时要掌控复杂的采购流程。采用本白皮书中的实践可帮助您的组织实现以下目标：

- 践行云财务管理
- 支出和使用量意识
- 资源成本效益
- 管理需求和供应资源
- 随着时间的推移不断优化

本白皮书面向的是技术和财务人员，例如首席技术官 (CTO)，首席财务官 (CFO)、架构师、开发人员、财务总监、财务规划师、业务分析师和运维团队成员。本白皮书不提供实施细节或架构模式，但提及了相关的适当资源。

成本优化

成本优化是在工作负载的整个生命周期中不断完善和改进的过程。本白皮书中的实践有助于构建和运营对成本敏感的工作负载，这些工作负载可在帮助实现业务成果的同时，最大限度降低成本并助力组织最大限度提高其投资回报。

设计原则

请考虑以下成本优化设计原则：

实施云财务管理：要获得财务上的成功并加速在云中实现业务价值，必须投资云财务管理。您的组织必须投入必要的时间和资源增强自身在这个新的技术和使用管理领域中的能力。与安全或运营能力类似，您的组织需要通过知识积累、计划、资源和流程来增强自身能力，从而成为一家具有成本效益的组织。

采用消费模型：仅为所用的计算资源付费，并可根据业务需求增加或减少使用量。例如，开发和测试环境通常只需要在每个工作日运行八个小时。您可以在不需要的时候停用这些资源，这样有可能节省 75% 的成本（40 小时对比 168 小时）。

衡量整体效率：衡量工作负载的业务产出及其交付成本。使用此数据了解您通过提高产出、增加功能和降低成本获得的收益。

不再把钱花在千篇一律的繁重工作上：AWS 会帮您料理繁重的数据中心运营工作，如安装、堆叠和驱动服务器。它还消除了使用托管服务管理操作系统和应用程序的运营负担。因此，您可以将精力集中在客户和业务项目而非 IT 基础设施上。

分析并划分支出属性：使用云，您可以更轻松地准确了解工作负载的成本和使用量，从而将 IT 成本透明地归属到收入来源和各个工作负载拥有者。这有助于衡量投资回报率 (ROI)，并让工作负载拥有者能够据此优化资源和降低成本。

定义

云中的成本优化包括五个重点方面：

- 践行云财务管理
- 支出和使用量意识

- 资源成本效益
- 管理需求和供应资源
- 随着时间的推移不断优化

与架构完善的框架中的其他支柱类似，考虑成本优化时也需要权衡取舍。例如，是要优化上市速度还是优化成本？在某些情况下，最好优化上市速度以便快速上市、交付新功能或按时完成任务，而不是优化预付成本。

设计决策有时是仓促作出的，并未结合数据进行考虑，而且人们往往会过度投入，而不是花时间进行基准测试以确定最合算的部署。过度补偿可能会导致部署过度预置且优化不足。但是，如果您必须将资源从本地环境“直接迁移”到云，然后再进行优化，这可能也算合理之举。

通过预先在成本优化策略中投入适量的精力，您可以确保始终如一地遵守最佳实践，避免不必要的过度预置，从而更轻松地实现云的经济优势。以下部分介绍了一些技巧和最佳实践，可帮助您开始并持续实施工作负载的云财务管理和成本优化。

践行云财务管理

借助云财务管理 (CFM)，企业可以在 AWS 上优化成本和使用量并进行扩展，从而实现业务价值和财务成功。

以下是云财务管理的最佳实践：

- 职能所有权
- 财务与技术人员合作
- 云预算和预测
- 对成本敏感的流程
- 对成本敏感的文化
- 量化通过成本优化实现的业务价值

职能所有权

建立成本优化部门：此部门负责营造和维续对成本敏感的文化。它可以是一个现有的个人、组织内的一个团队，也可以是整个组织中由关键财务、技术和组织利益相关者组成的新团队。

此部门（个人或团队）会排定成本管理和成本优化活动的优先级，并根据需要为这些活动投入一定比率的时间。相对于较大型企业中的全职部门，小型组织的这一部门在此方面花费的时间可能更少。

此部门需要采取多学科方法，并具备项目管理、数据科学、财务分析和软件/基础设施开发的能力。他们可以执行成本优化（集中式方法）、影响技术团队执行优化（分散式）或将两者相结合（混合式），从而提高工作负载的效率。可以对照成本优化目标（例如工作负载效率指标）来衡量此部门的执行和交付能力。

必须确保此部门获得高级管理层的支持。支持者即低成本云消费理念的倡导者，他们会为此部门提供升级支持，确保按组织确定的优先级开展成本优化活动。此部门及其支持者会共同确保组织在有效利用云资源并继续创造业务价值。

财务与技术人员合作

在财务与技术人员之间建立合作关系：由于缩短了审批、采购和基础设施部署周期，技术团队在云端的创新速度更快。这可能是对财务组织的一种调整，以前，他们习惯于执行耗时的资源密集型流程，以便在数据中心和本地环境中获取和部署资金，并且只在项目批准时进行成本分配。

在关键的财务和技术利益相关者之间建立合作关系，让他们就组织目标达成共识，并开发在云计算的可变支出模型中获得财务成功的机制。组织内的相关团队必须在云之旅的各个阶段参与成本和使用量讨论，包括：

- **财务领导：**首席财务官、财务总监、财务规划师、业务分析师、采购、供应商开发人员和应付账款负责人必须了解消费、采购选项和每月开票流程的云模型。由于云运营（如使用量的变化速率、即付即用的定价、分级定价、定价模型以及详细的计费和使用信息）与本地运营之间存在根本差异，财务组织必须了解云的使用对业务方面的影响，包括采购流程、激励跟踪、成本分配和财务报表。

- **技术领导：**技术领导（包括产品和应用程序拥有者）必须了解财务要求（如预算限制）和业务要求（如服务水平协议），如此才能实施工作负载并实现组织的预期目标。

财务与技术人员的合作可带来以下好处：

- 财务和技术团队几乎可以实时看到成本和使用量。
- 财务和技术团队建立了标准的操作程序来处理云支出差异。
- 在如何使用资本购买承诺折扣（例如预留实例或 AWS Savings Plans）以及如何使用云来发展组织方面，财务利益相关者充当战略顾问。
- 将现有的应付账款和采购流程用于云部署。
- 财务和技术团队协作预测未来的 AWS 成本和使用量，以调整/建立组织预算。
- 通过共通的语言以及对财务概念的一致理解，更好地进行跨组织沟通。

组织中应参与成本和使用量讨论的其他利益相关者包括：

- **业务部门拥有者：**业务部门拥有者必须了解云业务模式，从而为业务部门和整个公司提供指导。在需要预测增长和工作负载使用情况，以及评估不同购买选项（例如预留实例或 Savings Plans）时，这方面的云知识至关重要。
- **第三方：**如果您的组织使用第三方（例如顾问或工具），请确保他们与您的财务目标一致，并且可以通过参与模式和投资回报 (ROI) 证明一致性。第三方通常会帮助报告和分析其管理的任何工作负载，并且提供他们设计的任何工作负载的成本分析。

云预算和预测

建立云预算和预测：客户使用云来提高效率、速度和敏捷性，这导致成本和使用量的变化速度极快。随着工作负载效率的提高或者新工作负载和功能的部署，成本可以降低。或者，工作负载将扩展以服务更多的客户，这会增加云的使用量和成本。必须修改现有的组织预算流程，将这种变化因素考虑在内。

使用基于趋势（将历史成本用作输入）或者基于业务驱动因素（例如新产品发布或区域扩张）的算法，或者将趋势和业务驱动因素相结合，调整现有的预算和预测流程，使其更为灵活。

