Red Hat OpenShift Container Storage 4.8

Recovering a Metro-DR stretch cluster

クラスターおよびストレージ管理者の災害復旧タスク

Last Updated: 2023-02-21

Red Hat OpenShift Container Storage 4.8 Recovering a Metro-DR stretch cluster

クラスターおよびストレージ管理者の災害復旧タスク

Enter your first name here. Enter your surname here.
Enter your organisation's name here. Enter your organisational division here.
Enter your email address here.

法律上の通知

概要

このドキュメントでは、Red Hat OpenShift Container Storage で大規模な障害から復旧する方法を説明します。 This is a technology preview feature and is available only for deployments using local storage devices. Technology Preview features are not supported with Red Hat production service level agreements (SLAs) and might not be functionally complete. Red Hat does not recommend using them in production. These features provide early access to upcoming product features, enabling customers to test functionality and provide feedback during the development process.

目次

多様性を受け入れるオープンソースの強化

Red Hat では、コード、ドキュメント、Web プロパティーにおける配慮に欠ける用語の置き換えに取り組んでいます。まずは、マスター (master)、スレーブ (slave)、ブラックリスト (blacklist)、ホワイトリスト (whitelist) の 4 つの用語の置き換えから始めます。この取り組みは膨大な作業を要するため、今後の複数のリリースで段階的に用語の置き換えを実施して参ります。詳細は、弊社の CTO、Chris Wright のメッセージ を参照してください。

RED HAT ドキュメントへのフィードバックの提供

弊社のドキュメントについてのご意見をお聞かせください。ドキュメントの改善点があれば、ぜひお知らせください。フィードバックをお寄せいただくには、以下をご確認ください。

- 特定の部分についての簡単なコメントをお寄せいただく場合は、以下をご確認ください。
  1. ドキュメントの表示が **Multi-page HTML** 形式になっていて、ドキュメントの右上隅に **Feedback** ボタンがあることを確認してください。
  2. マウスカーソルで、コメントを追加する部分を強調表示します。
  3. そのテキストの下に表示される **Add Feedback** ポップアップをクリックします。
  4. 表示される手順に従ってください。
- より詳細なフィードバックを行う場合は、Bugzilla のチケットを作成します。
  1. Bugzilla の Web サイトに移動します。
  2. **Component** セクションで、**documentation** を選択します。
  3. **Description** フィールドに、ドキュメントの改善に関するご意見を記入してください。ドキュメントの該当部分へのリンクも記入してください。
  4. **Submit Bug** をクリックします。

第1章 概要

大規模な障害復旧のストレッチクラスターでは、完全または部分的なサイトの停止に直面しても回復性を提供することを考えると、アプリケーションとそのストレージのさまざまな復旧方法を理解することが重要です。

アプリケーションがどのように設計されているかによって、アクティブゾーンで再び利用できるようになるまでの時間が決まります。

サイトの停止に応じて、アプリケーションとそのストレージの復旧方法が異なる場合があります。復旧時間は、アプリケーションのアーキテクチャーによって異なります。復旧のさまざまな方法は以下のとおりです。

- RWX ストレージのあるゾーン対応の HA アプリケーションの復旧
- RWX ストレージを使用する HA アプリケーションの復旧
- RWO ストレージのあるアプリケーションの復旧
- StatefulSet Pod の復旧

第2章 ゾーンの障害について

このセクションでは、ゾーンの障害を、ゾーン内のすべてのOpenShift Container Platformノード、マスター、およびワーカーが2番目のデータゾーンのリソースと通信しなくなった（例: ノードの電源がオフになっている）障害と見なします。データゾーン間の通信が部分的にまだ機能している場合（断続的なダウン/アップ）は、復旧を成功させるために、クラスター、ストレージ、ネットワークの管理者がデータゾーン間の通信パスを切断する手順を実行する必要があります。

第3章 RWX ストレージのあるゾーン対応の HA アプリケーションの復旧

topologyKey: topology.kubernetes.io/zoneでデプロイされ、各データゾーンで1つ以上のレプリカがスケジュールされ、共有ストレージ（つまりRWX cephfsボリューム）を使用しているアプリケーションは、新しい接続に対して30～60秒以内にアクティブゾーンで回復します。ルーター Pod が失敗したデータゾーンでオフラインになると、**HAProxy** が接続を更新する短い一時停止です。

このタイプのアプリケーションの例は、Install Zone Aware Sample Application セクションを参照してください。

注記

サンプルアプリケーションをインストールするときは、OpenShift Container Platform ノード（少なくともOpenShift Container ストレージデバイスがあるノード）の電源をオフにしてデータゾーンの障害をテストし、ファイルアップローダーアプリケーションが利用でき、新しいファイルをアップロードできることを検証します。

