米国東リージョンにおける Amazon EC2 と Amazon RDS のサービス障害の概要 (参考和訳)

(注: この文書は、”Summary of the Amazon EC2 and Amazon RDS Service Disruption in the US East Region” の参考和訳です。公式の発表につきましては、以下をご参照ください。
http://aws.amazon.com/message/65648/)

先週 Amazon Elastic Compute Cloud (EC2)で発生した障害に関しましては、現時点で全ての影響のあったサービスを完全に復旧したことをご報告いたします。また、お客様に、今回の障害の詳細情報、サービス復旧の過程、今回のようなことを再び発生させないための予防措置につきましてお報せさせて頂きます。今回の障害では、沢山のお客様に影響を与えてしまったことを深く認識しております。そのため、実際に何が起こったのかについて詳細情報を共有させて頂くとともに、我々のお客様のためにどのようにサービスを改善していくのかお伝えしたいと考えております。

EC2 のお客様に先週影響を与えた事象は、米国東 (US East) リージョンにおける単一のアベイラビリティ・ゾーン内の Amazon Elastic Block Store (EBS) ボリュームの一部分が主に関連するものであり、その EBS ボリュームにおいて読み込み、書き込み操作が行えなくなりました。この文書では、それらのボリュームをスタックボリュームと呼びます。スタックボリュームを使用する EC2 インスタンスにおいても、スタックボリュームを読み書きしようとした際にスタックしてしまいました。当該のアベイラビリティ・ゾーンにおいて、スタックボリュームを復旧させ EBS クラスターを安定させるために、EBS のための全てのコントロール API 群（Create Volume、Attach Volume、Detach Volume、Create Snapshot 等。つまり、ボリュームの作成、アタッチ、デタッチ、スナップショット作成、等）を、その事象がおこっている期間において無効にしました。事象がおこった最初の日の 2 回に渡って、影響を受けた EBS クラスターは EBS API に影響を与えたために、米国東リージョン全体に渡り、API をコールした EBS において高いエラー率と遅延を発生させました。運用上における複雑に絡み合った課題が存在するものの、これらの事象は幾つかの相互に影響しあう根本原因によって発生したものであるので、今後似たような事象が発生した際にサービス継続を担保するための機会を我々は得ることができました。

EBS システムの概要

EBS のアーキテクチャを理解すると、今回の事象をより深くご理解いただくことができます。EBS は分散、複製ブロックデータストアであり、EC2 インスタンスから一貫性のある低レイテンシの読み込み、書き込みのために最適化されています。EBS サービスには 2 つのコンポーネントが存在します。一つは、EBS クラスターのセットであり、ユーザーデータを保持し、EC2 インスタンスへのリクエストを処理します。もう一つは、コントロール・プレーン・サービスのセ

ットであり、ユーザーリクエストの調整や、そのリクエストをリージョンにおける各アベイラビリティ・ゾーンの EBS クラスターに伝達します。

EBS クラスターは EBS ノードのセットで構成されています。これらのノードは、EBS ボリュームデータのレプリカを保持し、EC2 インスタンスへの読み書きのリクエストを処理します。EBS ボリュームデータは、耐久性、可用性の向上のため、複数の EBS ノードに複製されています。各 EBS ノードは、ピア・トゥー・ピアの素早いフェイルオーバー戦略を採用しており、もしコピーされたものの一つの同期がとれなくなった場合や、使用不可能になった場合には、積極的に新しいレプリカの割り当てを行います。EBS クラスターのノード群は、お互いに 2 つのネットワークを経由して接続されています。プライマリー・ネットワークは、高い帯域幅を持つネットワークであり、他の EBS ノード、EC2 インスタンス、EBS コントロール・プレーン・サービスとの全てのコミュニケーションにおいて通常の運用で用いられています。セカンダリー・ネットワークは、レプリケーションネットワークであり、帯域幅が比較的少ないバックアップ用のネットワークです。これにより、EBS ノードは、EBS クラスター内の他のノードと安定したコミュニケーションがとれ、データ複製のために追加の容量を提供します。このネットワークは、プライマリー・ネットワークからの全てのトラフィックを扱うように設計されておらず、むしろ EBS クラスターにおける EBS ノード間接続における信頼性を高めるものです。

