Technical White paper
RAID 保護とドライブ故障の Fast Recovery

RAID 保護は、すべての ETERNUS®ストレージシステムの中核を担います。お客様のアプリケーション要件に合ったレベルを選択するには、アプリケーションの要件を様々な角度で評価する必要があります。本書では、冗長性のなくなったモードでの動作、保護された状態にリカバリーするまでの時間、完全に復旧するまでの時間について考察します。Fast Recovery 機能は、RAID レベルの選択において長年抱かれてきた前提を変えます。
Fast Recovery 機能を導入すると、ほかの RAID レベルに比べてわずかな時間で保護された状態にリカバリーすることができます。

目次

図目次

表目次

1 はじめに

RAID は、長年にわたってストレージシステムにおけるデータ損失を防ぐ標準的な方法であり、ETERNUS ストレージシステム製品群では複数の形態を利用できます。ドライブが故障したときに、データが失われるのを防ぐ重要な要素には 3 つあります。1 つ目は故障が検出されたときの動作、2 つ目は保護された状態へのリカバリー、3 つ目は故障したドライブの交換時期およびストレージシステムで交換作業がどのように行われるかです。

ETERNUS ストレージシステムに Fast Recovery 機能を導入することにより、データの保護環境およびリカバリー環境が変わります。本書ではこの機能の価値を示し、RAID 保護の仕組みがストレージシステム製品の効果的なデプロイメントに与える影響について説明します。

RAID 技術については、富士通の資料および公開資料に多くの文書があるため、本書では詳しい説明は省いてあります。

2 定義

本書では、説明を十分に理解するためにいくつかの用語を定義しています。

- **完全保護** – 特定の RAID グループ構成で提供されるすべての保護機能が有効な場合の RAID グループの状態。

- **冗長性のなくなったモード** – アプリケーションからデータにアクセスできるが、保護されていないか、または完全保護に比べて低い保護レベルの RAID グループの状態。
- **リビルドモード** – アプリケーションからデータにアクセスできるが、故障は続いており、RAID グループを完全保護の状態に復旧する途中の状態。
- **コピーバックモード** – RAID グループは完全保護の状態だが、交換した新しいドライブに保護データを復旧している状態。
- **コピーバックレスモード** – もとのドライブで保存されていたデータを、交換する新しいドライブで格納せずに、リビルド対象に保護データを残すモード。
- **グローバルホットスペア**(GHS) – ストレージシステムの中の 1 つ以上のドライブで、RAID 構成に適したリビルド処理で故障したドライブの代わりにするために、すべての RAID グループから使用されます。
- **専用ホットスペア** – RAID グループを構成する 1 つのドライブ。アクティブデータを保持せず、グループ内の故障したドライブの代わりに使用します。
- **RAID6-FR** – RAID6 グループの特殊な形態。グループの通常運用で使用するすべてのドライブに専用ホットスペア相当の領域を持っています。ホットスペア領域は、グループ内のすべてのドライブに分配されています。有効なメンバーディスクの組み合わせは限られます。「xD」はデータドライブの数、「2P」は 2 つのパリティドライブ、「1HS」は 1 つの専用ホットスペアドライブを示します。有効な組み合わせは次のとおりです。

RAID6-FR 構成（RAID グループあたりの総ドライブ数順）	RAID グループあたりのユーザードライブ数	RAID グループあたりの総ドライブ数
(3D+2P)x2+1HS	6	11
(6D+2P)x2+1HS	12	17
(9D+2P)x2+1HS	18	23
(5D+2P)x4+1HS	20	29
(12D+2P)x2+1HS	24	29
(8D+2P)x3+1HS	24	31
(3D+2P)x6+1HS	18	31
(13D+2P)x2+1HS	26	31

表 1 – 有効な RAID6-FR RAID グループの組み合わせ

- **ユーザビリティ係数** – グループのドライブの総容量のうち、ユーザーデータの保持に使用される容量(表 3 を参照)。ユーザビリティ係数が高くなると、一定のユーザーストレージ量に必要なドライブ数が少なくなりコストは低くなります。ユーザビリティ係数が低くなると、一定のユーザーストレージ量に必要なドライブ数が多くなりコストは高くなります。

3 故障と保護の関係

RAID 保護されたセットの中でディスクドライブが故障した場合、動作に変化が起こります。故障したドライブに保持されていたデータ領域はもう使用できないため、アクセスには特別な処理が必要です。処理は RAID グループの構成によって異なります。

