川崎製鉄技報
KAWASAKI STEEL GIHO
Vol.26 (1994) No.3

光ディスクオートチェンジャを用いた大容量ファイルサーバの高信頼化技術

High Data Reliability Technology of Large File Server Using Optical Disk Autochanger

吉澤　功(Isao Yoshizawa)　桃尾　章生(Akio Momoo)　錦織　和彦(Kazuhiko Nishigori)

要旨：
光ディスクオートチェンジャを用いた大容量ファイルサーバシステムを開発した。本システムでは，ストレージデバイスとして光ディスクオートチェンジャを用いることにより，最大 200GB の容量を実現し，さらにハードディスクによるデータキャッシング機能を装備することにより，高速データアクセスを可能にしている。ファイルサーバシステムでは，その取り扱うデータの多さから高信頼性が要求されるが，本システムでは光ディスクオートチェンジャのデータ二重化機能，I/O エラー自動修復機能等を装備することにより，上記機能を持たない場合と比較して平均故障時間間隔 MTBF で約 74 倍と高信頼性を達成した。

Synopsis：
A large volume file server system using an optical disk auto changer has been developed by Kawasaki Steel Corp. This system has 200 GB disk space (maximum), and it is possible to transfer data fast by using a data cache system by the hard disk drive. File server system generally requires high data reliability because of its large data volume. This system uses some technologies to guarantee high reliability. One is the data mirroring function for optical disk auto-changers. Another is the auto I/O error recovery function. The MTBF (mean time between failure) of this system is 74 times longer than that of the other system which doesn't use these high reliability technologies.

本文は次のページから閲覧できます。

川崎製鉄技報
26 (1994) 3, 140-144

光ディスクオートチェンジャを用いた大容量ファイルサーバの高信頼化技術*

High Data Reliability Technology of Large File Server Using Optical Disk Autochanger

吉澤 功
Isao Yoshizawa
川鉄情報システム(株)
応用システム事業部
テレマーケティングシステム部 主任部員(課長)

桃尾 章生
Akio Momoo
川鉄情報システム(株)
基盤システム事業部
メディアシステム部 メディアシステムグループ 主任部員(課長)

錦織 和彦
Kazuhiko Nishigori
川鉄情報システム(株)
リエンジニアリング事業推進部

要旨

光ディスクオートチェンジャを用いた大容量ファイルサーバシステムを開発した。本システムでは，ストレージデバイスとして光ディスクオートチェンジャを用いることにより，最大200 GBの容量を実現し，さらにハードディスクによるデータキャッシング機能を装備することにより，高速データアクセスを可能にしている。ファイルサーバシステムでは，その取り扱うデータの多さから高信頼性が要求されるが，本システムでは光ディスクオートチェンジャのデータ二重化機能，I/Oエラー自動修復機能等を装備することにより，上記機能を持たない場合と比較して平均故障時間間隔MTBFで約74倍と高信頼性を達成した。

Synopsis:

A large volume file server system using an optical disk auto changer has been developed by Kawasaki Steel Corp. This system has 200 GB disk space (maximum), and it is possible to transfer data fast by using a data cache system by the hard disk drive. File server system generally requires high data reliability because of its large data volume. This system uses some technologies to guarantee high reliability. One is the data mirroring function for optical disk auto-changers. Another is the auto I/O error recovery function. The MTBF (mean time between failure) of this system is 74 times longer than that of the other system which doesn't use these high reliability technologies.

1 緒　言

近年，コンピュータ業界では，パーソナルコンピュータ，ワークステーションの高性能化，低価格化がすさまじい勢いで進行している。同時に，オープンシステム，ダウンサイジングをキーワードとして，コンピュータシステムの姿が大きく変化している。特にダウンサイジングの代表的な例として，コンピュータネットワーク上での資源の共有化が進んでおり，その中でもデータの共有化を行うファイルサーバの需要が増大している。特にCADや画像データ，帳票データ，図面管理などといった分野では，取り扱うデータも膨大であるために，従来のハードディスクでは対応しきれなくなってきている。

このような状況を踏まえて，光ディスクオートチェンジャを用いたファイルサーバの開発を行った。本ファイルサーバでは大容量のデータを扱うので，そのデータに対する信頼性の確保が重要な課題となる。

