Microsoft Azure 故障事件2020.1-2024.10

近5年 2020至今 Azure status history. https://status.azure.com/en-us/status/history/.

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	影响时长	事件描述	故障分析	故障原因	故障在系统中的影响	故障归因
1	Post Incident Review (PIR)-Storage - Impact to multiple services in Central US Tracking ID: 1K80-N_8		从2024年7月18日21:40 UTC至2024年7月19日22:00 UTC，Azure美国中部区域的多个服务由于Azure存储可用性问题出现了服务中断，影响持续时间为大约24小时。这次事件影响了虚拟机（VM）的可用性，进而对依赖Azure存储的多个Azure服务（包括Azure DevOps、SQL数据库、Cosmos DB、Azure IoT Hub等）造成了下游影响，导致服务不可用、连接问题和服务管理失败。	此次问题的根本原因是Azure存储单元的“允许列表”更新过程中的错误。虚拟机使用由Azure存储支持的持久性磁盘。为了确保安全，Azure存储单元仅接受来自虚拟机主机的特定网络地址范围的读写请求。由于虚拟机主机的动态变化，这些地址信息定期更新，并通过“允许列表”发布到所有存储单元。 2024年7月18日，在更新虚拟机主机地址时，负责生成该“允许列表”的服务器无法访问某些数据中心的源文件，导致生成的列表不完整。存储单元由于缺少虚拟机主机地址信息，开始拒绝虚拟机的磁盘请求，导致虚拟机可用性下降。这次更新没有遵循安全部署实践（SDP），比如可用区排序，从而导致整个区域受到广泛影响。	• **虚拟机**：虚拟机磁盘请求被拒绝，导致VM不可用或连接失败。 • **Azure服务**：包括Azure DevOps、Cosmos DB、SQL数据库、Azure IoT Hub、Microsoft Sentinel等服务出现连接问题和操作失败。 • **Microsoft云服务**：如Microsoft 365、Dynamics 365、Microsoft Entra，因依赖Azure服务也受到了影响。		**软件错误导致某些服务间的断联**（功能错误）
2	Post Incident Review (PIR) - Azure Resource Manager - Control plane issues in China North 3 Tracking ID: HVZN-VB0		2024年9月5日08:00至21:30（中国标准时间），持续13h30min，Azure中国区的客户在执行控制平面操作（例如创建、更新和删除资源）时遇到了显著的延迟问题，特别是在**中国北部3区**，影响最为严重。此次延迟导致了Azure Resource Manager（ARM）管理的各种服务出现超时和操作失败，影响了多个Azure服务，包括Azure Databricks、Azure Data Factory、Azure Kubernetes Service等。	此次故障的根本原因是**证书轮换**过程中，Azure Resource Manager跟踪后台操作的进程发生了多次崩溃。Azure Resource Manager是用于管理Azure资源的控制平面，当执行较长时间的操作（如创建虚拟机）时，这些操作需要后台跟踪和处理。 在证书轮换时，用于加密后台作业的证书也会被更新，所有正在处理的作业需要重新加密。然而，证书轮换暴露了更新过程中存在的潜在代码错误，导致相关进程崩溃，进而造成后台作业延迟并逐渐积压。	受影响的客户可能在以下服务中遇到操作失败、延迟或超时： 1. **Azure Databricks**：提交作业请求时可能遇到错误或失败。 2. **Azure Data Factory**：运行数据流活动或获取调试会话时，可能遇到内部服务器错误。 3. **Azure Database for MySQL**：创建、更新、删除数据库时，操作无法按预期完成。 4. **Azure Event Hubs**：读取或写入操作可能遇到响应缓慢。 5. **Azure Firewall**：更新策略时可能遇到延迟。 6. **Azure Kubernetes Service (AKS)**：集群管理操作（如扩展、更新、创建/删除集群）遇到延迟。 7. **Azure Service Bus**：读取或写入操作时可能遇到延迟。 8. **Microsoft Purview**：在创建、更新或删除资源时遇到延迟。 9. **Azure Resource Manager的其他服务**：如果使用的是中国北部3区中的资源组，服务管理操作可能失败。		**软件错误导致某些服务延迟升高**（更新代码错误）（服务可用性降低）
3	Post Incident Review (PIR) - Azure OpenAI Service - Errors in multiple regions Tracking ID: 4L44-3F0		2024年7月13日00:00 UTC至2024年7月19日22:00 UTC，Azure OpenAI (AOAI)服务的部分关键资源在清理操作期间被错误删除，导致跨多个区域的服务中断。此次事件影响了14个提供AOAI的区域，包括**Australia East、Brazil South、Canada Central、East US 2**等，客户在调用Azure OpenAI端点时可能遇到5xx错误。新的标准（按需付费）精调模型部署也在部分区域（如East US 2、Sweden Central等）无法使用，持续至2024年7月15日16:54 UTC。	Azure OpenAI服务依赖于一个内部自动化系统来管理Azure资源，包括用于托管OpenAI大型语言模型的GPU虚拟机和用于处理推理请求的Azure Machine Learning工作区。此次事件的根本原因在于自动化服务的全局配置文件中的不一致性。配置文件中的14个资源组被错误地标记为不再需要，尽管这些组中实际上仍包含关键资源，如GPU虚拟机和模型端点。 为了解决订阅限制并减少安全漏洞，团队在清理过程中删除了这些资源组，导致包含关键资源的子资源也被删除，进而触发了多区域的服务故障。由于此次更改是全局范围内发布的，没有采用逐区域的安全部署实践（SDP），导致事件在多个区域中同时发生。	• **受影响的服务**：Azure OpenAI服务在14个区域的客户可能遇到5xx错误或无法访问其OpenAI资源。 • **模型部署问题**：新的按需付费精调模型部署在4个区域（如East US 2、Sweden Central等）不可用。 • **下游影响**：包括虚拟机、Azure机器学习工作区等依赖这些资源的其他服务可能也受到了间接影响。		**软件错误导致服务失效**（人为维护出错）

4	Post Incident Review (PIR) - Network connectivity - Multiple services impacted in South Africa North and South Africa West Tracking ID: VT60-RPZ	2024年3月14日10:33 UTC至2024年3月15日11:00 UTC，使用**Azure南非北部**和**南非西部**区域的客户可能遇到网络连接失败、延迟增加或数据包丢失的问题。此次事件是非洲大陆多条海底光缆（包括WACS、MainOne、SAT3、ACE等）和部分陆地光纤线路的多次并发中断造成的，导致Azure在南非区域的多个服务受影响，包括Azure API管理、Azure虚拟桌面、Azure数据库和Microsoft Entra ID等服务。	Azure的南非区域依赖多条海底光缆和陆地光纤线路将其网络与全球其他地区连接。此次事件中，非洲西海岸多条关键的海底光缆发生了并发中断，再加上光缆中断后一个**Microsoft路由器的线路卡光纤模块**故障，导致备用网络路径的容量被耗尽，产生了网络拥堵，进而导致延迟和数据包丢失。 这次海底电缆的中断是由于**地震或海底滑坡**等自然灾害引发的，并且在此事件前的2月，东非海岸的光缆已经因类似的中断丢失了容量。这一系列事件导致南非Azure区域的网络冗余机制失效，并引发了广泛的服务影响。	• **网络连接问题**：客户在访问南非北部和南非西部区域资源时可能经历网络连接失败、延迟增加或数据包丢失。 • **Azure服务影响**：包括但不限于Azure API管理、Azure虚拟桌面、Azure SQL数据库、Azure Key Vault、Microsoft Entra ID等服务在南非区域受到影响，可能导致服务可用性下降、超时或请求失败。 • **Microsoft 365等服务影响**：由于Microsoft 365和其他Microsoft服务依赖Azure区域，部分服务也可能受此事件影响。	**网络连接中断导致服务降效** （延迟升高/丢包率上升）
