

2007年度

オープンソースソフトウェア活用基盤整備事業

第I期 テーマ型（調査）

「Linux ディスク冗長化機能の適用評価と
最適な適用方法の調査」

— 調査報告書 Executive Summary —

2008年 1月

独立行政法人 情報処理推進機構

1. 背景と目的

ディスク冗長化機能としては、従来、専用ハードウェアによる RAID (Redundant Array of Independent Disks) 機能が多く利用されてきたが、近年、ハードウェアの追加無しに実現可能で、より安価なソフトウェア RAID を利用する機会が増えており、基幹系システムで導入するケースも出てきている。今後は更に Linux カーネルに含まれて提供されるオープンソースソフトウェア(OSS)の md (Multiple Devices)や DM (Device Mapper)を用いる事例が増加することが予想される。RAID 機能では個々のディスクが故障した場合でも安定して処理を継続できるようにする必要があるので、エラー処理が重要である。しかしながら、一般的にエラー処理は通常処理に比べて実行頻度が低いため、体系的にテストを実施しない限り品質を高く維持することは困難である。

本調査の目的は、md の各 RAID レベルと DM の dm-mirror の機能、性能、品質およびソフトウェア RAID の問題点、復旧手順を明確化することである。調査対象は最新のコミュニティカーネル、および Red Hat Enterprise Linux 4.5 ディストリビューションのカーネルとした。なお、調査結果は Linux のソフトウェア RAID 開発コミュニティと、ソフトウェア RAID 機能の利用を考えるユーザに以下の利益をもたらす。

- 発見された致命的な問題をコミュニティに報告するとともに、開発した評価プログラムを公開することにより、Linux のソフトウェア RAID 機能の改善を図る。
- Linux のソフトウェア RAID の品質状況を明確化することにより、利用者が信頼性の高い機能や RAID レベルを選択することで未然に問題発生リスクを低減する。

この結果、システム障害の未然防止や、ダウンタイムの短縮が可能となり、ソフトウェア RAID を利用したシステムでの可用性向上が期待される。

2. 調査項目と結果

調査として機能評価、性能評価と品質評価の 3 つを実施した。品質評価では Linux カーネル内でディスク故障を模擬的に発生させるライブラリ (故障模擬ライブラリ)を開発し、ディスク故障発生時に動作するエラー処理が正しく実施されるか評価を行った。

(1) ディスク故障パターン

最初に、品質評価で使用するディスクの故障パターンに関する調査結果について説明する。SCSIとSATAの規格からOSに見えるエラーパターンを調査し、製品で発生した故障パターンに関するヒアリング結果と併せて、現実に発生するディスク故障パターンを整理した。この結果、模擬故障として使用する 8 種類のディスク故障パターンを決定した。

(2) 故障模擬ライブラリ

次に、品質評価テストプログラム実行時に、上記ディスク故障パターンに対応する模擬故障を Linux カーネル中で発生させるために開発した故障模擬ライブラリについて説明する。故障模擬ライブラリは SystemTap (Linux カーネル中に情報収集・解析用のフックをしかけて指定した処理を実行することができるツール)を使用して、カーネルの内部状態を書き換えて模擬故障を発生させる。報告書では Linux の SCSI 処理について説明した後、SystemTap を用いて、模擬故障をしかけた

めの処理の実装方法について報告する。

(3) 機能評価

機能評価では Linux のソフトウェア RAID 機能を利用、管理するにあたり、必要な基本機能が正しく動作しているか評価した。md や DM に関するドキュメントやインターネット上に公開されている情報から各 RAID ボリュームの設定手順・復旧手順・管理機能を調査し、「ソフトウェア RAID 設定手順書」に具体的な操作手順をまとめた。また実機上でそれらの機能が問題なく動作するか確認した。md は、調査対象の機能のほとんどが揃っているのに対して、DM は未サポートの機能が多く、DM が先行する md に未だ追いついていないことが確認された。

(4) 性能評価

性能評価では性能評価用のプログラムを開発し、ソフトウェア RAID の通常運用時の性能低下、復旧処理中のオーバーヘッド、および復旧動作にかかる時間を評価した。md の RAID1 と DM は write 時に多少のオーバーヘッドが見られたが、read 性能は RAID なしの時とほとんど変わらない結果が得られた。md の RAID5 は I/O サイズが小さい場合は性能が低下するものの、I/O サイズが大きい時は RAID なしの時と比べスループットが向上することが確認された。また、業務の負荷が高い場合には、縮退した RAID アレイの復旧にかかる時間が大きく延びるため、復旧を運用中に行う場合は、運用時の負荷を考慮して復旧作業を計画する必要があることがわかった。

(5) 品質評価

品質評価では故障模擬ライブラリを用いて、ディスク故障発生時に動作するエラー処理が正しく実施されるか評価した。評価の結果、md あるいは DM のバグと考えられる障害を、Red Hat Enterprise Linux 4.5 のカーネルで 2 種類、最新のコミュニティカーネルでは 4 種類検出し、エラー処理部分の品質向上が必要なことが確認された。ただし、補足調査の結果、コミュニティカーネルの md では、最も一般的であるメディアエラーに対しては、冗長系からデータを復元して書き戻すことにより、ディスクの代替セクタ割当てを行わせるなど、エラー処理の強化が進んでいることも確認された。

3. 結論

md は機能的に充実しているものの、操作には管理コマンドの多種多様なオプションやステータスファイルの内容に関する知識を必要とするため、初心者が使いこなすのは難しいと思われる。今後さらなる操作性や保守性の改善が必要と考える。DM は先行する md にまだ追いついておらず、今後の機能強化も重要であると考えます。

縮退状態では、エラーを正しく処理できないケースが存在するため、ログを確認する必要があることが判明した。ただし、一般的に、縮退状態でさらに故障が発生するとデータを復旧することは極めて困難となるため、縮退が発生した時点で、速やかにシステムを止めて故障ディスクの交換ができる運用を行うか、冗長性を高めて縮退状態になりにくい運用を行うべきである。

また、今回の調査では、最新のコミュニティカーネルの方が md あるいは DM のバグと考えられる

障害の種類が多いという結果が得られた。この原因としてはコミュニティにおける評価が弱く、エラー処理系が正しく動作しているか十分検証されないまま機能強化やバグ修正が行われているためと推測される。開発コミュニティの視点からは、本調査活動の一環として開発した模擬故障を用いた評価の枠組みを、コミュニティに認知してもらい、開発者に利用してもらうことが極めて重要であると考えられる。