東京リージョン(AP-NORTHEAST-1)で発生したAWS Direct Connectの事象についてのサマリー

Read in English

日本時間 2021 年 9 月 2 日に東京リージョン（AP-NORTHEAST-1）で発生した AWS Direct Connect サービスの中断に関する追加情報を提供いたします。午前 7 時 30 分(以下すべて日本時間)から、Direct Connect をご利用中のお客様は東京リージョンに向かうトラフィックについて断続的な接続の問題とパケットロスの増加を観測し始めました。この事象は、Direct Connect ロケーションから、顧客の Virtual Private Cloud（VPC）が存在する東京リージョンのデータセンターネットワークへのネットワークパスに沿ったネットワークレイヤーの 1 つでネットワークデバイスの一部に障害が発生したことが原因です。お客様は午後 12 時 30 分に復旧を観測しはじめ、午後 1 時 42 分に接続の問題は完全に解決されました。 アベイラビリティーゾーン間のトラフィック、リージョンへのインターネット接続、AWS Virtual Private Network (VPN) 接続 (一部のお客様が Direct Connect へのバックアップとして使用) など、他のすべてのネットワーク接続は影響を受けませんでした。他の AWS リージョンへの Direct Connect トラフィックも影響を受けませんでした。

2021 年 9 月 2 日午前 7 時 30 分に、AWSのエンジニアは、お客様が東京リージョンに接続する Direct Connect のパケットロスの増加について内部アラームを検知しました。Direct Connect サービスは、AWS がお客様と相互接続する Direct Connect ロケーションから、それぞれに多数の冗長ネットワークデバイスがある複数のネットワークレイヤーを介して AWS リージョンにトラフィックを転送することで、お客様のデータセンターと VPC 間のプライベート接続を提供します。これらのアラームは、Direct Connect ネットワークの単一レイヤー内のいくつかのデバイスの障害によって事象が引き起こされたことを特定しました。これらのデバイスはトラフィックを正しく転送していませんでしたが、障害が発生したネットワークデバイスを監視および削除する通常の自動プロセスを通じてネットワークから除外されていませんでした。代わりに、この自動プロセスは、通常よりもデバイスの故障率が高いことを検知し、エンジニアに調査して修復措置を講じるよう警告しました。エンジニアは、この警告を受けたとき、このレイヤに十分な冗長性があると判断し、正常なデバイスでトラフィックを処理できるように影響を受けているデバイスをサービスから除外しました。並行して、チームは故障の原因を調査しました。追加のデバイスを除外することで一時的な修復が得られましたが、その後いくつかの他のネットワークデバイスでも同じ障害が発生し、Direct Connect を利用中のお客様のネットワークに輻輳、接続の問題、またはパケットロスの増加が発生しました。エンジニアは、故障したデバイスをリセットして徐々にサービスに戻すなど、いくつかの緩和を試みましたが、障害は継続し、エンジニアは適切なキャパシティを維持して顧客への影響を完全に軽減することができませんでした。エンジニアは、障害を引き起こした可能性のある最近のデプロイメントも探しました。午後 12 時までに、エンジニアは、この障害が頻度の低いネットワークコンバージェンスイベントおよびファイバ切断に対するネットワークの反応時間を最適化するために導入された新しいプロトコルに関連している可能性があると考えました。この新しいプロトコルは何ヶ月も前に導入され、それ以来問題なく運用されていました。しかし、エンジニアは、障害が Direct Connect ネットワークのある層にあるネットワークデバイス上における、新しいプロトコルと新しいトラフィックパターンの相互作用に関連していると考えました。エンジニアは、単一のアベイラビリティーゾーンでこの新しいプロトコルを無効化して、持続的なリカバリの確立を確認しました。並行して、東京リージョン全体に展開する変更を準備しました。お客様は午後 12 時 30 分よりアプリケーションの復旧を確認し始め、午後 1 時 42 分に影響を受けたネットワークデバイスが安定した動作状態に復元され、Direct Connect サービスが通常の動作に戻りました。

