Amazon AWS 故障事件2020.1-2024.10

https://aws.amazon.com/cn/premiumsupport/technology/pes/

故障大类型分为：电力故障、系统硬件故障、系统软件错误、网络故障

序号	事件名称	事件描述	故障分析	故障原因	缓解与恢复过程	故障在系统中影响	故障归因
1	弗吉尼亚州北部（US-EAST-1）区域的Amazon Kinesis Data Streams服务事件摘要，2024 年 7 月 30 日	2024年7月30日，从下午2:45 PDT至晚上9:37 PDT，AWS北弗吉尼亚(US-EAST-1)区域的一些服务经历了**延迟增加和错误率上升的问题**。受影响的服务包括CloudWatch Logs、Amazon Data Firehose、Amazon S3事件框架、Amazon Elastic Container Service (ECS)、AWS Lambda、Amazon Redshift和AWS Glue。此次影响的根本原因是Amazon Kinesis Data Streams服务内部用于AWS服务的一个单元发生了性能退化。 Kinesis Data Streams在AWS服务架构中用于处理和存储来自多个来源的大量实时数据，例如CloudWatch Logs通过Kinesis来缓存日志流，再将数据持久化到CloudWatch存储中。Kinesis Data Streams使用分片化架构来确保高可用性和可扩展性。此次事件中，北弗吉尼亚区域一个仅供AWS内部使用的Kinesis单元发生故障，导致多个AWS服务直接或间接受到了影响。	该Kinesis单元的主机会定期发送状态信息给单元管理系统，以汇报每个分片的处理情况。然而，某些主机因处理过多的分片，状态信息体积过大，无法及时传输。这**导致管理系统延迟接收部分主机的状态信息，进而错误地判断这些主机不健康。** 管理系统错误地将这些主机的分片重新分配给其他主机，导致整个单元的分片重新分配频率大幅上升，进而导致Kinesis的数据传输出现问题，产生了延迟和错误。 随后，受影响的Kinesis单元中一个负责安全连接通信的组件超载，进一步影响了Kinesis数据流量的处理能力。	事件的根本原因是Kinesis单元管理系统在处理一个异常的工作负载时表现异常。具体而言，受影响的Kinesis单元内存在大量低吞吐量的分片（shards），这些分片的分布不均匀，导致部分主机承载了过多的分片。 在2024年7月30日上午9:09 PDT，该Kinesis单元开始例行部署，在此过程中，主机的上线与下线引发了问题，导致单元性能退化。 由于系统在部署时需要逐步将工作从下线的主机转移到其他主机，而单元管理系统没有有效地平衡这些低吞吐量分片的分布，导致少量主机承担了过多的负荷，进而导致状态信息传输不及时，触发了错误的健康监测判断，进一步导致负载重新分配加剧。	AWS工程团队在事件开始时立即介入调查，并采取措施减少内部工作负载的流量，以减轻管理系统的压力。通过增加处理容量并进行其他系统调整，最终在下午5:55 PDT，单元管理系统逐渐恢复，绝大多数请求在晚上7:21 PDT恢复正常。所有服务在晚上9:37 PDT恢复正常。	CloudWatch Logs：由于依赖Kinesis单元，CloudWatch日志请求处理出现延迟，直到7:21 PDT才大部分恢复正常，积压的旧日志在7月31日凌晨5:50处理完毕。 Amazon S3事件框架：从2:45 PDT到7:21 PDT事件处理延迟，积压的事件于8月1日凌晨2:38处理完毕。 Amazon Data Firehose：流数据的PutRecord和PutRecordBatch API调用失败率增加，延迟升高，但无数据丢失。 ECS和Lambda：使用awslogs日志驱动的ECS任务和Lambda函数日志处理受影响，某些ECS任务因无法健康检查被阻塞，Lambda客户无法获取函数执行日志。	**软件错误导致服务降效** **（分片体积设计不合理）→（延迟升高、错误率上升）**