- 故障率 – 装置が故障する確率を示しています。ここでは、ディスクドライブの故障を示します。ディスクドライブ内で起こる故障は様々で、RAID グループ構成を検討するうえで重要です。
- 冗長性のなくなった期間 – 故障したドライブに保持されていたデータの一部を RAID グループが再構築する期間。
- 保護のリカバリー期間 – RAID グループの障害保護のレベルが予期していたよりも低い状態である期間。この期間に、故障したドライブの内容をリビルドするために、ホットスペアが使用される場合があります。この間のホストアクセスのレスポンス時間は、通常よりも長くなります。
- 保護の復旧期間 – RAID グループを計画通りの構成に完全に復旧させるために必要な期間。この期間には、ドライブの交換時間、ストレージシステムに組み込む時間、計画通りの構成に復旧させる時間が含まれます。

3.1 RAID1 および RAID1+0 でドライブが故障したときの動作

RAID1 または RAID1+0 の場合、ミラーリングされた 2 つのドライブで並行して読み込みを行う代わりに、そのペアの残っている方のドライブのデータを読み込みます。ペアの両方のドライブに書き込みが指示される代わりに、片方のドライブに書き込まれます。RAID1+0 グループ内では、別のドライブが故障しても保護されますが、すでに故障しているドライブとペアになっているドライブが故障した場合、その 2 つ目の故障によってデータ損失が起こります。

適切なホットスペアドライブ(専用またはグローバルのどちらか)がある場合は、リビルドによってグループの保護状態がすぐに復旧されます。RAID1 または RAID1+0 のメンバーのリビルドでは、ペアの正常な方から交換したドライブ(ホットスペアまたは交換した新しいドライブのどちらか)にデータがコピーされます。リビルドの最大速度は、コピーで 1 つのドライブに書き込む速度によって決まります。

グループがコピーバックモードのグループとして構成されている場合、故障したドライブが交換されると、コピーバックが開始されます。前述のように、コピーバックにより、使用されているホットスペアドライブの内容が新しいドライブにコピーされ、その書き込み速度は、コピーで 1 つのドライブに書き込む最大速度に制限されます。

3.2 RAID5 でドライブが故障したときの動作

RAID5 の場合、リードアクセスには RAID グループ内の正常なドライブからデータを復元する必要があります。RAID5(4D+1P)グループでは、故障したドライブにアクセスできるようにデータを再構成するためには 4 つのドライブを読み込む必要があります。書き込みでは、すべての正常なドライブからの読み込みが行われ、データがストライプの中のどこに配置されているかによって 1 つまたは 2 つのドライブにライトバックすることが必要な場合があります。

適切なホットスペアドライブ(専用またはグローバルのどちらか)がある場合は、リビルドによってグループの保護状態がすぐに復旧されます。RAID5 グループのリビルドでは、すべての正常なドライブからの読み込みと交換したドライブへの書き込みが行われます。リビルド処理が完了する前に別のドライブが故障した場合は、データ損失が起こります。

RAID5 グループがコピーバックモードで構成されている場合、故障したドライブが交換されると、コピーバックが開始されます。コピーバックにより、使用されているホットスペアドライブの内容が新しいドライブにコピーされ、その書き込み速度はコピーで 1 つのドライブに書き込む最大速度に制限されます。

3.3 RAID6 でドライブが故障したときの動作

RAID5 の場合と同様に、RAID6 でのリードアクセスにも RAID グループ内の正常なドライブからデータを復元する必要があります。RAID6(4D+2P)グループの場合、故障したドライブにアクセスできるようにデータを再構成するためには 5 つのドライブを読み込む必要があります。同じように書き込みではすべての正常なドライブを読み込み、データがストライプの中のどこに配置されているかによって 2 つまたは 3 つのドライブに書き込みます。

適切なホットスペアドライブ(専用またはグローバルのどちらか)がある場合は、リビルドによってグループの保護状態がすぐに復旧されます。RAID6 グループのリビルドでは、すべての正常なドライブからの読み込みと交換したドライブへの書き込みが行われます。リビルド処理中に別のドライブが故障した場合でも、データを損失させずにリビルド処理をそのまま完了させることができますが、2 番目に故障したドライブにも対処する必要があります。リビルドの最大速度は、1 つのドライブに書き込む速度に制限されます。

