

スーパーコンピュータ「不老」における光ディスクライブラリを用いたコールドストレージシステムの構築

高橋 一郎¹⁾, 大島 聡史²⁾, 片桐 孝洋²⁾

1) 名古屋大学 情報推進部 情報基盤課

2) 名古屋大学 情報基盤センター 大規模計算支援環境研究部門

ichiro@itc.nagoya-u.ac.jp, ohshima@cc.nagoya-u.ac.jp,
katagiri@cc.nagoya-u.ac.jp

Building a cold storage system of the supercomputer "Flow" using the optical disk library

Ichiro Takahashi¹⁾, Satoshi Ohshima²⁾, Takahiro Katagiri²⁾

1) Information Infrastructure Division, Information Promotion Department, Nagoya University

2) High Performance Computing Division, Information Technology Center, Nagoya University

概要

名古屋大学情報基盤センターは、2020年7月のスーパーコンピュータのリプレイスで導入した光ディスクライブラリ装置を用いてコールドストレージシステムを構築し、利用者向けサービスを開始した。本稿では、コールドストレージシステムの概要、利用形態、運用アプリケーションソフトウェアについて紹介する。

1 はじめに

現在、スーパーコンピュータは数値シミュレーション、人工知能/機械学習、バイオインフォマティクスといった幅広い分野の研究に利用され、さらに実験データや観測データなど日々膨大なデータが扱われるよ

うになった。このような状況の中で、名古屋大学情報基盤センターでは2020年7月のスーパーコンピュータのシステム更新により、スーパーコンピュータとストレージシステム及び可視化システムを一新した。