可以使用 AWS Cost Explorer，基于应用至历史成本（基于趋势）的机器学习算法，预测每日（最多 3 个月）或每月（最多 12 个月）的云成本。

对成本敏感的流程

在组织流程中建立成本意识：必须在新的和现有的组织流程中建立成本意识。建议在可能的情况下重用和修改现有流程，这可以最大限度地减少对敏捷性和速度的影响。以下建议有助您在工作负载中建立成本意识：

- 确保变更管理包含成本度量，以量化变更对财务的影响。这有助于主动解决与成本相关的问题，并强调成本节省。
- 确保成本优化是您运营能力的核心组成部分。例如，您可以利用现有的事故管理流程来调查和确定成本及使用量异常（成本超支）的根本原因。
- 通过自动化或工具加快节省成本和实现业务价值。在考虑实施成本时，请在对话中加入 ROI 信息，以证明投入时间或资金的合理性。
- 扩展现有的培训和开发计划，在整个组织中开展成本意识培训。建议在其中加入持续的培训和认证。这样有助建立一个能够自我管理成本和使用量的组织。

通报成本和使用量优化进展：必须定期报告组织内成本和使用量的优化情况。可以设定专门的成本优化环节，或者将成本优化纳入工作负载的常规运营报告周期。AWS Cost Explorer 提供仪表板和报告。可以通过 AWS 预算报告根据配置的预算跟踪成本和使用量。

还可以将成本和使用量报告 (CUR) 数据用于 Amazon QuickSight，以提供包含更精细数据的高度定制的报告。

发布有关成本和使用量的通知，确保可以快速采取措施应对成本和使用量的变化。借助 AWS 预算，您可以针对目标发布通知。我们建议针对工作负载成本及使用量的增加和减少配置通知。

主动监控成本和使用量：建议在组织内部主动监控成本和使用量，而不仅仅是在出现异常或意外时。在整个办公室或工作环境中，一目了然的仪表板能确保关键人员可以访问所需信息，并凸显组织对成本优化的重视程度。通过可见的仪表板，您可以积极推动成功的结果并在整个组织中加以实施。

对成本敏感的文化

建立对成本敏感的文化：在整个组织中实施更改或计划，以建立对成本敏感的文化。建议先从小范围着手，然后随着能力的增强和组织对云的使用的增加，在更广泛的范围实施更大型的计划。

在对成本敏感的文化中，您可以在整个组织中以有机和分散的方式执行最佳实践，从而扩展成本优化和云财务管理。与严格的自上而下的集中式方法相比，只需要很少的工作就可以在整个组织中培养起较高的能力水平。

文化上的细微改变可对您当前和将来的工作负载的效率产生重大影响。这种情况的示例包括：

- 以游戏方式展示整个组织的成本和使用量。这可以通过一个公开可见的仪表板或比较各个团队的规范化成本和使用量报告（例如，每个工作负载的成本、每笔交易的成本）来完成。
- 认可成本效率。公开或私下奖励自愿或主动实现的成本优化成果，并从错误中吸取教训，以免日后重蹈覆辙。
- 创建自上而下的组织要求，确保工作负载按预定义的预算运行。

及时了解新发布的服务：您或许可以实施新的 AWS 服务和功能，以提高工作负载的成本效率。请定期查阅 AWS 新闻博客、AWS 成本管理博客和 AWS 最新更新，了解有关新服务和功能版本的信息。

量化通过成本优化实现的业务价值

量化通过成本优化实现的业务价值：除了报告通过成本优化节省的费用外，建议量化其创造的附加价值。成本优化的效益通常以每项业务成果降低的成本来量化。例如，当您购买 Savings Plans 时，可以量化按需 Amazon Elastic Compute Cloud (Amazon EC2) 节省的成本，这可以降低成本并维持工作负载输出水平。当闲置的 Amazon EC2 实例终止，或者未连接的 Amazon Elastic Block Store (Amazon EBS) 卷删除时，可以量化削减的 AWS 支出成本。

通过量化成本优化带来的业务价值，您可以了解组织取得的全部效益。由于成本优化是一项必要的投资，因此量化业务价值之后，您就可以向利益相关者说明投资回报。如果能够量化业务价值，在未来的成本优化投资中，就可以从利益相关者那里得到更多支持，并获得一个框架来衡量组织成本优化活动的成果。

然而，成本优化带来的好处绝不仅仅在于降低或规避成本。考虑捕获其他数据来衡量效率提升值和业务价值。改进的示例包括：

- **执行成本优化最佳实践：**例如，资源生命周期管理降低了基础设施和运营成本，并为实验创造了时间和意想不到的预算。这提高了组织的敏捷性，并带来新的创收机会。
- **实施自动化：**拿 Auto Scaling 来说，它可以最小的工作量确保弹性，并通过消除手工的产能规划工作来提高员工工作效率。有关运营弹性的更多详细信息，请参阅 Well-Architected 运营可靠性支柱白皮书。
- **预测未来的 AWS 成本：**预测使得财务利益相关者可以与其他内部和外部组织的利益相关者一同设定期望，并有助于提高组织的财务可预测性。AWS Cost Explorer 可用于预测成本和使用量。

资源

请参阅以下资源，了解用于预算和预测云支出的 AWS 最佳实践的更多信息。

- 使用预算报告通报预算指标
- 使用 AWS Cost Explorer 进行预测
- AWS 培训
- AWS 认证
- AWS 云管理工具合作伙伴

支出和使用量意识

了解组织的成本和驱动因素对于有效管理成本和使用量以及识别降低成本的机会至关重要。在组织中，通常会有多个团队运行多个工作负载。这些团队可能在不同的部门，每个部门都有其自己的收入来源。将资源成本分摊到工作负载、各个组织或产品拥有者可以推动更高效的资源使用模式，减少浪费。准确的成本和使用量监控能够帮助您了解各部门和产品如何盈利，并让您能够针对组织内的资源分配做出更明智的决策。组织中各层级的人员都了解使用量是推动变化的关键，因为使用量变化会导致成本变化。

考虑采用多元方法来了解您的使用量和支出情况。您的团队必须收集数据、分析，然后报告。要考虑的关键因素包括：

- 治理
- 监控成本和使用量
- 停用

治理

为了管理云中的成本，您必须通过以下治理领域来管理使用量：

制定组织策略：执行治理的第一步是按照组织要求来针对云的使用制定策略。这些策略定义组织如何使用云以及如何管理资源。策略应涵盖与成本或使用量有关的资源和工作负载的所有方面，包括资源生命周期内的创建、修改和停用。

策略应该简单易懂，能够在整个组织中有效实施。从广泛的、高层级的策略开始，例如允许在哪个地理区域使用，或者一天中应该运行资源的时间。逐步为各组织部门和工作负载细化策略。常见策略包括可以使用哪些服务和功能（例如，测试/开发环境中性能较低的存储区），以及哪些类型的资源可供不同团队使用（例如，开发账户中最大的资源规模是中等）。

确立方向性目标和执行性目标：为组织制定成本和使用量的方向性目标及执行性目标。方向性目标为组织提供有关预期结果的指引和方向。执行性目标则提供要实现的具体可衡量的结果。方向性目标的一个示例是：在略微（非线性）增加成本的情况下，显著提升平台使用量。执行性目标的一个示例是：在成本增长不到 5% 的情况下，将平台使用量提升 20%。另一个常见的方向性目标是每 6 个月提高一次工作负载的效率。与之相关的执行性目标是，工作负载的每项输出成本每 6 个月降低 5%。