5	Post Incident Review (PIR) - Services impacted by power, BIOS, and Virtual Machine issues - East US Tracking ID: 2LZ0-3DG	2023年9月16日07:24至19:00 UTC，使用**虚拟机（VM）的Azure东部美国区客户遇到连接问题。这起事件由一个可用区内的数据中心的多个扩展单元断电**引发，导致虚拟机节点重启。虽然大多数节点成功重启，但部分节点未能自动恢复，进而影响了多个依赖这些虚拟机的服务，包括**SQL数据库**、**Service Bus**和**Event Hubs**。受影响的客户经历了服务中断或性能下降，恢复工作从16:30 UTC开始，至19:00 UTC完全恢复。	故障由**不间断电源（UPS）整流器**故障引发，该UPS为三个静态转换开关（STS）供电。这些开关负责在独立冗余电源之间无缝切换。当UPS整流器故障后，STS切换到备用UPS，但由于主UPS恢复并短暂提供不稳定的电源，STS再次切换回主UPS，导致主UPS彻底故障。由于多次电源切换的逻辑限制，STS未能及时切换回冗余电源，导致部分扩展单元短暂断电1.9秒。 断电后，一些虚拟机未能成功重启，这是由于**BIOS软件中的已知漏洞**导致部分节点无法连接到PXE服务器完成引导。最初的故障排除误导了工程团队，认为问题与网络或PXE服务器相关，导致恢复延迟。此外，团队尝试强制重启节点，但受制于内部审批流程，进一步拖延了恢复进度。	• **虚拟机（VM）**：部分虚拟机离线，导致应用和服务中断，恢复工作从16:30 UTC逐步开始，至19:00 UTC完成。 • **SQL数据库**：依赖虚拟机的SQL数据库出现中断，使用自动故障转移的客户经历了大约8小时的停机。 • **Service Bus和Event Hubs**：非区域冗余部署的客户服务性能下降或中断。	**电力故障导致服务失效** （部分虚拟机离线、数据库连接中断）
6	Post Incident Review (PIR) - Services impacted after power /cooling issue - Australia East Tracking ID: VVTQ-J98	2023年8月30日08:41至2023年9月3日20:00 UTC，**澳大利亚东部区域**的Azure服务因电力波动导致冷却系统故障，影响了多个Azure、Microsoft 365和Power Platform服务。此次事件始于电力波动导致13台冷却系统的冷却机组停止运行，其中5台冷却机组无法手动重启，导致数据中心温度升高，部分计算和存储单元自动关闭。尽管大部分服务在2023年8月30日22:40 UTC恢复，但部分存储、SQL数据库和Cosmos DB服务恢复延迟，完整缓解于9月3日20:00 UTC。	此次故障由**电力波动**引发，原因是距离澳大利亚东部区域约18英里的电力基础设施受到**雷击**。电压下降导致多个数据中心的冷却系统冷却机组关闭，其中一些自动重启失败。由于水循环系统中的水温过高，最后5台冷却机组无法重启，导致两个数据大厅温度上升，部分网络、计算和存储基础设施自动关闭，以防止硬件损坏。 在冷却系统恢复之前，数据中心团队关闭了受影响的数据大厅中的剩余基础设施，以防止温度进一步升高。随着冷却系统恢复，基础设施逐步恢复上线，但存储和SQL数据库等服务恢复时间较长，特别是受影响的存储单元和SQL数据库中的一个租户环的恢复最为复杂。	• **Azure服务**：包括Azure Active Directory、SQL数据库、Cosmos DB、Azure Kubernetes Service、Azure Service Bus、Azure虚拟机等多个Azure服务受到影响，客户可能经历服务不可用、延迟增加或请求失败。 • **SQL数据库**：大多数区域冗余的SQL数据库未受到影响，但部分使用代理模式连接的客户因缺乏区域冗余的连接网关而遭遇中断。 • **Cosmos DB**：区域冗余账户和多区域写入账户保持正常运行，但非区域冗余账户则因基础设施关闭而失去可用性。 • **Azure Kubernetes Service (AKS)**：服务的控制平面依赖的SQL数据库未配置区域冗余，导致AKS服务受到影响。	**电力故障导致服务降效、失效** （设备温度升高，数据大厅温度上升，系统物理设备关闭，服务不可用或降效）
7	Post Incident Review (PIR) - Azure Monitor - Logs data access issues Tracking ID: XMGF-5Z0	2023年7月6日23:15 UTC至2023年7月7日09:00 UTC，Azure Monitor Log Analytics和Microsoft Sentinel的部分数据未能成功摄取。此外，平台日志通过诊断设置发送到客户目的地（如Log Analytics、Storage、Event Hub和Marketplace）时也出现部分数据丢失。此次事件是由于Microsoft内部服务部署中的一个错误引发的，导致遥测管理控制平面超负荷，影响了所有地区的客户，尤其对Sentinel的安全运营中心（SOC）功能产生了影响。	2023年7月3日，**Azure容器应用服务**的一次代码部署开始，通过常规的**安全部署实践（SDP）**进行，首先在Azure的金丝雀和预生产环境中发布。此版本的配置错误导致服务无法正常启动，服务进入每5到10秒自动重启的循环。每次重启时，服务会触发遥测代理重新获取配置，导致代理也频繁重启。 由于遥测代理的频繁重启，数百个主机的代理不断请求遥测控制平面，远超预期负载，最终耗尽了控制平面的容量，影响到其他依赖该平面的服务，导致数据摄取和路由失败。遥测控制平面作为单点故障的风险已知，Azure Monitor团队正努力消除此风险。	• **Azure Monitor Log Analytics和Microsoft Sentinel**：部分日志数据未能成功摄取，导致查询结果不完整或为空，影响检测、分析、狩猎查询等功能。 • **平台日志**：通过诊断设置发送到客户目的地的日志数据未能成功路由，影响了日志的存储和分析。 • **安全运营中心（SOC）**：Sentinel用户可能在事件调查时遇到不完整的数据，导致安全事件分析受限。	**软件错误导致部分服务失效** （配置错误）→（管理平面过载）→（服务失效）

8	Post Incident Review (PIR) - Azure Resource Manager - West Europe Tracking ID: RNQ2-NC8	2023年3月23日02:20至07:30 UTC期间，Azure资源管理器（ARM）在**西欧地区**的资源管理操作出现问题，影响了使用**Azure CLI**、**Azure PowerShell**和**Azure门户**的用户以及依赖ARM进行资源管理的Azure服务。在西欧地区的ARM API调用失败率高达50%，约占全球请求的3%。主要受影响的是靠近西欧地区的用户和工作负载，VPN用户和某些托管网络上的用户也可能受到影响。	该事件由**正反馈循环**导致ARM Web API层**饱和**引发。问题源于一次高频锁竞争，该竞争阻塞了API请求路径，导致CPU负载显著增加，后台工作线程无法正常处理。这导致长时间运行的异步操作（如数据库和Web请求）的延迟增加，超时问题引发客户和内部服务的**重试请求**，进而加剧负载，最终饱和了ARM Web API层的可用CPU容量。 事件的触发因素是最近引入的一个**缓存机制**，用于减少复杂功能标志解析的时间。该机制在西欧部署后，被某个**内部服务**的独特调用模式触发了高频锁竞争。该服务执行的每日缓存刷新任务产生了大量请求，导致正反馈循环，最终导致ARM Web API层的CPU资源耗尽。	• 受影响的客户在**西欧区域**的资源管理请求失败率高达50%。 • 使用**Azure CLI**、**Azure PowerShell**和**Azure门户**执行的资源管理操作可能会失败。 • 依赖ARM API的Azure服务，如**Azure存储**和**Azure SQL数据库**等，在内部操作中也可能受到影响。	**软件错误导致服务降效** （部分服务请求失败率达50%）