スーパーコンピュータ「不老」の全体構成を、Fig.1に示す。

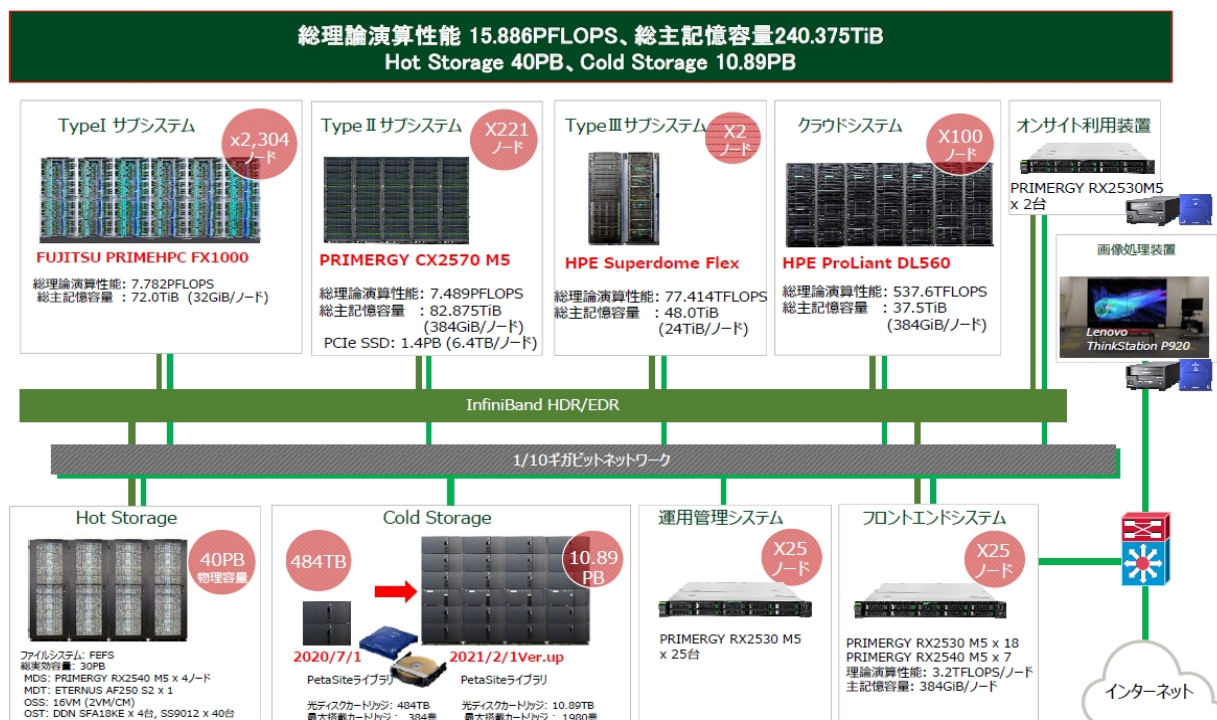


Fig.1 スーパーコンピュータ「不老」

新ストレージシステムでは、大規模データを高速に蓄積するためのホットストレージに加えコールドストレージとしては国内のスパコンで初となるソニー社製のオプティカルディスク・アーカイブ（以後 ODA と呼ぶ）PetaSite 拡張型ライブラリーシステム（以後 ODA ライブラリー装置と呼ぶ）を 2020 年 7 月と 2021 年 2 月の 2 フェーズで導入を進めている。Phase2 の ODA ライブラリー装置の初期搭載カートリッジ数は 1092 巻、記憶容量は 6PB である。

このコールドストレージの ODA ライブラリー装置では、多様化する大容量データの受け渡しに対応するために、ライブラリー内でのカートリッジの利用以外に、ライブラリー外へのカートリッジの搬出利用や、持ち込みカートリッジのライブラリー内への収納利用にも対応した。Fig.2 に、ODA カートリッジとデータの流れを示す。

本稿では、2020 年 7 月に導入した Phase1 の ODA ライブラリー装置を使ったコールドストレージサービスの概要、利用形態、運用アプリケーションについて簡単に紹介する。

2 特徴

コールドストレージシステムの特徴は、次のとおりである。

- 1 カートリッジを 1 ボリュームとして利用者に割当てて利用する。利用申請は、10 カートリッジ単位である。
- ODA ライブラリー内のカートリッジの記録フォーマットは、外付けの ODA 単体ドライブと同じフォーマットである。
- ODA ライブラリー装置内で記録したカートリッジを取出して、外付けの ODA 単体ドライブで利用できる。また、外付けの ODA 単体ドライブで記録したカートリッジを、ODA ライブラリー装置に収納して利用できる。
- ODA ライブラリー装置内のカートリッジは固体番号以外に名前(ボリューム通番)を付けて管理する。
- カートリッジの所有者情報やカートリッジの記録情報を DB を使って管理する。
- 利用申請情報を基に利用者が使用できるカートリッジやログインノードを限定する。
- ベンダー提供の SDK を使って独自のストレージ管理ソフトウェアを開発して、様々な利用形態に対応する。
- サービス終了時に、利用者にカートリッジを返却できる。
- ODA ライブラリー装置はフロントエンドサーバを兼ねている。

