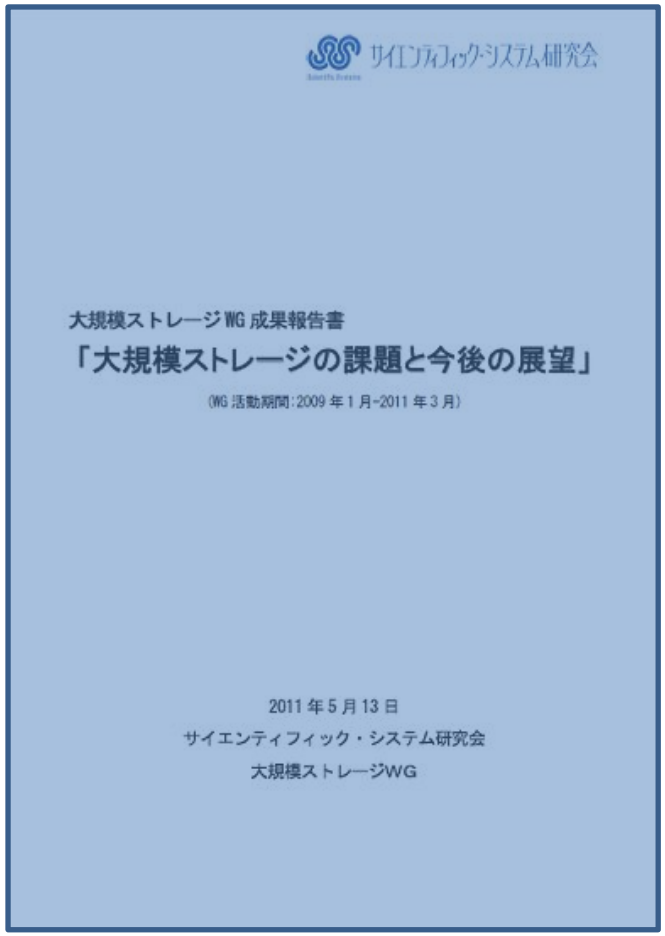
ーSS研大規模ストレージWG成果報告ー

大規模ストレージシステムの課題と今後の展望

SS研 大規模ストレージWG まとめ役
宇宙航空研究開発機構 藤田直行
fujita@chofu.jaxa.jp



ハンドアウト無し

成果報告書　目次

1. はじめに
2. 大規模ストレージシステムの動向
3. 各会員システムの概要
4. 大規模ストレージにおける課題とその対策
5. **ファイルシステム健康診断**
6. 今後に向けての提言

付録

3.1　システムの概要
3.2　今後のシステムの強化ポイント

4.1　**大容量化**
4.2　**高速化**
4.3　運用
4.3.1　**バックアップ**
4.3.2　**リビルド**
4.3.3　**データ移行**
4.3.4　**高可用性**

2.1　**ストレージ技術**
2.2　**高性能ファイルシステム**
2.3　**HSMシステム**
2.4　海外スーパーコンピュータシステムにおけるストレージシステム

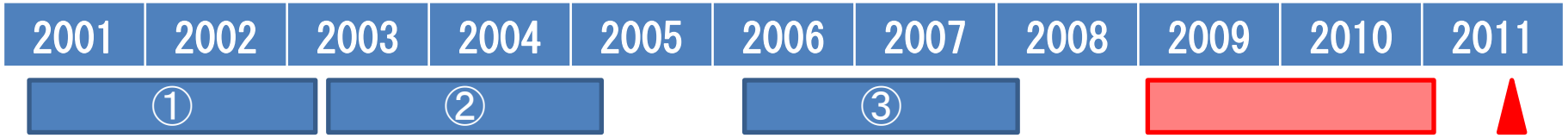
Agenda

- はじめに
- 大規模ストレージシステムの動向
- 大規模ストレージシステムの課題と対策
  - 大容量化
  - 高速化
  - バックアップ
  - リビルド
  - データ移行
  - 高可用性
- ファイルシステム健康診断
- おわりに

Agenda

- はじめに
- 大規模ストレージシステムの動向
- 大規模ストレージシステムの課題と対策
  - 大容量化
  - 高速化
  - バックアップ
  - リビルド
  - データ移行
  - 高可用性
- ファイルシステム健康診断
- おわりに

■SS研でのストレージに関するWG



①ネットワーク時代の統合ストレージマネージメントWG
(活動期間:2001年5月～2003年4月)

②ストレージを中心としたシステムマネージメントWG
(活動期間:2003年5月～2005年4月)

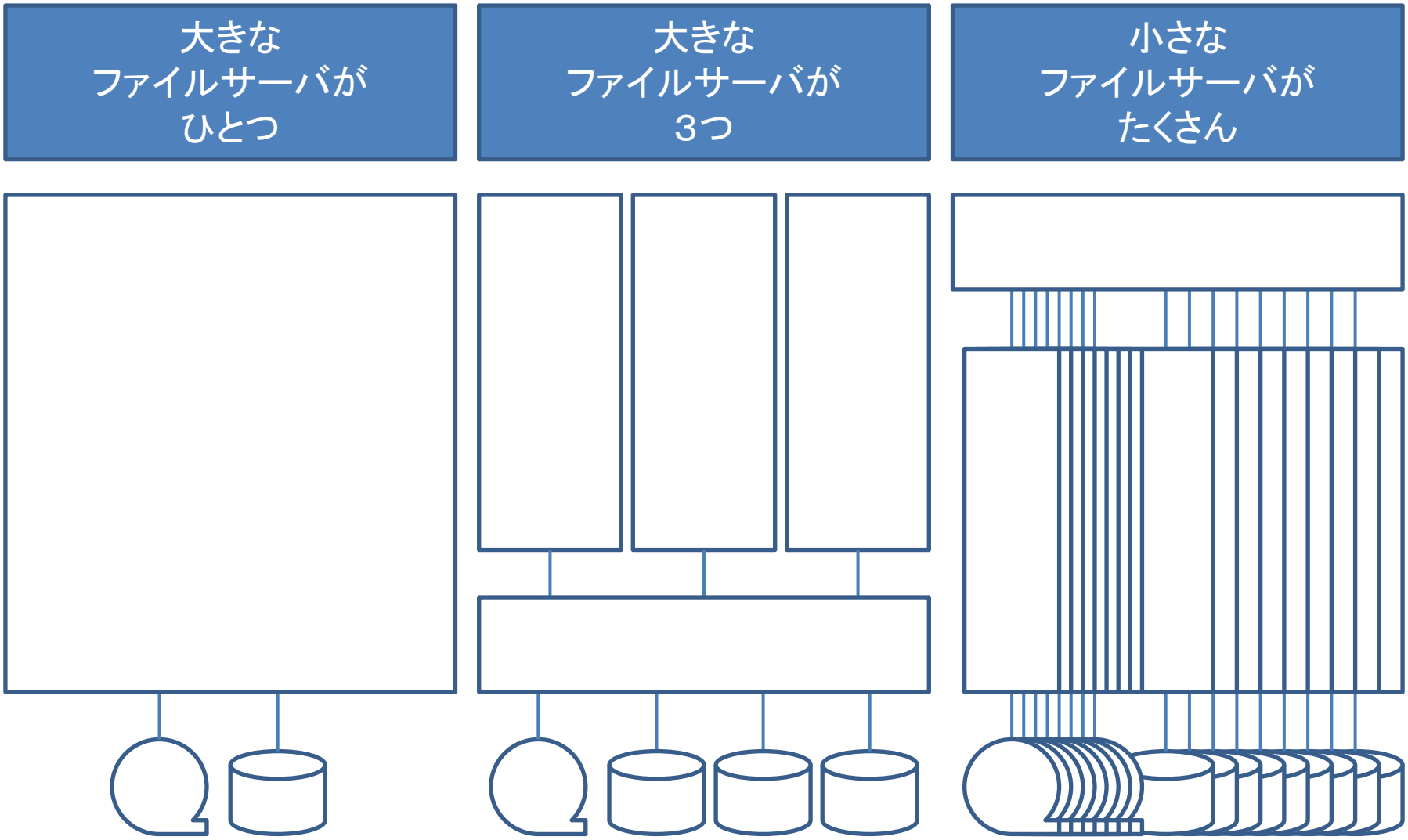
③データマネジメントを意識したストレージソリューションWG
(活動期間:2006年2月～2008年4月)