9	Post Incident Review (PIR)- Azure Storage - West Europe Tracking ID: R 36-P80	2023年3月6日03:50 UTC至17:55 UTC期间，部分使用Azure Storage的客户在对位于**西欧地区**的存储资源执行请求时遇到了比预期更多的**限流**问题。受此影响的服务包括Azure Automation、Azure Arc enabled Kubernetes、Azure Bastion、Azure Batch、Azure Container Apps、Azure Data Factory (ADF)、Azure ExpressRoute、Azure HDInsight、Azure Key Vault (AKV)、Azure Logic Apps、Azure Monitor、Azure Synapse Analytics等，导致这些服务的请求失败或性能下降。	Azure Storage使用**限流机制**来确保存储账户的使用保持在发布的存储账户限制内，同时保护存储服务的扩展单元不超过其资源限制。当某个扩展单元达到限制时，系统会对负载过高的存储账户进行限流以平衡负载。此次事件是由于**限流算法**的一次**配置更新**引发的。该配置在部署过程中触发了部分扩展单元的异常限流，导致一些存储账户受到过度限流，进而影响相关Azure服务。 尽管该配置更新遵循了常规的**安全部署实践**，但在部署到西欧部分扩展单元时，由于负载特性与测试场景不同，触发了意外的限流问题。	• 受影响的客户在**西欧地区**的存储请求出现更高的限流。 • **依赖Azure Storage的服务**，如Azure Automation、Azure Batch、Azure Data Factory等，出现间歇性请求失败和性能下降问题。 • **Azure Key Vault**的创建、读取、更新和删除操作受限，部分请求受到影响。	**软件错误导致服务降效** （配置错误、软件设计错误）
10	Post Incident Review (PIR)- Multi-service outage- Asia-Pacific Area Tracking ID: VN11-JD8	2023年2月7日20:19 UTC至2023年2月9日04:30 UTC 部分部署在**东南亚和东亚区域**的Azure客户在访问和管理资源时遇到问题。此次故障源自东南亚区域中的一个**可用性区域（AZ）**发生的冷却系统故障，导致基础设施关闭以保护数据和硬件。这导致了资源不可用，且涉及的服务未能按预期进行故障转移。	## 1.东南亚区域的冷却系统故障 在2023年2月7日15:17 UTC，东南亚某个AZ经历了**电压波动**，导致一部分冷却设备（冷却器）故障。虽然电力系统按预期管理了波动，但冷却器未能自动重新启动，甚至在执行手动重启操作时也未恢复正常。这最终导致了数据中心温度持续上升，并于20:19 UTC自动关闭了计算、网络和存储基础设施以保护数据和硬件。 问题出在冷却器的**压缩机控制卡**，它在电压波动后停止响应，无法正常重启。虽然冷却器制造商的技术人员进行了修复，但当时冷却水温度已经超出冷却器允许的重启阈值，导致**热锁定**。为了恢复温度，数据中心不得不关闭部分基础设施以减少负载，最后在2023年2月8日14:00 UTC时冷却器全部恢复正常，温度也回到了正常范围。 ## 2. ARM控制平面及BCDR服务的意外影响 尽管部分服务部署在**区域冗余配置**中未受到重大影响，但**Azure资源管理器（ARM）控制平面受到了意外的冲击，尤其是其依赖于受影响区域中的Cosmos DB实例**。由于配置错误，该Cosmos DB实例未按预期进行区域冗余，这导致了ARM服务的元数据无法访问，最终影响了整个东南亚区域的ARM控制平面。 为缓解影响，工程师将东南亚区域的ARM流量重定向到其他区域，但这引发了东亚区域的流量过载，进而影响了东亚区域的ARM服务。 另外，**业务连续性和灾难恢复（BCDR）**服务（如Azure站点恢复、Azure备份、使用地理冗余存储的存储账户）也未能按预期执行冗余切换，进一步加剧了影响。	• 部署在东南亚区域的服务无法访问，包括：Azure App Services, Azure Kubernetes Service, Cosmos DB, SQL Database, Storage等。 • 区域冗余服务未按预期故障转移，包括Azure站点恢复、Azure备份等。 • SQL数据库客户在数据库不可用期间经历了**长达数小时**的停机，部分客户手动进行区域切换或使用自动故障转移机制恢复服务。 ## 故障响应 1. **02月7日15:17 UTC**: 发生电压波动，部分冷却器停止运行。 2. **02月7日20:19 UTC**: 基础设施过热，自动关闭计算、存储资源。 3. **02月8日04:10 UTC**: 为减轻影响，ARM控制平面流量从东南亚重定向到东亚区域。 4. **02月8日14:00 UTC**: 数据中心温度恢复正常，冷却器重新启动。 5. **02月8日16:30 UTC**: 大部分服务恢复正常。 6. **02月9日04:30 UTC**: 最后一个SQL数据库Tenant Ring恢复，故障完全解决。	**电力故障导致服务失效** **软件错误导致服务失效**

11	Post Incident Review (PIR)- Service management issues - East US 2 Tracking ID: BS81-390	2023年1月31日05:55 UTC到2023年2月1日00:58 UTC期间，东美国2区域（East US 2 region）中的一部分用户在进行虚拟机（VM）的服务管理操作时（如创建、删除、更新、扩展、启动、停止等）遇到了错误通知。受影响的范围仅限于该区域三个可用区（Availability Zones）中的一个物理可用区（AZ-02），其他两个可用区的虚拟机没有受到影响。	故障发生在东美国2区域中的一个可用区（AZ-02）。每个可用区都被进一步划分为多个分区，以确保一个分区的故障不会影响其他分区的处理。虚拟机的部署请求由可用区内的某个分区处理，而网关服务负责将流量路由到正确的分区。 **问题原因：** 1. **资源耗尽问题**：2023年1月31日05:55 UTC时，AZ-02中的一个分区由于底层数据存储的内部资源限制耗尽，出现了数据访问问题，导致第一个分区出现故障。 2. **分区不可用**：在调查第一个分区问题的过程中，11:43 UTC时同一可用区的第二个分区也出现了类似问题并且变得不可用。然而，即使分区不可用，缓存服务仍然持有该分区的数据，新部署在此期间仍然可以成功。 3. **缓存服务重启问题**：12:04 UTC时，缓存服务重启后无法检索数据，因为目标分区已经宕机。此时，由于资源创建策略的配置问题，所有分区在网关服务中都需要创建新的资源。缓存服务缺少数据后，这一策略导致了所有新的虚拟机创建请求被阻塞，从而进一步导致更多失败和服务延迟，影响了下游服务。	• **分区的资源耗尽**：第一个分区由于底层数据存储中的资源限制耗尽，导致数据访问失败，触发了部分虚拟机服务管理操作的错误。 • **缓存服务的重启失效**：第二个分区故障后，虽然缓存最初维持了数据，但缓存服务重启后无法恢复数据，而系统的资源创建策略要求从头创建新资源，导致了更广泛的虚拟机创建阻塞。 • **下游服务依赖故障**：由于虚拟机管理操作的中断，依赖于虚拟机服务的多个Azure下游服务也相继出现故障。	软件错误导致服务降效，以及一些列的服务失效 （存储资源耗尽）→ (资源耗尽)
12	Post incident Review (PIR) - Enrolling new certificates / Provisioning new resources - Azure Public / Government /China Tracking ID: YTGZ-1Z8	从2022年11月2日00:42 UTC到11月3日05:55 UTC期间，部分Azure服务的客户在尝试创建新资源时遭遇失败。受影响的服务包括：应用网关（Application Gateway）、堡垒机（Bastion）、容器应用（Container Apps）、数据库服务（MySQL弹性服务器、Postgres弹性服务器等）、ExpressRoute、HDInsight、Open AI、SQL托管实例（SQL Managed Instance）、流分析（Stream Analytics）、VMware解决方案以及VPN网关（VPN Gateway）。这些服务在创建新资源时需要生成新的证书，而处理新证书请求的证书注册机构（RA）在此期间发生了服务降级，导致无法在这些服务中创建新资源。	在2022年11月1日23:56 UTC到11月2日00:52 UTC期间，Azure内部的一个证书授权机构（CA）经历了短暂的服务降级。同时，证书注册机构（RA）收到了一波证书续期请求，导致这些请求在RA端积压。虽然证书授权机构恢复了服务，开始处理积压的请求，但由于RA内部存在一个性能问题，新的请求速率超过了RA处理队列中请求的速度。 **具体原因分析：** 1. **性能缺陷**：这一性能问题是由于前一个月引入的功能增强中无意间引入的。RA处理请求的速度低于新请求的进入速度，导致RA服务出现自动限流（throttling），从而进一步导致了更多的重试请求，形成恶性循环。 2. **测试漏洞**：这个潜在的性能缺陷在部署和健康监控过程中未被捕捉到，这是因为测试过程中未覆盖到这种特定的负载和条件。	**故障产生原理：** 1. **证书请求积压**：由于证书授权机构（CA）短暂降级，RA端的证书请求开始积压。积压的请求量在CA恢复后过于庞大，超出了RA的处理能力。 2. **性能缺陷触发限流**：RA的性能缺陷导致处理速率下降，新请求不断涌入。系统自动限流，进一步导致请求重试，增加了RA的负载，RA无法恢复到正常状态。 3. **测试缺陷导致问题未能提前发现**：这一系列问题在部署前未被发现，是因为测试过程中未考虑到这种特定的高负载场景和自动重试的相互作用。	**软件错误导致服务降效** （证书请求和请求重试机制堆积）