云工作负载的一个常见方向性目标是提高工作负载效率，即随着时间的推移降低工作负载每项业务成果的成本。建议为所有工作负载实施此目标，并设定执行性目标，例如每 6 至 12 个月将效率提高 5%。通过在成本优化中增强能力以及发布新服务和服务功能，可以在云中实现这一目标。

账户结构：AWS 拥有一个父级对多个子级的账户结构，通常称为主（父级，之前称为付款人）账户-成员（子级，之前称为关联）账户。最佳实践是，无论组织规模或使用情况如何，始终至少有一个主账户和一个成员账户。所有工作负载资源应仅驻留在成员账户内。

对于您应该拥有多少 AWS 账户这一问题，没有标准答案。评估当前和未来的运营和成本模型，以确保您的 AWS 账户结构反映了组织的方向性目标。有些公司出于业务原因会创建多个 AWS 账户，例如：

- 需要在组织部门、成本中心或特定工作负载之间实施管理和/或财务和计费隔离。
- AWS Service Limits 设置为特定于特殊工作负载。
- 工作负载和资源之间必须进行隔离和分离。

在 AWS Organizations 内，整合账单会在一个或多个成员账户与主账户之间创建结构。通过成员账户，您可以按团队隔离和区分成本和使用量。常见做法是每个组织部门（如财务、营销和销售）、每个环境生命周期（如开发、测试和生产）或每个工作负载（工作负载 a、b 和 c）具有单独的成员账户，然后使用整合账单将这些关联账户汇总在一起。

通过整合账单，您可以将多个成员 AWS 账户的付款整合至一个主账户下，同时仍可查看每个关联账户的活动。由于成本和使用量在主账户中汇总，因此，您可以最大限度地提高服务量折扣，并最大限度地利用承诺折扣（Savings Plans 和预留实例）来获得最高折扣。

AWS Control Tower 可以快速设置和配置多个 AWS 账户，确保治理符合您组织的要求。

组织团队和角色：制定策略后，可以在组织内创建用户的逻辑组和角色。这样，您就可以分配权限并控制使用量。从高层级的人员分组开始，这通常与组织部门和岗位角色（例如，IT 部门的系统管理员或财务主管）相一致。这些组将执行相似任务并需要相似访问权限的人员集结在一起。角色定义组必须做什么。例如，IT 部门的系统管理员需要创建所有资源的权限，而分析团队成员仅需要创建分析资源。

控制 – 通知：实施成本控制的第一步通常是进行相关设置，以便在发生成本或使用量超出策略的事件时触发通知。这样，您就可以迅速采取行动，并验证是否需要采取纠正措施，而不会限制工作负载或新活动或对它们产生负面影响。了解工作负载和环境限制后，可以强制实施治理。在 AWS 中，通知是通过 AWS 预算执行的，因此您可以定义 AWS 成本、使用量和承诺折扣（Savings Plans 和预留实例）的月度预算。可以在总成本级别（如所有成本）创建预算，也可以在更细粒度的级别创建预算，其中只包含特定的维度，如关联的账户、服务、标记或可用区。此外，还可以将电子邮件通知附加到预算中，如果当前或预测的成本或使用量超出定义的百分比阈值，将触发该通知。

控制 – 强制实施：在第二步中，您可以通过 AWS Identity and Access Management (IAM) 和 AWS Organizations 服务控制策略 (SCP)，在 AWS 中强制实施治理策略。借助 IAM，您可以安全地管理对 AWS 服务和资源的访问。您可以使用 IAM 控制谁能创建和管理 AWS 资源、可创建的资源类型以及可在何处创建。这样可以最大限度地减少创建不必要的资源。使用先前创建的角色和组，并分配 IAM 策略以强制实施正确的使用量。SCP 用于集中管控组织中所有账户的最大可用权限，以确保您的账户始终在访问控制准则允许的范围内。SCP 仅在启用了所有功能的组织中可用，并且您可以将 SCP 配置为默认情况下拒绝或允许对成员账户执行操作。有关实施访问管理的更多详细信息，请参阅 Well-Architected 安全性支柱白皮书。

控制 – 服务配额：治理也可以通过管理服务配额实现。通过确保为服务配额设置最低开销并进行准确维护，您可以最大限度地减少组织要求以外的资源创建。要实现这一点，您必须了解要求的改变速度、了解正在进行的项目（资源的创建和停用），以及影响可以实施的配额更改速度的因素。必要时，服务配额可用于增加配额。

AWS 成本管理服务与 AWS Identity and Access Management (IAM) 服务集成在一起。您可以结合使用 IAM 服务与成本管理服务，在账单控制台中控制外界对您的财务数据以及 AWS 工具的访问。

跟踪工作负载生命周期：确保跟踪工作负载的整个生命周期。这样可以确保在不再需要工作负载或工作负载组件时，可以将其停用或对其进行修改。这在发布新服务或功能时尤其有用。现有的工作负载和组件看起来仍在使用中，但是应该停用以将客户重定向到新服务。注意工作负载的先前阶段 – 在工作负载进入生产之后，可以停用以前的环境或大幅降低其容量，直到再次需要它们为止。

AWS 提供了许多可用于实体生命周期跟踪的管理和治理服务。您可以使用 AWS Config 或 AWS Systems Manager 提供一份详尽的 AWS 资源和配置清单。建议集成现有项目或资产管理系统来跟踪组织内的活动项目和产品。将当前系统与 AWS 提供的丰富事件集和指标结合起来，您就可以构建大量生命周期事件的视图并主动管理资源，以减少不必要的成本。

有关实施实体生命周期跟踪的更多详细信息，请参阅 Well-Architected 卓越运营支柱白皮书。

监控成本和使用量

支持团队详细了解工作负载，从而针对成本和使用量采取行动。要进行成本优化，首先必须详尽了解成本和使用量明细，能够对未来的支出、使用量和功能进行建模和预测，并实施足够的机制以使成本和使用量符合组织的目标。以下是监控成本和使用量时必须完成的步骤：

配置详细的数据源：在 Cost Explorer 中启用每小时粒度，并创建成本和使用量报告 (CUR)。这些数据源最确切地反映了整个组织中的成本和使用量。CUR 提供所有收费 AWS 服务的每日或每小时使用量粒度、费率、成本和使用属性。CUR 中的所有可能维度包括：标记、位置、资源属性和账户 ID。

使用以下自定义项配置 CUR：

- 包括资源 ID
- 自动刷新 CUR
- 每小时粒度
- 版本控制：覆盖现有报告
- 数据集成：Athena（Parquet 格式和压缩）

使用 AWS Glue 准备分析数据、使用 Amazon Athena 执行数据分析、使用 SQL 查询数据。您也可以使用 Amazon QuickSight 构建复杂的自定义视图，并在整个组织内分发。

确定成本归属类别：与您的财务团队和其他利益相关者合作，以了解必须在组织内部如何分摊成本的要求。必须将工作负载成本分摊至整个生命周期，包括开发、测试、生产和停用。了解组织如何对学习、员工培养和创意构思进行成本归类。这有助于将用于此目的的账户正确分配给培训和开发预算，而不是一般的 IT 成本预算。

确定工作负载指标：了解如何根据业务成功来衡量工作负载的输出。每个工作负载通常有一组表示性能的主要输出。如果您的工作负载复杂且包含许多组件，则可以对列表进行优先级排序，或者为每个组件定义和跟踪指标。与团队合作，了解要使用哪些指标。此部分将用于了解工作负载的效率，或每项业务输出的成本。