新しいプロトコルを無効にすることでこの事象は解決しましたが、エンジニアリングチームは根本原因を特定するために引き続き取り組んできました。この事象は、ネットワークデバイスのオペレーティングシステム内の潜在的な問題によって引き起こされたことを確認しました。このバージョンのオペレーティングシステムでは、ネットワークのフェールオーバー時間を改善するために使用される新しいプロトコルが有効になります。新しいオペレーティングシステムとプロトコルは、以前から実稼働環境で正常に実行されています。AWS は、オペレーティングシステムを変更し新しいプロトコルを AWS ネットワークに導入するために、制御され、自動化され、テストされ、計測された手順を使用します。この手順は、専用のラボでの一連のストレステストから始まり、ネットワークデバイスの有効パケットと無効パケット(不正な形式のパケットなど)の両方に対する回復力を検証します。ラボテストで特定された異常は、新しいコードが本番環境にリリースされる前に診断され、根本原因が特定され、修復されます。この包括的なテストでも、ラボ環境ですべてのトラフィックとパケットの順列をテストすることはできません。したがって、AWS は、ネットワークデバイスのオペレーティングシステムの変更を段階的で制御された方法で本番環境にリリースするデプロイ手順を使用します。この手順では、アップグレードされたデバイスが実稼働トラフィックを処理することができますが、アップグレードされたデバイスからアップグレードされていないデバイスに容易に切り戻すことができる特定の場所で個々のデバイスをアップグレードします。この段階的な実稼働への展開手順では、アップグレードされたデバイスは、パフォーマンスの問題や機能エラーについて広範囲に監視されます。このアップグレードプロセスは何度も正常に使用されており、この最新のデバイスオペレーティングシステムのアップグレードでも遵守されました。新しいプロトコルとオペレーティングシステムは、2021 年 1 月に初めて実稼働環境に導入されました。過去 8 か月間にわたって、この新しいプロトコルとオペレーティングシステムは、すべての AWS リージョンで徐々に実稼働環境にリリースされ、潜在的な問題を示すことなく Direct Connect のサービスを提供してきました。過去数日間で、エンジニアはネットワークオペレーティングシステムの欠陥を特定し、問題を引き起こす非常に特殊なパケット属性とコンテンツのセットが必要であると判断しました。これらの条件は非常に特殊で稀であるものの、継続的にこのシグニチャに一致し得たカスタマートラフィックによってこのイベントが発生しました。悪意のある振る舞いがあったとは考えておりません。AWS 東京リージョンでこの問題が発生した新しいプロトコルを無効にしました。また、この変更を他のすべての AWS リージョンに慎重に適用するためにお客様に影響が出る前にこの問題を検出して修復するための拡張方法も開発しました。この問題によるお客様へのさらなる影響はないと確信しています。

AWS のサービスが日本のお客様と多くのビジネスにとって重要であるかを理解しており、今回の事象による影響を心からお詫び申し上げます。AWS は、高い可用性でサービスを運営してきた長年の実績があり、お客様の信頼を維持し、お客様とビジネスに必要な可用性を達成するために全力を尽くします。

Summary of AWS Direct Connect Event in the Tokyo (AP-NORTHEAST-1) Region

We would like to provide additional information about the AWS Direct Connect service disruption that occurred in the Tokyo (AP-NORTHEAST-1) Region on September 2, 2021. Beginning 7:30 AM JST, Direct Connect customers began to experience intermittent connectivity issues and elevated packet loss for their traffic destined towards the Tokyo Region. This was caused by the failure of a subset of network devices on one of the network layers along the network path from Direct Connect edge locations to the Datacenter network in the Tokyo Region, where customers' Virtual Private Clouds (VPCs) reside. Customers started seeing recovery by 12:30 PM JST and by 1:42 PM JST, connectivity issues were fully resolved. All other forms of network connectivity, including traffic between Availability Zones, internet connectivity to the Region, and AWS Virtual Private Network (VPN) connectivity (which some customers use as a back-up to Direct Connect) were not impacted. Direct Connect traffic to other AWS Regions was also not impacted.