13	Post Incident Review (PIR) - Canonical Ubuntu issue impacted VMs and AKS Tracking ID: 2TWN-VTO	从2022年8月30日06:00 UTC到8月31日16:00 UTC期间，启用了**自动升级（Unattended-Upgrades）的Ubuntu 18.04（bionic）虚拟机（VMs）用户遇到了DNS解析错误**的问题。此次问题影响了所有Azure区域，包括公共云和主权云。受此影响的下游Azure服务包括Azure Kubernetes服务（AKS）、Azure监控（Azure Monitor）、应用洞察（Application Insights）、日志分析（Log Analytics）和Microsoft Sentinel。具体影响包括AKS客户可能遇到Pod创建失败（如ImagePullBackoff），因为kubelet无法解析容器注册表的DNS名称。	问题源于2022年8月30日06:00 UTC时发布的Canonical Ubuntu安全更新。Azure上运行Ubuntu 18.04且启用了自动升级的虚拟机开始下载并安装新软件包，包括**systemd版本237-3ubuntu10.54**。该版本的systemd引入了一个**竞争条件bug**（race-condition bug），导致虚拟机的DNS配置丢失。由于启用了自动升级，安全更新默认每天自动下载并应用。这些安全更新因其重要性，没有经过Azure的**安全部署实践（Safe Deployment Practices, SDP）**流程。因此，此次更新未经过充分的测试和验证。	**故障产生原理：** 1. **systemd更新引发竞争条件bug**：Ubuntu发布的systemd版本引入了竞争条件错误，导致DNS配置丢失。 2. **DNS解析失败**：Azure上的Ubuntu 18.04虚拟机失去DNS解析能力，网络连接中断。 3. **下游服务受影响**：多个Azure下游服务（如AKS、Azure Monitor等）依赖这些受影响的虚拟机，因此也遇到了类似的DNS解析问题，导致服务中断或警报延迟。	**软件错误导致连接中断** （系统自动更新导致DNS配置丢失，影响网络连接）

14	Post Incident Review (P|R) - Datacenter power event - West US 2 Tracking ID: MMXN-RZO	从2022年8月27日02:47 UTC到8月28日02:00 UTC期间，部分客户在尝试访问位于**西美国2区域（West US 2）的数据资源时遇到了访问失败的问题。事件最初是由该区域的公用电力中断引发的，尽管大多数数据中心的备用电力系统按设计运行以防止影响，但两个数据中心**的部分备用电力系统出现故障，导致客户受到影响。大部分客户在8月27日07:00 UTC前恢复正常，但部分资源需要手动恢复，直到8月28日02:00 UTC才完全恢复。	此次事件的触发点是8月27日02:47 UTC时发生的一次**公用电力中断**，影响了西美国2区域的多个数据中心。该区域由10多个数据中心组成，分布在三个可用区（Availability Zones）。具体情况如下： 1. **电力故障起因**：高压静电线的故障导致230千伏输电线路上出现电压激增，导致两个变电站的断路器打开，从而导致大面积的公用电力中断，影响了整个西美国2区域。 2. **备用电力系统表现**：大多数数据中心的备用电力系统按设计切换到了电池供电，然后迅速转为发电机供电。但是，在**两个数据中心**，由于**两个独立且不相关的备用电力系统问题**，部分服务器未能成功切换到发电机供电，导致影响： • 在第一个数据中心，少量服务器架的**不间断电源（RUPS）**系统在切换到发电机时失败，导致服务器短暂断电。 • 在第二个数据中心，约12%的**主UPS系统**的电池出现故障，无法在切换到发电机时支撑负载，导致服务器断电，直到UPS故障清除并重新上线。	• **高压静电线故障**导致电压激增，影响变电站的电力供应，进而导致整个西美国2区域内的所有数据中心失去公用电力。 • **备用电力系统问题**：虽然大多数数据中心成功切换到备用电源，但两个数据中心的不同UPS系统故障，导致部分服务器断电，进而影响到客户资源的可用性。 • **Azure服务依赖故障**：多个依赖受影响基础设施的Azure服务如存储、虚拟机等也因此受到了波及，导致更多的客户服务中断。	**电力故障导致服务失效** （多个数据中心的物理设备断电，承载的服务失效）
15	Post Incident Review (PIR) - Wide Area Network - Multiple Regions Tracking ID: YKDK-TT8	2022年6月29日02:40 UTC至20:14 UTC期间，部分客户因**Microsoft广域网（WAN）位于伦敦地区的路由器问题，经历了间歇性网络故障和数据包丢失问题。问题源于伦敦地区的某个WAN路由器发生了部分硬件故障，导致该路由器处理的流量中约0.4%的数据包路由不正确。这些问题影响了通过该路由器的客户流量，尤其是那些位于英国西部（UK West）、英国南部（UK South）、欧洲西部（Europe West）和欧洲北部（Europe North）**区域的流量。	经过调查，问题被归结为位于英国伦敦的一个WAN路由器的部分硬件故障。具体表现为该路由器的某个线卡（line card）出现问题，导致约0.4%的数据包在传输过程中丢失。这对流向英国西部、英国南部、欧洲西部和欧洲北部区域的部分客户流量造成了间歇性的网络故障。 由于此问题是间歇性发生的，被称为“灰色故障”（gray failure），因此系统的自动警报未能立即识别到问题所在。实际上，这台故障路由器的140个物理端口中只有4个端口发生了故障，这导致路由器间歇性地部分失效，但未生成任何可操作的错误信息。 对于使用Azure VPN网关的客户，网络流量在到达Azure VPN网关之前会经过这些WAN路由器。因此，WAN核心路由器的间歇性数据包丢失导致了Azure VPN的连接中断。	**故障产生原理：** 部分硬件故障：伦敦区域的WAN路由器的一部分硬件（4个物理端口）出现了问题，导致约0.4%的数据包被错误处理或丢失。 间歇性问题：由于该问题是间歇性发生且影响范围较小，导致自动警报未能及时识别出问题的严重性。 Azure VPN连接问题：Azure VPN网关客户的网络流量通过这些路由器时，数据包丢失导致连接中断或性能下降。	**硬件故障导致连接中断** （底层物理设备故障）→ (丢包升高)
16	Post Incident Review (P|R) - Azure Software Load Balancer - Multiple Regions Tracking ID: YVTL-RSO	从2022年6月28日05:26 UTC至7月1日04:00 UTC，部分客户在连接软件负载均衡器（SLB）后面的资源时，经历了间歇性网络故障和数据包丢失。 受影响的客户在连接其资源时，经历了间歇性的连接失败和数据包丢失。	软件负载均衡器（SLB）是Azure数据中心和集群内管理网络流量的软件路由和负载均衡服务。SLB的核心是多路复用器（MUX）服务器，这些服务器组成一个SLB的扩展单元。每个扩展单元托管许多虚拟IP地址（VIPs），通过这些VIPs将流量路由到后方的目标IP地址（DIPs）。MUX通过定期探测DIPs的健康状况来动态监控和分配流量。	经过调查，问题的根本原因是最近一次代码部署中引入的变更，导致某些MUX实例存储后端DIPs的顺序发生了变化。此变更仅影响拥有超过15个DIPs的负载均衡器。DIPs的顺序对流量如何选择后端服务器至关重要，因为它决定了具体的连接如何分配给后端服务器。这种顺序的变化影响了流经多个MUX的流量，主要体现在两个场景中： Azure防火墙客户：Azure防火墙是一个有状态防火墙，依赖于内部的负载均衡器来扩展处理流量。当有超过15个DIPs时，流量被固定分配给特定的防火墙实例进行处理。然而，由于负载均衡器问题，同一流的数据包被分配到了不同的实例，导致数据包丢失和流量中断。 服务终端或私有链接终端连接：由于代码变更，IPv6数据包路由头中的流标识符发生了变化。当流量通过多个MUX时，新MUX将连接转发到不同的后端服务器，导致该请求被拒绝，从而引发数据包丢失。	软件错误导致连接中断 （代码错误）→ (数据包丢失)