2	弗吉尼亚州北部（US-EAST-1）区域的AWS Lambda服务事件摘要，2023 年 6 月 13 日	2023年6月13日，从上午11:49 PDT开始，AWS北弗吉尼亚（US-EAST-1）区域的客户在调用Lambda函数时遇到了**较高的错误率和延迟。**由于Lambda功能的退化，其他AWS服务如Amazon STS、AWS管理控制台、Amazon EKS、Amazon Connect和Amazon EventBridge等也受到了影响。Lambda函数调用的性能在下午1:45 PDT开始恢复，所有受影响的服务在下午3:37 PDT完全恢复。 Lambda使用了分片化架构，每个单元由多个子系统组成，以处理客户代码的函数调用。Lambda前端负责接收和路由客户函数调用，而Lambda调用管理器负责管理计算资源的分配。此次事件的根本原因是Lambda前端在应对流量增加时，触发了**一个潜在的软件缺陷**，导致计算资源虽然成功分配，但Lambda前端未能充分利用这些资源，从而造成函数调用的错误率上升和延迟增加。	AWS Lambda的架构是基于多个子系统的单元化设计。每个单元包括前端和调用管理器等关键组件。Lambda前端在处理请求时，触发了该系统的一个潜在缺陷，当服务流量超过了某个容量阈值后，系统未能正确利用已分配的计算资源。 由于前端无法利用这些资源，函数调用失败率增加。Lambda通过异步或流式事件源触发的函数调用也因此积压，无法及时处理。 事件触发后，工程团队立即着手调查，并通过缩减Lambda前端计算容量来缓解问题，从而避免再次触发该缺陷。	事件的触发是在Lambda前端应对北弗吉尼亚区域流量增加时，前端计算容量达到一个之前未触及的阈值。这触发了一个潜在的软件缺陷，导致Lambda执行环境被成功分配但未能被前端使用，致使Lambda无法为新的请求分配计算资源，进而导致错误率上升和延迟增加。 该缺陷影响了Lambda某个单元内的调用，而其他Lambda单元不受影响。	在11:49 PDT问题爆发后，工程团队在12:26 PDT确认了潜在缺陷的存在并发现其影响了计算资源的分配。随后，团队立即通过缩减Lambda前端的计算容量来避免触发该问题。在下午1:30 PDT，新的Lambda函数调用开始恢复，1:45 PDT，Lambda函数调用已全面恢复。所有异步事件的处理也在下午3:37 PDT前恢复正常。	Amazon STS：从11:49 PDT到2:10 PDT，Amazon STS服务出现了较高的错误率，特别是SAML联合登录过程中，新的身份验证会话受到影响，而现有的会话不受影响。 Amazon EventBridge：在11:49 PDT到1:45 PDT之间，EventBridge事件路由到Lambda时的延迟增加，最大延迟达到801秒。 Amazon EKS：在新集群的创建过程中出现了错误率上升和延迟增加，现有EKS集群未受影响。 AWS管理控制台：北弗吉尼亚区域的AWS管理控制台从11:48 PDT到2:02 PDT无法正常访问，用户看到“控制台不可用”或“504超时”错误页面。 Amazon Connect：从11:49 PDT到1:40 PDT，由于Lambda事件的影响，Amazon Connect的呼叫、聊天和任务处理失败，座席登录和使用Connect功能也受到影响。 AWS Support Center：支持中心的功能从11:49 PDT到2:38 PDT受影响，特别是前11分钟内，创建、查看和更新支持案例的请求可能会失败，呼叫和聊天功能不可用，但通过网页和邮件创建案例的功能不受影响。	**软件错误导致服务降效** （软件缺陷）→（延迟升高、错误率上升）

3	弗吉尼亚州北部(US-EAST-1)区域的AWS 服务事件摘要，2021 年 12 月 7 日	2023年12月的一次故障影响了AWS内部网络的通信，导致多项AWS服务出现延迟和错误，进而影响了使用这些服务的客户。AWS的内部网络托管着许多基础服务（如监控、内部DNS、授权服务以及部分EC2控制平面），与主要AWS网络相连。此次事件**源于一次自动的扩展活动**，该活动导致大量客户端的连接行为超出了内部网络和主要AWS网络之间的路由设备的承载能力，最终**造成通信延迟和网络拥堵**。 由于网络拥堵，内部监控系统的数据不可用，阻碍了AWS运维团队及时发现问题并采取行动。拥堵也导致DNS解析错误，团队在早期工作中试图缓解内部DNS流量问题，改善网络负载，然而这一措施并未完全解决拥堵问题。最终通过将部分流量隔离、禁用某些高流量服务并增加网络容量，AWS团队在下午1:34 PST时显著改善了拥堵，下午2:22 PST时所有网络设备完全恢复正常。	事件的根本原因是AWS一次内部服务的扩展活动引发了大量客户端连接行为，这超出了内部网络和主要AWS网络之间路由设备的处理能力，导致网络拥堵。 