■大規模ストレージWG
(活動期間:2009年1月～2011年3月)

■大規模ストレージWGメンバ

			氏名	機関/所属(2011年3月31日現在)
会員	担当幹事		松澤　照男	北陸先端科学技術大学院大学
	推進委員	(まとめ役)	藤田　直行	宇宙航空研究開発機構
			八代　茂夫	高エネルギー加速器研究機構
			竹内　康雄	神戸大学
			八木　雅文	国立天文台
			早戸　良成	東京大学宇宙線研究所
			黒川　原佳	理化学研究所
賛助会員(富士通)	推進委員	(まとめ役)	栄元　雅彦	富士通(株)
			荒木　純隆	富士通(株)
			長屋　忠男	富士通(株)
			住元　真司	富士通(株)
			甲斐　俊彦	富士通(株)
	オブザーバ		酒井　憲一郎	富士通(株)
			阿部　孝之	富士通(株)
			飯島　敏治	富士通(株)
			川村　寛	富士通(株)
			坂口　吉生	富士通(株)
			塚原　知宏	富士通(株)
			松井　秀司	富士通(株)

■ファイルシステムの変化 (昨年度の懇談会資料より)



“世界最高性能のファイルシステムの提供を開始”

Agenda

- はじめに
- 大規模ストレージシステムの動向
- 大規模ストレージシステムの課題と対策
  - 大容量化
  - 高速化
  - バックアップ
  - リビルド
  - データ移行
  - 高可用性
- ファイルシステム健康診断
- おわりに

大規模ストレージシステムの動向

- ホストインターフェース
- ディスクドライブ
- ストレージシステムとファイルシステムの統合
  (将来技術)
- 高性能ファイルシステム
- HSMシステム

■ホストインターフェース

ホストインターフェース種	現状 2010年 [bit/s]	次世代 2011～2012年 [bit/s]	次々世代 2013年～ [bit/s]	備考
ファイバーチャネル	8G	16G	？	将来のロードマップが未定
FCoE	―	10G	40/100G	将来有望
iSCSI	1G	10G	40/100G	オーバーヘッド大