RAID6 グループがコピーバックモードで構成され、ホットスペアドライブが最初のリカバリーに使用されている場合、故障したドライブが交換されるとコピーバックが開始されます。コピーバックにより、使用されているホットスペアドライブの内容が新しいドライブにコピーされ、その書き込み速度はコピーで 1 つのドライブに書き込む最大速度に制限されます。

大容量の NL-SAS ドライブでは、リビルドとコピーバックに多くの時間がかかり、さらにホストトラフィックがアクティブな場合は、リビルドとコピーバックの完了に 1 日以上かかることがあります。次にドライブ故障があっても RAID6 構成によって保護され、故障したドライブの交換が 1 日以内に行われる保守契約を結んでいる場合は、これらの RAID グループにホットスペアドライブを使用することはお薦めしていません。通常、故障したドライブは最初のリビルドが完了する前に交換できますが、リビルドはコピーバックが始まる前に完了する必要があるため、リビルドとコピーバックの実行中は、システムの性能が長時間低下した状態になります。

3.4 RAID6-FR でドライブが故障したときの動作

RAID6-FR は RAID6 グループの新しい形態で、複数の RAID6 グループをサポートする追加のドライブが 1 つ含まれています。セットのすべてのドライブの予備領域を使用して、ドライブ 1 つ分と同等の領域が提供されます。ローテーションする割り当てスキームを使用してセット内のサブグループをデュアルパリティで保護しながら、すべてのドライブにデータが渡されます。

RAID6-FR グループのドライブが故障すると、すぐにすべての正常なドライブの予備領域へのリビルドが始まります。このリビルドの速度は、1 つのドライブへの書き込み速度で制限されないため、ほかの RAID 構成よりも迅速にリビルド作業が進みます。

表 2 に、標準的な RAID6 構成と多くの RAID6-FR 構成の相対的なリビルド速度を示します。RAID6-FR のリビルド速度の方がはるかに速い主な理由は、復元されるデータ領域が RAID グループのすべての正常なドライブから提供され、1 つのドライブへの書き込み速度というボトルネックが排除されるためです。これにより、グループの保護状態が低い時間は構成計画時よりも大幅に短縮します。相対リビルド速度は、RAID6(3D+2P)で必要とされる速度(MB/s)を基準とした比率です。この速度は、リビルドの対象ドライブの書き込み速度によって決まります。リビルドにかかる時間は、ドライブの種類やサイズ、RAID6-FR 構成、リビルド実行中のシステムのトラフィック量によって異なります。

RAID 構成 （リビルド速度順）	速度 （ホストトラフィックなし）	速度 （ホストトラフィックあり）
RAID6(3D+2P)	1.0	0.5
RAID6-FR(13D+2P)x2+1HS	1.1	0.3
RAID6-FR(9D+2P)x2+1HS	1.3	0.4
RAID6-FR(6D+2P)x2+1HS	1.4	0.5
RAID6-FR(3D+2P)x2+1HS	1.6	0.8
RAID6-FR(3D+2P)x6+1HS	3.5	1.0

表 2 – 相対リビルド速度

ホストトラフィックがあると、確実にリビルド速度は低下します。これは、通常、リビルドの優先度は低く、ユーザーからの要求に対する処理が優先されるからです。また、RAID6-FR 構成ではリビルド速度が速いので、ストレージシステムが計画より低い保護レベルにさらされる時間が大幅に短縮されることは明らかです。

4 保護にかかる費用

データ損失に対する保護は無償ではありません。各 RAID 構成で提供される様々なレベルの保護には費用がかかります。保護費用を知る 1 つの方法として、ドライブの総容量のうちのどれくらいの容量をユーザーデータに使用できるようにするかを考える方法があります。つまり、総ドライブ数に対するユーザードライブ数の比率を見るということです。この比率と、ドライブが故障したときの損失データに対する費用とを比較検討する必要があります。

表 3 に、ほぼ同数のユーザードライブに対するユーザビリティのレベルとそれに関連する保護レベルを RAID 構成ごとに示します。

- 保護レベル 0 – ドライブが故障するとデータ損失が起こります。データは全く保護されません。
- 保護レベル 1 – グループのドライブが 1 つ故障してもデータは保護されて損失しませんが、もう 1 つドライブが故障するとデータが損失します。
- 保護レベル 2 – グループの 2 つのドライブが故障しても、データは保護されて損失しません。