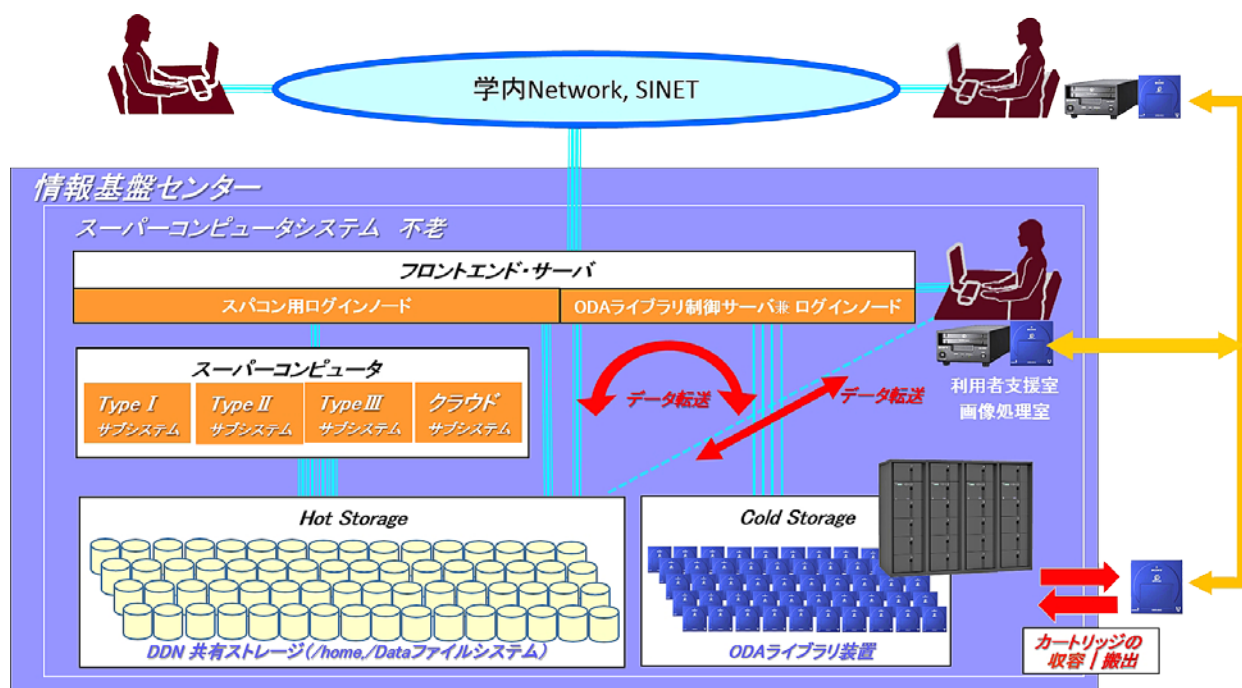


Fig.2 ODA カートリッジとデータの流れ

3 システム構成

スーパーコンピュータ「不老」は、特徴と役割の異なる4種類の計算サブシステム (Type I サブシステム、Type II サブシステム、Type III サブシステム、クラウドシステム) とフロントエンドサーバ、ホットストレージ、コールドストレージ、および周辺装置群で構成される。

3.1 ホットストレージシステム

ホットストレージは共有ストレージで、各計算サブシステムとフロントエンドサーバ群に InfiniBand EDRx8 で接続し、ミドルウェアに富士通 FEFS ファイルシステムを使って高速な IO 性能を提供する。

3.2 コールドストレージシステム

コールドストレージシステムは、ODA ライブラリ装置と ODA ライブラリ装置を制御する ODA ライブラリ制御サーバで構成される。

ODA ライブラリ装置は、ODA カートリッジを格納するキャビネット、ドライブ、キャビネット内の ODA カートリッジを識別してドライブにセットするロボット機構、ODA カートリッジをライブラリ内に収容および搬出する機構で構成される。この ODA ライブラリ装置は、Fibre Channel でフロントエンドサーバを兼ねている ODA ライブラリ制御サーバに接続されている。

表1とFig.3、Fig.4に、コールドストレージシステムの構成とハードウェアの仕様を示す。また、表2に、フロントエンドサーバを兼ねている ODA ライブラリ制御サーバの仕様を示す。

	Phase1	Phase2
ODA制御サーバ数	1	4
ODAライブラリ数	1	4
機種名	PetaSite拡張型ライブラリシステム	
総物理容量	484TB	10.89PB (2.72PB×4ライブラリ)
総スロット数	88巻	1980巻 (495×4ライブラリ)
総ドライブ数	5	20 (5×4ライブラリ)
ロボット機構	1	4 (1×4ライブラリ)

表.1 コールドストレージシステム諸元

項 目	仕 様
CPU	Intel Xeon Gold6248(2.5GHz/20コア)×2 理論演算性能 3.2TFLOPS
主記憶	384GiB (32GiB DDR4×12)
内蔵ストレージ	2TB SSD, 19.2 SSD
インターコネクト	InfiniBand EDR×1
SANインターフェース	FC16Gbps×4
ネットワークインターフェース	10GTwinax×2, 1000BASE-T×2