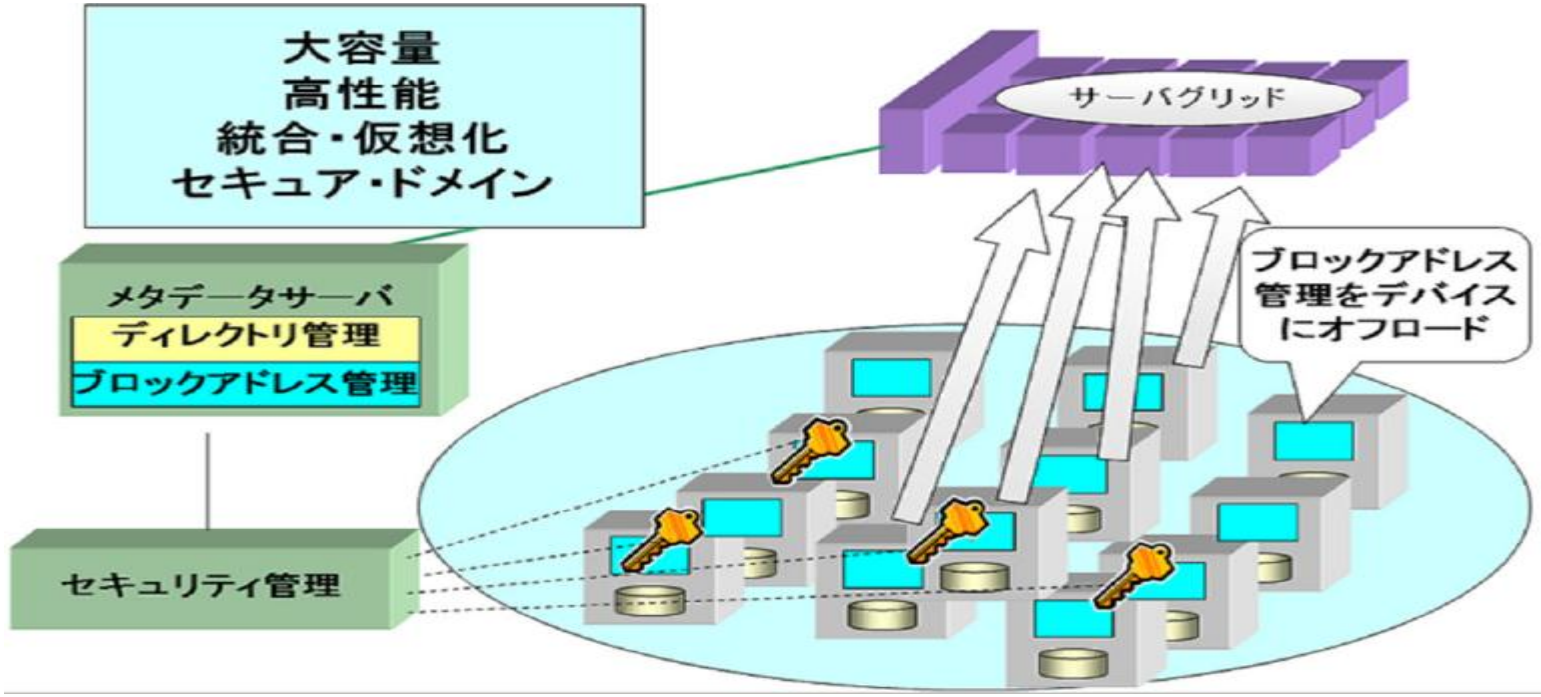
■ディスクドライブ

種別	サイズ 回転数	ｲﾝﾀｰﾌｪｰｽ	現状 2010年	次世代 2011～ 2012年	次々世代 2013年～	備考
オンライン ディスク	3.5” 15,000rpm	FC	600GB	←	―	将来供給 停止
	2.5” 10,000rpm	SAS	600GB	900GB	容量増	主流へ
ニアライン ディスク	3.5” 7,200rpm	SATA/SAS	2TB	3TB	4TB	容量単価 で優位
	2.5” 7,200rpm	SAS	―	1TB	容量増	
SSD	―	SAS	200GB	400GB (SLC)/ 800GB (MLC)	容量増	性能重視 の 領域では 将来有望

SLC:Single Level Cell、素子に1bitのみ記録する。高速で書き込み回数上限値が多い。
MLC:Multi Level Cell、素子に2bit以上を記録する。容量対価格では優位であるが書き込み回数上限値が少ない。

■ストレージシステムとファイルシステムの統合(将来技術)

- OSD(Object Storage Device)の特徴
  - ✓ SNIA(Storage Networking Industry Association)が標準化を検討中
  - ✓ データをオブジェクトとして管理
  - ✓ オブジェクト単位にセキュリティのためのアクセスキーやQoS、リテンション等のアトリビュート情報があり、これらの制御をストレージ側が自律的に行う
  - ✓ アクセスノードおよびストレージノードをスケーラブルに拡張できる
  - ✓ アトリビュート情報の設定により、ストレージノードによる自律的なバックアップやマイグレーションが可能であり、オンラインでのストレージマイグレーションも可能



■高性能ファイルシステム

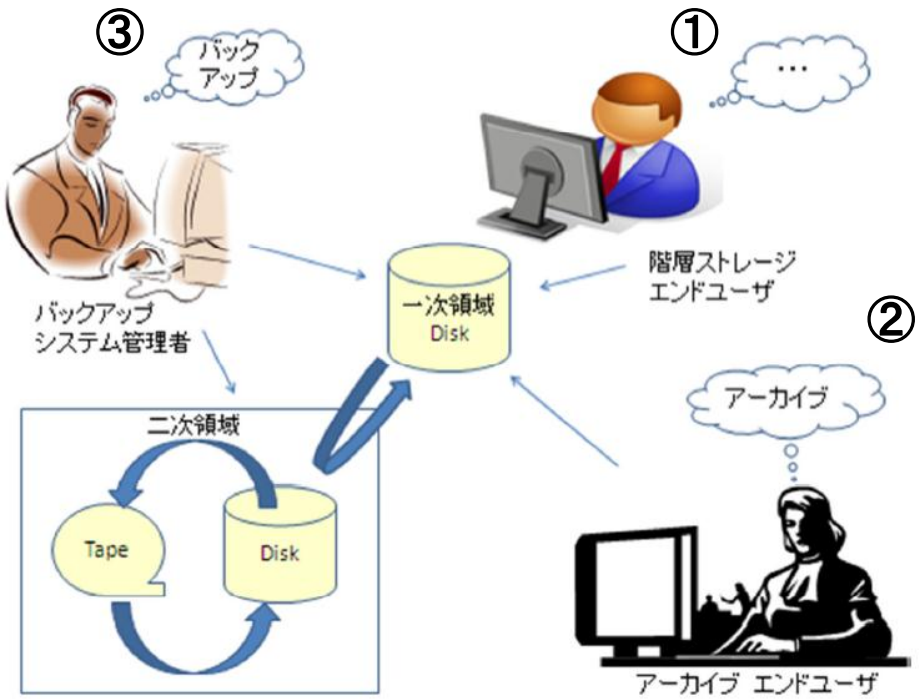
ベンダ	ファイルシステム名	備考
富士通	SRFS:NFSベース	ローカルFS+SRFSの形態で利用 大規模SMPサーバ利用で数千ノードまでサポート
日立	HFS	高性能ファイルシステム：東大T2K
ＮＥＣ	GFS LXFS:Lustreベース	SAN共有ファイルシステム Lustre動作を保証
HP	CFS:Lustreベース	独自に改造、サポート
ＩＢＭ	GPFS	独自クラスタファイル
SGI	CXFS	SAN共有ファイルシステム
CRAY	Lustre, PanFS, NFS, Storenext, GPFS	主としてXT系でサポート
Panasas	PanFS、pNFS	ブレード型HWと一体となったファイルシステム&ストレージ
DDN	Lustre	自社製ストレージ向けに最適化提供&サポート

■今後の動向
・クラスタファイルシステムへの移行が進む
・性能確保のためHDD数やサーバ数の増大が予想され、設置面積・コスト・電力が課題

■HSMシステム

■利用用途
①階層ストレージ
②アーカイブ
③バックアップ

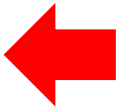
■製品動向
大規模HPCユーザの
HSMソフトウェアとしては、
・SAM-FS
・HPSS ➡ GHI:GPFS+HPSS
・PetaServe
で大半を占めるが、ここ数年機能面の画期的な進歩は無い



■今後の動向
上述②、③の用途が多く、①の用途は衰退していくと予想

Agenda

- はじめに
- 大規模ストレージシステムの動向
- 大規模ストレージシステムの課題と対策
  - 大容量化
  - 高速化
  - バックアップ
  - リビルド
  - データ移行
  - 高可用性
- ファイルシステム健康診断
- おわりに