17	Post Incident Review (PIR)-Datacenter cooling event - East US 2 Tracking ID: NMB2-NDO	2022年6月7日02:41至14:30 UTC期间，部分客户在连接位于东美国2区域（East US 2）某个可用区（Availability Zone, AZ）中的资源时遇到了困难。此次事件影响了该区域三大可用区之一中的部分存储和计算资源，进而影响了依赖这些资源的Azure服务。虽然许多支持可用区的服务没有受到影响，但五个服务（Application Insights、Log Analytics、Managed Identity Service、Media Services、NetApp Files） 由于尚未完全支持可用区，因此受到了区域性影响。	此次事件的根本原因是东美国2区域某个可用区的数据中心中发生了一次意外的电力振荡。问题源于不间断电源（UPS）模块的异常行为，导致电力系统中出现了电气瞬变，这进一步导致空气处理单元（AHUs）检测到潜在故障，并自动关闭以防止损坏。	电气瞬变和冷却系统中断：UPS模块的异常振荡导致整个数据中心的电力分配系统中出现了电气瞬变，影响了包括机械冷却系统的电力供应。空气处理单元（AHU）自我保护性地关闭，导致数据中心的冷却系统中断。虽然计算、网络和存储基础设施未受到直接电力中断的影响，但由于冷却系统的关闭，数据中心的温度上升，触发了一部分IT基础设施的自我保护关闭。 保护性关机：随着温度上升，部分计算、存储和网络设备被自动关闭以防止硬件损坏并保护数据完整性。尤其是存储基础设施受到了严重影响，导致依赖这些存储的服务出现可用性问题。 存储基础设施和虚拟机影响：受热保护关机和网络连接问题影响的存储单元中，有八个存储扩展单元（包括LRS/GRS冗余存储账户）受到重大影响。这些扩展单元中的虚拟机（VMs）使用了受影响的标准硬盘（Standard HDD）存储，导致虚拟机无法访问其虚拟磁盘。当VM与存储的输入/输出操作在120秒内无法完成时，系统会将虚拟机标记为失去连接，并触发临时关闭。任何依赖这些虚拟机运行的工作负载，包括Azure自身的服务和客户的第三方服务，都会受到影响。	**电力故障导致服务失效**
18	RCA - Service Management Operation Errors Across Azure Services in East US 2 Tracking ID: Y 5-9C0	从2022年4月8日12:25 UTC至2022年4月9日14:40 UTC，东美国2区域（East US 2）的客户可能经历了服务管理错误、延迟和超时问题。受影响的服务包括Azure虚拟机（VMs）、虚拟机规模集（VMSS）、Azure数据工厂（ADF）、Azure Databricks、Azure Synapse、Azure备份、Azure站点恢复（ASR）以及Azure虚拟桌面（AVD）。这些问题主要表现为GET和PUT请求的错误，并且对启用自动扩展的服务可能产生了数据面影响。	问题的根本原因在于计算资源提供者（CRP）网关服务遇到了严重的吞吐量下降，主要原因是与其相关的**分配器服务（Allocator service）**在某个可用区发生故障，导致重试风暴。具体分析如下： 分配器服务的故障：一个与CRP相关的分配器服务在某个可用区发生故障，触发了重试风暴。尽管分配器服务通过重启故障区的实例得以恢复，但在恢复过程中，工作积压暴露了.NET CLR和垃圾回收器（Garbage Collector，GC）的问题。 .NET垃圾回收器过载：由于积压的工作量以及服务的自动重试，CRP网关的处理能力进一步下降，导致大量传入调用失败。平时每分钟约有25,000次调用的服务，由于重试机制，每分钟的调用次数激增到150,000次，进一步加剧了负载问题。 垃圾回收器的性能瓶颈：在高负载下，.NET CLR的垃圾回收器处理了大量的堆内存分配和释放（heap churn），导致异常增多。同时，垃圾回收器的过载引发了.NET运行时进程范围锁的不利交互，进一步降低了系统的吞吐量。		网络故障导致服务降效 （网关服务故障导致吞吐量下降）
19	RCA - Azure SQL DB and Cosmos DB Unavailable Tracking ID: SL1P-TSZ	从2022年2月12日11:45 UTC到2022年2月15日11:43 UTC期间，部分使用SQL数据库和Cosmos DB的客户遇到了数据库不可用的问题，并可能在连接到数据库实例时出现错误。此次问题主要影响了六个区域内使用特定代代硬件托管的SQL和Cosmos DB资源的客户。	问题的根本原因与Azure的最新一代硬件上运行的SQL和Cosmos DB资源的网络控制平面连接中断有关。最新一代硬件采用了优化的网络控制平面和数据平面，旨在提升性能并减少延迟。该控制平面使用TLS证书进行加密认证，证书定期轮换并在短时间内过期以确保安全。 由于远程过程调用（RPC）机制中的竞态条件，部分节点未能正常更新已轮换的TLS证书，导致证书过期后连接失败。虽然Azure计算平台的常规维护在1月已通过代码更新解决了这一问题，但SQL和Cosmos DB环境并未在第一时间进行更新，计划于2月下旬推送更新。2月12日，大量已轮换的TLS证书同时过期，导致SQL和Cosmos DB节点的连接中断。	**总结：** 此次事件由最新一代硬件的网络控制平面未能及时更新TLS证书引发，导致数据库实例的连接中断。通过部署紧急代码更新和修复，Azure团队成功恢复了服务并防止了类似问题的再次发生	**软件错误导致服务失效** （数据库服务连接中断）
20	RCA - Azure Resource Manager - Issues with management and resource operations Tracking ID: 8V39-P9Z	从2022年1月13日09:00 UTC到1月14日20:00 UTC期间，部分使用Azure资源管理器（ARM）进行部署、修改或删除Azure资源的客户遇到了延迟、超时和失败的问题。影响最严重的时间段分别为1月13日15:30 UTC开始的5小时，以及1月14日00:00 UTC开始的8小时，受影响的区域包括但不限于：西美国、西美国2、南中美国、北欧、西欧、东亚和东南亚。	问题的根本原因是2022年1月6日开始推送的一次代码修改，暴露了ARM平台处理长时间运行操作（“任务”）的基础设施中一个潜在的缺陷。此次代码修改引发了极小部分任务执行异常，但每次异常都会禁用一小部分任务执行基础设施。随着时间的推移，任务执行逐渐从已经部署新代码的区域转移到备份配对区域。 在最初的16小时内，备份配对区域正常执行任务，未对客户造成影响。然而，当新代码部署到备份配对区域后，问题开始在这些区域蔓延，导致任务排队、延迟增加和超时。在某些情况下，任务执行的延迟时间过长，导致最终任务失败，客户在这些情况下会看到操作失败。 由于任务执行基础设施的实施方式，累积的失败未在监控系统中及时可见，导致工程师最初误判问题原因，并尝试了一些无效的缓解措施。结果，1月14日00:00 UTC开始的第二个影响期持续了大约8小时。	**总结：** 此次事件由代码修改引发的任务执行基础设施缺陷导致，初期未能通过监控系统识别，影响了多个Azure区域的资源管理操作。通过回滚有问题的代码，Azure团队最终解决了问题，并计划改进监控系统以防止类似问题再次发生。	**软件错误导致服务失效** （代码缺陷）

21	Azure Cosmos DB - East US Tracking ID: 9VT8-HPG	从2022年1月4日12:30 UTC到2022年1月5日07:41 UTC，部分拥有Azure Cosmos DB账户的客户在东美国（East US）区域遇到了连接和服务可用性错误。由于该区域的一个Cosmos DB集群不可用，因此客户在该区域的数据库的所有新旧连接均可能出现错误或超时。 **（典型的级联故障事件）**	Azure Cosmos DB依赖Azure Service Fabric作为底层平台来提供集群的容错性。Service Fabric使用环形拓扑结构，每个节点通过与邻近节点建立租约关系来检测故障。特定的节点群组（称为仲裁者）负责确定其他节点的集群成员身份。如果某个节点未能在租约超时之前刷新租约，邻近节点会报告该节点不可用，并由仲裁者决定该节点是否应退出集群。这一检查通过定时回调完成。 在此次事件中，某个节点的定时回调以远高于预期的频率多次触发，导致该节点错误地报告其邻居节点为不可用。按照设计，仲裁者在未能收到健康状态的通知后，信任了这些错误信息。此过程持续进行，最终导致节点的法定数量（quorum）丧失，整个集群下线。 集群恢复正常是在故障节点手动重启后，作为缓解措施的一部分进行的。	**缓解措施：** 手动重启节点：Azure工程师团队手动重启了触发错误回调的故障节点，使集群重新上线并恢复正常运行。 改进系统设计：将分析和优化触发定时回调的机制，以确保未来不会再次发生类似的误报问题。 **总结：** 此次事件的根本原因是Azure Service Fabric定时回调机制中的问题，导致集群中某个节点错误地报告邻居节点为不可用，最终引发了集群下线。通过手动重启故障节点，Azure团队成功恢复了服务可用性。	**软件错误导致连接中断** （软件机制设计问题）→ (集群下线)
