Ipomoea-01 大規模共通ストレージシステムの運用

前田 光教[1)], 宮嵜 洋[1)], 佐藤 孝明[1)], 福沢 秋津[1)], 中張 遼太郎[1)],
山田 新[1)], 山本 和男[1)], 中島 研吾[2)],塙 敏博[2)]

1) 東京大学 情報システム部 情報基盤課
2) 東京大学 情報基盤センター

maeda@cc.u-tokyo.ac.jp

The operation of Ipomoea-01 Large-scale Common Storage System

Mitsunori Maeda[1)], Hiroshi Miyazaki[1)], Takaaki Sato[1)], Akitsu Fukuzawa[1)],Ryotaro Nakahari[1)], Hajime Yamada[1)], Kazuo Yamamoto[1)], Kengo Nakajima[2)], Toshihiro Hanawa[2)]

1) Information Technology Group, Information Systems Department, The University of Tokyo
2) Information Technology Center, The University of Tokyo

概要

2022 年 6 月より運用を開始した Ipomoea-01 大規模共通ストレージシステムに関する導入の経緯とシステムの概要に加え、利用者が Ipomoea-01 を利用するための制度について説明する。

1 はじめに

東京大学情報基盤センター[1](以下、本センター)では、大規模共通ストレージシステム(Ipomoea-01)[2]を導入した。現在、本センターでは、「計算・データ・学習」融合スーパーコンピュータシステム(Wisteria/BDEC-01)[3]ならびに大規模超並列スーパーコンピュータシステム(Oakbridge-CX(OBCX))[4]を運用している。また、2022 年 3 月までメニーコア型大規模スーパーコンピュータシステム(Oakforest-PACS(OFP))[5]を運用していた。他に、本センターを含め全国 11 研究機関で構成・運用されるデータ活用社会創成プラットフォーム(mdx)[6]が本センターで稼働している。このように、本センターでは複数のシステムを運用しており、それぞれストレージが独立しているため、複数のシステムを利用する利用者には不便であることから、OFP の運用終了を機に Ipomoea-01 の導入を決定した[7]。OFP のファイル移行を考慮し 2022 年 1 月に導入して 6 月から正式サービスを開始している。3 年後には次号機 Ipomoea-02 を追加導入して 6 年での更新を予定している。

図 1. Ipomoea-01 の外観

2 導入の背景

スーパーコンピュータの処理能力の向上に伴い、扱うデータ量も増加の一途をたどっている。特に、「計算・データ・学習」の融合を目指す新しい分野では、大量の観測データ、パラメータスタディの結果ファイルなどを処理する必要がある。本センターでは従来ストレージは各システムに附属して導入され、各システムのストレージは独立していた。近年は本センターの利用者も目的や手法に応じて複数のシステムを同時に利用する事例が増加している。また、システムがリプレースされる場合には大量のデー

タをバックアップする必要があった。個別のワークロードのデータ量が増加していることから、このような状況は利用者に多大な不便を強いることになり、本センターの全システムからアクセス可能な共通ストレージの導入が強く求められていた。

3 Ipomoea-01 の設計方針

Ipomoea-01 の設計に際しては、以下の点に配慮した。

3.1 ストレージ通信ネットワークには RDMA に対応した 400Gbps の Ethernet を採用

東京大学柏キャンパスと柏 II キャンパスに複数のシステムが設置されており、最大で 10 km 距離が離れている。これらのシステムから安定したアクセスを可能にするため、スーパーコンピュータのストレージシステムで一般的に用いられる InfiniBand ではなく、RDMA に対応した Ethernet(RoCE v2)を採用した。また、今後導入されるスーパーコンピュータシステム、次世代 Ipomoea-02 との接続を考慮して、調達時点で最新の規格である 400Gbps に対応したネットワークスイッチを導入した。なお、Ipomoea-01 を構成するサーバ等は最大 100Gbps で接続され、冗長化されている。