大規模ストレージシステムの課題と対策

- 大容量化
- 高速化

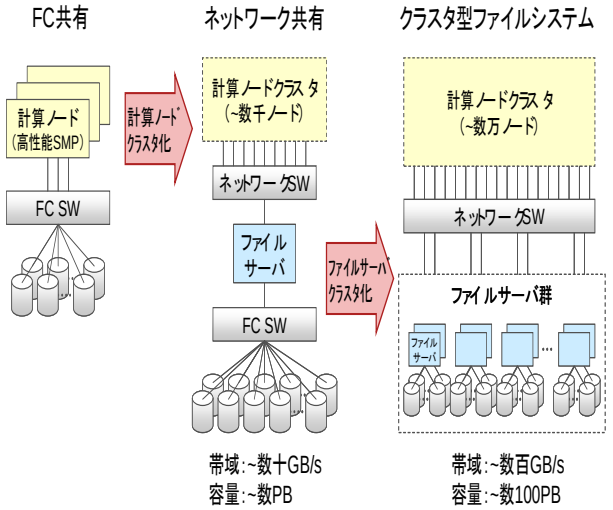
→ ハードウェア / ソフトウェア での対策

- バックアップ
- リビルド
- データ移行
- 高可用性

運用 → 課題の再整理と対策

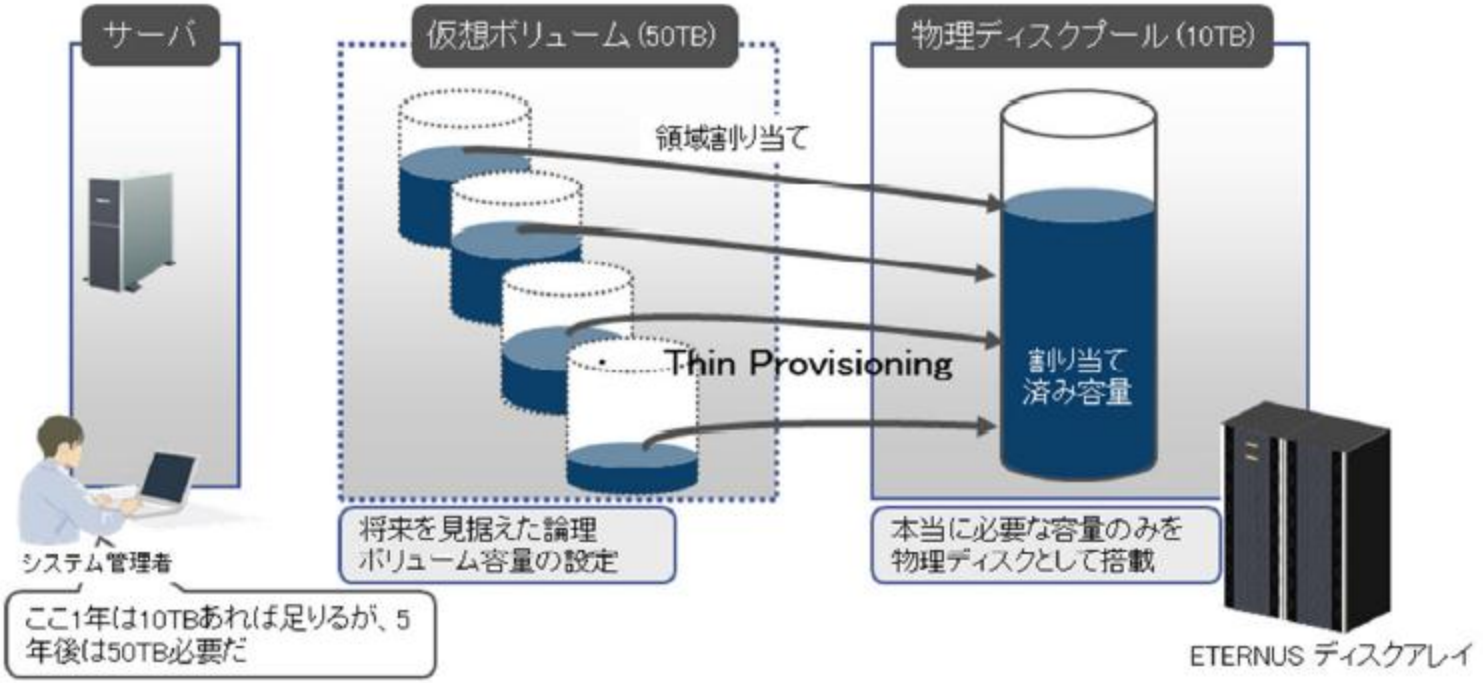
■大容量化

■課題
✓ディスクドライブ容量は1.5～2倍/1.5～2年。
✓システム要求は10～100倍/更新時。
✓総コア数・メモリ数の増加に伴い、
　ファイルサーバはクラスタ化へ。
✓設置面積・消費電力・空調能力・コストが増大。



■ハードウェアによる対策
✓高密度実装による設置面積改善
✓高効率電源,コントローラ小型化,冷却用ファンの改善による電力削減
✓SSDの大容量化による容量当たりの消費電力改善
✓その他、Thin Provisioning、De-duplication、MAID
✓テープ媒体の利用によるコスト削減
■ソフトウェアにおける課題
➢大規模化：TB/s級の性能、EB級のファイルシステム容量が必要
➢スケーラビリティ：ストレージのサーバ台数に比例した性能・容量のスケールアップ
➢使い易さ：動的なハードウェア増設、運用状態の監視