为成本和使用量分配组织含义：在 AWS 中实施标记以将组织信息添加到您的资源中，然后将其添加到成本和使用量信息中。标签是键值对 — 键是定义的，必须在整个组织中唯一，值则对于一组资源唯一。键值对的一个示例是键为 Environment，值为 Production。生产环境中的所有资源都有这个键值对。借助标记，您可以使用有意义、相关的组织信息对成本进行分类和跟踪。您可以应用代表组织类别（例如成本中心、应用名称、项目或拥有者）的标签，标识工作负载和工作负载的特征（例如测试或生产），以在整个组织中分摊成本和使用量。

当您将标记应用于 AWS 资源（如 EC2 实例或 Amazon S3 存储桶）并激活标记后，AWS 会将此信息添加到成本和使用量报告。您可以在带标签和无标签的资源上运行报告并执行分析，以更好地遵守内部成本管理策略，并确保准确归属。

跨组织账户创建和实施 AWS 标记标准之后，您将能够一致且统一地管理和治理 AWS 环境。在 AWS Organizations 中使用标记策略定义有关如何在 AWS Organizations 账户的 AWS 资源上使用标签的规则。借助标记策略，您可以采用标准化方法轻松标记 AWS 资源。

AWS 标签编辑器可用于为多个资源添加、删除和管理标记。

AWS Cost Categories 可用于向成本分配组织含义，而无需在资源上添加标签。您可以将成本和使用量信息映射到唯一的内部组织结构。您可以定义类别规则，以使用账单维度（例如账户和标签）对成本进行映射和分类。除了标记之外，这还提供了另外一个级别的管理能力。您还可以将特定账户和标记映射到多个项目。

配置账单和成本优化工具：要修改使用量和调整成本，组织中的每个人都必须能够访问其成本和使用量信息。建议所有工作负载和团队在使用云时都配置以下工具：

- **报告：**汇总所有成本和使用量信息。
- **通知：**当成本或使用量超出定义的限值时触发通知。
- **当前状态：**配置显示当前成本和使用量水平的仪表板。仪表板应位于工作环境中的显眼位置（类似于操作仪表板）。
- **趋势分析：**能够以所需的粒度显示成本和使用量在指定时间段内的变化。
- **预测：**能够显示预计的未来成本。
- **跟踪：**对照配置的方向性目标或执行性目标显示当前的成本和使用量。
- **分析：**可让团队成员在所有可能的维度执行详尽至每小时粒度的自定义和深入分析。

您可以使用 AWS 本机工具（如 AWS Cost Explorer、AWS 预算以及 Amazon Athena 和 QuickSight）提供此功能。您还可以使用第三方工具，但是，必须确保为此工具花费的成本给组织带来价值。

根据工作负载指标分配成本：成本优化旨在以最低的价格实现业务成果，这只能通过按工作负载指标分配工作负载成本（按工作负载效率衡量）来实现。通过日志文件或其他应用程序监控来监

控定义的工作负载指标。将此数据与工作负载成本（可通过查看具有特定标签值或账户 ID 的成本获得）相结合。建议每小时进行一次分析。如果有一些静态成本要素（例如，全天候运行的后端数据库）且请求率不同（例如，使用量高峰在上午 9 点至下午 5 点，晚间的请求数量很少），则效率通常会变化。了解静态成本和可变成本之间的关系有助于您将精力集中在优化活动上。

停用资源

当您管理项目、员工和技术资源的列表之后，随着时间的推移，您将能够识别不再使用的资源或不再具备拥有者的项目。

在资源生命周期内跟踪资源：停用不再需要的工作负载资源。一个常见的示例是用于测试的资源，在测试完成后，可以将其删除。通过标签跟踪资源（并在这些标签上运行报告）可帮助您确定要停用的资产。使用标记是跟踪资源的一种有效方法，它通过标记资源的功能或资源的已知可停用日期来跟踪资源。然后，可以在这些标签上运行报告。功能标记的示例值是“featureX 测试”，用于根据工作负载生命周期标识资源的用途。

实施停用流程：在整个组织中实施标准化流程，以识别和删除未使用的资源。该流程应该定义执行搜索的频率以及删除资源的流程，以确保满足所有组织要求。

停用资源：搜索未使用资源的频率和工作量应反映潜在的节省额，因此，与成本较高的账户相比，对成本较低的账户进行分析的频率应该更低。搜索和停用事件可由工作负载中的状态更改触发，比如产品生命周期结束或被更换。搜索和停用事件也可由外部事件触发，如市场条件发生变化或产品终止。

自动停用资源：使用自动化技术可以减少或消除停用流程中的相关成本。将工作负载设计为执行自动化停用将减少工作负载在其整个生命周期内的总成本。您可以使用 AWS Auto Scaling 来执行停用流程。还可以使用 API 或开发工具包来实施自定义代码，以自动停用工作负载资源。

资源

请参阅以下资源，详细了解有关支出意识的 AWS 最佳实践。

- AWS 标记策略
- 激活用户定义的成本分配标签

- AWS 账单和成本管理
- 成本管理博客
- AWS 多账户计费策略
- AWS 开发工具包和工具
- 标记最佳实践
- Well-Architected 实验室 – 成本基础知识
- Well-Architected 实验室 – 支出意识

资源成本效益

为工作负载使用合适的服务、资源和配置是节省成本的关键。创建具有成本效益的资源时，请考虑以下几点：

- 选择服务时评估成本
- 选择正确的资源类型、规模和数量
- 选择最佳定价模型
- 制定数据传输计划

您可以使用 AWS 解决方案架构师、AWS 解决方案、AWS 参考架构和 APN 合作伙伴来帮助您根据所学知识选择架构。

选择服务时评估成本

确定组织需求：在为工作负载选择服务时，了解组织的优先要务至关重要。确保在成本和其他 Well-Architected 支柱（例如性能和可靠性）之间取得平衡。完全成本优化的工作负载是最符合组织需求的解决方案，但不一定是成本最低的。与组织内的所有团队会面以收集信息，例如产品、业务、技术和财务。

分析所有工作负载组件：对工作负载中的所有组件进行全面分析。确保在分析成本与工作负载在其生命周期内可能节省的成本之间取得平衡。必须确定组件的当前影响以及未来的潜在影响。例

如，如果拟议资源的成本为每月 10 USD，在预测的负载下不会超过每月 15 USD，则花一天的时间将成本降低 50%（每月 5 USD）可能会超过系统使用寿命内的潜在收益。使用更快、更有效的基于数据的预估可为该组件带来最佳的总体结果。

工作负载可能会随时间变化，如果工作负载架构或使用量发生变化，原本合适的服务集可能未必仍是最优之选。为甄选服务进行分析时，必须考虑工作负载当前和未来的状态以及使用量水平。为将来的工作负载状态或使用量实施服务可以减少或消除未来进行更改所需的工作量，从而降低总体成本。

AWS Cost Explorer 和 CUR 可以分析概念验证 (PoC) 或运行环境的成本。您也可以使用 AWS 简单月度成本结算器或 AWS 定价计算器估算工作负载成本。

托管服务：托管服务消除了维护服务的运营和管理负担，让您可以专注于创新。此外，由于托管服务在云级别运行，因此可以提供更低的单位事务或服务成本。

利用节省下来的时间，您的团队将能够专注于解决技术债务、创新和增值功能。例如，您可能需要尽快将本地环境“直接迁移”到云，然后再进行优化。值得探索的是，通过使用消除或减少许可证成本的托管服务，您可以节省多少成本。