そこで本論文では，光ディスクオートチェンジャを用いた大容量のデータを扱うファイルサーバを構築する上での高信頼化技術[1]について論じる。

* 平成6年5月11日原稿受付

2 ファイルサーバの概要

2.1 前提条件

冒頭にも述べたようにオープンシステム，ダウンサイジングという要求を満たすホストコンピュータの機能としては，

① 標準UNIXの搭載

② 既存のネットワークと接続可能 (Ethernet, TCP/IP, NFS)

③ 他のアプリケーションの変更を必要としない

といったことが考えられた。

また，ファイルサーバのストレージとしては階層化大容量ストレージ技術を生かし，高速アクセスを可能にした最大容量50 GBの超高速オートチェンジャシステムKJ 50を使用することにした。

2.2 ファイルサーバの特徴

光ディスクオートチェンジャを使ったファイルサーバは，すでにいくつか製品化されている。そこで，今回の開発では，他社製品との差別化を図っていく必要がある。よって，本ファイルサーバには以下の機能を付加することとした。

(1) 高速アクセス

他社のファイルサーバは，光ディスクオートチェンジャの低速性からデータの自動倉庫的な使用方法が普通である。しかし，本ファイルサーバは階層化大容量ストレージ技術により，高速にアクセスすることが可能となる。

（2） 光ディスクのリムーバブル性の維持

光ディスクオートチェンジャは他製品に見られる大容量のディスク的な使い方も考えられるが，本ファイルサーバでは専用のドライバを開発し，光ディスクのリムーバブル性を積極的に活用できるようにする。

（3） 二重化

ファイルサーバとして信頼性を高めるために，光ディスクの二重化を行う。これにより二重化された光ディスクのうち，片方が使用不能となっても，もう片方のディスクにより，データが保存されている。また，使用不能となった光ディスクをオンラインで交換することにより，エラー発生時の自動修復機能を付加する。

（4） MMI（Man Machine Interface）機能

ファイルサーバシステムの管理，運用をできるだけ容易に行うためにGUI（graphical user interface）機能を活用し，システムの管理・運用を行うユーティリティを作成する。

2.3 ハードウェア構成

本ファイルサーバはUNIXマシンをCPUとし，最大8台のKJ 50を接続することが可能である。KJ 50はジュークボックス型光ディスクオートチェンジャとこの光ディスクオートチェンジャの制御を受け持つ階層化モジュールとで構成されている。**Fig. 1**に本ファイルサーバのハードウエア構成を示す。

使用している光ディスクは，1枚当たり1 GBの容量を持つものである。光ディスクオートチェンジャは本光ディスクを50枚格納することができる。したがって，本ファイルサーバは二重化された場合に最大200 GBの大容量が実現される。**Table 1**に光ディスクオートチェンジャ，光ディスクの主な仕様を示す。

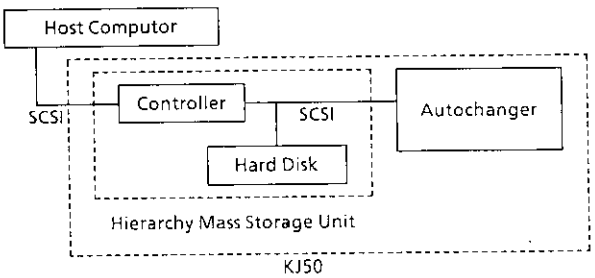

Fig. 1 Structure of hardware

Table 1 Specifications of optical disk and optical disk changer

Optical Disk	
Recording capacity (formatted)	1 000 Mbytes
Tracks per surface	19 968 tracks
Sector per track	17-32 sectors
Bytes per track	1 024 bytes
Optical Disk Autochanger	
Capacity	50 Gbytes
Installed drive	2 units
Total average cartridge exchange time	14.0 s
Average cartridge exchange time	8.0 s