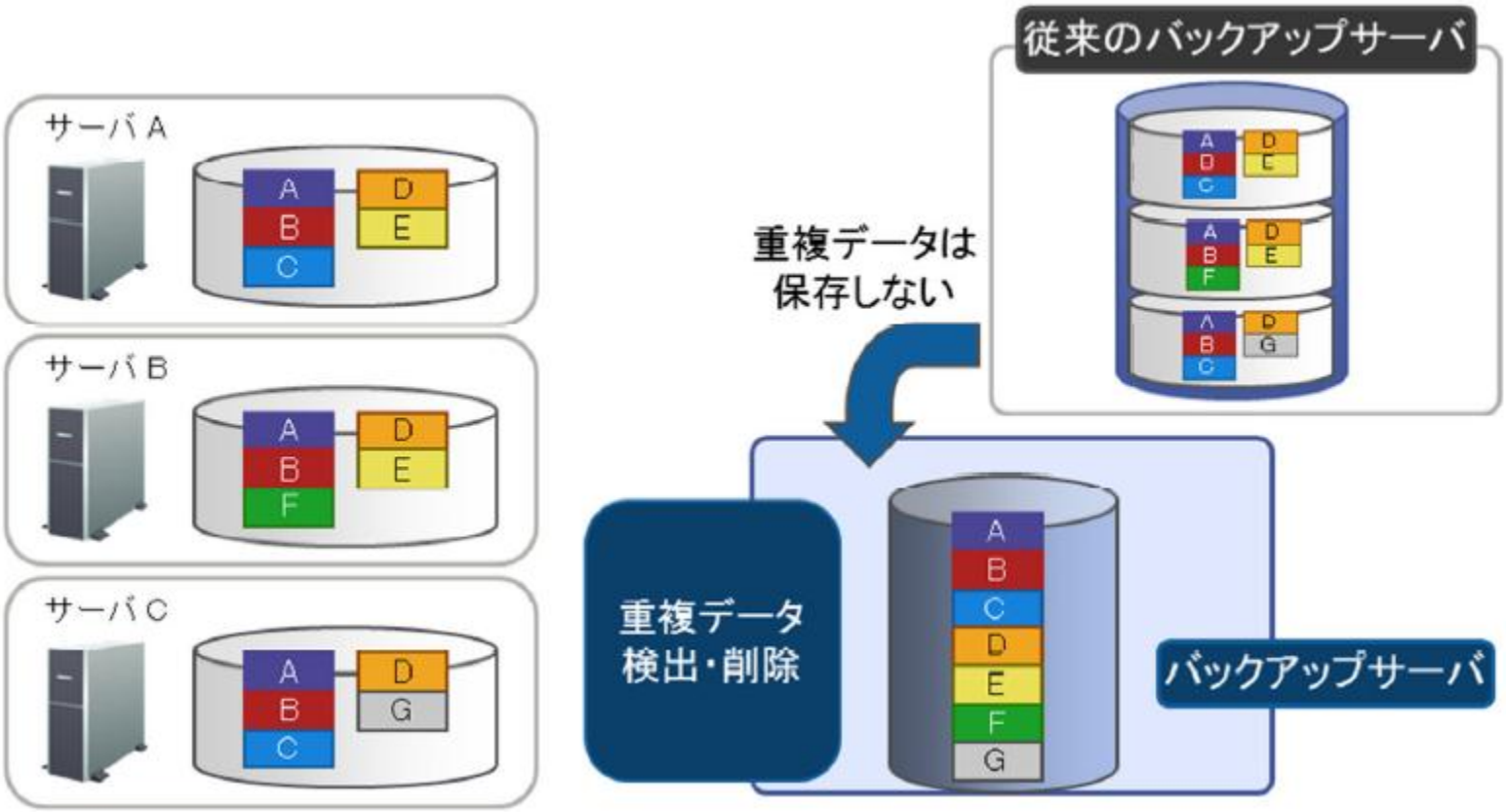
■Thin Provisioning



■物理ディスクをプール化することにより、ボリュームごとによる利用率の違いを気にすることなく、全体で物理ディスクを有効利用できる

■仮想化により、事前の困難な容量設計が不要となるため、設計時の検討時間が短縮されます

■De-duplication



■運用状態の監視

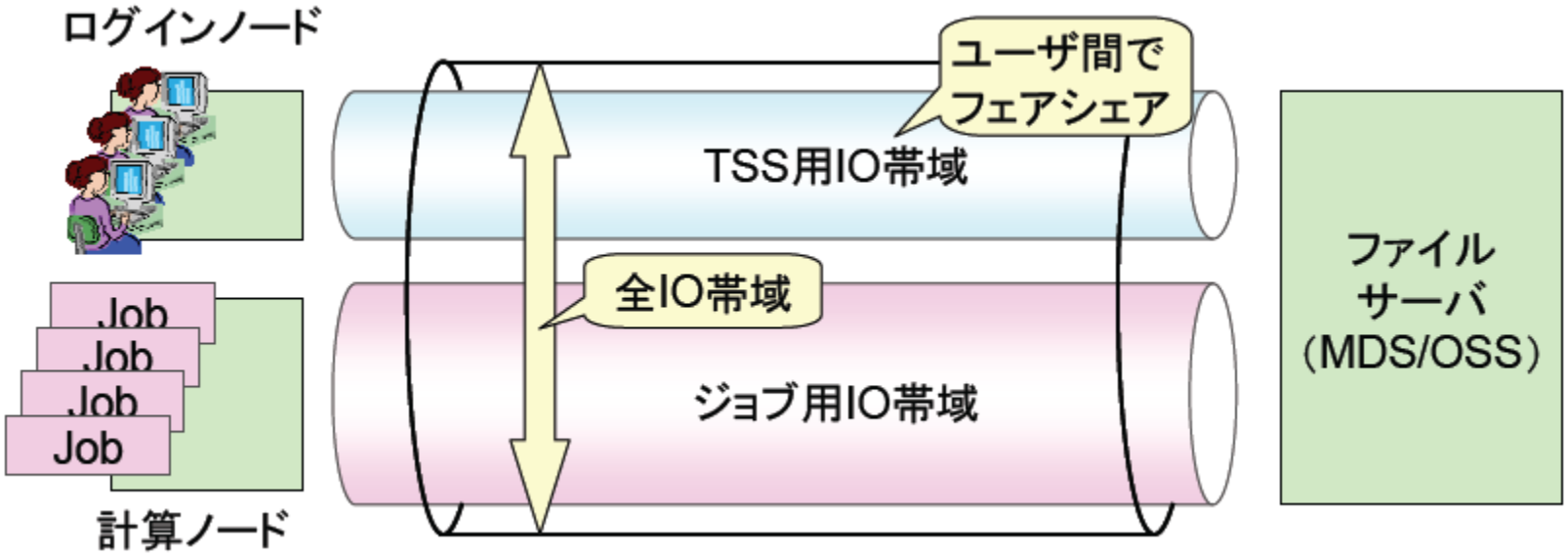
対象		内容	アクション
負荷	サーバ	負荷の偏り(CPU，メモリ)	・統計情報から原因調査 ・QoS設定変更 ・ジョブ運用ポリシィ見直し
	ネットワーク	通信量の偏り	
	ディスク	IO量の偏り	
ディスク容量		空き容量の偏り	・格納先の優先度を調整
統計情報		クライアント・ユーザ単位に取得 ・ディスクIO量(IOサイズ別) ・メタ操作数(open回数等)	・悪さをしているユーザを特定・指導