第4章 RWX ストレージを使用する HA アプリケーションの復旧

topologyKey: kubernetes.io/hostname を使用している、またはトポロジ構成をまったく使用していないアプリケーションは、同じゾーンにあるすべてのアプリケーションレプリカに対する保護がありません。

注記

これは、Pod 仕様の podAntiAffinity および **topologyKey: kubernetes.io/hostname** でも発生する可能性があります。これは、この非アフィニティルールがホストベースであり、ゾーンベースではないためです。

これが発生し、すべてのレプリカが障害のあるゾーンにある場合、RWXストレージを使用するアプリケーションはアクティブゾーンで復旧するのに6～8分かかります。この一時停止は、障害が発生したゾーンのOpenShift Container Platformノードが**NotReady**（60秒）になり、デフォルトの Pod エビクションタイムアウトが期限切れ（300秒）になるためです。

第5章 RWO ストレージを使用したアプリケーションの復旧

RWO ストレージ(ReadWriteOnce)を使用するアプリケーションには、この Kubernetes の問題 で説明されている既知の動作があります。この問題のため、データゾーンに障害が発生した場合、RWOボリュームをマウントしているそのゾーンのアプリケーション Pod（例: **cephrbd**ベースのボリューム）は6〜8分後に **Terminating** ステータスのままになり、手動の介入なしではアクティブゾーンに再作成されません。

ステータスが **NotReady** の OpenShift Container Platform ノードを確認します。OpenShift コントロールプレーンとの通信を妨げる問題がある場合があります。この通信の問題であっても、永続ボリュームに対して IO 操作を実行している可能性があります。

2 つの Pod が同じ RWO ボリュームに同時に書き込む場合は、データ破損のリスクが発生します。**NotReady**ノード上のプロセスが終了するか、終了できるようになるまでブロックされるようにするには、何らかの対策を講じる必要があります。

- 帯域外管理システムを使用してノードの電源をオフにし、確認を行うことは、プロセスを確実に終了させる例です。
- 障害が発生したサイトのノードがストレージと通信するために使用するネットワークルートを使用しないことも解決策になります。

注記

障害が発生した1つまたは複数のゾーンにサービスを復元する前に、永続ボリュームを持つすべての Pod が正常に終了したことを確認する必要があります。

Terminating Pod がアクティブなゾーンで再作成されるようにするには、Pod を強制的に削除するか、または関連付けられた PV でファイナライザーを削除します。これら 2 つのアクションのいずれかが完了すると、アプリケーション Pod がアクティブゾーンで再作成され、その RWO ストレージが正常にマウントされます。

Pod の強制削除

強制削除は、Podが終了したというkubeletからの確認を待ちません。

```
$ oc delete pod <PODNAME> --grace-period=0 --force --namespace <NAMESPACE>
```

<PODNAME>

Pod の名前です。

<NAMESPACE>

プロジェクトの namespace です。

関連付けられた PV のファイナライザーの削除

Terminating Pod によってマウントされる Persistent Volume Claim (PVC) に関連付けられた PV を見つけ、**oc patch** コマンドを使用してファイナライザーを削除します。

```
$ oc patch -n openshift-storage pv/<PV_NAME> -p '{"metadata":{"finalizers":[]}}' --type=merge
```

<PV_NAME>

Pod の名前です。

関連付けられた PV を見つける簡単な方法として、Terminating Pod を記述することができます。複数割り当ての警告が表示される場合は、PV 名が警告に含まれるはずです（例: pvc-0595a8d2-683f-443b-aee0-6e547f5f5a7c）。

```
$ oc describe pod <PODNAME> --namespace <NAMESPACE>
```

<PODNAME>

Pod の名前です。

<NAMESPACE>

プロジェクトの namespace です。
出力例:

```
[...]
Events:
  Type     Reason                  Age   From                     Message
  ----     ------                  ----  ----                     -------
  Normal   Scheduled               4m5s  default-scheduler        Successfully assigned openshift-storage/noobaa-db-pg-0 to perf1-mz8bt-worker-d2hdm
  Warning  FailedAttachVolume      4m5s  attachdetach-controller  Multi-Attach error for volume "pvc-0595a8d2-683f-443b-aee0-6e547f5f5a7c" Volume is already exclusively attached to one node and can't be attached to another
```

第6章 STATEFULSET POD の復旧

ステートフルセットの一部である Pod には、RWO ボリュームをマウントする Pod と同様の問題があります。詳細は、Kubernetes リソースの StatefulSet considerations を参照してください。

6-8 分後にアクティブなゾーンで再作成するために StatefulSet の Pod 部分を取得するには、RWO ボリュームを持つ Pod と同じ要件（つまり、OpenShift Container Platformノードの電源をオフにするか通信を切断する）で Pod を強制的に削除する必要があります。