2.4 ソフトウェア構成

ファイルサーバのソフトウェア構成を**Fig. 2**に示す。この中で新たに開発する部分は，以下のとおりである。

2.4.1 階層化モジュール用ドライバ

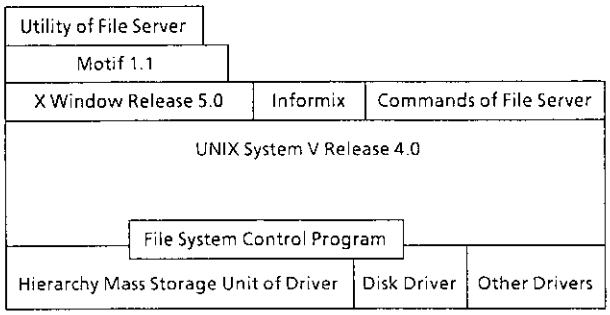

Fig. 2 Structure of software

本ファイルサーバの特徴である光ディスクのリムーバビリティやKJ 50の機能を効果的に利用していくために，本開発では階層化モジュール用のドライバを作成した。本ドライバにより，本ファイルサーバ独自の機能（光ディスクのリムーバビリティなど）が可能となる。

2.4.2 管理コマンド，デーモン

光ディスクへのファイルシステム作成，ミラーリングの設定といった光ディスクオートチェンジャや階層化モジュールの管理を行うために，専用のコマンド，デーモンの開発を行った。これにより，コマンドレベルでのファイルサーバ管理を容易にした。

2.4.3 ユーティリティ

ファイルサーバの操作性向上のために，X Window System，Motifといったグラフィックスライブラリを用いたGUI指向のユーティリティを作成した。多数の光ディスクを容易に管理していくには，GUIを用いたユーティリティは非常に有効である。本ユーティリティにより管理される対象は，階層化モジュール，光ディスクオートチェンジャに格納されている光ディスク，システムの設定などである。これによりシステム管理，保守をドライバやコマンドを意識することなく容易に行うことができる。**Fig. 3**に本ユーティリティの画面例を示す。

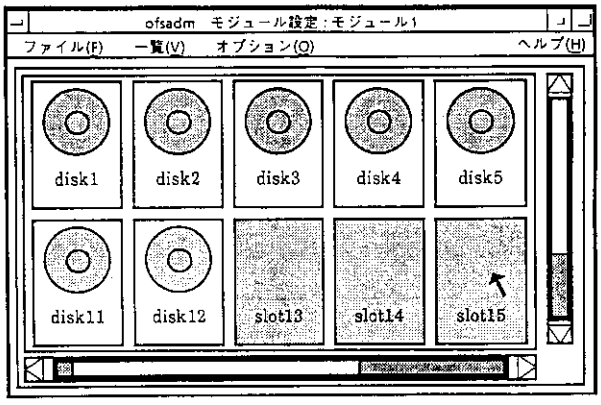

Fig. 3 Image of GUI

2.5 ファイルサーバ管理

2.5.1 階層化モジュールの制御

本ファイルサーバで用いているKJ 50の最大の特徴は，独自のキャッシュ・アルゴリズムによる高速性である。その方法を以下に述べる。

（1） データ書き込み時

ホストから転送されたデータは，階層化モジュールのハードディスクに書き込まれるために，高速に転送が可能である。ハードディスクに書き込まれたデータは，ホストの空き時間に光ディスクオートチェンジャ側に自動的に転送し光ディスクにデータを書き込む。これにより，ハードディスク並の書き込み速度が得られる。

（2） データ読み込み時

光ディスクオートチェンジャに格納されたデータは，いつアクセスされたか，何回アクセスされたかによって優先順位が付けられ，順位の高いデータがハードディスクに格納される。これにより，階層化モジュール部にホストが要求したデータがあった場合，高速なデータ読み込みが可能となる。

どちらの場合も，光ディスクオートチェンジャでの光ディスク交換時間（Table 1）を緩和することができるために，通常の光ディスクオートチェンジャよりも圧倒的に高速にアクセスができる。

2.5.2 光ディスクオートチェンジャ制御

現在どの光ディスクがドライブに挿入されているか，どのタイミングで光ディスクを交換するのかといった光ディスクオートチェンジャの管理をホストが行うには，負荷が大き過ぎる。KJ 50では，階層化モジュールがこれらの制御を行っている。