RAID 構成（ユーザビリティ係数順）	RAID グループ数	ユーザードライブ数	ドライブ数合計	ユーザビリティ係数	保護レベル
RAID0(4D)	6	24	24	1.00	0
RAID6-FR (13D+2P)x2+HS	1	26	31	0.84	2
RAID6-FR (12D+2P)x2+HS	1	24	29	0.83	2
RAID6-FR (9D+2P)x2+HS	1	18	23	0.78	2
RAID6-FR (8D+2P)x3+1HS	1	24	31	0.77	2
RAID5(4D+1P)+GHS	6	24	31	0.77	1
RAID6-FR (6D+2P)x2+HS	2	24	34	0.71	2
RAID6-FR (5D+2P)x4+HS	1	20	29	0.69	2
RAID6(4D+2P)+GHS	6	24	37	0.65	2
RAID6-FR (4D+2P)x5+1HS	1	20	31	0.65	2
RAID6-FR (4D+2P)x2+1HS	2	16	26	0.62	2
RAID6-FR (3D+2P)x6+HS	1	18	31	0.58	2
RAID6-FR (3D+2P)x2+HS	4	24	44	0.55	2
RAID1+0(4+4)+GHS	6	24	49	0.49	1*

表 3 – ユーザビリティと保護の関係

(1*: 故障した 2 つ目のドライブが、最初に故障したドライブとペアになっているドライブでない場合は、RAID1+0 グループでは 2 つ目のドライブの故障は保護されます。)

グローバルホットスペアは、通常、ドライブ数を減らしユーザビリティ係数を向上させるために、複数の RAID 構成間で使用されていることに注意してください。グローバルホットスペアを使用して保護されているグループで 1 つ目のドライブが故障しても、すぐにリビルドを開始します。ほかのグループでも故障する可能性はありますが、使用する領域がないのでリビルドは先延ばしになり、ホットスペアが交換される前に別のドライブが故障した場合、データを損失する危険性があります。

5 保護のリカバリー

保護の仕組みの重要な要素は、冗長性のなくなった期間と、構成時に計画した保護の度合いに近い状態に復元するために必要な時間です。故障したドライブを交換するまでは、構成で計画された保護レベルではありません。できる限り早くもとの保護レベルへの復元を完了し、後続の故障によるデータ損失から守る必要があります。ドライブが故障してからリビルドが完了するまでの時間は、様々な要因によって大きく変わります。要因には次のものがあります。

- **RAID 構成** – 多くの RAID 構成(RAID1、RAID1+0、RAID5 および RAID6)では、1 つの交換ドライブやホットスペアドライブに対してリビルドが行われ、リビルドの速度は 1 つのドライブの書き込み速度に制限されます。RAID6-FR では、リビルド処理中にグループのすべての正常なドライブを使用でき、そのためはるかに速い速度でリビルドし、冗長性のなくなった期間を減らすことができます。
- **ドライブのサイズと速度** – 故障したドライブのサイズと速度によって、リビルドの速度とリビルドが完了するまでにかかる時間が決まります。大容量で遅いドライブの場合、リビルドには長時間かかることがあります。
- **ホストトラフィック** – システムのホストトラフィックのレベルも、リビルドの速度に影響します。通常、リビルドより、ホストからの要求に対する処理が優先されるため、ホストトラフィックが多いと冗長性のなくなった期間も長くなります。最大リビルド速度、つまり冗長性のなくなった期間が最も短いのは、ホストトラフィックがほとんどないときです。この場合、RAID 構成とドライブの種類によって冗長性のなくなった期間が決まります。

5.1 コピーバックし、構成を完全に復旧する際の考慮点

故障したドライブはタイムリーに交換する必要があり、それは RAID6-FR においても例外でないことを認識することは重要です。
Fast Recovery は、さらなる故障に対して完全な保護を提供しますが、故障したドライブが交換されると、RAID6-FR グループにそのドライブを組み込んで復旧を完了させます。この作業には、グループに再び導入される 1 つのドライブの内容をリビルドする必要があります。この作業の速度は再び導入される 1 つのドライブのデータの書き込み速度に制限されます。この時間は、図に示されている RAID5 グループのリビルド時間とほとんど同じです。ホストトラフィックがほとんどないときに比べ、ホストトラフィックがあるときははるかに長い時間がかかります。ホストトラフィックが少なければこの作業に丸一日かかることはなく、リビルド速度は最速です。RAID6-FR の非常に優れた特徴の 1 つは、再導入処理中でもグループは完全に保護され、故障してもデータを損失せずに復旧できることです。