ノードがデータを複製している際に他のノードへの接続を失ったとき、そのノードは他のノードが落ちたと仮定します。耐久性を担保するために、データを複製する先の、新しいノードを探す必要があります（これは、再ミラーリングと呼ばれます）。再ミラーリングのプロセスの一部において、EBS ノードはその EBS クラスター内において十分なサーバースペースを持った他のノードを探索し、そのサーバーと接続し、そしてボリュームデータを伝達します。正常稼働しているクラスターにおいては、新しいレプリカの場所を探す作業は、ミリ秒で実施されます。データが再ミラーリングされている際に、そのデータのコピーを持っている全てのノードは、新たなノードが完全にそのデータを保持したことを確認するまでデータを保持し続けます。これにより、顧客データの喪失を防ぎ、さらなるレベルの防壁を提供することになります。また、顧客のボリュームのデータが再ミラーリングされる時、システムが新しいプライマリーレプリカ（書き込み可能な）を特定するまで、そのデータへのアクセスはブロックされます。これは、あらゆる障害可能性において、EBS ボリュームデータの一貫性のために必要となります。このような状況は、ボリュームへ I/O 操作を行おうとする EC2 インスタンスの視点からすると、ボリュームはスタックした状態に見えます。

EBS クラスターに加えて、ユーザーリクエストを受け入れ、それを適切な EBS クラスターに伝達するコントロール・プレーン・サービスが存在します。EC2 リージョン毎に、1 セットの EBS コントロール・プレーン・サービスが存在しています。コントロール・プレーン・サービス自体は、可用性と耐障害性のために、アベイラビリティ・ゾーンをまたがって分散されています。これらのコントロール・プレーン・サービスは、EBS クラスターに対して、EBS クラスターが各ボリュームのプライマリ・レプリカを選択する際に（一貫性のために、各ボリュームのプライマリ・レプリカは必ず一つである必要があります）、管理者として振る舞います。コントロール・プレーン・サービスを構成するために複数のサービスが存在していますが、この文書では、これらをまとめて、”EBS コントロール・プレーン”と呼びます。

主な障害

4 月 21 日 12:47AM (PDT) に、米国東リージョンの単一のアベイラビリティ・ゾーンにおいて、通常の拡張作業の一環としてネットワーク設定の変更が行われました。 変更作業の目的は、プライマリー・ネットワークの容量増強でした。変更作業における標準的な手順の一つとして、増強を行うためにプライマリーEBS ネットワーク内の冗長なルーターからトラフィックを移動させます。このトラフィックの移動が正常に行われず、プライマリー・ネットワーク内の別のルーターに移る代わりに、より容量の小さい予備の EBS ネットワークへルーティングされてしまいました。プライマリー・ネットワークのトラフィックが意図的に移動されてしまったものの、移動先であるセカンダリー・ネットワークはそのトラフィック量を受け入れることができなかったため、影響を受けたアベイラビリティ・ゾーン内の EBS クラスターの一部にとっては、プライマリーとセカンダリー・ネットワークの機能が失われたことを意味しました。結果として、アベイラビリティ・ゾーン内の多数の EBS ノードが所属するクラスター内の別ノードから完全に分離された状態になりました。通常起こりうるネットワークの接続不良の状況と異なり、この変更作業によってプライマリーとセカンダリーのネットワーク双方への接続が同時に失われてしまったため、影響を受けたノード間では完全に通信が行えなくなりました。

ネットワーク接続に関する問題が発生したことにより、ある EBS クラスター内の多数の EBS ノードが自身のレプリカと通信できなくなりました。その後、誤ったトラフィック移行作業がロールバックされ、ネットワーク接続が回復しましたが、これらのノードが一斉に、再ミラーリングを行うための容量を EBS クラスター内で検索し始めました。繰り返しになりますが、通常であれば、この検索自体は数ミリ秒で完了します。今回は、大量のボリュームが同時に影響を受けたため、EBS クラスター内の未使用領域があっという間に使い果たされてしまい、多くのノードがスタック状態のままになり、クラスター内で使用できる容量を繰り返し探し続けることになりました。結果として再ミラーリング・ストーム（再ミラーリングの嵐）の状態が発生し、ノードが新しいレプリカを作成するのに必要な容量をクラスター内で検索し続け、多数の EBS ボリュームが実質的にスタックした状態になってしまいました。この時点で、影響を受けたアベイラビリティ・ゾーン内における 13%のボリュームがスタックしていました。

ここまでの説明のような現象が発生した後、障害の発生した EBS クラスターは EBS コントロール・プレーンにも影響を及ぼしはじめました。影響下のアベイラビリティ・ゾーンの EBS クラスターが再ミラーリング・ストームに陥り、容量を使い果たしてしまうと、そのクラスターでは Create Volume API コールが発行できなくなりました。EBS コントロール・プレーン、特に Create Volume の API コールはタイムアウトが長く設定されているため、API コールが遅延したことによってリクエストの処理に利用できるスレッドが枯渇していきました。EBS コントロール・プレーンはリージョンごとにリクエストの処理に使えるスレッド・プールを持っています。キューイングされた大量のリクエストによってスレッドが使い果たされてしまうと、EBS コントロール・プレーンは API リクエストを処理することができなくなり、リージョン内の別アベイラビリティ・ゾーンに対する API コールも失敗するようになりました。4 月 21 日 2:40AM (PDT) に、障害が発生していたアベイラビリティ・ゾーンで新しく発行された Create Volume リ