■高速化

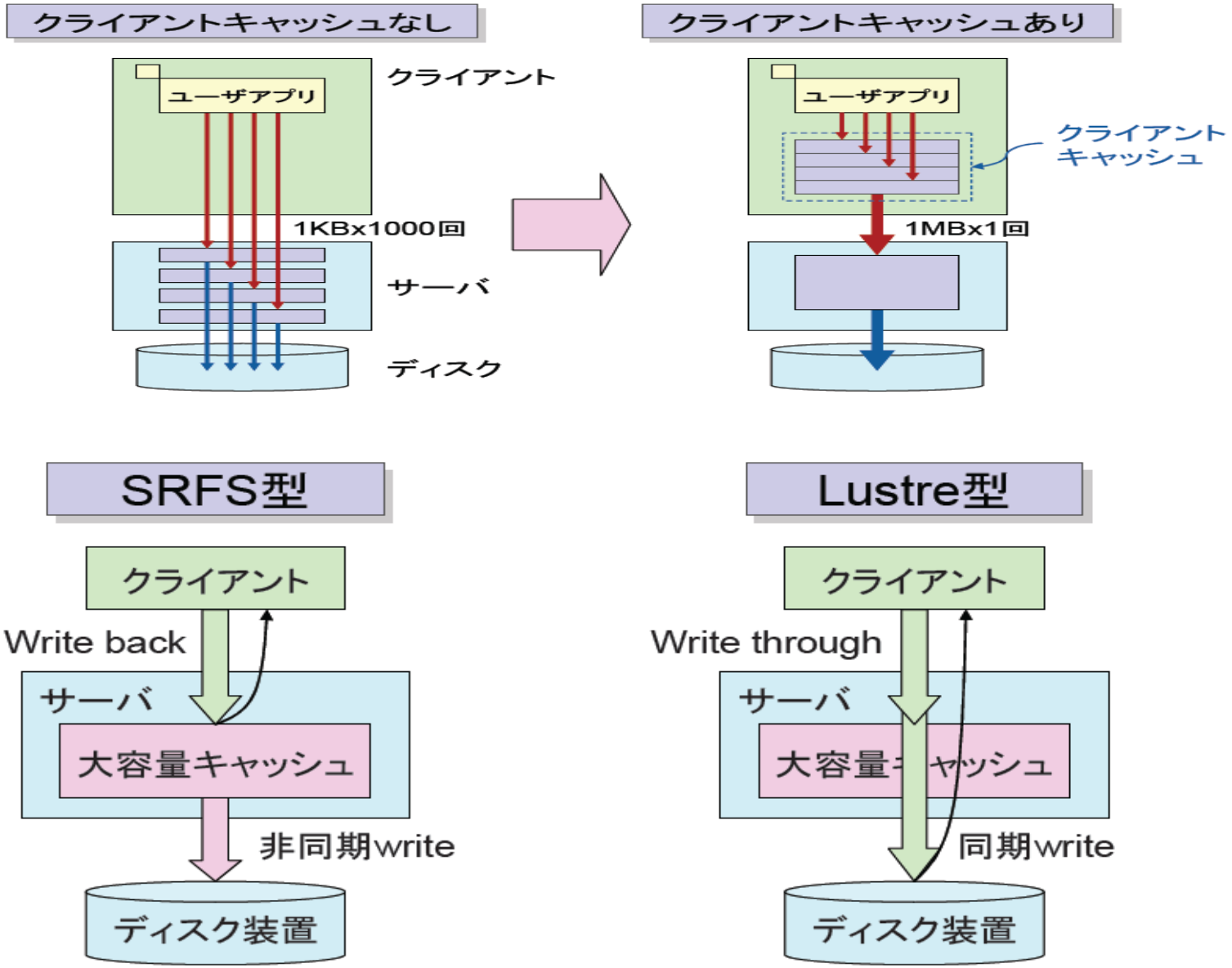
■課題
✓非定常計算の増大等により、IOの高速化が求められている。

■ハードウェアによる対策
✓ストレージの性能はドライブ種とその数によってほぼ決まるため、
　高速ドライブと分散処理による高速化が基本となる。
✓SSDは小IOサイズのリード処理でオンラインディスクの50倍高速。
✓SSDは大IOサイズライト処理1多重ではオンラインディスクと同程度。
　多重アクセスの場合は4倍高速。
■ソフトウェアにおける課題(報告書にいくつかの実験結果の記載有り)
➢メタデータアクセス性能の改善：
　・高速SMPサーバの利用、
　・メタデータアクセスの緩和
　・メタデータアクセスの分散化
➢TSSレスポンス保障
➢小サイズIO・多数ファイル
　・クライアントキャッシュ
➢ディスク性能越え
　・サーバキャッシュ
➢運用ポリシ選択
➢輻輳防止

■TSSレスポンス保障

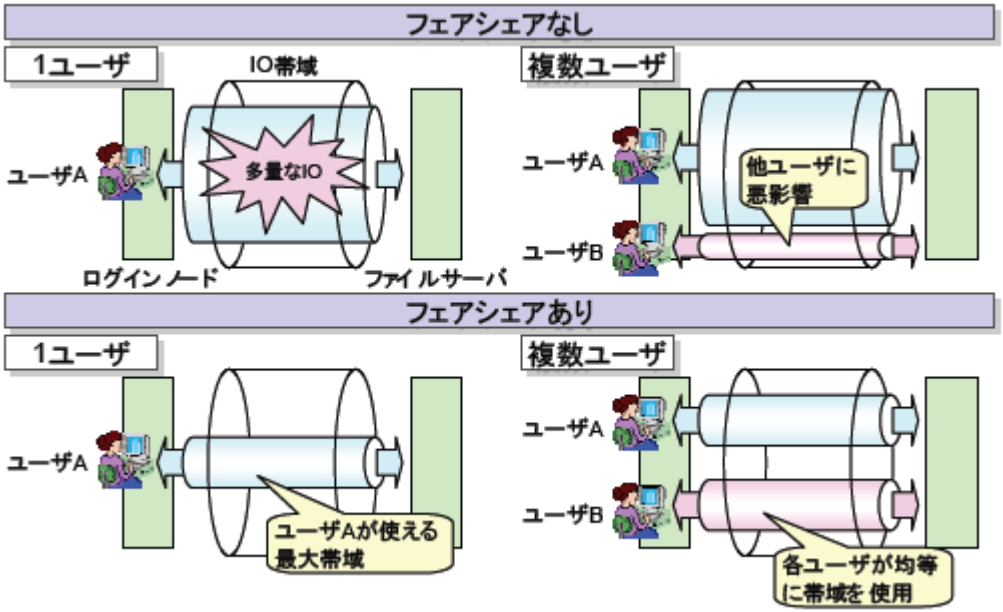

ログインノード
Job
Job
Job
Job
計算ノード
ユーザ間で
フェアシェア
TSS用IO帯域
全IO帯域
ジョブ用IO帯域
ファイル
サーバ
(MDS/OSS)