3.2 並列ファイルシステムに Lustre ファイルシステムを採用

本センターで運用中のシステムは、いずれも Lustre ファイルシステム、または Lustre ファイルシステムを拡張した FEFS(Fujitsu Exabyte File System)を用いている。そこで、Ipomoea-01 でも Lustre ファイルシステムを用いることとした。mdx を除いた全てのシステムでは、OS 上で共通したユーザ ID、グループ ID で管理されており、Lustre ファイルシステムとして各システムから Ipomoea-01 のストレージを直接マウントすることも技術的には可能であると考えられる。ただし、運用の柔軟性を考慮し、現時点では Ipomoea-01 のストレージを一旦 NFS エクスポートし、それを各システムのログインノードでマウントしている。この際には、必要に応じてシステム毎にユーザ ID のマッピングを変更することも可能である。

3.3 ストレージ容量やファイル数に配慮

アーカイブ用途が多いことを想定し、ストレージのアクセス性能よりも、ストレージ容量を優先した。また、データ公開やデータ共有の用途を踏まえ、比較的小さなサイズのファイルが多数格納されることを配慮してメタデータサーバの構成を決定した。

4 システム概要

4.1 ストレージ

Ipomoea-01 のストレージは、並列ファイルシステムで構成された DDN 社製 Lustre ベースの DDN EXAScaler で、ストレージ容量は 25.9 PB である。MDS(メタデータサーバ)として NVMe SSD 3.84TB を 17 本搭載する DDN 1U Server を 4 台と MDT(メタデーターゲット)として DDN SFA200NVX を 1 台で構成する。また、OSS(オブジェクトストレージサーバ)/OST(オブジェクトストレージターゲット)は DDN ES7990X を 1 台と SS9012 を 4 台の 1 セットとして 5 セットで構成する。1 セットあたり NL-SAS-HDD 18TB を 383 本搭載し、それぞれ 100Gbps で接続しネットワークスイッチを介することで転送速度は 125GB/s に及ぶ。(表 1)

表 1. システム全体諸元

項目	機器諸元
ファイルシステム	DDN EXAScaler (Lustre ベース)
ストレージ容量	25.9PB
i-node 数上限	168 億
ストレージデータ転送速度	125GB/s
MDS + MDT	DDN 1U Server × 4 台 + DDN SFA200NV × 1 台
メタデータ格納デバイス	NVMe SSD 3.84TB × 17
OSS + OST	DDN ES7990X × 5 台 + SS9012 × 20 台
ファイルデータ格納 デバイス	NL-SAS-HDD 18TB × 383 (OSS × 1 台 + OST × 4 台 あたり)

4.2 ログインノード兼接続用サーバ

Ipomoea-01 のログインノードは接続用サーバを兼ねている。接続用サーバとは、本センターの各システムから NFS により Ipomoea-01 に接続する機能を有するサーバで、富士通製 PRIMERGY RX2530 M6 4 台をそれぞれ 100Gbps でネットワークスイッチに接続している。オペレーティングシステムは Red Hat Enterprise Linux 8 で稼働する。Ipomoea-01 のログインノードで、ストレージに保存するファイルを編集、操作等ができる他、本センターの各システムからもファイル操作等が可能である。

4.3 管理サーバ群

管理サーバ群はログインノード兼接続用サーバと同型の富士通製 PRIMERGY RX2530 M6 で、運用管理サーバ 2 台、認証サーバ 2 台、利用支援 Web サーバ 2 台、セキュリティログ保存サーバ 2 台の 8 台で構成し、それぞれ 100Gbps でネットワークスイッチに接続している。オペレーティングシステムは Red Hat Enterprise Linux 8 で稼働する。また、管理サーバ群で共有可能なネットワークストレージ(富士通製 ETERNUS HX2100)を備え、管理サーバ群接続用スイッチを介して接続する。

4.4 ネットワークスイッチ

ネットワークスイッチは、400Gbps を 32 ポート有する NVIDIA 社製 MELLANOX SN4700 2 台で構成する Ethernet ネットワークスイッチである。並列ファイルシステム、ログインノード兼接続用サーバ、管理サーバ群をそれぞれ 100Gbps で接続し、外部接続用ネットワークとは総バンド幅 200Gbps で接続する。(図 2)

ログインノード兼接続用サーバを介し NFS マウントできるよう、OBCX の外部接続ルータと 40Gbps を 2 本で、Wisteria の外部接続ルータとは 100Gbps を 2 本で接続する。