クエストを無効にする変更を行い、2:50AM までの時点で、Create Volume リクエスト以外の API に関する遅延やエラー発生率は正常値に復帰しました。

本障害の早い段階で、この EBS クラスターの障害をより深刻なものとしてしまった要因は 2 つあります。まず、ノードが新しいノードを発見するのに失敗している間、十分な間隔を置かずに何度も検索を継続してしまいました。また、EBS ノードのプログラム内にレース・コンディション（競合状態）があり、非常に低い確率ながら、EBS ノードが同時に大量のレプリケーション・リクエストを処理しようとする際にノードを停止させてしまい得るものでした。正常に稼働している EBS クラスターでは、仮に発生するとしても、ごく少数のノードがクラッシュするだけのものですが、今回の再ミラーリング・ストームにおいては、極めて大量の接続要求が生じていたため、より頻繁に障害につながってしまう結果となりました。このバグによってノードが失われていった結果、さらに多くの EBS ボリュームにおいて再ミラーリングが必要な状況になりました。この現象によって、さらに多くのボリュームがスタックした結果、より多くのリクエストが再ミラーリング・ストームに追加されました。

5:30AM (PDT) の時点までに、リージョン内で EBS の API コールのエラー発生率と遅延が再び上昇しました。あるボリュームが再ミラーリングを行おうとする場合、EC2 インスタンス、ボリュームのデータを保持している EBS ノード、そして制御を行う EBS コントロール・プレーンの間でネゴシエーションを行い、複数のデータコピーのうち一つだけがプライマリーのレプリカとして指定され、すべてのアクセスの送信先として EC2 インスタンスから認識される必要があります。このネゴシエーション処理によって、EBS ボリュームの一貫性が保証されます。前述の「競合状態」によって多くの EBS ノードが停止し続けたため、EBS コントロール・プレーンにおけるネゴシエーション処理の件数が増加しました。データの再ミラーリングが正常に行われなかったため、システムが再試行を繰り返し、新規リクエストが流入し続けたため、処理件数が増加していきました。この処理負荷によって EBS コントロール・プレーンの処理能力が低下し、再びリージョン内の EBS API コールへ影響を及ぼしました。8:20 AM (PDT) に、EBS チームは障害の発生している EBS クラスターと EBS コントロール・プレーンの間のすべての通信を遮断する作業を開始しました。この作業によって当該アベイラビリティ・ゾーン内での EBS API コールが全て実行されなくなりましたが、リージョン内の別アベイラビリティ・ゾーンについては EBS API コールの遅延とエラー発生率は正常に復帰しました。（API コールが実行されなくなったアベイラビリティ・ゾーンの復旧作業については次のセクションで解説します。）

障害の発生していた EBS クラスターにおいても大多数の EBS ボリュームは依然として正常に稼働していましたので、影響範囲を広げずにクラスターを復旧させることが重要でした。
11:30AM (PDT) に EBS チームが作業を行い、クラスター内のノード間の重要な通信を妨げることなく、障害の発生しているクラスター内の EBS サーバーが他のサーバーへ無意味にアクセスすることを防止するための変更を加えました。（この時点では、新しいノードへ接続できたとしても、いずれにせよ容量に余裕がありませんでした。）この処置が行われてからはクラスター内の障害が進行して新しいボリュームがスタックすることはなくなりました。この変更が加えられる時点までに、競合状態によってサーバーに障害が発生した結果として、当該アベイラビリティ・ゾーン内の 5%の EBS ボリュームが追加的にスタックしてしまっていました。この変更

が実施されるまでの合計としては、アベイラビリティ・ゾーン内でスタックしたボリュームは全ボリューム数の 13%でした。

4 月 21 日の正午まで、障害の生じていたアベイラビリティ・ゾーン以外のアベイラビリティ・ゾーンにおいても、新しく EBS タイプの EC2 インスタンスを起動する際にエラー発生率の上昇がみられました。この現象は、障害発生時刻から 21 日の正午まで約 11 時間にわたって継続していました。前述の経過中で多くの API が使えなかった特定の期間を除くと、EBS タイプの EC2 インスタンスの起動は不可能ではありませんでしたが、高いエラー発生率と遅延が生じていました。EBS タイプのインスタンスの起動処理は、ボリュームを新しいインスタンスにアタッチするために必要な特定の API による影響を受けていました。当初、警告機能の粒度が荒かったため、EBS クラスターの障害による一般的なエラーによって、その API と EC2 起動に関するエラーが見つかりづらい状況でした。11:30AM (PDT) にコントロール・プレーンへの修正が行われ、EBS タイプの EC2 インスタンスの起動におけるエラー発生率は速やかに低下し、正午までにはほぼ正常値に復帰しました。