表.2 ODA ライブラリ制御サーバ諸元

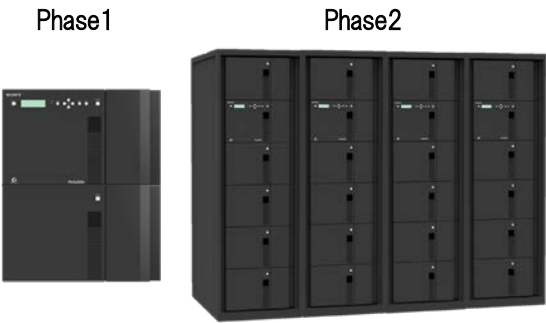


Fig.3 ODA ライブラリ装置

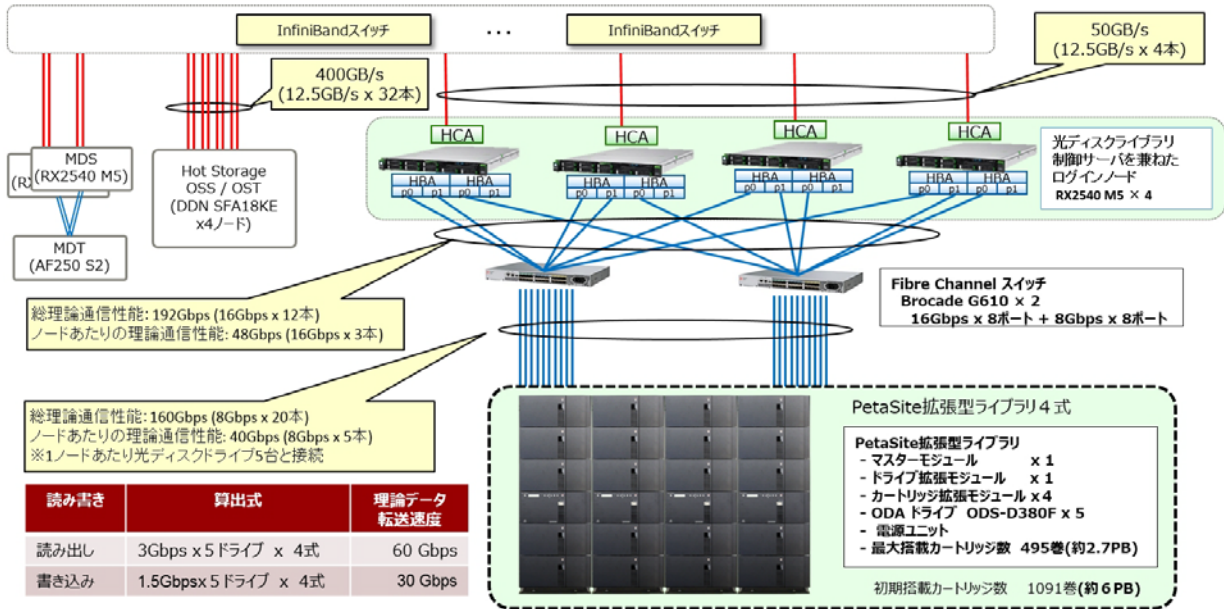


Fig.4 Phase2のコールドストレージシステム構成

3.3 ODA 単体ドライブ

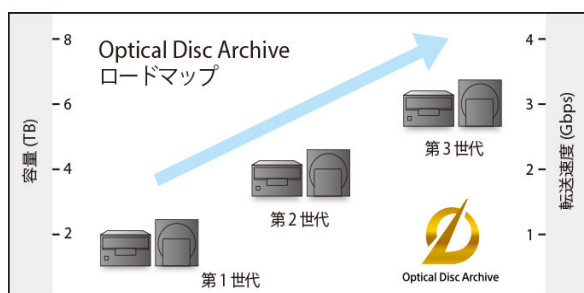
情報基盤センター内の利用者支援室と画像処理室のサーバには、ODA カートリッジの ODA ライブラリ装置外での利用や可搬型利用に対応するために ODA 単体ドライブが USB 接続され、Linux 環境と Windows 環境で利用できる。Linux 環境では、InfiniBand EDRx8 でホットストレージに接続され、高速にファイル転送を行うことができる。また Windows 環境では、ベンダー提供の Windows Explorer 相当の GUI を使ったファイル操作が行えるソフトウェアが利用できる。Fig.5 に、ODA 単体ドライブユニットを示す。



Fig.5 ODA 単体ドライブユニット

型番	ODC5500R	
記憶容量	5.5TB	
記録型メディア	追記型(Write Once)	
読み出し回数	100万回以上	
動作温度	5~55℃	
保存耐久性	100年以上 (SONY社加速試験による推定)	

表.3 ODA カートリッジ諸元



	第1世代	第2世代	第3世代
容量	~1.5TB	3.3TB	5.5TB
読み出し	~1.1Gbps	2Gbps	3Gbps
書き込み (w/Verify)	~440Mbps	1Gbps	1.5Gbps