由于内部网络和主要AWS网络之间通信设备的性能问题，拥堵情况恶化，导致DNS解析错误和其他服务的通信延迟。	事件的根本原因是AWS一次内部服务的扩展活动引发了大量客户端连接行为，这超出了内部网络和主要AWS网络之间路由设备的处理能力，导致网络拥堵。 由于内部网络和主要AWS网络之间通信设备的性能问题，拥堵情况恶化，导致DNS解析错误和其他服务的通信延迟。	团队最初通过重定向DNS流量来缓解拥堵，减轻网络设备负载。虽然这一措施帮助恢复了一些服务，但问题并未完全解决。 随后，团队采取了一系列缓解措施，包括隔离高流量源、禁用部分服务以及增加网络设备容量，逐步恢复正常通信。 到下午1:34 PST，网络拥堵明显缓解，2:22 PST时所有网络设备完全恢复正常。	AWS控制平面：许多AWS服务的控制平面（如EC2、RDS、EMR、Workspaces等）依赖于内部网络，受此事件影响，EC2实例启动和管理API的错误率和延迟上升，直到下午2:40 PST才完全恢复。 Route 53 API：从7:30 AM PST到2:30 PM PST，Route 53的API功能受限，客户无法修改DNS条目，但已有的DNS解析未受影响。 AWS管理控制台：从7:33 AM PST到2:22 PM PST，北弗吉尼亚地区的AWS管理控制台访问出现错误，影响了用户登录和操作。 Amazon STS：由于内部网络问题，STS的身份验证服务出现延迟，特别是通过OpenID Connect (OIDC) 的第三方身份提供商登录受到影响，直到下午4:28 PST才完全恢复。 CloudWatch监控：整个事件期间，CloudWatch监控数据的延迟增加，部分监控数据丢失，客户难以实时了解其应用程序的影响。 API Gateway和EventBridge：API Gateway在事件初期受到影响，服务器无法与内部网络通信，直到下午4:37 PST逐步恢复。EventBridge的事件传递功能也在事件初期受到影响，下午6:40 PST恢复正常。 容器服务（ECS、EKS、Fargate）：虽然现有的容器实例未受影响，但新容器的启动因EC2 API问题受到影响。Fargate的API错误率在5:00 PM PST时恢复正常，但部分客户仍遇到容量不足问题，特别是“4 vCPU”任务的启动受限。 Amazon Connect：因API Gateway问题，Amazon Connect无法处理电话、聊天和任务请求，直到下午4:41 PST完全恢复正常。	**软件错误导致连接中断** （自动化的扩展活动）→ (内部网络通讯拥塞)
4	东京(AP-NORTHEAST-1)区域的AWS Direct Connect事件摘要，2021 年 9 月 2 日	2021年9月2日，AWS Direct Connect在东京（AP-NORTHEAST-1）区域发生了服务中断。从上午7:30 JST开始，Direct Connect客户经历了间歇性的连接问题和数据包丢失，影响了通往东京区域的流量。故障原因是**网络路径中一个网络层的一部分设备失效**，这些设备负责将Direct Connect边缘位置的流量转发至东京区域的数据中心网络。在12:30 JST，客户开始恢复正常连接，1:42 JST时问题完全解决。此事件没有影响到其他网络连接方式（如可用区间通信、互联网连接、AWS VPN连接等），并且Direct Connect通往其他AWS区域的流量也未受到影响。	AWS Direct Connect通过多层冗余网络设备来确保客户的私有连接。此次事件中的问题出现在某一网络层中，当多个设备同时失效时，流量未能正常转发，**导致拥堵和连接中断。** 该故障是由一个新的网络协议和网络设备操作系统中的潜在问题触发的。尽管AWS在引入新协议和操作系统时进行了广泛的实验室压力测试，但某些特定的流量和数据包组合在实验室环境中未能被完全模拟出来。因此，尽管此协议和操作系统已在生产环境中运行了8个月，问题仍然潜伏，直到特定的客户流量触发了此问题。	故障源于Direct Connect网络中的一个网络层的设备失效，这些设备负责将客户的流量从Direct Connect边缘位置转发到东京区域的VPC。设备的失效导致了数据包丢失和连接问题。 虽然AWS的自动化系统检测到了这些设备的异常，但由于设备未按照正常流程被移出网络，导致问题加剧。工程师在收到告警后，手动移除了故障设备并开始恢复流量，但由于更多设备出现相同的故障，导致网络拥堵和持续的连接问题。 