ホストコンピュータからある光ディスクへアクセスする場合，通常は先に述べたようにコントロール部のハードディスクのキャッシュを用いている。しかし，直接光ディスクに対してアクセスする場合（たとえば，読み込み時にキャッシュにヒットしなかった場合）には，オートチェンジャにおける光ディスク交換などのトランザクションが発生する。例えば，光ディスクオートチェンジャのドライブに光ディスクが挿入されていて，別の光ディスクを読み込む場合を考える。ホストコンピュータからのアクセス要求が階層化モジュールに出されると，階層化モジュールはドライブに挿入されている光ディスクを取り出し，目的の光ディスクをドライブに挿入する。そしてホストコンピュータのアクセス要求に答える。これにより，ホストコンピュータは通常のハードディスクと同様のアクセス方法でデータを読み書きすることができる。

また，KJ 50をもっと柔軟に利用するために，階層化モジュールには四つのモードがある。

（1） スタート：通常状態のモード。

（2） ストップ：システム終了のためのモード。スタートと対。

（3） ノーマル：階層化モジュールが光ディスクオートチェンジャの制御を行うモード。

（4） デバッグ：ホストコンピュータから直接光ディスクオートチェンジャを制御するモード。

このような階層化モジュールの機能を十分に活用するために，本開発では，階層化モジュール専用のドライバを開発した。

2.5.3 ファイルシステムの拡張

（1） ボリューム

ファイルサーバの利用目的は，大容量のデータ空間の提供である。また，本ファイルサーバの特徴である光ディスクのリムーバブル性を失わせないために，次のような方法を採用した。

光ディスクへの最小管理単位は片面の500 MBとし（ドライブが光ディスクの片面ずつしかアクセスできないため），カーネルに組み込まれたデバイス管理システムを利用して1枚の光ディスクを仮想的に1 GBのデバイスとする。この仮想デバイス（以降ボリューム）は最大2 GB（光ディスク2枚）まで拡張することができる。したがって，ファイルシステムの構築や，マウント，アンマウントといったオペレーションは，すべてボリュームに対して行われる。

（2） ミラーリング

本ファイルサーバでは，システムの信頼性を向上させる方法として，光ディスクの二重化を行っている。このデバイス管理システムの機能を使って，2枚の光ディスクに対してミラーリングを設定したボリュームを作成することにより，光ディスクの二重化を行うことができる。**Fig. 4** にミラーリングの概念を示す。ただし，同一の光ディスクオートチェンジャに格納された光ディスクでミラーリングを設定すると，書き込み時に光ディスク交換などのオーバーヘッドが大きくなる。そこで，本ファイルサーバでは光ディスクオートチェンジャ単位でのミラーリングの設定を行うことにした。例えば，2台の光ディスクオートチェンジャが接続されている場合，**Fig. 5** に示すように光ディスクオートチェンジャ1と光ディスクオートチェンジャ2とでミラーリングの設定をしたとする。このミラーリングの設定により光ディスクオートチェンジャ1の光ディスク1にボリ