22	Microsoft Graph - intermittent 400-level errors accessing Graph APIs - Mitigated Tracking ID: PLT7-RTZ	从2021年11月12日02:00 UTC到2021年11月15日17:00 UTC，位于北美和亚太地区（APAC）的部分客户在尝试访问Microsoft Graph APIs时，遇到了间歇性的400级错误（如400 Bad Request）。这些错误可能影响了客户的API调用，导致API请求无法如预期完成。	**初步故障根因：** 此次问题的根本原因是最近一次为改进Microsoft Graph API基础设施所进行的更新引入了一个配置问题。该配置问题发生在Microsoft Graph API接口与其底层的Internet信息服务（IIS）驱动之间，导致某些API调用无法正常完成，进而引发了间歇性的400级错误。	**缓解措施与改进：** 配置修复：工程师团队识别并修复了API接口与IIS驱动之间的配置问题，恢复了Microsoft Graph API的正常运行。 **总结：** 此次事件由Microsoft Graph API与底层IIS驱动之间的配置问题引发，导致API调用失败和间歇性的400级错误。问题修复后，API服务恢复正常，Microsoft将采取措施改进配置流程，确保未来避免类似问题。	软件错误导致连接中断 （配置错误）
23	Microsoft Azure Portal - Issues while trying to create an application - Mitigated Tracking ID: 4M8X-VTZ	从2021年9月3日15:00 UTC到2021年9月9日01:24 UTC期间，部分使用Microsoft账户（MSA）登录Azure门户的客户在尝试创建应用程序时遇到了问题。此次问题不影响使用Azure AD租户登录的用户。	问题的根本原因是处理请求的容量不足。由于系统无法处理高并发的请求量，导致使用Microsoft账户的用户在创建应用程序时遇到问题。 **缓解措施与改进：** 容量扩展：Azure团队会对相关系统的容量进行扩展，以确保足够的资源来处理所有客户请求。 监控改进：提升监控系统的能力，确保未来在容量问题出现之前能够及时预警并采取措施。	**总结：** 此次问题由容量不足导致，影响了使用**Microsoft账户（MSA）**登录Azure门户的客户在创建应用程序时的体验。Azure团队正在通过扩展容量和改进监控措施，确保未来避免类似问题的发生。	硬件故障导致服务失效 （容量不足）
24	RCA - Service management operation failures - North Europe Tracking ID: 0 JL-9SG	从2021年7月7日21:19 UTC到2021年7月8日12:10 UTC，北欧区域的部分客户在进行依赖计算资源的服务管理操作时，可能间歇性地遇到了错误。受影响的服务管理操作包括虚拟机和磁盘管理的请求。	**故障根因：** 问题的根本原因是磁盘管理服务前端的请求队列达到了限制，导致部分请求被拒绝。这引发了与虚拟机和磁盘管理相关的服务管理请求出现间歇性失败。 **触发原因：**问题由平台更新触发。此次平台更新导致磁盘管理服务的副本之间进行批量故障转移，意外地引起某些调用的高延迟，导致服务前端的请求队列积压。 后果：请求积压进一步导致后续调用出现更高的延迟和失败。大多数失败的请求在重试后成功，但首次请求间歇性失败。	**缓解措施与改进：** 问题缓解：通过调整和优化平台更新流程，防止磁盘管理服务的前端请求队列超限，避免类似问题再次发生。 深入调查：正在进一步调查导致高延迟和队列积压的具体事件顺序，以防止此类情况在未来的更新中重现。 **总结：** 此次事件由磁盘管理服务前端的请求队列达限导致，触发点为一次平台更新引发的批量故障转移。Azure团队正在通过改进更新流程和深入调查问题的具体原因，防止类似问题的再次发生。	软件错误导致服务降效、失效 （平台更新错误导致服务请求达到上限）
25	RCA -Azure Resource Manager - Degraded Performance managing resources Tracking ID: 1V9K-PSZ	从2021年6月29日22:24 UTC到2021年6月30日14:30 UTC期间，部分Azure客户在访问Azure门户和其他Microsoft及Azure服务时遇到了间歇性错误。此次影响波及多个服务和区域，影响程度各异。大多数服务在2021年6月30日14:30 UTC之前完全恢复。在此期间，部分重试请求可能成功。	问题源于Azure资源管理器（ARM）服务与其依赖的后端存储之间的连接配置。此次问题是由一次配置更新引发的： 配置更新故障：在进行后端存储配置更新的维护操作后，部分机器无法连接到更新后的存储端点。 自动重启问题：随着时间的推移，部分机器在正常重启时加载了新的配置，但新配置中包含一个问题，导致这些机器与存储端点的连接失败。 服务性能下降：随着受影响的机器增多，导致部分区域的服务性能逐渐下降。由于此次配置更改被认为不会产生影响，因此以并行方式推送，导致多个区域受到影响。	**缓解措施与改进：** 回滚故障配置：Azure团队通过识别并修复配置问题，恢复了存储端点的正常连接，逐步恢复受影响的服务。 改进发布流程：Azure团队计划优化配置变更的发布流程，确保未来即使进行非预期影响的更改时，也采取更稳妥的逐步发布策略，以避免广泛影响。 **总结：** 此次事件由Azure资源管理器与后端存储的配置更新问题引发，导致部分服务和区域的性能下降和间歇性访问错误。通过回滚故障配置，Azure团队成功恢复了服务，并计划优化配置管理流程以防止类似问题的再次发生。	软件错误导致服务降效 （配置更新错误）

26	RCA - Error notifications for service management operations - West US 2 Tracking ID: LL1H-9CZ	从2021年6月14日22:00 UTC到2021年6月15日11:15 UTC期间，**西美国2（West US 2）**区域的一部分客户在执行服务管理操作（如创建、更新、删除）时遇到了错误。此次问题影响了多个服务。2021年6月15日09:20 UTC，缓解措施被应用，随着积压的服务管理请求负载减少，服务逐渐恢复，11:15 UTC时所有受影响服务完全恢复。	此次问题由多个因素共同导致： 后端访问控制服务的高CPU使用：西美国2区域专门用于处理服务管理请求的某个后端访问控制服务经历了一段异常的内部流量峰值，导致该服务的CPU使用率异常升高，进而导致服务请求超时。 代码缺陷：承载该服务的资源因驱动程序中的代码缺陷而不可用。该缺陷在特定的高负载条件下被触发，进一步加剧了问题并延长了修复时间。 容量预测不足：由于此后端服务无法自动扩展，Azure依赖于压力测试来预测容量需求。然而，此次事件的压力测试未能考虑西美国2区域的特定配置，导致未能及时应对流量激增的需求。	**缓解措施：** 网络规则调整：2021年6月15日06:39 UTC，Azure团队引入了有针对性的网络规则，屏蔽了某些内部流量，以减少后端服务的负载。 配置更改与驱动程序修复：2021年6月15日06:51 UTC，团队对基础设施应用了配置更改，并移除了导致问题的驱动程序。 增加容量：2021年6月15日09:20 UTC，团队为内部基础设施增加了额外容量，稳定了客户面对的服务，最终缓解了问题。 **总结：** 此次事件由后端访问控制服务的高CPU使用率、驱动程序缺陷和容量预测不足共同导致，影响了西美国2区域的服务管理操作。通过引入网络规则、修复驱动程序并增加基础设施容量，Azure团队成功恢复了服务。	软件错误导致服务失效 硬件故障导致服务CPU使用率升高，过载至降效
27	RCA -Issues accessing the Azure portal and other Microsoft services Tracking ID: KN22-39Z	从2021年5月20日06:52 UTC到16:20 UTC，部分Azure客户在访问Azure门户及其他Microsoft和Azure服务时可能遇到了间歇性错误。此次问题影响了多个服务和区域，影响程度和恢复时间因服务而异。大多数服务在16:20 UTC时已完全恢复。	问题源于Azure区域中的一系列瞬时名称解析问题。具体影响时段如下： 06:52 UTC到07:10 UTC：影响欧洲区域。 09:00 UTC到09:30 UTC：影响印度区域。 15:53 UTC到16:20 UTC：主要影响欧洲（尤其是英国）区域。 这些名称解析问题是由最近一次对Azure边缘DNS服务器进行的代码部署中的代码回归引发的。代码回归引入了锁争用问题，当此问题被触发时，导致部分边缘服务器上的进程进入暂停状态，暂时停止处理流量。虽然这些进程会自动恢复并重新开始服务，但在恢复期间，出现了间歇性查询丢失和服务性能下降。此次问题的触发概率较低，且问题在部署逐步完成几天后才开始显现。	回滚部署：为解决问题，Azure团队通过**安全部署流程（SDP）**将有问题的部署回滚到之前已知的正常状态。回滚首先在受影响的区域进行，随后在全球范围内展开。 监控与验证：在16:20 UTC进行缓解后，Azure团队继续监控平台，确保服务在回滚之前和过程中保持稳定。 **总结：** 此次事件由边缘DNS服务器上的代码回归引发的锁争用问题导致，进而导致名称解析故障，影响了多个区域的服务。通过回滚有问题的部署，Azure团队成功恢复了服务，并确保了服务的稳定性。	软件错误导致连接中断