障害の発生したアベイラビリティ・ゾーンの EBS のリカバリーについて

4 月 21 日 12:04PM （PDT）、障害は一箇所のアベイラビリティ・ゾーンに抑えこみ、劣化した EBS クラスタは安定しました。この為、他のアベイラビリティ・ゾーンでの API 使用は問題なく、新たなボリュームがスタックすることは無くなりました。この時点で我々は、リカバリー処理に焦点を絞りました。障害が発生したアベイラビリティ・ゾーン内のおよそ 13%のボリュームがスタック状態であり、EBS の API はその特定のアベイラビリティ・ゾーンで利用不可となっていました。優先事項は、ストレージ容量を追加し、スタックしたボリュームが新たなレプリカを作成できる容量を作ることでした。

EBS チームは容量追加作業において、2 つの課題に直面しました。1 つ目はノードに障害が発生すると、EBS クラスタは全てのデータレプリカの再ミラーリング処理が終わるまでこのノードを再利用することはありません。これはクラスターに障害が発生した場合、データの復旧を可能にするためです。お客様のボリュームが依存するノードが修復可能なのを確かめるまで障害が発生した容量は使えませんので、大量に新たな容量をクラスターに追加する必要がありました。このため、米国東リージョン内で新たなサーバー容量を物理的に移動させ、劣化している EBS クラスターに追加をするため多大な時間を必要としました。2 つ目はノード間で新たな容量を参照するための通信を抑えこむために修正したところ（上記ステップでクラスタを安定させるために行ったのですが）、新しい容量をクラスタに追加するのに困難が生じました。既存のサーバーに新たな負荷を与えないようにしながら、EBS チームはネゴシエーション処理を新規ビルドしたサーバーに適用させるため慎重に修正を加え処理をさせる必要がありました。これらの処理が想定していたより時間が掛かってしまいました。4 月 22 日 02:00AM（PDT）、EBS チームは大量の新たな容量を追加することに成功しレプリケーションのバックログの作業を開始。4 月 22 日 12：30PM （PDT）、9 時間に渡って EBS ボリューム復旧が行われ、障害の発生したアベイラビリティ・ゾーンの全体の 2.2%以外が復旧しました。復旧した EBS ボリュームはレプリケーション済みだったのですが、EC2 インスタンスからすると、スタックボリュームの少数が復帰したと認識されませんでした。これはボリュームが EBS コントロール・プレーンとの通

信を待ち状態であったためです。コントロール・プレーンと接続されてはじめて、EC2 インスタンスと安全に再接続し、書き込み可能なコピーとして認識されるのです。

クラスターに必要な容量が追加された後、EBS チームは EBS コントロール・プレーンから障害の発生したアベイラビリティ・ゾーンへの API 使用復旧と、スタックし続けているボリュームへのアクセス復旧のため作業をしました。劣化した EBS ノードと EBS コントロール・プレーン間での大量のステート変更の残務処理が溜まっている状態でした。復旧したボリュームと EBS コントロール・プレーンへ影響を与えないように、これらを徐々に適用していきました。最初に、障害の発生したアベイラビリティ・ゾーンへの API アクセスを有効にする試みを行いましたが、EBS コントロール・プレーンに負荷をかけないように、ステート変更の残務処理の処理速度を慎重に調整しました。また、同じリージョン内の他のアベイラビリティ・ゾーンに影響を与えないために、残務処理を行うための別の EBS コントロール・プレーンの準備を開始しました。残務処理の処理速度を調整するスロットルは、システム全体を安定させるにはあまりに粗い作りであったことが分かってきましたので、4 月 22 日の夕方から 4 月 23 日早朝にはスロットルをより最適化するための作業をしました。土曜日にはスロットルの最適化と専用 EBS コントロール・プレーンの準備が完了しました。この EBS コントロール・プレーンへのトラフィックの初期テストは好調で、4 月 23 日 11:30 AM （PDT）には残務処理の実施を徐々に進めました。3:35PM （PDT）、この EBS コントロール・プレーンから障害の発生したアベイラビリティ・ゾーンへのアクセスが完成しました。EBS コントロール・プレーンが調整してくるのを待機していたボリュームは、アタッチしていたインスタンスから利用出来るようになりました。4 月 23 日 6:15 PM（PDT）、障害の発生したアベイラビリティ・ゾーンの EBS への API アクセスが利用可能となりました。