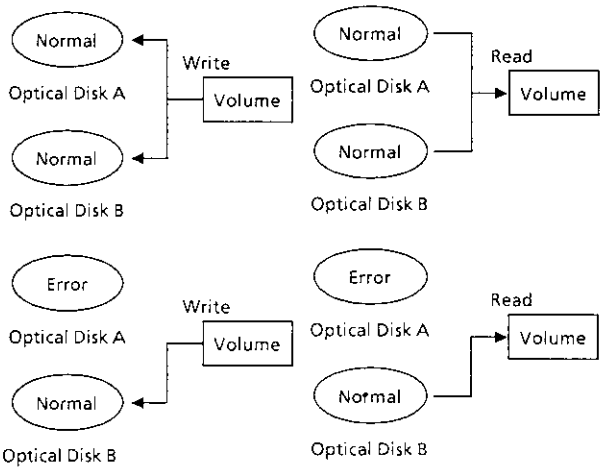

Fig. 4　Read/write of mirroring

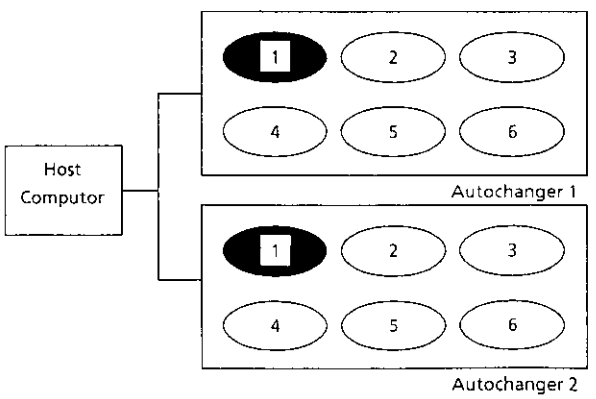

Fig. 5　Setting of mirror

ュームを作成すると，自動的に光ディスクオートチェンジャ2の光ディスク2とミラーリングを行う。これにより，光ディスク交換などが起こった場合でも，それぞれの光ディスクオートチェンジャが一度ずつ光ディスクの交換を行えばよいので，オーバヘッドを最小限におさえることができ，アクセスを早くできる。また，ミラーリングされたボリュームを自動的に作成できるので，管理を簡単に行うことができる。

3 I/Oエラー自動復旧機能

本ファイルサーバでは，データの信頼性の向上のためにI/Oエラー自動復旧機能を付加した。この機能は，ミラーリングしている光ディスクにエラーが発生した場合，それを検出し，再びミラーリングした状態に回復させる機能である。

3.1 KJ 50のエラー回復処理

KJ 50は光ディスクオートチェンジャやキャッシュ用ハードディスクの動作異常に関しては，再試行等によりエラーのリカバリーを行う。

この他に，異常状態で問題となるのは光ディスクでの読み取り／書き込み異常である。これは，光ディスク表面がほこり等によって汚れた場合に発生することがある。このような場合の対応として，KJ 50は以下のように処理ができるようになっている。

（1） 光ディスクがダメージを受けた多くの場合，まずそれは読み取り／書き込み時に何回かのリトライを行い，光ディスクドライブ自身が，ディスクのダメージを検出する。このような状態が発生すると，KJ 50ではどのディスクがダメージを受けたかをロギングして，このロギングデータをホストコンピュータから読み取る。

（2） ホストコンピュータでは，適当な時間間隔で，（1）で述べたロギングデータを読み取り，光ディスクの異常の有無を確認する。何らかの異常が検出された場合，KJ 50に対してディスクリカバリコマンドを発行することにより，そのディスクのリカバリを行う。

（3） KJ 50は（2）のリカバリコマンドを受け取ると，該当する光ディスクの内容を予備の光ディスクにコピーし，異常な光ディスクと予備光ディスクを交換する。

この処理により，異常のあったディスクと正常なディスクをディスクのデータも含めて交換ができ，その結果，ユーザーデータに影響を与えずに異常ディスクの処理ができる。

3.2 光ディスクの故障時の復旧手順

Fig. 4に示したようにミラーリングしている光ディスクの一方が故障した場合，もう一方の光ディスクにアクセスすることにより正常な状態を保つことができる。ここで，故障した光ディスクを新しい光ディスクと交換し，再びミラーリングの状態にすることにより耐故障性が向上する。ミラーリングの状態に再び戻すためには，以下の手順をとる。

（1） 故障した光ディスクを新しい光ディスクに交換する。

（2） ボリュームとして定義された二つの仮想デバイスを切り離す。

（3） 再び二つの仮想デバイスを結合する。これにより新しい光ディスクに正常な光ディスクの内容がコピーされミラーリングされた状態に戻る。

仮想デバイスの切り離し，結合は，市販されているソフトウェアに用意されたコマンドを使っている。

3.3 自動復旧機能

ファイルサーバのシステム管理者の負荷を軽減するため，光ディスクの状態監視，ミラーリングの復旧を自動化した。光ディスクの状態は，KJ 50に用意されたコマンドにより知ることができ，光ディスクの各面に対して以下の5種類の状態で示される。

正常

ERROR　：エラーが発生した。このディスクは使用できない。

WARN　：エラーが発生したがリトライで回復できた。このディスクの使用は続けられる。

MAINT　：光ディスクの表面にゴミが付いているような場合でメンテナンスが必要。データのバックアップが望ましい。

UNMATCH：キャッシュディスクから光ディスクへデータを書き戻した時にエラーが発生した。このディスクはKJ 50をメンテナンスモードに切り替えて排出する必要がある。