通常，可以设置托管服务的部分属性，以确保容量足够。您必须设置和监控这些属性，以便最大限度地减少多余容量，并最大限度地提高性能。您可以使用 AWS 管理控制台或 AWS API 和开发工具包修改 AWS Managed Services 的属性，以使资源需求匹配不断变化的要求。例如，您可以增加或减少 Amazon EMR 集群（或 Amazon Redshift 集群）上的节点数量，以扩展或缩减集群。

您还可以在 AWS 资源上打包多个实例，以实现更高的密度使用量。例如，您可以在单个 Amazon Relational Database Service (Amazon RDS) 数据库实例上预置多个小型数据库。随着使用量的增长，您可以遵照快照和还原流程将其中一个数据库迁移到专用 RDS 数据库实例。

在托管服务上预置工作负载时，您必须了解调整服务容量的要求。这些要求通常是时间、工作量和对正常工作负载运营的任何影响。预置的资源必须留出时间来进行任何更改，并预置必要的开销以允许这样做。通过使用与系统和监控工具（如 Amazon CloudWatch）集成的 API 和开发工具包，可以将修改服务所需的持续工作量减少至接近零。

Amazon Relational Database Service (RDS)、Amazon Redshift 和 Amazon ElastiCache 提供托管数据库服务。Amazon Athena、Amazon Elastic Map Reduce (EMR) 和 Amazon Elasticsearch 提供托管分析服务。

AWS Managed Services (AMS) 是代表企业客户和合作伙伴运营 AWS 基础设施的服务。它提供了一个安全且合规的环境，您可以将工作负载部署到其中。AMS 使用具有自动化功能的企业云运营模型，可以满足组织要求，更快地迁移到云中并降低持续的管理成本。

无服务器或应用程序级服务：您可以使用无服务器或应用程序级服务，如 AWS Lambda、Amazon Simple Queue Service (Amazon SQS)、Amazon Simple Notification Service (Amazon SNS) 和 Amazon Simple Email Service (Amazon SES)。这些服务剔除了管理资源的需要，并提供代码执行、排队服务和消息传递功能。另一个好处是，它们可以根据使用量扩展性能和成本，从而实现有效的成本分配和归属。

有关无服务器的更多信息，请参阅 Well-Architected 无服务器应用程序剖析白皮书。

分析工作负载的使用量随时间的变化情况：随着 AWS 发布新的服务和功能，适用于您的工作负载的最佳服务可能会发生变化。所需的工作量应反映出可能带来的好处。工作负载审核频率取决于您的组织要求。如果工作负载的成本很高，则尽早实施新服务可最大限度地节省成本，因此提高审核频率可能是有利的。审核的另一个触发因素是使用模式发生变化。使用量发生重大变化可能表明备用服务更加理想。例如，为获得更高的数据传输速率，直接连接服务可能比 VPN 便宜，并且会提供所需的连接。预测服务变更的潜在影响，因此您可以监控这些使用量水平触发器，并更快地实施最具成本效益的服务。

许可成本：使用开源软件可以消除软件许可成本。随着工作负载规模的扩展，这可能对工作负载成本产生重大影响。将许可软件能够带来的好处与总成本进行比较，确保拥有最优化的工作负载。对许可中的任何更改及其对工作负载成本的影响建模。如果供应商更改了数据库许可证的成本，请调查这会如何影响工作负载的整体效率。考虑供应商的历史定价公告，了解其产品中的许可更改趋势。许可成本也可以独立于吞吐量或使用量进行扩缩，例如按硬件扩缩的许可证（CPU 绑定许可证）。应避免使用这些许可证，因为成本会迅速增加，而且无法取得相应的结果。

可以使用 AWS License Manager 管理工作负载中的软件许可证。您可以配置许可规则并强制执行必要的条件，以帮助防止违反许可的行为，还能降低因许可证过期而产生的成本。

选择正确的资源类型、规模和数量

通过选择最佳的资源类型、规模和资源数量，可以以最低成本的资源满足技术要求。合理调整大小活动需要考虑工作负载的所有资源、各项资源的所有属性以及合理调整大小操作中涉及的工作

量。合理调整大小可以是一个由使用模式变化和外部因素（如 AWS 价格下降或新 AWS 资源类型）触发的迭代过程。如果合理调整大小所需的成本超过工作负载整个生命周期中可能节省的成本，则合理调整大小也可以是一次性的。

AWS 中有许多不同的方法：

- 执行成本建模
- 基于指标或数据选择规模
- 自动（基于指标）选择规模

成本建模：对工作负载及其每个组件执行成本建模，以了解资源之间的平衡，并在给定的具体性能水平下，确定工作负载中每个资源的正确规模。对不同预测负载下的工作负载执行基准测试活动，并比较成本。建模工作应该反映可能带来的好处，例如花费的时间与组件成本或预计可节省的成本成正比。有关最佳实践，请参阅 AWS 架构完善的框架的性能效率支柱白皮书的“审核”部分。

AWS Compute Optimizer 可协助对正在运行的工作负载进行成本建模。它根据历史使用量为计算资源提供合理调整大小的建议。这是计算资源的理想数据源，因为它是一项免费的服务，并且会利用机器学习根据风险等级提出多个建议。您还可以将 Amazon CloudWatch 和 CloudWatch Logs 与自定义日志一起用作数据源，用于其他服务和工作负载组件的合理调整大小操作。

以下是成本建模数据和指标的建议：

- 监控必须准确反映最终用户体验。为时间段选择正确的粒度，并仔细选择最大值或第 99 个百分位值而不是平均值。
- 为覆盖任何工作负载周期所需的分析时间段选择正确的粒度。例如，如果执行为期两周的分析，您可能会忽略高利用率的月度周期，这可能导致预置不足。

基于指标或数据的选择：根据工作负载和资源特征选择资源规模或类型，例如计算、内存、吞吐量或写入密集型资源。通常使用成本建模、工作负载的上一个版本（例如本地版本）、文档或关于工作负载的其他信息源（白皮书、发布的解决方案）进行选择。

基于指标自动选择：在工作负载中创建一个反馈循环，此循环使用正在运行的工作负载中的活动指标来对该工作负载进行更改。您可以使用托管服务（如 AWS Auto Scaling），将其配置为代您执行合理调整大小的操作。AWS 提供 API、开发工具包和功能，让您可以轻松修改资源。您可以

对工作负载进行编程以停止和启动 EC2 实例，从而允许更改实例大小或实例类型。这带来双重好处：既合理调整了大小，又几乎消除了进行更改所需的所有运营成本。

某些 AWS 服务内置了自动类型或大小选项，如 S3 智能分层。S3 智能分层会根据您的使用模式，自动在两个访问层之间移动数据：频繁访问和非频繁访问。

选择最佳定价模型

执行工作负载成本建模：考虑工作负载组件的要求并了解潜在的定价模型。定义组件的可用性要求。确定工作负载中是否存在执行功能的多个独立资源，以及工作负载随着时间推移的需求情况。使用默认的按需定价模型和其他适用模型比较资源成本。考虑资源或工作负载组件的任何潜在更改。

定期执行账户级别的分析：定期执行成本建模可确保能够跨多个工作负载进行优化。例如，如果总体上按需使用多个工作负载，则变更的风险较低，并且实施基于承诺的折扣可降低总体成本。建议每两周到一个月定期执行一次分析。这样您可以进行少量调整性采购，因此定价模型的覆盖范围会随着工作负载及其组件的变化而不断变化。