障害の発生した API アクセスを再開したことにより、リージョン内の全アベイラビリティ・ゾーンでの API 利用が可能になりました。残り 2.2%のボリューム復旧には手動で行う必要があり、EBS チームは障害の初期段階からこれらのボリュームをデータロスから守るため念の為にスナップショットを S3 に作成していました。EBS チームはスナップショットからボリュームを復旧させるためのプログラムを用意しテストを行い、夜中に徐々に適用していきました。4 月 24 日 12:30PM （PDT）、この方法で復旧できるボリュームはすべて復旧し、障害の発生したボリュームの 1.04%以外が復旧されました。EBS チームはマシントラブルなどによるスナップショットの作成が不可能であった残りのボリュームを解析し 03:00 PM （PDT）には復旧作業を開始。結果として、障害の発生したアベイラビリティ・ゾーンのボリュームのうち 0.07%を、障害前と一貫した状態に復旧することができませんでした。

Amazon Relation Database Service (RDS) への影響

障害の発生した EBS は EC2 インスタンスに影響を与えた上に、Relational Database Service (RDS) にも影響を与えました。RDS はデータベースとログストレージに EBS を利用しています。これによって障害の発生したアベイラビリティ・ゾーンでの一部の RDS に影響が出てしまいました。

RDS インスタンスを利用する際に、単一のアベイラビリティ・ゾーン (以後、single-AZ) 、もしくは、レプリケーションが行われるマルチプル・アベイラビリティ・ゾーン (複数のアベイラビリティ・ゾーン、以後、multi-AZ) を、お客様が選ぶことができます。Single-AZ のデータベースインスタンスの場合、アベイラビリティ・ゾーンでの障害時に影響を受けてしまいます。この度スタックしたボリュームが single-AZ の RDS だったものが影響を受けました。障害の発生したアベイラビリティ・ゾーンでピーク時 45%の single-AZ のインスタンスが"スタック"I/O となりました。該当する EBS よりも RDS への影響が比較的に多くなってしまいましたが、これは、RDS が複数の EBS ボリュームを利用しているからです。複数 EBS ボリュームを利用することでデータベースの I/O 容量を増加できるのですが、single-AZ のデータベースインスタンスが用いるボリュームの内一つでも"スタック"I/O が発生した場合、オペレーションが出来なくなってしまいます。EBS の復旧が進むにつれスタックしていた single-AZ データベースインスタンスの割合も減っていき、24 時間後には 41.0%、36 時間後には 23.5%、48 時間後には 14.6%とし、最終的に週末には作業が終了しました。大部分のデータベースインスタンスを復旧することが出来たのですが、障害の発生したアベイラビリティ・ゾーン内の 0.4%の single-AZ データベースインスタンスが、復旧できない EBS ボリュームを保持することとなりました。これらのデータベースインスタンスで、自動バックアップ（デフォルト時オン）を用いていたお客様は、point-in-time データベースリストアでリカバリーをすることが可能です。

RDS multi-AZ は別々のアベイラビリティ・ゾーンに存在する 2 台のデータベースレプリカにデータを同期し冗長化を提供します。プライマリ・レプリカの障害時にも RDS は自動的にそれを判断しセカンダリレプリカへフェイルオーバーします。この度、"スタック" I/O が発生し米国東リージョン内の 2.5%の multi-AZ データベースで自動的にフェイルオーバーが発生しませんでした。主な原因としてネットワーク障害が起き（プライマリーとセカンダリーでの通信が不可になり）、プライマリ・レプリカでの"スタック"I/O が新たなバグとして生じました。このバグが依存する状態で我々のモニタリングエイジェントが自動的にセカンダリレプリカへフェイルオーバー処理をさせるのはデータ紛失の恐れもありマニュアル操作が必要となりました。こちらのバグに関する修正についても鋭意作業しています。

障害を避けるための防止策

この障害の発端はネットワークの設定の変更でした。我々は将来に向けてこのよう事態の発生を防ぐために変更プロセスの監査を行い、自動化を進める予定でおります。また、このような障害にも耐えうるソフトウェアとサービスを構築することに専念してまいります。この障害に対して施した多くの作業は、EBS サービスが将来起こりうる同様の障害に直面したときからも守るための糧となります。

我々は、今後に向けて幾つかの対策を施し、クラスターが再ミラーリングのストームに巻き込まれることから防ぐようにします。まず、余剰分の容量を持つことで、デグレードした EBS クラスターが大量の再ミラーリングのリクエストを吸収し、再ミラーリングのストームを避けることが出来ます。大量のリカバリー処理をする際に、どの程度の容量が必要になったかをより理解しましたので、これまでの容量プランニングと監視を修正し、巨大なスケールでの障害時