On September 2, 2021 at 7:30 AM JST, internal alarms alerted AWS engineers to elevated packet loss for Direct Connect customers connecting to the Tokyo Region. The Direct Connect service provides private connectivity between a customer's data center and their AWS VPCs by forwarding traffic from the edge locations where AWS interconnects with customers, to the AWS Region through multiple network layers - each with many redundant network devices. These alarms identified that the impact was caused by the failure of several devices in a single layer of the Direct Connect network. While these devices were not correctly forwarding traffic, they were not being removed from the network through the normal automated processes that monitor and remove failed network devices. Our automation instead noticed a higher rate of failed devices than normal and alerted engineers to investigate and take remediation action. When engineers were alerted, they determined that there was enough redundancy at this layer and began removing the impacted devices from service so that traffic could be handled by other healthy devices. In parallel, the team investigated the cause of the failure. While the removal of additional devices provided temporary remediation, several other network devices subsequently began to experience the same failure, resulting in network congestion, connectivity issues, or elevated packet loss for Direct Connect customers. Engineers attempted several mitigations, such as resetting failed devices and slowly bringing them back into service, but the failures continued and the engineers were unable to maintain adequate healthy capacity to fully mitigate the customer impact. Engineers also looked for any recent deployments that may have triggered the failure. By 12:00 PM JST, engineers suspected that the failure may be related to a new protocol that was introduced to optimize the network's reaction time to infrequent network convergence events and fiber cuts. This new protocol was introduced many months prior and this change had been in production since then without any issues. However, engineers suspected that the failure was related to the interaction of this new protocol and a new traffic pattern on the network devices at this layer of the Direct Connect network. Engineers started disabling this new protocol in a single Availability Zone to monitor and establish sustained recovery, while in parallel preparing the change to be deployed across the Tokyo Region. Customers started reporting recovery to their applications by 12:30 PM JST and by 1:42 PM JST affected networking devices were restored to a stable operational state and the Direct Connect service returned to normal operations.

While disabling the new protocol resolved the event, engineering teams have continued working to identify the underlying root cause. We have now confirmed that this event was caused by a latent issue within the network device operating system. This version of the operating system enables a new protocol which is used to improve the failover time of our network. The new operating system and protocol have been running successfully in production for multiple months. We use a controlled, automated, tested, and instrumented procedure for changing the operating system and introducing the new protocol to the AWS network. This procedure starts with a series of stress tests in a dedicated lab to validate the resiliency of the network device to both valid and invalid (i.e., malformed) packets. Any anomalies identified in lab testing are diagnosed, root causes identified, and remediated before the new code is released to production. Even with this comprehensive testing, it is not possible to test every traffic and packet permutation in a lab environment. Therefore, AWS uses a deployment procedure that releases network device operating system changes to production in a slow and controlled fashion. This procedure upgrades individual devices in specific places where the upgraded devices can be exposed to production traffic but where traffic can easily fail away from the upgraded devices to non-upgraded devices. During this gradual production deployment, the upgraded devices are extensively monitored for performance issues and functionality errors. This upgrade process has been used many times successfully and was followed with this most recent device operating system upgrade. The new protocol and the operating system were first deployed to production in January 2021. Over the last 8 months, this new protocol and the operating system have been gradually released to production in all AWS Regions and has been serving Direct Connect customer traffic without any indication of the latent issue. Over the last several days, engineers have been able to identify the defect in the network operating system and determined that it requires a very specific set of packet attributes and contents to trigger the issue. While these conditions are very specific and unlikely, this event was triggered by customer traffic that was able to consistently generate packets that matched this signature. We have no reason to suspect malicious intent. We have disabled the new protocol that triggered this issue in the AWS Tokyo Region. We have also developed an enhanced way to detect and remediate this issue before customer impact, as we carefully apply this change to all other AWS Regions. We are confident that there will be no additional customer impact from this issue.

We understand how critical AWS services are for our customers and many businesses in Japan, and we sincerely apologize for the impact that this event may have caused. We have a long track record of operating our services with high levels of availability and will do everything possible to maintain our customers' trust and help them achieve the availability they need for their customers and businesses.