表.4 ODA メディアと ODA ドライブの諸元

3.4 ODA とは

ODA は、デジタルデータの長期保存 (ISO/IEC 16963 を参考にした加速試験による保存寿命 100 年以上) を目的とした、大容量光ディスクストレージシステムである。データを格納するカートリッジに複数枚の堅牢性の高い業務用光ディスクを格納し、1 つのボ

リュームとして利用することができる。

この光ディスクアーカイブソリューションが提供する光ディスクは、WORM (Write Once Read Many) メディアで、物理的にもデータ書き換えができないライトワンスメディアである。ソニー独自の技術により、8 チャンネル光学ドライブを搭載したドライブユニットを使って、8 個のレーザーヘッドでディスク両面を同時に高速に読み書きでき、On-the-Fly ベリファイ記録機能を標準搭載している。

スーパーコンピュータ「不老」では、この ODA の最新の第 3 世代の「単体ドライブユニット」と「ライブラリ装置」が利用できる (Fig.3、Fig.5 参照)。表 3、表 4 に、ODA カートリッジと ODA ドライブの諸元を示す。

3.5 ODA カートリッジの利用上の制限事項

ODA ライブラリ装置で ODA カートリッジを取扱う場合の留意点を、以下に示す。

- ・記録メディアは、Write Once 記録方式である。
- ・同一カートリッジに対しての多重アクセス (カートリッジ内のコピーを含む) は行えない。
- ・記録操作は、シーケンシャル記録のみ。
- ・カートリッジの記録実効容量は、約 5TB、作成可能ファイル数は約 80 万ファイル。
- ・ディレクトリおよびファイル階層は、最大 64 階層まで。
- ・同一ディレクトリ内に、大文字小文字でファイル名を重複させることはできない。
- ・Linux コマンドの利用制限がある。

3.6 ODA ドライブの性能評価

1 ブロック 1 GB のバイナリーデータを 1GB~100 GB のファイルサイズでシーケンシャル Read 又はシーケンシャル Write を行って 1 秒当りのスループットを測定した。ドライブ当りの最大データ転送速度は、読み出し最大 380MB/s、書き込み最大 208MB/s の高速アクセス性能を計測した。ファイルサイズが大きい方が Read/ Write 性能が出ることと、ブロックサイズの大きさが Read/Write 性能に大きく影響しないことがわかった。なお、今回の性能測定には次の点は考慮していない。

- ・光ディスクの媒体への記録は、ディスクの内側から外側に向かって記録され、内側は遅く、外側は性能がよくなる。再生も同様となる。
- ・カートリッジ内でディスクチェンジが発生した場合は、数十秒程度かかる。

4 運用アプリケーション

以下に、コールドストレージサービスを運用管理するためのソフトウェアについて述べる。

4.1 ベンダー提供 SDK

次のような基本機能を有する SDK を導入ベンダーに開発依頼した。

- ・カートリッジの登録情報を、DB を使って管理する機能。
- ・カートリッジに記録したファイル情報を DB で管理し、閲覧および検索する機能。
- ・キャビネットとドライブ間のカートリッジを搬送する機能。
- ・ドライブの状態を取得する機能。
- ・カートリッジ単位で使用者を限定する機能。
- ・カートリッジをライブラリ装置外に搬出およびライブラリ装置内に収納する機能。

4.2 センター開発機能

4.1 のベンダー提供の SDK を使って、次の機能を持つアプリケーションを作成した。

- ・利用者認証を行い、カートリッジをキャビネットから取出し、ロボットアームを制御してドライブに取り付け、マウント操作を行う機能。
- ・利用者認証を行い、カートリッジをアンマウントしてドライブを開放し、ロボットアームを制御してキャビネットに返却する機能。
- ・ジョブ・スケジューラ **PBS pro** を使って、カートリッジにアクセスするジョブを投入・制御する機能。
- ・ロボットアームを排他制御する機能。
- ・ドライブを制御する機能。
- ・カートリッジの ODA ライブラリ外での利用を考慮して、指定したパス名配下のすべてのディレクトリ情報とファイル情報をフルパス形式でファイルに格納し、ディレクトリ構造の **tree** 表示やキーワードを使って検索する機能。
- ・カートリッジの管理(新規登録, 変更, 取消し, ライブラリ内に収納, ライブラリ外に搬出)を行う機能。