において追加の安全な容量を持つことにします。容量バッファは既に飛躍的に増加させており、この数週間内で必要な新規の容量を持つ予定でおります。また、EBS サーバノードにおけるリトライのロジックを修正し、再ミラーリングのストームにクラスターが巻き込まれることから防ぎます。大量のインタラプションが発生した場合に、リトライロジックをより積極的に譲歩させて、再ミラーするために無駄に新しいノードを探索に行くのではなく、既存のレプリカを使って再接続を試みるようにしていきます。この変更を実施していく中で、このロジックの修正によって再ミラーリングのストームの根本原因に対処できるという確信を得ております。最後に、EBS ノード障害を引き起こした競合状態のソースを発見するに至りました。我々は修正方法がわかったので、今後数週間の間にテストを行ったうえで我々のクラスターにデプロイしていきます。これらの変更は障害再発に対して、3 つの別々の防護策を提供します。

マルチプル・アベイラビリティ・ゾーン(複数のアベイラビリティ・ゾーン)への影響

EC2 はリージョンとアベイラビリティ・ゾーンという可用性を実現する上で、2 つの重要な構成要素を提供しています。リージョンは、意図的にインフラストラクチャを完全に分離して展開しているものです。多くのお客様は非常に高いレベルの耐障害性を達成するために EC2 のリージョンを複数利用しております。しかしながら、我々はお客様のいかなるデータもリージョン間でレプリケートしないために、その場合はお客様ご自身のアプリケーションを通じて行っていただく必要があります。また各リージョンを管理するのに、異なった API のセットを使う必要があります。リージョンはパワフルな可用性の構成要素をお客様に提供しますが、独立性の利点を得るためには一部アプリケーション作成者の貢献が必要になります。リージョン内においては、お客様が高い耐障害性のあるアプリケーションを容易に構築できるためにアベイラビリティ・ゾーンというものを提供しています。アベイラビリティ・ゾーンは、お客様に高速で低遅延なネットワーク接続やデータのレプリケート、そして一貫した管理用 API のセットを提供しつつ、高度に独立するように構築されている、物理的にも論理的にも分離されたインフラストラクチャです。例えばあるリージョン内で稼働している場合において、お客様は EBS スナップショットを取得し、如何なるアベイラビリティ・ゾーン内においてもリストアすることが出来、また同一の API を使って EC2 や EBS リソースとしてプログラムから操作することが出来ます。このような疎結合な仕組みを提供することで、お客様が高い耐障害性のあるアプリケーションを容易に構築することが出来るようになります。

今回の障害において、大きく分けて 2 つの影響がありました。1 つ目として、影響を受けた EBS ボリュームが”スタック”したため、該当アベイラビリティ・ゾーンで稼働しているアプリケーションに影響が出ました。EBS サービスのアーキテクチャによって、この稼働しているインスタンスへの影響は該当アベイラビリティ・ゾーンだけに限定されておりました。結果的に、マルチプル・アベイラビリティ・ゾーンの利点を生かしているアプリケーションを構築されているお客様にこの障害による甚大な可用性の被害は出ておりません。木曜日の時点で、幾つかのお客様から該当のアベイラビリティ・ゾーン以外でも EBS ボリュームが”スタック”しているとの報告を受けました。私共の監視システムでは EBS コントロール・プレーン分及びボリュームでの再ミラーリングのストームは明らかに該当アベイラビリティ・ゾーン内だけで影響を及ぼ

しており、同一リージョン内の残りのアベイラビリティ・ゾーン内にある既存 EBS ボリュームに対しての大きな影響はありませんでした。
正常なアベイラビリティ・ゾーンにおいて我々の想定よりもやや多い”スタック”したボリュームが発見されましたが、それでも極めて少ない数です。言い換えますと、該当リージョン内の影響のあったアベイラビリティ・ゾーン以外における、”スタック”したボリュームのパーセンテージのピークは 0.07%以下でした。これらの”スタック”したボリュームの数も調査しました。このわずかに高い数の”スタック”したボリュームは、前述した EBS コントロール・プレーンの遅延とエラー率から影響を受けており、正常な再ミラーリングからの復旧の遅れが原因です。ボリュームの再ミラーリングは常に実施されており、常にそういうボリュームは存在しているのが現状です。
また、下記で記述する内容を実施することによって、EBS コントロール・プレーンをより強固にし、このようなわずかに高いレートですら、同様の障害から防いでくれると確信しています。

