



# Red Hat OpenShift Data Foundation 4.9

## Recovering a Metro-DR stretch cluster

Red Hat OpenShift Data Foundation での大規模な障害からアプリケーションとそのストレージを復旧する方法に関する説明



## Red Hat OpenShift Data Foundation 4.9 Recovering a Metro-DR stretch cluster

---

Red Hat OpenShift Data Foundation での大規模な障害からアプリケーションとそのストレージを復旧する方法に関する説明

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

Copyright © 2022 | You need to change the HOLDER entity in the en-US/Recovering\_a\_Metro-DR\_stretch\_cluster.ent file |.

The text of and illustrations in this document are licensed by Red Hat under a Creative Commons Attribution–Share Alike 3.0 Unported license ("CC-BY-SA"). An explanation of CC-BY-SA is available at

<http://creativecommons.org/licenses/by-sa/3.0/>

. In accordance with CC-BY-SA, if you distribute this document or an adaptation of it, you must provide the URL for the original version.

Red Hat, as the licensor of this document, waives the right to enforce, and agrees not to assert, Section 4d of CC-BY-SA to the fullest extent permitted by applicable law.

Red Hat, Red Hat Enterprise Linux, the Shadowman logo, the Red Hat logo, JBoss, OpenShift, Fedora, the Infinity logo, and RHCE are trademarks of Red Hat, Inc., registered in the United States and other countries.

Linux<sup>®</sup> is the registered trademark of Linus Torvalds in the United States and other countries.

Java<sup>®</sup> is a registered trademark of Oracle and/or its affiliates.

XFS<sup>®</sup> is a trademark of Silicon Graphics International Corp. or its subsidiaries in the United States and/or other countries.

MySQL<sup>®</sup> is a registered trademark of MySQL AB in the United States, the European Union and other countries.

Node.js<sup>®</sup> is an official trademark of Joyent. Red Hat is not formally related to or endorsed by the official Joyent Node.js open source or commercial project.

The OpenStack<sup>®</sup> Word Mark and OpenStack logo are either registered trademarks/service marks or trademarks/service marks of the OpenStack Foundation, in the United States and other countries and are used with the OpenStack Foundation's permission. We are not affiliated with, endorsed or sponsored by the OpenStack Foundation, or the OpenStack community.

All other trademarks are the property of their respective owners.

## 概要

このドキュメントでは、Red Hat OpenShift DataFoundation で大規模な障害から復旧する方法を説明します。Recovering a Metro-DR stretch cluster is a technology preview feature. 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

前書き .....	3
第1章 ゾーンの障害について .....	4
第2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 .....	5
第3章 RWX ストレージを使用する HA アプリケーションの復旧 .....	6
第4章 RWO ストレージを使用したアプリケーションの復旧 .....	7
第5章 STATEFULSET POD の復旧 .....	9



## 前書き

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は異なります。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 のリカバリー](#)

## 第1章 ゾーンの障害について

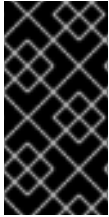
このセクションでは、ゾーン障害は、ゾーン内のすべての OpenShift Container Platform、マスターノード、およびワーカーノードが 2 番目のデータゾーンのリソース (たとえば、電源がオフになっているノード) と通信しなくなった障害と見なされます。データゾーン間の通信がまだ部分的に機能している (断続的にアップまたはダウンしている) 場合、回復を成功させるには、クラスター、ストレージ、およびネットワーク管理者がデータゾーン間の通信パスを切断する必要があります。



## 第2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 ReadWriteMany (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。



### 重要

サンプルアプリケーションをインストールする場合は、OpenShift Container Platform ノード (OpenShift Data Foundation デバイスを使用するノード) の電源をオフにし、file-uploader アプリケーションが利用可能であることを検証するためにデータゾーンの失敗をテストし、新しいファイルをアップロードします。

## 第3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、またはトポロジー設定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ー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、ReadWriteMany (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回復するのに 6-8 分の時間がかかります。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4章 RWO ストレージを使用したアプリケーションの復旧

ReadWriteOnce (RWO) ストレージ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は、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。ノードが OpenShift コントロールプレーンと通信できない問題が生じる可能性があります。ただし、ノードは永続ボリューム (PV) に対して I/O 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** ノードのプロセスが終了するか、または終了するまでブロックされていることを確認します。

ソリューションの例:

- 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によってストレージとの通信に使用されるネットワークルートを無効します。



### 注記

障害のあるゾーンまたはノードにサービスを復元する前に、PV が指定されたすべての Pod が正常に終了していることを確認します。

**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 を強制的に削除

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-ae0-6e547f5f5a7c**)。

```
$ 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-ae0-6e547f5f5a7c" Volume is already exclusively attached to one
node and can't be attached to another
```

## 第5章 STATEFULSET POD の復旧

ステートフルセットの一部である Pod には、ReadWriteOnce (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 [StatefulSet の考慮事項](#) で参照されます。

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 ノードの電源オフまたは通信が切断されている) で Pod を強制的に削除する必要があります。