使用 AWS Cost Explorer 建议工具寻找享受承诺折扣的机会。

要为 Spot 工作负载寻找机会，请查看总体使用量的小时视图，并确定使用量或弹性的定期变化周期。

定价模型：AWS 有多种定价模型，您可以符合组织需求、最具成本效益的方式支付资源费用。以下部分介绍了各种采购模式：

- 按需
- Spot
- 承诺折扣 – Savings Plans
- 承诺折扣 – 预留实例/容量
- 地理选择
- 第三方协议和定价

按需：这是默认的即付即用定价模式。当您使用资源（如按需使用 EC2 实例或 DynamoDB 等服务）时，可以按小时支付固定费用，并且无需长期承诺。您可以根据应用程序的需求增加或减少资源或服务的容量。按需模式有小时费率，但是根据服务的不同，可以以 1 秒为单位计费（例如 AWS Lambda 或 Linux EC2 实例）。推荐以下应用程序使用按需模式：具有定期出现峰值的短期工作负载（例如为期四个月的项目）或无法中断的不可预测工作负载。按需模式还适用于要求运行时不间断，但运行时间不够长、无法享受承诺折扣（Savings Plans 或预留实例）的工作负载，例如预生产环境。

Spot：Spot 实例是备用的 EC2 计算容量，可提供按需价格高达 90% 的折扣，无需长期承诺。借助 Spot 实例，您可以在相同的预算下显著降低运行应用程序的成本或扩展应用程序的计算容量。不同于按需模式，如果 EC2 需要恢复容量，或者 Spot 实例的价格超出您配置的价格，可以中断 Spot 实例并触发 2 分钟的警告。平均而言，Spot 实例的中断时间在 5% 以下。

如果有队列或缓冲区，或者有多个资源独立处理请求（例如 Hadoop 数据处理），Spot 是理想之选。通常，这些工作负载无状态且灵活，具备容错能力，例如批处理、大数据和分析、容器化环境和高性能计算 (HPC)。非关键工作负载（例如测试和开发环境）也适合使用 Spot。

Spot 还可集成到多个 AWS 服务，如 EC2 Auto Scaling 组 (ASG)、Elastic MapReduce (EMR)、Elastic Container Service (ECS) 和 AWS Batch。

如果需要回收 Spot 实例，EC2 会通过 CloudWatch Events 传递的 Spot 实例中断通知以及实例元数据发送一段两分钟警告。在这两分钟内，您的应用程序可以使用利用这段时间保存其状态、耗尽运行的容器、上传最终日志文件，或从负载均衡器中删除自己。两分钟结束时，您可以选择休眠、停止或终止 Spot 实例。

在工作负载中采用 Spot 实例时，请考虑以下最佳实践：

- **将最高价格设置为按需费率：**这样可以确保按当前的即期价格（最便宜的价格）付款，并且支付的费用永远不会高于按需费率。可通过控制台和 API 获得当前及历史费率。
- **在尽可能多的实例类型之间保持灵活性：**在实例类型的系列和规模方面都要灵活，以提高达到目标容量要求的可能性、获得可能的最低成本，并最大限度减小中断的影响。
- **灵活安排工作负载的运行位置：**可用容量可能因可用区而异。这样做可提高实现目标容量的可能性（因为会利用多个备用容量池），并最大限度降低成本。

- **连续性设计**：针对无状态和容错性设计工作负载，这样即使一些 EC2 容量被中断，也不会影响工作负载的可用性或性能。
- 我们建议将 Spot 实例与按需和 Savings Plans/预留实例相结合，以最大限度优化工作负载成本并提高性能。

承诺折扣 – Savings Plans：AWS 通过保留或承诺使用一定数量的资源并为您的资源收取折扣价，为您提供多种降低成本的方法。通过 Savings Plan，您可以在一年或三年的时间里保证每小时的花费，从而享受所有资源的折扣价。Savings Plans 提供 EC2、Fargate 和 Lambda 等 AWS Compute 服务的折扣。当您做出承诺时，您每小时支付该承诺金额，并且其会以折扣费率从您的按需使用费中扣除。例如，您承诺每小时 50 USD，按需使用费每小时 150 USD。考虑到 Savings Plans 定价，特定使用费都可享受 50% 的折扣。因此，50 USD 的承诺额涵盖 100 USD 的按需使用费。您将支付 50 USD（承诺）和剩余的 50 USD 按需使用费。

Compute Savings Plans 是最灵活的方案，提供高达 66% 的折扣。它们自动跨可用区、实例大小、实例系列、操作系统、租约、区域和计算服务应用。

Instance Savings Plans 不太灵活，但提供更高的折扣率（高达 72%）。它们自动跨可用区、实例大小、实例系列、操作系统和租约应用。

有三种付款选项：

- **无预先付款**：没有预先付款；然后，每个月按逐渐减少的每小时费率为当月的总小时数支付费用。
- **部分预先付款**：相比无预先付款，折扣率更高。为部分使用量预先支付费用；然后，每个月按逐渐减少的每小时费率为当月的总小时数支付费用。
- **全额预先付款**：预先支付整个期间使用量的费用，在承诺涵盖的剩余期限内不会产生任何其他费用。

您可以在工作负载中任意组合使用这三种购买选项。

Savings Plans 先应用到购入它的账户中的使用量（从最高折扣率到最低折扣率），然后应用到所有其他账户的合并使用量（从最高折扣率到最低折扣率）。

建议在没有使用量或资源的账户（例如主账户）中购买所有 Savings Plans。这可以确保 Savings Plans 适用于所有使用情况下的最高折扣率，最大限度提高折扣金额。

工作负载和使用量通常会随时间而改变。建议随着时间的推移持续购买少量的 Savings Plans 承诺使用量。这样可确保您保持较高的覆盖率，以最大限度提高折扣，并且计划始终与工作负载和组织要求紧密匹配。

由于折扣可能发生变化，请不要在账户中设置目标覆盖率。覆盖率低并不一定意味着潜在节省费用高。您的账户的覆盖率可能很低，但如果使用量由小型实例组成，并且使用许可的操作系统，则节省的费用可能仅为几个百分点。相反，在 Savings Plan 建议工具中跟踪和监控可能节省的费用。经常查看 Cost Explorer 中的 Savings Plans 建议（执行定期分析），并继续购买承诺使用量，直到估计的节省额低于组织所需的折扣为止。例如，跟踪并监控您的潜在折扣是否保持在 20 % 以下（如果高于该值，则必须购买）。

监控利用率和覆盖率，但仅用于检测更改。不要以特定的利用率或覆盖率作为目标，因为这并不一定与节省额成比例。确保购买 Savings Plans 可增加覆盖率，如果覆盖率或使用率降低，则确保对其进行量化并了解相关情况。例如，您将工作负载资源迁移到较新的实例类型，这会减少现有计划的利用率，但是性能效益胜于节省的成本。

承诺折扣 – 预留实例/承诺：与 Savings Plans 类似，预留实例为做出最低资源运行量承诺的用户提供高达 72% 的折扣。预留实例可用于 RDS、Elasticsearch、ElastiCache、Amazon Redshift 和 DynamoDB。Amazon CloudFront 和 AWS Elemental MediaConvert 也会在您做出最低消费承诺时提供折扣。预留实例当前可用于 EC2，但是，Savings Plans 提供相同的折扣级别，同时具有更大的灵活性，且没有管理开销。