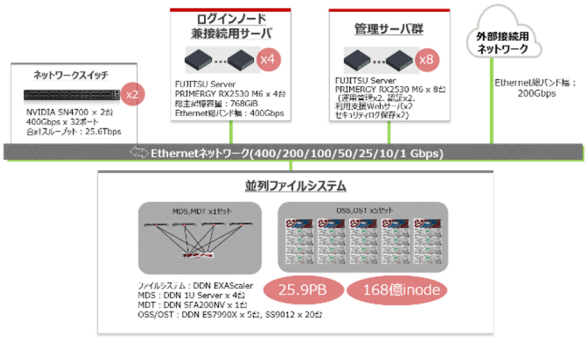

図 2. システム構成概要

5 利用制度

本センターのスーパーコンピュータシステムのいずれかにユーザ ID を有する場合は、申込不要で規定のディスク容量を無償で利用できる。ユーザ ID がない場合でも有償での利用が可能である。

表 2. Ipomoea-01 利用負担金表

ディスク容量	大学・公共機関等	企業
1TB	600 円/1 か月	720 円/1 か月
[10TB まで 1TB ごと]	[350 円/1 か月]	[420 円/1 か月]
10TB	3,750 円/1 か月	4,500 円/1 か月
[100TB まで 1TB ごと]	[250 円/1 か月]	[300 円/1 か月]
100TB	26,250 円/1 か月	31,500 円/1 か月
[1000TB まで 1TB ごと]	[200 円/1 か月]	[240 円/1 か月]
1000TB	206,250 円/1 か月	247,500 円/1 か月
[以降 1TB ごと]	[175 円/1 か月]	[210 円/1 か月]

5.1 ユーザ ID を有する場合

利用者ごとの領域には 5TB のディスク容量が無償で利用できる。また、登録されているスーパーコンピュータで付与されているグループ

のディスク容量の 15%を各プロジェクトのグループごとの領域として無償で利用できる。

(例) 以下の 2 グループに所属している場合（ユーザ ID：UserA）
Wisteria/BDEC-01: ProjectA（100TB、2022 年 12 月まで）
Oakbridge-CX: ProjectB（20TB、2023 年 3 月まで）
→ /home/UserA/：5TB（2023 年 3 月まで）
/work/ProjectA/：15TB（2022 年 12 月まで）
/work/ProjectB/：3TB（2023 年 3 月まで）

無償分を超えて利用したい場合は利用負担金(表 2)でディスク容量を追加することもできる。ただし、ディスク容量追加の金額算出時には有償で追加済みのディスク容量を含む。

(例) 既に 1TB・12 か月を有償で追加済みで、更に 10TB・12 か月を追加する場合
→ ([350 円/1 か月] × 9TB × 12 か月) + ([250 円/1 か月] × 1TB × 12 か月) = 40,800 円

5.2 ユーザ ID がない場合

利用申込手続きをすることで利用することができ、利用申込ディスク容量に対する利用負担金(表 2)が発生する。

5.3 ファイル Web 公開サービス

Ipomoea-01 上のファイルを Web 公開することができる。利用者のホームディレクトリ(/home 配下)を対象とし、意図しない公開とならないよう代表者からの希望制としている。申込後に Web 公開に必要なファイル(.htaccess)を該当ディレクトリに配置し、利用者自身で認証や公開範囲を設定する。

5.4 OFP サービス終了に伴う対応

2022 年 3 月にサービス終了した OFP の利用者向けにファイル移行サービスと 2022 年 11 月 30 日までの無償利用措置を実施した。OFP からの移行ファイルの対象を希望制とすることで移行期間の短縮を図った。詳細は次項で述べる。

OFP からのファイル移行サービスを希望した利用者は 2022 年 11 月 30 日までは利用者ごとの領域に 5TB の無償分のディスク容量が付与される。グループ代表者がファイル移行サービスを希望したグループは 2022 年 11 月 30 日まではグループごとの領域に 5TB の無償分のディスク容量が付与される。ただし、2022 年度も同プロジェクトコードで本センターのスーパーコンピュータへの登録がある場合は、登録されているスーパーコンピュータにおけるディスク容量の 15%が無償分のディスク容量として付与される。移行した容量が無償分を超えている場合もあり 2022 年 12 月 1 日以降も利用するには有償申込を必要とする。2022 年 12 月 1 日以降も利用する場合のディスク容量と利用負担金(表 2)は同様である。