光ディスクへのアクセス時にエラーが発生すると，光ディスクのステータスはその状況に応じたものに変わる。このステータスを調べることによってファイルサーバシステムで使用する全ての光ディスクの状態を一定時間ごとに監視する。もし，エラーが生じているならば，その情報をメッセージとしてシステム管理者へ知らせ，さらにミラーリングに使用しているならば再びミラーリングの状態へ戻すように動作する。

ミラーリングの復旧を行うためには，まず障害の発生した光ディスクを新しい光ディスクと交換しなくてはならない。これを自動化するために光ディスクオートチェンジャの50枚の光ディスクの中の1枚を通常は使用せずに予備の光ディスクとして用意することにした。そして予備の光ディスクと指定したスロット番号の光ディスクを交換するためのコマンドをKJ 50に持たせた。

光ディスクの状態監視，ミラーリングの自動復旧を一定時間ごとに自動的に行うためにこの機能をデーモンとして作成した。

この自動復旧機能の処理の流れ図を**Fig. 6**に示す。このデーモンでは，光ディスクの状態が，先に述べたERRORまたはMAINTの状態であるものについて自動的にミラーリングの復旧を行う。光ディスクの状態がUNMATCHである時はシステムを運用しながらのディスクの交換を行うことができないので，エラーメッセージを表示するのみである。また，障害が発生した光ディスクがミラーリングされていない場合，予備の光ディスクが用意されていない場合は，その旨を知らせるメッセージを表示する。

4 評　価

ファイルサーバにおいてミラーリングの自動復旧を行うことによりどれほどの信頼性の向上がなされるかを考察する。ここで評価の指標として，MTBF（mean time between failures：平均故障間隔）[2]を用いる。システムのMTBFは，故障率λの逆数である。

$$MTBF = \frac{1}{\lambda} \quad \cdots\cdots(1)$$

また，システムが二重化してある場合のMTBFは

$$MTBF = \frac{3}{2\lambda} \quad \cdots\cdots(2)$$

さらに，二重化システムにおいて故障が起こったとき，システムを止めずに故障した部品をある時間内に修理でき再び二重化した状態

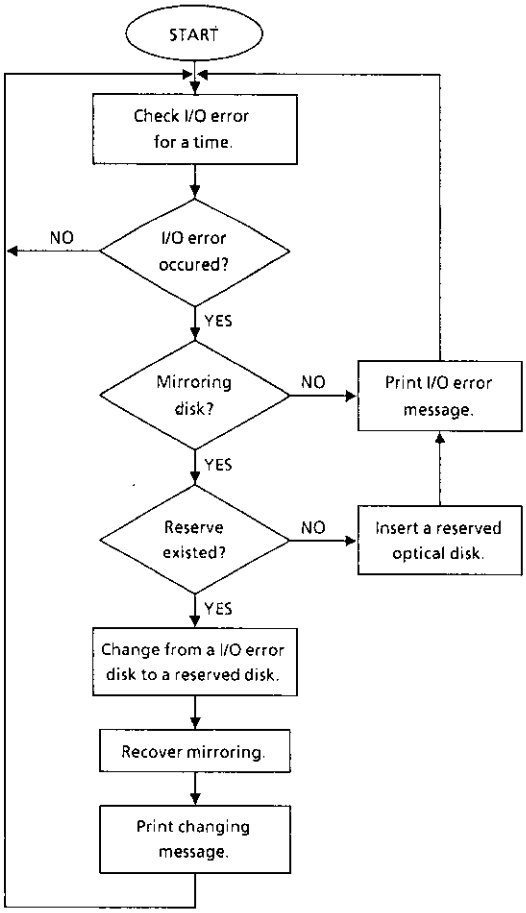

Fig. 6　Flow chart for I/O error automatic recovery

にできるならば，この場合のMTBFは

$$MTBF = \frac{\mu + 3\lambda}{2\lambda^2} \quad \cdots\cdots(3)$$

で表示される。ここで，修復率μは平均修理時間（MTTR：mean time to repair）の逆数で表される。

$$\mu = \frac{1}{MTTR} \quad \cdots\cdots(4)$$

信頼性については，CPU，ハードディスク，光ディスクオートチェンジャなどのファイルサーバシステムの各要素の信頼性を含めてシステム全体の信頼性として評価すべきである。しかし，ここではオートチェンジャなどの信頼性は考慮せずに光ディスク媒体の信頼性のみについて着目して評価する。光ディスクをミラーリングせずに使用した場合，ミラーリングした場合，さらに自動リカバリ機能を使用した場合のそれぞれにおけるMTBFを算出してみる。