■クライアント/サーバ　キャッシュ

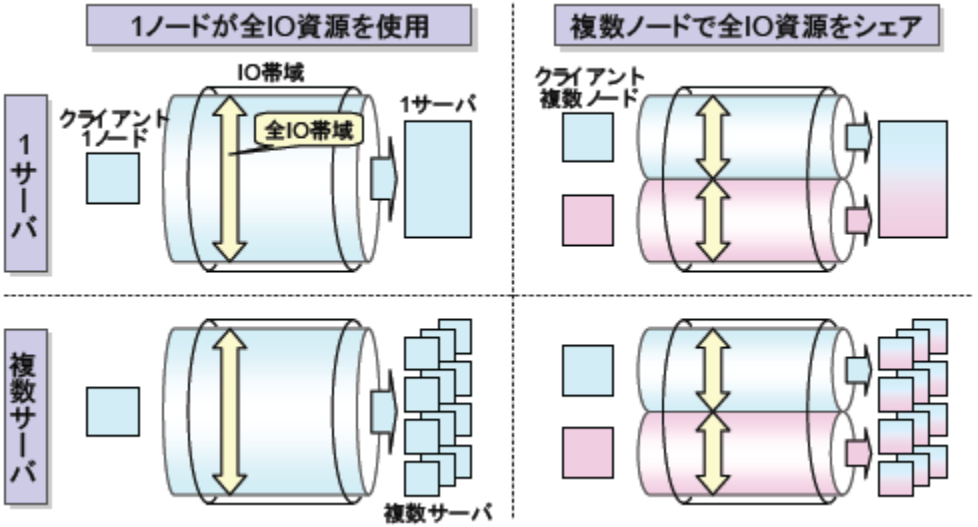


■運用ポリシ選択

フェアシェア



総ｽﾙｰﾌﾟｯﾄ優先



■バックアップ

■課題

物理バックアップにおける、バックアップデータ量の抑制

■対策

明確な対策は無い。以下の対策の組み合わせ、サイト独自の最適なバックアップ運用

✓論理バックアップ(スナップショット)

　＋処理時間、運用性

　－保全が完全ではない

✓De-Duplicationの導入

　＋保全性、運用性

　－新規システム導入によるコスト面

✓世代数/頻度の削減、対象の限定、ユーザ自身による実施

　・運用面での対策

　・信頼性やユーザへのサービスレベル低下とのトレードオフ

■リビルド

■課題

リビルド処理時間の長期化

例)ETERNUS DX80,ディスク単体容量1TB(7200rpm),RAID6(4D+2P)

- 無負荷時　　：約18.5 時間
- 業務IO 100[IOPS]時：約81.5 時間

※リビルド中はレスポンスが15%悪化する

■対策

根本解決策は見つかっていない。運用の工夫により影響範囲を狭めることが現実解。

✓コピーバックが排除可能な保守の仕組み(現状大変困難)

✓ファイルシステムの構築による影響の局所化

■データ移行

■課題

データ移行期間の短縮化

例)WG参加機関の次期システムへの移行期間試算

・0.7ヶ月間～6.5年間 ➡ 0.7ヶ月間～7.3ヶ月間

■対策

観点	対策
ハードウェア	・移行時のネットワーク帯域を増強する。 ・移行用テープドライブを増強する。 ・新システムでのDe-Duplication装置を適用する。
ソフトウェア	・媒体の物理移動／リパックで対応する。 ・NFS V4を活用する。
運用	・データ量を削減する。 （不要データの削減を促す。ディスク課金をする。） ・テープ→ディスクの移行を、ディスク→ディスクに変更し、データ移行の速度を向上させる。 ・現システム運用中にテープ媒体を次世代媒体に変更し、テープ巻数を削減する。
次世代ﾃｸﾉﾛｼﾞｰ	・OSD(Object Storage Device)を適用する。

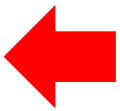
■高可用性

■課題
常にどこかのハードウェアが故障している状況が想定される
数百台規模のファイルサーバ・ストレージ装置で、
継続的サービスを実現する。

■対策
✓ハードウェアの二重化
✓ソフトウェアによる交換・リカバリ制御

Agenda

- はじめに
- 大規模ストレージシステムの動向
- 大規模ストレージシステムの課題と対策
  - 大容量化
  - 高速化
  - バックアップ
  - リビルド
  - データ移行
  - 高可用性
- ファイルシステム健康診断
- おわりに



ファイルシステム健康診断(1/3)

- 背景

  大規模ストレージの計測に適したファイルシステム性能測定の共通ツールが存在せず、性能把握やシステム間の測定結果の比較ができない。

- 目的

  (1)導入システムの健康診断

  (2)運用中の定期健康診断

- 診断モデル

  (1)導入システムの健康診断

  • 最大スループットバンド幅性能

  • 最小スループットバンド幅性能

  • メタデータアクセス

  (2)運用中の定期健康診断

  • 最大スループットバンド幅性能

  (3)精密検査

  • 本診断の対象外ではあるが、精密検査が可能なように考慮されている。

  • キャッシュの影響制御等、診断ツール内のスクリプトをカスタマイズする。

ファイルシステム健康診断(2/3)

- 診断項目

(1)大規模データ転送

(スループット性能:プロセス当たりのファイル量一定、大IO 長)

(2)データ量一定

(スループット性能:小ファイルサイズ、小IO 長)

(3)メタデータアクセス

(レスポンス性能)

診断が一時間程度で完了すること。

- 診断クラス

クラス	規模	論理 ｽﾙｰﾌﾟｯﾄ (ｵｰﾀﾞｰ)	最大 ﾌﾟﾛｾｽ数	最大 ﾃﾞｨｽｸ 使用量	最大 ﾌｧｲﾙ数	ｼｽﾃﾑ例
XL	超大規模	100GB/s	1000	1TB	1,000,000	次世代ﾏｼﾝ
L	大規模	10GB/s	100	100GB	100,000	ｾﾝﾀﾏｼﾝ
M	中規模	1GB/s	10	10GB	10,000	ｾｸｼｮﾝﾏｼﾝ
S	小規模	100MB/s	5	5GB	5,000	個人ﾏｼﾝ

ファイルシステム健康診断(3/3)

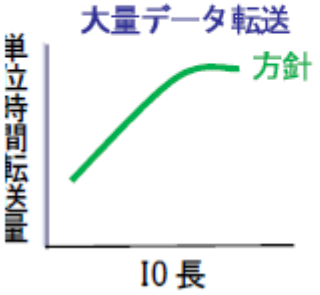
- 診断ツール
  - オープンソースソフトウェア(OSS)であること
  - 導入が簡単であること
  - MPI によるプロセス並列に対応していること
  - 必要な測定ができるツールであること

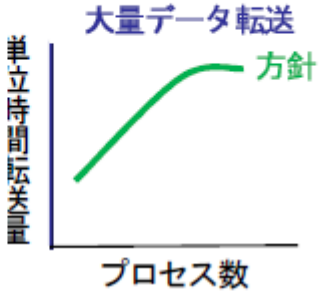
名称	Version	説明
OpenMPI	1.4.2	最も普及しているOSSのMPIソフトウェア。 但し、測定対象システムに既にMPI環境が整備されている場合は不要。
Mdtest	1.8.3	ファイルの操作(create,stat,unlink等)の応答性能を評価するツール。
IOR	2.10.2	データ転送性能を評価するツール。IO長やファイルサイズ等が指定可能。

■診断例(JAXA)

・システム設計方針

大ファイル・大 IO 長、多数プロセスでのバンド幅を重視したシステム。
アプリケーションからの IO 長は MB 単位を想定しファイルシステムブロックサイズは 1MB に設定。