お客様がマルチプル・アベイラビリティ・ゾーン(multi-AZ)の利点を活かすことでこの障害の影響を避けることが出来る一方で、EBS コントロール・プレーンでリージョンを超えて EBS ボリュームの生成と操作に対して影響があったことをご報告させて頂きます。EC2 の 1 つの利点は問題のあったリソースを瞬時に入れ替えることが出来る点です。EBS コントロール・プレーンが性能低下していた際、もしくは、稼働していなかった際には、お客様が影響のあったボリュームを異なるアベイラビリティ・ゾーンにある EBS ボリュームや EBS ブートの EC2 インスタンスに移行するのは困難でした。このような事態の再発防止は我々の最優先事項です。

お客様に数々の疎結合の仕組みを提供しているとはいえ、我々のデザイン・ゴールは、アベイラビリティ・ゾーンを完全に独立して稼働しているのと変わらない状態することです。我々の EBS コントロール・プレーンは、各ゾーンにおける障害に対する耐久性を持ちながら、ユーザがマルチプル・アベイラビリティ・ゾーンにあるリソースにアクセスできるように設計されています。この障害から、デザイン・ゴールを完全にするために、我々がより投資をしていなくてはいけない事を学びました。ある 1 つのアベイラビリティ・ゾーンがマルチプル・アベイラビリティ・ゾーンを超えた EBS コントロール・プレーンへ影響を与えないために、我々は 3 つの施策を実施します。1 つ目として、単一のアベイラビリティ・ゾーンのクラスタで処理に時間のかかりすぎるリクエストによってスレッド枯渇が発生するのを防ぐために、タイムアウトのロジックを改善することを行います。この変更によって、4 月 21 日の午前 12 時 50 分から午前 2 時 40 分までの API への影響を防げていた事でしょう。2 番目に起こった API への影響の原因対策として、EBS コントロール・プレーンを機能拡張し、よりアベイラビリティ・ゾーンを意識できる作りにして、オーバー容量時に負荷を賢く配分するようにします。これは我々のシステムで、既に実現している他のスロットル機能に似ています。さらに、EBS コントロール・プレーンが持つ機能を、各 EBS クラスタサービスに持ち込むことができると考えています。EBS コントロール・プレーンから多くの機能を移し、これらのサービス (サポートしている EBS クラスターと同じアベイラビリティ・ゾーンで動く) を EBS クラスタ毎に配置することで、EBS コントロール・プレーンにおいて、今までよりもさらにアベイラビリティ・ゾーンの独立性を提供できるようになります。

マルチプル・アベイラビリティ・ゾーン利用をより簡単にするために

我々は、お客様がより簡単にマルチプル・アベイラビリティ・ゾーンを利用し、そのメリットを得られるようにしていきます。第一に、Amazon Virtual Private Cloud （VPC）を含む全ての我々のサービスを、マルチプル・アベイラビリティ・ゾーンで提供するようにします。現在、VPC は単一のアベイラビリティ・ゾーンでのみ利用できます。我々は早急にマルチプル・アベイラビリティ・ゾーンで VPC が利用できるように、ロードマップを修正します。これによって VPC のお客様は、現在 VPC を利用していない EC2 のみのお客様がおこなっているように、マルチプル・アベイラビリティ・ゾーンを利用した可用性の高いアプリケーションを構築できるようになります。

今回の障害からも、信頼性の高い Multi-AZ をより簡単に設計・運用して頂けるように、我々がより良い仕事をする必要があることを認識しています。あるお客様のアプリケーション（あるいはデータベースのようにクリティカルなコンポーネント）が単一のアベイラビリティ・ゾーンにだけに配置されている場合には、他のコンポーネントがマルチプル・アベイラビリティ・ゾーンにまたがって配置されていたとしても、クリティカルな単一障害点が単一のアベイラビリティ・ゾーンに存在することになります。このようなケースでは運用上の問題が発生したときに、アプリケーションに深刻な影響を及ぼします。その一方で、複数アベイラビリティ・ゾーンに配置された堅牢なアプリケーションでは運用を継続できます。我々はお客様に対して、特定のアベイラビリティ・ゾーンがまるごと消失したとしても、アプリケーションの可用性に影響を与えないような、Multi-AZ アプリケーションを作成するためのより良いツールの提供を目指します。我々は、お客様が共通のデザインパターンを使ってアプリケーションロジックの設計が出来るよう、お手伝いしなければならないことを認識しています。本件では、あるお客様は深刻な影響を受けた一方で、あるお客様はリソース自体には影響があったものの彼らのアプリケーションには影響が見られませんでした。