（1）　ミラーリングしない場合

光ディスクのMTBFを10年と仮定すると，オートチェンジャの中には50枚の光ディスクが格納されているので，オートチェンジャ全体でのMTBFは

$$MTBF = 1/50\lambda = 1\,752\ (時間) = 0.2\ (年) \quad \cdots\cdots(5)$$

光ディスクをミラーリングせずに使用した場合のジュークボックス全体のMTBFが（5）式の値である。

（2）　ミラーリングを使用した場合（自動リカバリなし）

次に，光ディスクをミラーリングして使用した場合について考える。自動リカバリ機能を使用していない場合はシステム管理者がミラーリングが壊れたことをエラーメッセージによって知り，コマンドを入力することによりディスクの交換をしてミラーリングを復旧させなければならない。ここで，システム管理者が毎日システム運用チェックを行っており，エラーが起こった場合には平均12時間でミラーリングの復旧が行えると仮定した時のMTBFは（7）式のように14.8年となり，ミラーリングしないものに比べ約74倍に向上する。

$$MTTR = 1/\mu = 12\ (時間) \quad \cdots\cdots(6)$$

$$MTBF = (\mu + 3 \times 50\lambda)/2\lambda^2$$
$$= 1.30 \times 10^5\ (時間) = 14.8\ (年) \quad \cdots\cdots(7)$$

（3）　ミラーリングを使用した場合（自動リカバリあり）

次に，光ディスクをミラーリングして使用し，それぞれのジュークボックスに予備の光ディスクが用意され，自動的に光ディスクの交換，ミラーリングの復旧を行う場合について考察する。この時に予備の光ディスクが故障した光ディスクと交換された後，次に故障するまでに予備の光ディスクがまた新しく用意されているという条件のもとで考える。故障した光ディスクを交換しミラーリングの状態を回復するまでの時間は，その実測値の平均から

$$MTTR = 1/\mu = 5.91\ (時間) \quad \cdots\cdots(8)$$

である。したがってこの場合のMTBFは（9）式の値となる。

$$MTBF = (\mu + 3 \times 50\lambda)/2\lambda^2$$
$$= 2.62 \times 10^5\ (時間) = 29.9\ (年) \quad \cdots\cdots(9)$$

光ディスクの媒体のみに着目した場合，光ディスクのミラーリングによってその信頼性が向上し，さらに自動復旧機能を用いた場合は平均修復時間の短縮がなされ，単にミラーリングを行っただけの場合よりもさらに良い結果が得られることがわかる。

システム運用上の観点からみると，光ディスクの状態の監視，異常時の復旧を自動的に行うので，システム管理者が運用状況をチェックしなければならない頻度が減り，負荷を軽減できるという効果がある。

5　結　　言

今回開発したファイルサーバシステムの高信頼化技術について報告した。本技術の中でも，特にI/Oエラー自動復旧機能を利用することにより光ディスクの状態を自動監視でき，異常があった場合はその旨をエラーメッセージとして表示することができる。また，ミラーリングの状態が壊れた場合は，これを自動的に復旧することができる。最大200 GBの大容量の光ディスクの管理を行う負荷が軽減され，ミラーリングの自動復旧によってMTBFを約74倍の引き延ばすことができ，その結果，システムを長期運用する上での信頼性の向上がなされる。

また，本機能により，ディスクのミラーリングの自動復旧を行う上では，予備の光ディスクが常に用意されていることによってその効果がある。現在は光ディスク50枚の内の1枚を予備の光ディスクとして用意するようになっているが，これを2枚あるいは3枚にすることも考えられる。今後，実際の運用において光ディスクのエラー発生率を調べ，自動復旧機能の評価を行いたい。

参　考　文　献

1)　D. P. Siewiorek and R. S. Swarz:　"Reliable Computer Systems," (1992), [Digital Press]
2)　P. K. Lala:　「フォールト・トレランス入門」, (1988), [オーム社]