在调查的过程中，工程师发现问题可能与一个几个月前引入的新协议有关。该协议用于优化网络在应对网络收敛事件和光纤切割事件时的反应速度。虽然这个协议已成功运行了几个月，但在特定流量模式下，与新协议的交互引发了故障。	工程团队首先移除了失效的网络设备并试图恢复正常流量，同时开始排查可能的原因。工程师怀疑该问题与新协议有关，并开始在某个可用区禁用此协议，观察恢复效果。 通过禁用新协议，网络逐步恢复稳定，至12:30 JST时，客户的连接逐渐恢复，1:42 JST时所有受影响的设备恢复正常。	停用新协议：此次事件发生后，AWS工程团队在东京区域禁用了触发问题的新协议，并计划将该更改扩展到其他区域以防止类似问题的发生。 协议缺陷修复：工程团队确认了网络设备操作系统中的潜在问题，并指出该问题需要非常特定的流量条件才能触发。虽然这种条件罕见，但他们已经开发了增强的检测和修复手段，确保可以在客户受到影响之前解决类似问题。 全球范围的防范措施：AWS正在对其他区域的设备进行同样的修复，确保未来不会再出现类似的问题。AWS对内部网络设备的操作系统和协议有着严格的部署流程，未来也将继续通过逐步部署和监控来保证稳定性。	**软件故障导致连接中断** （网络协议缺陷）→（系统中网络层设备失效）→ (流量未能正常转发，连接中断)

5	弗吉尼亚州北部(US-EAST-1)区域的Amazon Kinesis事件摘要，2020 年 11 月 25 日	2020年11月25日，AWS北弗吉尼亚（US-EAST-1）区域发生了与Amazon Kinesis相关的服务中断。Amazon Kinesis是一项实时流数据处理服务，广泛用于其他AWS服务。这次事件的触发点是凌晨2:44 AM PST开始的前端服务器容量增加操作，该操作在3:47 AM PST完成。事件发生时，Kinesis的“前端”服务器集群负责流数据的验证、路由和分片的分配。事件的初始表现为5:15 AM PST时，Kinesis记录读写错误报警触发。	每个Kinesis前端服务器需要与集群中的其他服务器通信，并维护分片映射表。当增加新的容量时，前端服务器需逐步了解新的集群成员并创建新的通信线程。在此次扩展中，线程数超过了操作系统允许的上限，导致服务器无法正确处理数据请求，出现高错误率。 故障的根源在于线程数超出操作系统限制，导致缓存构建失败，使前端服务器无法正确路由流数据请求。	事件的根本原因是Kinesis前端服务器集群的**操作系统线程数量超过了配置限制**。每台前端服务器都需要创建大量线程与其他前端服务器通信，在此次容量增加操作后，服务器线程数超出操作系统允许的最大线程数，导致缓存（shard-map）的构建失败，无法将请求正确路由到后端集群。 此外，由于前端服务器资源用于缓存构建和处理请求，过快重启服务器会导致资源竞争，进一步**加剧错误率**。	每个Kinesis前端服务器需要与集群中的其他服务器通信，并维护分片映射表。当增加新的容量时，前端服务器需逐步了解新的集群成员并创建新的通信线程。在此次扩展中，线程数超过了操作系统允许的上限，导致服务器无法正确处理数据请求，出现高错误率。 故障的根源在于线程数超出操作系统限制，导致缓存构建失败，使前端服务器无法正确路由流数据请求。	Amazon Cognito：由于Cognito使用Kinesis来处理API访问模式数据流，Kinesis的中断导致Cognito的缓冲区出现问题，用户池和身份池API出现高错误率和延迟，影响用户身份验证和AWS临时凭证获取。错误率在7:01 AM PST显著增加，Cognito团队通过部署缓解措施于12:15 PM PST恢复正常。 CloudWatch：CloudWatch的PutMetricData和PutLogEvents API在5:15 AM PST开始出现错误率和延迟增加，导致部分警报进入“数据不足”状态。Kinesis恢复后，CloudWatch于10:31 PM PST恢复正常。 Lambda：由于Lambda函数调用依赖于CloudWatch进行指标发布，在6:15 AM PST开始，由于CloudWatch不可用，Lambda的本地缓冲导致内存竞争并出现调用错误。工程团队在10:36 AM PST缓解了此问题。 CloudWatch Events 和 EventBridge：这些服务在事件开始时API错误率上升，Kinesis恢复后逐步清理事件积压，影响于4:15 PM PST得到解决。	**软件错误导致服务失效** （线程数超过配置限制）→（资源争用，错误率上升）