预留实例提供相同的定价选项：无预先付款、部分预先付款和全额预先付款，并且期限相同：一年或三年。

可以在区域或特定可用区购买预留实例。在可用区中购买时，它们提供容量预留。

虽然 EC2 具有可转换的 RI，但应该对所有 EC2 实例使用 Savings Plans，因其增加了灵活性并降低了运营成本。

应使用相同的流程和指标来跟踪和购买预留实例。建议不跟踪整个账户中 RI 的覆盖率。同时，建议不要监控或跟踪利用率，而应在 Cost Explorer 中查看利用率报告，并使用表中的净节省额列。如果净节省额是很大的负数，则必须采取措施补救未使用的 RI。

EC2 队列：EC2 队列可用于定义目标计算容量，然后为队列指定实例类型以及按需和 Spot 实例的余额。EC2 队列将自动启动最低价格的资源组合，以满足定义的容量。

地理选择：在架构解决方案时，最佳实践是设法将计算资源放在更接近用户的位置，以提供更低的延迟和强大的数据主权。对于全球用户，您应该使用多个位置来满足这些需求。您应该选择尽可能降低成本的地理位置。

AWS 云基础设施围绕区域和可用区构建。区域是指全球范围内的某个物理位置，每个区域由多个可用区组成。可用区由一个或多个分散的数据中心组成，每个都拥有独立的配套设施，其中包括冗余电源、联网和连接。

每个 AWS 区域都在当地市场条件下运营，每个区域的资源定价可能不同。选择特定区域来运行解决方案组件或整个解决方案，以便您可以在全球范围内以尽可能低的价格运行。您可以使用 AWS 简单月度成本结算器来估算各区域中工作负载的成本。

第三方协议和定价：当您在云中使用第三方解决方案或服务时，确保定价结构与成本优化结果保持一致非常重要。定价应与其带来的结果和价值成比例。这方面的一个例子是可带来一定百分比节省额的软件，节省额（结果）越高，其价格也就越高。除非您能提供特定账单每一部分的结果，否则与账单成比例的协议通常不会与成本优化保持一致。例如，如果您使用的其他服务没有带来任何好处，提供 EC2 相关建议并收取整个账单一定比例费用的解决方案将会增加。另一个示例是根据所托管资源的成本按一定百分比收费的托管服务。实例越大并不一定意味着需要更多的管理工作，但会收取更多费用。确保这些服务定价安排包括成本优化计划或服务中的功能，以提高效率。

制定数据传输计划

云的优势之一在于它是一种托管的网络服务。不再需要管理和操作一组交换机、路由器和其他相关的网络设备。云中网络资源的耗用和付费方式与您为 CPU 和存储器付费的方式相同—只需为使用的资源付费。要在云中优化成本，必须高效利用网络资源。

执行数据传输建模：了解数据传输在您的工作负载中发生的位置、传输的成本以及相关的收益。因此，您可以做出明智的决定来修改或接受架构决策。例如，您可能有一个多可用区配置，可以在可用区之间复制数据。您可以对结构成本建模，并确定这是可接受的成本（类似于在两个可用区中支付计算和存储费用），以实现所需的可靠性和弹性。

对不同使用级别的成本进行建模。工作负载的使用量可能随时间而变化，不同的服务可能在不同的级别上更具有成本效益。

使用 AWS Cost Explorer 或成本和使用量报告 (CUR) 来了解数据传输成本并对其建模。配置概念证明 (PoC) 或测试您的工作负载，并在实际的模拟负载下运行测试。您可以根据不同的工作负载需求对成本进行建模。

优化数据传输：针对数据传输进行架构，可确保您最大限度地降低数据传输成本。这可能涉及使用内容分发网络来定位更靠近用户的数据，或者使用从您的本地设施到 AWS 的专用网络链接。您还可以使用 WAN 优化和应用程序优化来减少组件之间传输的数据量。

选择服务以降低数据传输成本：Amazon CloudFront 是一个全球内容分发网络，可提供低延迟、高传输速度的数据。它在世界各地的边缘站点缓存数据，从而减少资源负担。通过使用 CloudFront，您可以减少向全球大量用户分发内容的管理工作，同时将延迟降到最低。

AWS Direct Connect 允许您建立到 AWS 的专用网络连接。与基于互联网的连接相比，这可以降低网络成本、增加带宽并提供更一致的网络体验。

您可以通过 AWS VPN 在专用网络和 AWS 全局网络之间建立安全的专用连接。它是小型办公室或业务合作伙伴的理想之选，因其提供快速简便的连接，并且是完全托管的弹性服务。

VPC 终端节点允许通过专用网络在 AWS 服务之间建立连接，可用于减少公共数据传输和 NAT 网关成本。网关 VPC 终端节点不按小时收费，支持 Amazon S3 和 Amazon DynamoDB。接口 VPC 终端节点由 AWS PrivateLink 提供，有小时费和每 GB 使用成本。

资源

请参阅以下资源，详细了解有关资源成本效益的 AWS 最佳实践。

- AWS Managed Services：企业变革之旅视频
- 使用 Cost Explorer 分析成本
- 获取预留实例建议
- 合理调整大小入门建议
- Spot 实例最佳实践
- Spot 队列
- 预留实例如何工作

- AWS 全球基础设施
- Spot 实例顾问
- Well-Architected 实验室 – 资源成本效益

管理需求和供应资源

在迁移到云时，您仅为所需内容付费。您可以在需要时供应与工作负载需求匹配的资源，这样就无需进行昂贵且浪费的超额配置。还可以通过限流、缓冲区或队列来修改需求，以满足需求并以更少的资源达成目标。

您应该在即时供应的经济优势和预置需求之间取得平衡，以协调资源故障、高可用性和预置时间。根据您的需求是固定的还是可变的，您可以计划创建指标和自动化，这样即使在扩展时也能够尽可能减少对环境的管理工作。修改需求时，必须知道工作负载可以允许的可接受最大延迟。

在 AWS 中，您可以采用多种不同的方法来管理需求和供应资源。以下部分介绍了如何使用这些方法：

- 分析工作负载
- 管理需求
- 基于需求的供应
- 基于时间的供应

分析工作负载：了解工作负载的需求。组织需求应指出工作负载对于请求的响应时间。响应时间可用于确定是否管理了需求，或者资源的供应是否会改变以满足需求。

分析应包括需求的可预测性和可重复性、需求的变化速率以及需求的变化量。确保在足够长的时间内执行分析，以纳入任何季节性变化，例如月末处理或假期高峰。

确保分析工作反映实施扩展的潜在好处。查看组件的预期总成本，以及在工作负载生命周期内增加和减少的使用量和成本。

您可以将 AWS Cost Explorer 或 Amazon QuickSight 与 CUR 或应用程序日志一起使用，以对工作负载需求进行可视化分析。

管理需求

管理需求 – 限流：如果需求源具有重试功能，可以实施限流。限流会告诉需求源，如果当前无法处理请求，则应稍后再试。需求源将等待一段时间，然后重新尝试请求。实施限流的优势是可限制最大资源量和工作负载成本。在 AWS 中，可以使用 Amazon API Gateway 实施限流。有关实施限流的更多详细信息，请参阅 Well-Architected 可靠性支柱白皮书。

管理需求 – 基于缓冲区：与限流类似，缓冲区会延迟请求处理，从而允许以不同速率运行的应用程序有效通信。基于缓冲区的方法使用队列来接受来自产生方的消息（工作单元）。然后消息将由使用方读取并处理，这样消息就能够以满足使用方业务需求的速率运行。无需担心产生方必须处理数据持久性和反向压力等限流问题（因为使用方运行缓慢，导致产生方运行缓慢）。

在 AWS 中，您可以从多个服务中进行选择，以便实施缓冲方法。Amazon SQS 是一项托管服务，提供允许单个使用方读取单个消息的队列。Amazon Kinesis 提供允许众多使用方读取相同消息的流。