6 OFP からのファイル移行

6.1 予備調査

Ipomoea-01 導入前の 2021 年 12 月に OFP 利用者に対し運用終了に伴うファイル移行希望の予備調査を実施した。この調査でファイル移行に 3 ヶ月を要することが算出され、OFP のストレージとファイル移行に必要な最小構成機器を 4 月から 1 か月間残すことにした。

6.2 ファイル移行構成

ファイル移行は OFP と Ipomoea-01 を NFS マウントし rsync を実行することとした。OFP はプリポストノード 8 台、Ipomoea-01 はログインノード 4 台を移行用サーバに活用し rsync を多重起動する。移行対象ファイルを順次移行するが、ファイルサイズやファイル数が大きいものは分析した上で多重度と調整しながら移行する。

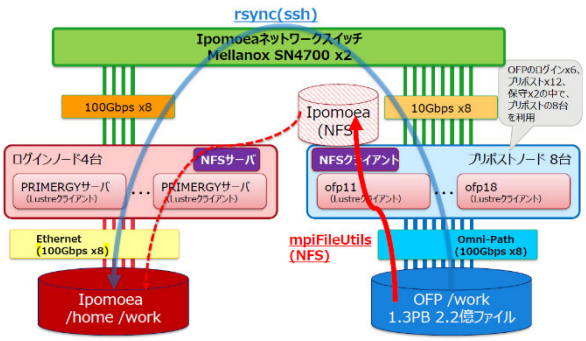

図 3. ファイル移行構成図

6.3 ファイル移行

2022 年 1 月にファイル移行希望の本調査を実施し、その下旬からファイル移行を開始するが、OFP は 3 月までサービスを継続しており利用者のファイルが更新されるため、サービス終了した 4 月に rsync にて差分データの最終同期を実施した。移行前に取得済みのファイルリストからファイルの同一性(チェックサム)を比較することで移行完了の確認とする。ただし、ファイルリストの全数ではなく、無作為にファイルを抽出している。

表 3. ファイル移行対象データ

	ユーザ	グループ	合計
ディレクトリ数	113	35	148
ユーザ数	96	-	-
グループ数	76	35	-
サイズ(GiB)	1,224,573	108,427	1,332,999
ファイル数	222,402,820	1,134,035	222,536,855
ファイル数(1MiB 未満)	173,563,418	589,902	174,153,320

7 利用状況

Ipomoea-01 のディスク使用率は 2022 年 10 月現在 4%程度に留まっている。また、OFP から移行したファイルが大きく、ディスク使用量の 70%近い容量が無償分を超えた利用者で締めているため、12 月からの有償サービスが始まるまでにさらに使用率が下がる可能性がある。一方、OBCX のサービスが 2023 年 5 月末まであり、OFP の後継機を 2024 年 4 月に導入する予定であるため、本センターの利用者が Ipomoea-01 にデータ移行することを想定しており、今後使用率が高くなると予測する。

8 まとめ

本稿では、2022 年 6 月より運用を開始した Ipomoea-01 大規模共通ストレージシステムについて導入と利用状況を報告した。まだ 4 ヶ月が経過したばかりであり利用は多いとは言えない。一方で、複数システムを利用している利用者がシームレスなデータ利用とシステム間のデータ移行時の不便を解消できることは間違いなく、今後の利用拡大に期待する。

参考文献

[1] 東京大学情報基盤センター http://www.cc.u-tokyo.ac.jp/

[2] Ipomoea-01 システム https://www.cc.u-tokyo.ac.jp/supercomputer/ipomoea01/service/

[3] Oakbridge-CX スーパーコンピュータシステム http://www.cc.u-tokyo.ac.jp/supercomputer/obcx/service/

[4] Wisteria/BDEC-01 スーパーコンピュータシステム https://www.cc.u-tokyo.ac.jp/supercomputer/wisteria/service/

[5] Oakforest-PACS スーパーコンピュータシステム https://www.cc.u-tokyo.ac.jp/supercomputer/ofp/service/

[6] mdx: データ活用社会創成プラットフォーム https://mdx.jp/

[7] 東京大学情報システム部「大規模共通ストレージシステム（第 1 世代）運用に関するお知らせ」、東京大学情報基盤センタースーパーコンピューティングニュース、Vol.24 No.1 (2022 1 月). https://www.cc.u-tokyo.ac.jp/public/news.php#VOL24