6 まとめ

本書では、RAID6-FR の機能は 1 つ目のディスク故障に対するリカバリー時間をほかの RAID 構成で必要な時間のわずか **3 分の 1** の時間にまで短縮することを示してきました。これにより、2 つ目のディスク故障によるデータ損失の可能性が減少します。リカバリー時間を短縮することにより、通常のホストレスポンス時間は従来のリカバリー処理の場合に比べて大幅に短くなります。さらに、RAID6-FR ではドライブの交換処理中も完全に保護されるため、データ損失からより確実に守ります。

ホストトラフィックが多い場合はリカバリー時間が長くなりますが、RAID6-FR ではトラフィックがある場合でも、リカバリー時間はほかの RAID 構成に比べて格段に短いままです。ほかの場合と同じように、リカバリー時間は RAID グループを構成するドライブのサイズや速度の性能の影響を受けます。

本書では、アプリケーション環境の特定要件を満たすリカバリー時間とストレージシステムのコストを、バランスよく決定するための情報を示してきました。容量が少ない高速のドライブと比較すると、大容量 NL-SAS ドライブで RAID6-FR を構成した場合に、リカバリー時間を大幅に短縮できることは明らかです。つまり、ドライブ故障が発生した場合に、できるだけ長い時間データを保護できます。