クラウドにおけるアーキテクチャのベストプラクティスを、よりお客様、パートナーにご理解して頂くために、無料の Webinar を 5 月 2 日から（日本語では 5 月 10 日から）開始します。最初に扱うトピックは、耐障害性のあるアプリケーションの設計、クラウドのためのアーキテクチャ、Web ホスティングのベストプラクティスです。来る数週間のうちにさらに多くの一連のトピックを追加し、継続的に頻繁に行っていきます。実施する Webinar は、世界中のお客様に対して、複数のタイムゾーンで数回ずつ行います。我々は詳細な Q&A のための時間を Webinar に用意します。お客様またはパートナー向けのフォローアップディスカッションも用意する予定です。それらの Webinar は、一連の AWS クラウドのためのアーキテクチャのベストプラクティスホワイトペーパーと同様に、新しい AWS ウェブサイト内のアーキテクチャセンター(英語)から入手できます。また、Multi-AZ レベルの自動バランシングを実現している S3、SimpleDB および Multi-AZ RDS のように、お客様がアプリケーション上で面倒な作業をすることなく複数アベイラビリティ・ゾーンによる利益を得られるような追加サービスを、継続して提供できるようにします。

リカバリーの高速化

我々は、EBS クラスターにおけるボリューム回復の透明性、制御、自動化の向上にも力をいれます。我々は EBS クラスターを管理運用するための様々なツールを持っていますが、チームがクラスターの回復のために用いた、より詳細な制御や、スロットリングを、EBS ノードに直接構築していきます。ボリューム回復のために必要であった様々な種別のリカバリモデルを、自動化できるようにしていきます。これによって回復プロセスの多くの時間を節約することになります。また、クラスター運用が性能低下しているときにボリューム機能の保護のために何ができるかを調査します。“スタック”ボリュームのスナップショット作成する機能もここに含まれるでしょう。もしお客様がこれを使えれば、リージョン内の他のアベイラビリティ・ゾーンにアプリケーションを回復させることがより簡単になります。

障害発生時のコミュニケーションとサービスヘルスツールの改善

今回の障害にける技術的な洞察および改良に加えて、我々はお客様とのコミュニケーションに対して改善の必要性を認識しました。我々はコミュニケーションをより頻繁に行い、かつ多くの情報を含むようにしたいと考えています。サービス停止の間、お客様は出来る限り詳細な挙動や、修復にどれだけの期間がかかるのか、そして再発させないために我々が何をしているのかを知りたいと願っていることを理解しています。シニアリーダーシップチーム全体を含むほとんどの AWS チームは、事件の調整やトラブルシュートおよび回復に直接関与しました。障害が発生した当初、根本的な原因の特定よりも、お客様のためにどのように運用上の問題として解決するかを考える事に集中していました。我々は、解決に注力し、問題に対しては注力しないことがお客様にとって正しいことあり、それが我々のサービス回復とお客様のより早期の状態回復の助けになると考えていました。我々は、新しい情報があり、その情報が正確であると確信したときにお客様にお報せするよう心がけていました。サービスが正常状態に戻った際に、データ収集と分析を行い、この詳細報告(post mortem)を行うことを予定していました。とは言うものの、我々はこの領域について改善できると考えています。我々は今回の障害が発生してからの一連の流れの中で、より定期的なアップデートを行うように変更しました。また、今後も同様の頻度でのアップデートを続ける計画です。加えて、本件のようなトラブルにおいてはデベロッパーサポートチームを拡張して配置し、より早く意味ある情報を提供するように組織できるよう改善作業を開始しています。また、お客様自身で、利用リソースが影響をうけたかどうかを、より簡単に理解できるよう、インスタンスに障害が発生したかどうか API を通じてお客様自身で見ることができるようなツールを開発中です。

影響を受けたお客様へのサービスクレジットについて

障害のあった時間帯に、米国東リージョン内の影響を受けたアベイラビリティ・ゾーンにおいて、アタッチされた EBS ボリューム、動作中の RDS データベースインスタンスがあったお客様に対しては、そのお客様のリソースやアプリケーションが影響を受けたかどうかに関わらず、EBS ボリューム、EC2 インスタンス、RDS データベースインスタンスの 10 日分の利用の 100%

相当額を提供します。お客様はこれを受け取るために特に何もする必要はありません。自動で次回の AWS の請求に反映されます。お客様は AWS Account Activity ページをログインすればその対象かどうか確認できます。

おわりに

最後になりましたが、お客様にあらためてお詫びを申し上げます。我々は AWS のサービスがお客様のビジネスにとってどれほど重要であるか存じ上げております。そして我々は本件から学び、サービス全体を改善するために可能な全てを行って参ります。我々は、本件のあらゆる側面における詳細な理解と、サービスとプロセスの改善の意思決定のために、これからの数週間にわたって時間を費やす所存でございます。

AWS チーム