28	RCA - Authentication errors across multiple Microsoft services Tracking ID: LN01-P8Z	从2021年3月15日19:00 UTC到2021年3月16日09:37 UTC，部分客户在执行身份验证操作时遇到了错误，影响了依赖**Azure Active Directory (Azure AD)**进行身份验证的Microsoft服务和第三方应用。虽然Azure AD的缓解措施在3月15日21:05 UTC已完成，但一些服务的恢复时间有所不同，具体如下： 22:39 UTC 15 March 2021：Azure资源管理器恢复。 01:00 UTC 16 March 2021：大多数区域的Azure密钥保管库（Key Vault）恢复。 01:18 UTC 16 March 2021：Azure存储的配置更新应用于第一个生产租户。 01:50 UTC 16 March 2021：Azure门户功能完全恢复。 04:04 UTC 16 March 2021：Azure存储的配置更改应用于大多数区域。 04:30 UTC 16 March 2021：剩余的Azure Key Vault区域（西美国、中央美国和东美国2）恢复。 09:25 UTC 16 March 2021：Azure存储恢复完成，事件完全缓解。	Azure AD使用加密密钥支持OpenID等身份标准协议进行签名操作。为了保持安全，系统定期自动移除不再使用的密钥。最近几周内，由于跨云迁移的复杂性，一特定密钥被标记为保留。然而，系统中的一个自动化错误忽略了该保留状态，导致该密钥被错误移除。 Azure AD根据互联网身份标准协议，将签名密钥的元数据发布到全局位置。当3月15日19:00 UTC发布新元数据后，使用这些协议的应用开始接收更新的元数据，并不再信任使用已移除密钥签署的令牌，导致终端用户无法访问这些应用。 服务的遥测数据识别了问题，并且在19:35 UTC，工程团队停止了正在进行的基础设施变更。21:05 UTC，密钥元数据被回滚至先前的状态。 应用程序需要拾取回滚后的元数据并刷新其缓存，个别应用程序的恢复时间因缓存的处理方式不同而有所变化。一部分Azure存储资源因缓存的元数据导致恢复较慢，工程团队部署了更新，强制刷新缓存。此过程完成后，事件于3月16日09:37 UTC宣布完全缓解。	**缓解措施：** 回滚元数据：在21:05 UTC，Azure AD团队回滚了签名密钥元数据，以确保应用程序能够重新信任正确的密钥。 缓存刷新：针对受缓存影响的应用，强制刷新缓存，确保所有服务逐步恢复正常。 多阶段防护措施：Azure AD正在进行多阶段的安全部署流程（SDP）改进，第一阶段已为添加新密钥提供了保护，但移除密钥的部分将在第二阶段完成，计划于年中完成。该改进将防止类似事件再次发生。 **总结：** 此次事件由自动化错误移除加密密钥引发，影响了Azure AD及其依赖的多个Microsoft服务和第三方应用的身份验证。通过回滚元数据和刷新缓存，服务逐步恢复，Azure团队也正在实施额外的防护措施，以防止未来类似问题的发生。	软件错误导致连接中断 （某些服务不可被访问）

29	RCA-Connectivityl ssues-UK South Tracking ID: CSDC-3Z8	从2020年9月14日13:30 UTC到9月15日00:41 UTC，英国南部（UK South）区域的部分客户在连接Azure服务时遇到问题。使用可用区冗余（Availability Zones）且配置了区域冗余（Zone Redundancy）的客户在此次事件中未出现服务中断。某些情况下，客户可能遇到服务管理操作的影响，但**区域冗余存储（ZRS）**在整个事件期间保持可用。	此次事件的根本原因是英国南部一个数据中心的冷却设备出现了问题，导致服务中断。具体情况如下： 维护操作导致错误关闭：在9月14日进行的一次设施维护活动中，冷却系统的水塔补水泵通过建筑自动化系统（BAS）被错误地关闭。该错误在13:30 UTC被发现，此时该问题已开始影响到下游机械系统，导致支撑这些系统的电力基础设施也被关闭。 2N冗余设计失效：尽管Microsoft数据中心采用了2N冗余设计，即拥有全冗余、镜像系统以保护系统免受中断影响，但此次的级联故障影响了支撑机械系统的两侧电力基础设施，导致冷却设备无法正常运行。 自动保护机制触发：当内部系统检测到热事件后，自动化系统开始关闭网络、存储和计算基础设施的部分资源，以保护硬件和数据的持久性。由于部分资源的连接问题，某些设备无法自动关闭，团队通过手动干预完成了这些资源的关闭。 问题诊断与修复：团队花费了约120分钟时间诊断根本原因并开始修复冷却设备。15:45 UTC，冷却系统恢复运行，16:30 UTC受影响数据中心的温度恢复到正常操作范围。	**网络与服务恢复：** 16:30 UTC，网络恢复工作开始，团队通过重启网络交换机恢复其状态，优先恢复Azure管理基础设施、存储集群和计算集群的正常运行。 23:32 UTC，网络恢复完成，随后存储和计算集群的连接也恢复正常。团队进一步采取措施将剩余不健康的服务器重新上线。 **总结：** 此次事件由数据中心冷却系统的维护错误引发，导致冷却设备关闭，并进而触发了级联故障，影响了英国南部区域的一部分Azure服务。通过冷却系统的修复和网络设备的恢复，服务逐步恢复正常。	电力故障导致服务失效 （电力故障导致设备自动关闭）
30	RCA - Network Latency Issue-West Europe Tracking ID: 8KLC-1T8	2020年9月3日09:21至17:32 UTC期间，部分客户在访问西欧（West Europe）区域托管的资源时，可能遇到了间歇性延迟或连接问题。在此期间，重试请求可能会成功。	此次问题由两起独立事件的相继发生引发： 光纤传输问题：在故障开始的约4小时前，靠近数据中心的某些**本地活动（可能是施工）**导致光纤电缆之间的数据包在传输过程中被损坏。这些损坏的数据包被检测到并丢弃，自动化网络系统将受影响的链路下线并通知现场进行修复。这是标准操作流程，且初步安全检查确认没有相关影响。 光纤切断事件：在09:21至09:26 UTC之间，距离数据中心约5公里处的另一条路径上发生了一次重大光纤切断事件，导致该路由的50%容量受到影响。单独来看，这一事件也不会对西欧区域的整体网络流量造成影响。 虽然每个事件单独发生时不会产生明显影响，但结合起来，这两起事件导致数据中心之间的9条链路承载了不平衡的流量，链路变得拥塞并出现数据包丢失（影响了不到2%的总容量）。经过这些拥塞链路的连接会遇到数据包丢失和延迟。由于连接会分散在可用的链路上，因此重试请求可能不会受到影响，并且能够成功。 缓解延迟的主要原因在于需要值班工程师识别出链路中断的多个原因，并找到最佳方式来减少拥塞和重新平衡流量。在最初的响应阶段，多个并发警报导致工程师的操作将拥塞从一个链路转移到了另一个链路，但未能彻底解决问题。	**缓解措施与改进：** 重新平衡流量：通过识别并重新平衡各链路上的流量，最终减少了网络拥塞并恢复了正常连接。 改进事件响应：由于初期响应过程中警报过多导致操作混乱，Azure团队计划改进警报管理和响应流程，以确保未来能够更快、更准确地识别问题。 **总结：**此次问题是由光纤损坏和切断事件的结合引发，导致了西欧区域部分链路的网络拥塞和数据包丢失。虽然每个事件单独发生时影响有限，但组合效应导致了连接问题。通过重新平衡流量	2次独立的网络故障导致连接中断
31	RCA -Virtual Machines - Virtual machine unexpected restarts Tracking ID: 8S8J-9T8	从2020年7月7日07:24 UTC到7月17日21:16 UTC期间，部分使用虚拟机（VMs）的客户可能遇到了间歇性的连接失败，尝试访问某些虚拟机时出现问题。此外，这些虚拟机可能出现了意外重启的现象。	问题的根本原因是一次操作系统更新部署任务中包含了一个代码配置错误。该错误导致了一部分集群上之前修复的已解决bug被意外回滚。这一错误在执行高磁盘I/O工作负载的虚拟机宿主节点上表现为系统死锁，导致运行在这些节点上的虚拟机重启。	**缓解措施：** 停止部署：在发现问题后，Azure团队立即停止了正在进行的部署任务。 开发并部署修复：团队开发了一个新的部署任务，包含检测补丁需求的代码修复。该修复部署到所有受影响的集群，缓解了虚拟机重启和客户影响。 加速缓解：在全面修复的同时，团队还通过识别受影响的宿主节点并重新附加补丁，为部分客户加速了问题的解决。 **总结：**此次事件由操作系统更新中的代码配置错误导致，回滚了一些已解决的bug，进而在执行高I/O负载的虚拟机上引发了系统死锁和意外重启。通过停止部署并推送修复补丁	软件错误导致服务失效 （代码配置错误）→ (系统死锁和重启)