7 最小リビルド時間および通常のリビルド時間

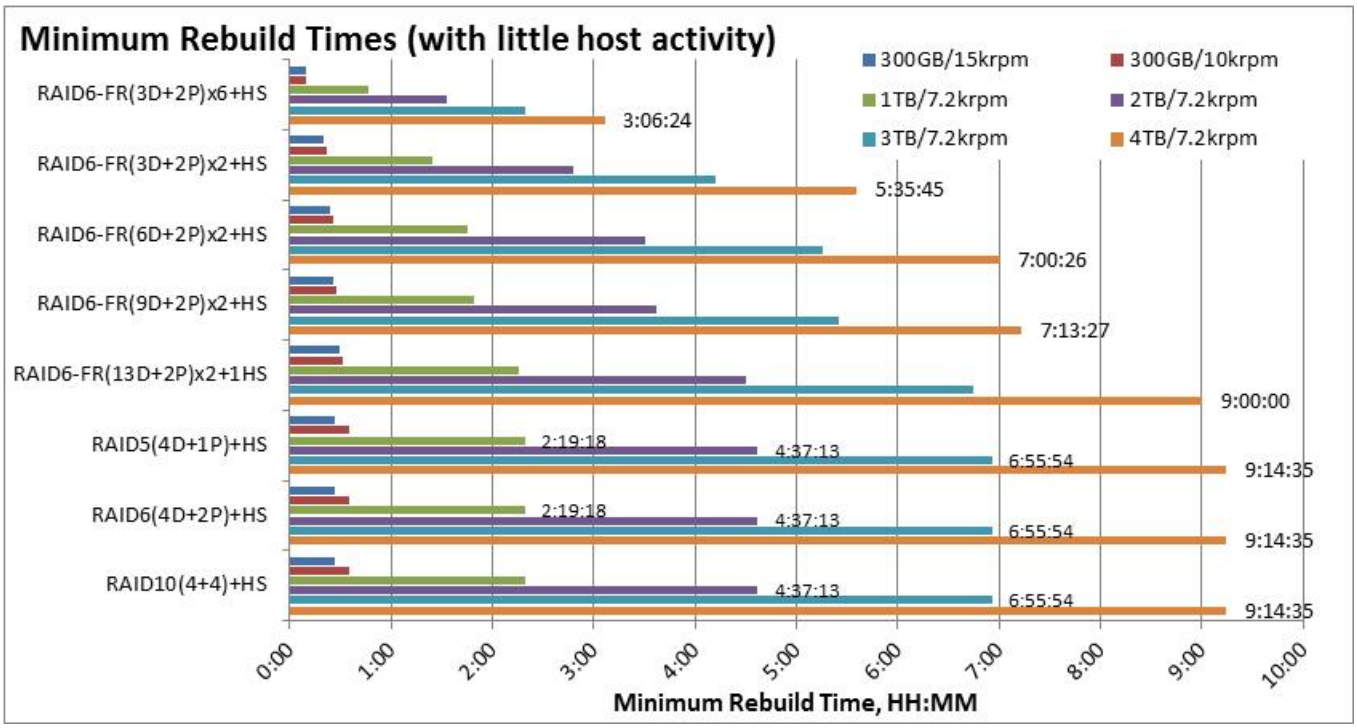

図 1 – ホストトラフィックが非常に少ない場合の最小リビルド時間

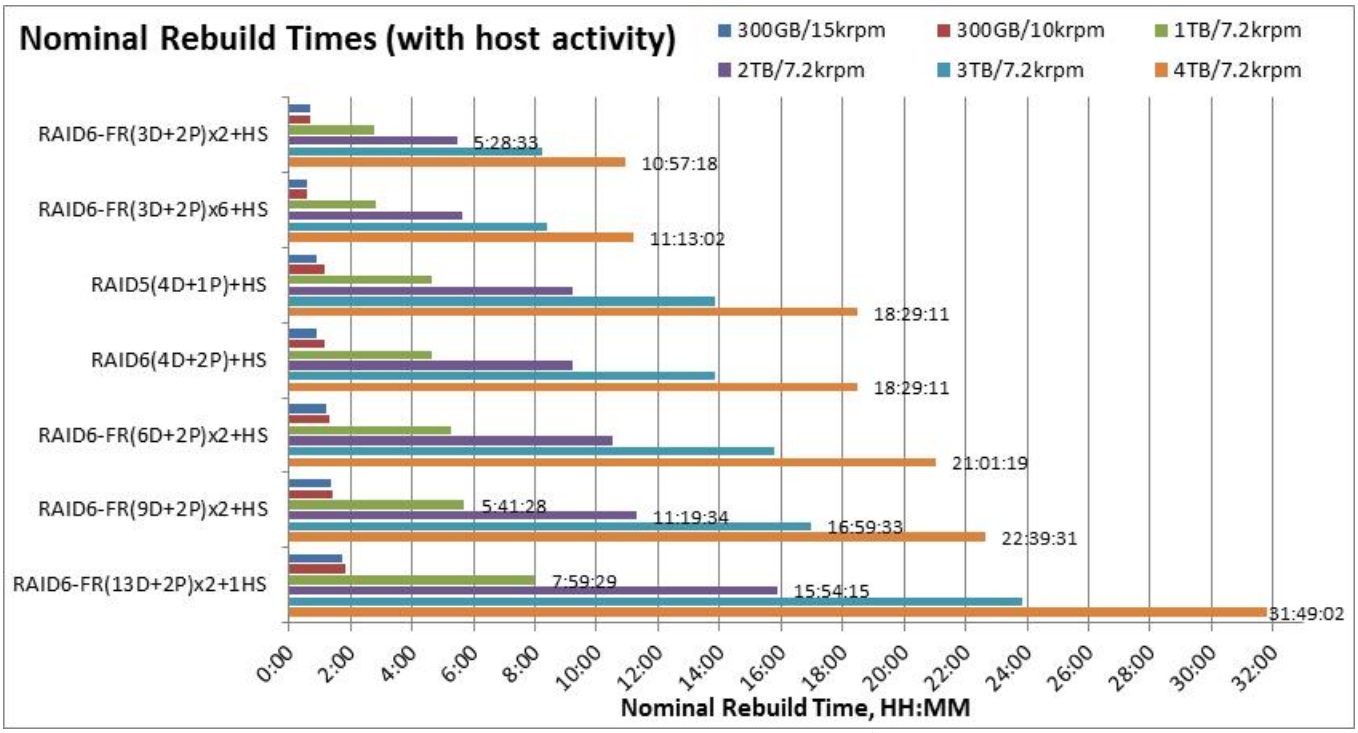

図 2 – ホストトラフィックがある場合の通常のリビルド時間

ホストが動作している、より通常の状態に近い状況では、ホストが動作していない最適な場合に比べてリビルド時間はおよそ 2 倍になります。RAID6-FR 構成を使用する主な利点は、リビルド時間が標準的な RAID 構成のわずか 30%～50%であるということです。これにより、冗長性のなくなった期間が大幅に短縮し、ドライブが使用できる場合でも、交換するドライブを組み込むより前に 2 番目の故障に対する保護を提供します。