4.3 利用方法

ライブラリ内のカートリッジを利用する方法を、以下に示す。

- 1) 使用するカートリッジが格納されているフロントエンドサーバ (ライブラリ制御装置) に **PuTTY** 等のターミナルソフトを使って公開鍵認証でログインする。
- 2) DB 検索コマンドを使って、所有するカートリッジ

の一覧や空きスペース情報およびカートリッジに格納されているファイル情報を検索して、使用するカートリッジを特定する。

- 3) ODA ライブラリ内のカートリッジへのアクセスは、センターが提供する次の 2 種類のバッチジョブ投入コマンドを使ってジョブを投入して、ファイル操作を行う。

- ・インターラクティブジョブで実行
- ・バッチジョブで実行

インターラクティブジョブは、ジョブ投入後、一定時間内に対話形式でコマンドを入力してカートリッジへのアクセスを行う。バッチジョブを利用する場合は、まず、カートリッジへのアクセス手順のコマンド列を記載したテキストファイル(バッチファイル)を作成する。そして、そのファイルの名前をジョブ投入コマンドのオペランドに指定してバッチジョブを投入してカートリッジへのアクセスを行う。

実際のカートリッジをアクセスするための前処理としては、センター提供の **odaalloc** コマンドを使って、次の一連の処理を順に実行してカートリッジをマウントする。

- ・カートリッジの所有者の確認
- ・ドライブの状態の確認
- ・キャビネットからドライブにカートリッジを転送
- ・マウントポイント作成
- ・ボリュームのマウント操作

ボリュームのマウント処理が完了すれば、**ls**, **mkdir**, **cd**, **cp**, **tar**, **rm** 等の **Linux** のファイル操作コマンドを使って、ホットストレージとの間でデータ転送やファイル操作を行うことができる。また、フロントエンドサーバには、カートリッジへの書き出し用の中間ファイルとして、**SSD** のローカルストレージも利用できる。

作業が終了したら、カートリッジの後処理として、センター提供の **odafree** コマンドを使って次の一連の処理を順に実行してカートリッジをアンマウントする。

- ・カートリッジの所有者の確認
- ・ドライブの状態の確認
- ・アンマウント操作
- ・カートリッジ内の記録情報を DB の反映
- ・ドライブの開放
- ・カートリッジをキャビネットに戻す
- ・マウントポイントの消去

5 課題と今後の予定

本稿では、2020 年 7 月の導入した Phase1 の ODA ライブラリ装置と ODA 単体ドライブユニットを使ったコールドストレージサービスについて紹介した。

今回導入した PetaSite 拡張型ライブラリーシステムでは、アーカイブを目的とした商用ストレージ管理ソフトウェアを導入せずに、独自のストレージ管理ソフトウェアを開発することにより柔軟な利用形態を利用者に提供することを目指した。今回、ODA ライブラリ装置をスパコンに接続したのは初の試みなので、ソフトウェアを開発するに当たり、想定外の問題が発生した。現在、データ転送テストやドライブの負荷テストを終え、コールドストレージサービスの利用を開始したところである。

問題点としては、システム当りに搭載するドライブ数が 5 ドライブと少ない点である。この少ないドライブを、どのように割当てて有効に利用するかが課題である。

今後、利用者のご意見やご要望を取り入れて、2021 年 2 月の Phase2 の本稼動に向けてシステムの整備を進める予定である。

本稿で紹介した事例やアイデアがコールドストレージを管理されている方々や利用されている方々の何らかの参考になれば幸いである。

謝辞

コールドストレージシステムを構築するにあたり、ハードウェア及びソフトウェアのサポートいただいたソニービジネスソリューション株式会社、富士通株式会社の方々に感謝いたします。

参考資料

[1] オプティカルディスク・アーカイブのご紹介

https://www.youtube.com/watch?v=TAYQ_FiJrAk

[2] Introduction of PetaSite Library Systems

<https://www.youtube.com/watch?v=pPbt0sARido>

[3] オプティカルディスク・アーカイブ

カートリッジ耐久性実験（水災害,煮沸試験）

<https://www.youtube.com/watch?v=UnIrreBvndg>

<https://www.youtube.com/watch?v=nGzrGgDMupw>