使用基于缓冲区的方法进行架构时，请确保架构工作负载以在所需时间内处理请求，并且您能够处理重复的工作请求。

动态供应

基于需求的供应：利用云的弹性来供应资源以满足不断变化的需求。利用 API 或服务功能，以编程方式动态改变架构中云资源的数量。这使您能够在架构中扩展组件，并在需求高峰期间自动增加资源数量以保持性能，也可以在需求量降低时减少容量以降低成本。

Auto Scaling 可帮助您调整容量以维持稳定、可预测的性能，并确保成本最低。它是一项完全托管的免费服务，与 Amazon EC2 实例和 Spot 队列、Amazon ECS、Amazon DynamoDB 与 Amazon Aurora 集成。

Auto Scaling 提供自动资源发现功能，以帮助您在工作负载中找到可以配置的资源，它具有内置的扩展策略来优化性能、成本或者在两者之间取得平衡，并提供预测性扩展来协助应对定期出现的峰值。

Auto Scaling 可以实施手动、计划或基于需求的扩展，您还可以使用来自 Amazon CloudWatch 的指标和警报触发工作负载的扩展事件。典型的指标可以是标准 Amazon EC2 指标，例如 CPU

利用率、网络吞吐量和 ELB 观察到的请求/响应延迟。如果可能，应该使用指示客户体验的指标，通常是来自工作负载中的应用程序代码的自定义指标。

当构建基于需求的方法时，请注意两个重要事项。首先，了解您必须以多快的速度预置新资源。其次，了解供应和需求之间的差额将发生变化。您必须准备好应对需求变化的速度，并准备好应对资源故障。

Elastic Load Balancing (ELB) 通过在多种资源之间分配需求来帮助您扩展规模。随着实施的资源越来越多，您可以将它们添加到负载均衡器中以满足需求。AWS ELB 支持 EC2 实例、容器、IP 地址和 Lambda 函数。

基于时间的供应：基于时间的方法可以协调资源容量以满足可预测或时间明确定义的需求。此方法通常不依赖资源的利用水平。基于时间的方法可以确保资源在需要的特定时间可用，并且提供时不会因启动流程和系统或一致性检查而发生延迟。使用基于时间的方法，您可以在繁忙时段提供额外的资源或增加容量。

您可以使用计划的 Auto Scaling 实施基于时间的方法。工作负载可以在定义的时间按计划扩展或缩减（例如办公时间开始时），从而确保用户就位或需求出现时资源可用。

您可以利用 AWS API 和开发工具包以及 AWS CloudFormation，在需要时自动预置和停用整个环境。此方法非常适合仅在定义的办公时间或时间段运行的开发或测试环境。

您可以使用 API 来扩展环境中的资源大小（纵向扩展）。例如，可以通过更改实例大小或分类纵向扩展生产工作负载。这可以通过停止和启动实例，以及选择不同的实例大小或分类来实现。这种技巧也可以应用于其他资源，如 EBS 弹性卷，您可以在使用时对其进行修改以增加大小、调整性能 (IOPS) 或更改卷类型。

当构建基于时间的方法时，请注意两个重要事项。首先，使用模式的一致性如何？ 其次，如果模式发生更改会产生什么影响？ 您可以通过两种方式提高预测的准确性：监控工作负载和使用商业智能。如果您发现使用模式发生重大更改，可以调整时间，以确保提供覆盖范围。

动态供应：您可以使用 AWS Auto Scaling，或者通过 AWS API 或开发工具包在代码中加入扩展。这样省去了手动更改环境的操作成本，因而工作负载的总体成本得以降低，而且执行速度变得更快。这将确保工作负载资源在任何时候都最符合需求。

资源

请参阅以下资源，详细了解有关管理需求和供应资源的 AWS 最佳实践。

- API 网关限流
- Amazon SQS 入门
- Amazon EC2 Auto Scaling 入门

随着时间的推移不断优化

在 AWS 中，您可以检查新服务并将其实施到工作负载中，以随着时间的推移不断优化。

审核和实施新服务

AWS 会不断发布新服务和功能，建议您时常审视现有的架构决策以确保它们始终具有成本效益。随着需求的变化，一定要主动停用不再需要的资源、组件和工作负载。考虑采取以下措施，以帮助随着时间的推移不断优化：

- 制定工作负载审核流程
- 审核和实施服务

制定工作负载审核流程：为确保工作负载始终最具成本效益，您必须定期对其进行审核，以了解是否有机会实施新的服务、功能和组件。为确保整体成本尽可能低，此过程必须与潜在的节省额成比例。例如，与占总支出 5% 的工作负载相比，应更经常、更彻底地审核占总支出 50% 的工作负载。考虑任何外部因素或波动。如果工作负载服务于特定的地理位置或市场领域，并且您已预测出该领域会出现的变化，则提高审核频率可能会节省成本。审核时要考虑的另一个因素是实施更改的工作量。如果测试和验证变更的成本很高，则审核的频率应该降低。

考虑维护过时和旧式组件及资源的长期成本，以及无法在其中实施新功能的事实。当前的测试和验证成本可能会超过预计的收益。但是，随着时间的推移，工作负载和当前技术之间的差距会增大，进行更改的成本可能会大幅增加，导致成本升高。例如，迁移到新的编程语言的成本当前可能不具成本效益。然而，五年之后，熟练使用该语言的人员的成本可能会增加，并且由于工作负载的扩展，迁移到新语言的系统规模更大，这其中涉及的工作量甚至高于以前。

将工作负载分解成多个组件，分配组件的成本（估算即可），然后在每个组件旁边列出因素（例如工作量和外部市场）。使用这些指示信息来确定每个工作负载的审核频率。例如，您可能觉得 Web 服务器的成本高、变更的工作量小、外部因素多，因而审核频率很高。而中央数据库的成本可能中等、变更的工作量很大、外部因素较少，因而审核频率也为中等。

审核工作负载和实施服务：为实现新 AWS 服务和功能的优势，必须对工作负载执行审核流程，并根据需要实施新服务和功能。例如，您可以审核工作负载，并使用 Amazon Simple Email Service (SES) 替换消息传递组件。这省去了运行和维护实例队列的成本，同时以更低的成本提供所有功能。

总结

成本优化和云财务管理是一项持续的工作。您应定期与财务和技术团队合作，审核架构方法，更新组件选项。

AWS 致力于在您构建高度灵活、响应迅速的自适应部署时，帮助您最大限度地降低成本。要真正优化您的部署成本，请充分利用本白皮书中讨论的相关工具、技巧和最佳实践。

贡献者

本文档的贡献者包括：

- Philip Fitzsimons，Amazon Web Services Well-Architected 高级经理
- Nathan Besh，Amazon Web Services Well-Architected 成本主管
- Levon Stepanian，Amazon Web Services
- Keith Jarrett，成本优化部门业务拓展主管
- PT Ng，Amazon Web Services 商业架构师
- Arthur Basbaum，Amazon Web Services 业务开发经理
- Jarman Hauser，Amazon Web Services 商业架构师

延伸阅读

有关更多信息，请参阅：

- AWS 架构完善的框架

文档修订

日期	描述
2020 年 4 月	更新以加入 CFM、新服务，同时与 Well-Architected 集成。
2018 年 7 月	更新以反映对 AWS 的更改，并加入与客户审核的过程中了解到的信息。
2017 年 11 月	更新以反映对 AWS 的更改，并加入与客户审核的过程中了解到的信息。
2016 年 11 月	首次发布