32	RCA - Azure Resource Manager - Failures creating or deleting resources Tracking ID: DLZG-7CO	从2020年6月4日07:45 UTC到16:57 UTC期间，所有公共Azure区域的部分客户在尝试通过Azure资源管理器（ARM）创建或删除服务资源时，可能遇到了部署失败的情况。这次问题是由于底层的网络问题导致的。虽然相关的网络资源实际上已成功创建或删除，但ARM未收到部署状态通知，因此客户看到的服务创建或删除操作失败。这次问题还可能影响某些资源的GET或READ操作，影响用户比例不到0.01%。	问题的根本原因是最近的一次ARM部署中包含了一个配置文件错误。该配置文件保存了ARM用于操作状态查询调用的URL端点，文件中针对网络资源的URL端点设置不正确。由于这个错误的配置，ARM对网络服务管理操作的状态查询失败，客户在尝试创建或删除网络资源时就会看到失败提示。 测试未发现问题：此次错误配置文件在部署到生产环境之前未被发现，因为在测试时使用的是健康的配置文件。更新在生产环境中使用了最新的配置文件，而没有进行充分的测试，导致该问题的发生。	缓解措施： 故障发现与修复：问题于13:00 UTC被定位为配置文件的错误端点，团队随后更正了配置文件中的URL端点。 分批部署修复：团队首先在单个区域中验证修复成功，然后按批次将修复部署到其他区域。每一批次完成后，都会进行验证。到16:57 UTC，所有区域的修复部署完成，确认问题得到缓解 **总结：**此次事件的根本原因是ARM的配置文件中包含了错误的网络资源URL端点，导致客户在创建或删除网络资源时遇到失败。通过修正配置文件并安全地重新部署。	软件错误导致服务失效 （代码配置错误）→ (网络中断)→ (服务失效)
33	RCA- Multiple Services - Central India Tracking ID: SLN3-HDO	从2020年5月18日12:41 UTC到5月19日08:30 UTC，部分客户在连接中印度（Central India）区域的资源时遇到了困难。多个存储和计算扩展单元下线，影响了虚拟机及其他依赖这些资源的Azure服务。	此次问题由区域电力供应商的电力问题引发。具体细节如下： 5月18日11:25 UTC，区域电力供应商发生电力问题，导致中印度数据中心转为发电机供电。尽管数据中心的基础设施系统正常切换至发电机供电，但其中**两个托管室的包裹式空气处理单元（PAHU）**未能按设计正常运行，导致这些房间的空气温度超过了操作阈值。 现场团队立即触发警报并尝试恢复PAHU系统，但初步努力未能成功。到13:22 UTC，工程师开始关闭计算、网络和存储资源，以保护数据中心设备免受热损坏。	**缓解措施：** 恢复PAHU系统：工程师首先调查了PAHU为何全部关闭的问题。到16:31 UTC，工程师通过绕过故障组件并依次重启每个PAHU单元，使温度回到了安全操作范围内。 服务恢复：在温度恢复正常后，工程师开始准备恢复设备。由于部分网络设备和服务器是手动关闭的，因此需要手动重新上电。网络设备首先恢复，存储集群在网络可用后自动恢复。在存储和网络恢复后，依赖的计算硬件也开始恢复，虚拟机及其他Azure服务的恢复过程于5月19日08:30 UTC完成。目前，所有托管室已恢复至电力供应商供电，且所有PAHU单元均处于自动控制状态。 **总结：**此次事件由中印度区域电力供应问题引发，导致数据中心的部分托管室的冷却系统（PAHU）未能正常运行，进而引发一系列基础设施的自动关闭。通过工程师团队恢复冷却系统、手动重新上电并逐步恢复存储、网络和计算资源，服务逐步恢复到正常状态。	电力故障导致服务失效 （电力故障影响冷却系统）→ (集群自动关闭设备)
34	Issues accessing resources in the Azure Portal - Mitigated Tracking ID: PMN6-7D8	从2020年4月29日18:41 UTC到2020年4月30日11:00 UTC，部分客户在访问Azure门户时可能遇到问题，并收到“访问被拒绝（Access Denied）”的错误消息。尽管如此，客户可以使用编程方法（如PowerShell或Azure CLI）通过资源ID列出其订阅中的资源并访问这些资源。	问题的根本原因是最近的一次部署任务引入了一个软件漏洞，影响了**基于角色的访问控制（RBAC）**信息的同步。RBAC信息对于Azure资源管理器（ARM）等资源管理服务至关重要，而这些服务正是Azure门户用于显示资源的基础。由于同步失败，Azure门户调用失败，资源未能如预期显示。	缓解措施与改进： 部署回滚：在识别到问题后，Azure团队立即停止了受影响的部署任务，并着手修复该软件漏洞。 系统监控改进：未来，团队计划加强RBAC同步相关的监控，确保类似问题能尽早被发现并解决。 **总结：**此次问题由部署过程中引入的RBAC同步问题引发，导致部分客户在访问Azure门户时遇到了“访问被拒绝”的错误。虽然Azure门户受到影响，但客户仍可通过编程方式访问其资源。通过修复该软件漏洞，Azure团队解决了这一问题并确保了系统的后续稳定性。	软件错误导致连接中断 （某些服务访问被拒绝）
35	RCA - Service Management /Authentication Errors - Azure China Tracking ID: SND4-L80	从2020年3月5日21:03 CST (UTC+8) 到3月6日16:03 CST，部分Azure中国区域的客户在对这些区域托管的资源执行服务管理操作时遇到了失败。客户在尝试访问Azure门户或其他Azure资源时，也可能遇到身份验证失败。	问题的根本原因与传输层安全（TLS）证书的验证机制有关。客户在连接Azure服务时，需要验证Azure服务的TLS证书，该验证依赖于访问在线证书状态协议（OCSP）服务和证书吊销列表（CRL）。 Azure中国使用外部的证书颁发机构（CA），该CA托管了OCSP和CRL端点。 在此次事件中，这些端点在中国的客户无法访问。问题源于一个电信提供商，影响了Azure及CA的其他客户。无法访问OCSP和CRL端点导致客户的证书验证失败，进而无法连接Azure服务。 这些OCSP和CRL端点通过**内容分发网络（CDN）**在多个位置进行了镜像，但由于网络路径问题，客户仍无法访问这些端点。	**缓解措施：** DNS记录更新：Azure工程师通过更新DNS记录，将客户的请求重新路由到可访问的备用OCSP和CRL端点，从而缓解问题。 网络排查：由于问题涉及多个公司在网络路径中的协调，故排查和解决过程耗时较长。 部分服务的提前缓解：部分Azure服务通过提前部署最新的CRL到其服务器，实现了更快的缓解。 **总结：**此次事件的根本原因是由于**外部证书颁发机构（CA）**的OCSP和CRL端点在中国的客户不可访问，导致客户无法进行TLS证书验证，从而无法连接到Azure服务。Azure团队通过更新DNS记录重新路由请求，成功缓解了问题，并保证后续服务的正常运行。	软件错误导致连接中断

36	RCA-SQL Database and dependent services - Service Availability Issues Tracking ID: 5TYQ-DCO	从2020年1月24日22:00 EST到2020年1月25日08:15 EST，Azure政府区域的部分客户在访问Azure SQL数据库和数据仓库资源或其依赖的服务时遇到了失败。具体表现为，客户在尝试建立新连接时可能遇到错误或超时，但已经建立的连接池继续正常工作。一些管理操作（如故障切换到地理冗余区域）也受到了影响。	**故障根因：** 问题的根本原因是SQL数据库和数据仓库的连接需要通过一组负载均衡的前端节点（网关）。工程师确定最近的一次维护活动未能成功完成，导致这些网关持有了错误的证书配置，阻止了与SQL资源的连接。	**缓解措施：** 部署更新：Azure工程师通过向受影响区域部署更新来缓解错误配置。 部分恢复：部署于2020年1月25日03:30 EST左右在部分区域开始生效，系统逐步恢复。 完全恢复：到2020年1月25日08:15 EST，所有区域完全恢复。 **总结：** 此次事件是由于网关的错误证书配置导致Azure SQL数据库和数据仓库的连接失败。通过工程师团队的更新部署，问题得以逐步缓解并最终完全恢复。	软件错误导致连接中断 （代码配置错误）→（数据库、数据仓库服务连接失败）