・診断

設計方針通り、1 ファイル 1MB 以上での多数プロセスによるアクセスに適切。注意点は以下。
－同時にファイル stat を行うジョブは 50 プロセス未満とすべき
－ファイル／ディレクトリの create/remove は 100 プロセス同時実行がピークで以降は頭打ち

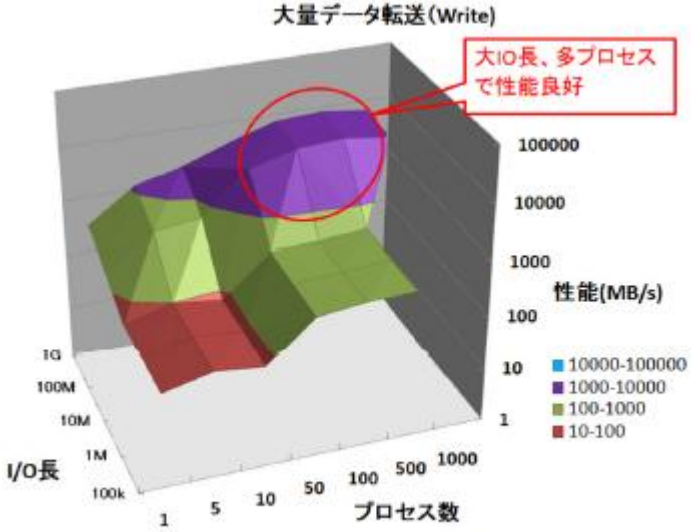


図 5.7.1 宇宙航空研究開発機構(調布) 大規模データ転送

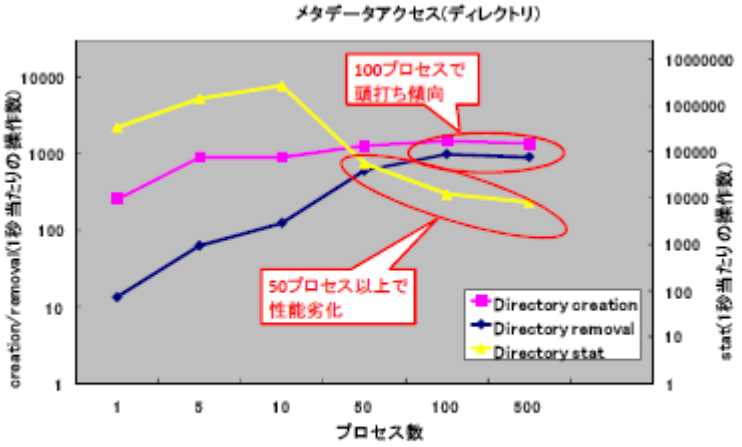
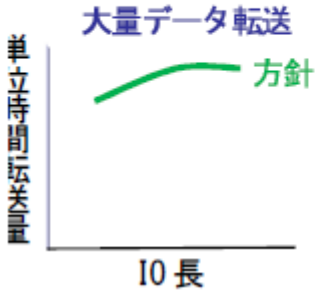


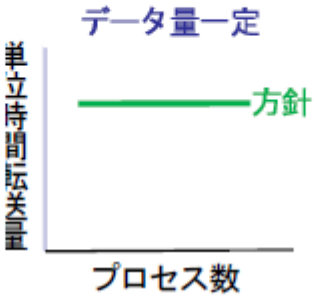
図 5.7.5 宇宙航空研究開発機構(調布)
メタデータアクセス(ディレクトリ)

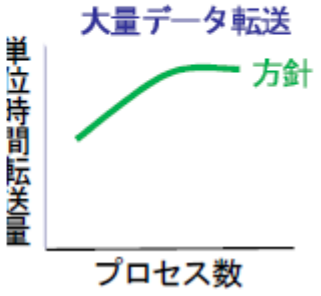
■診断例(理研)

・システム設計方針

小頻度大 IO から多頻度小 IO まで幅広く高いレンジで安定したバンド幅性能を出す。
ブロックサイズは 64KB に設定。







・診断

ーデータアクセスは設計方針通りの性能を発揮。プロセス数や IO 長を意識せず使用可能。
ーファイル操作系ではファイル stat を同時に 50 超発行するジョブでは性能劣化がある点に注意。

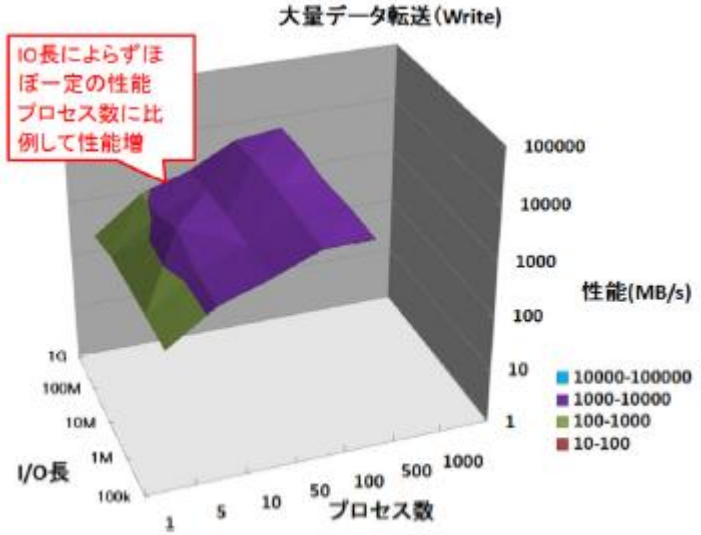


図 5.7.6 理化学研究所 大規模データ転送

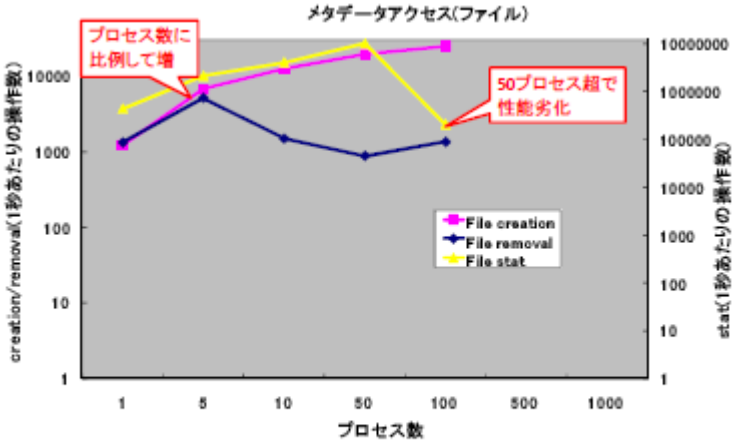
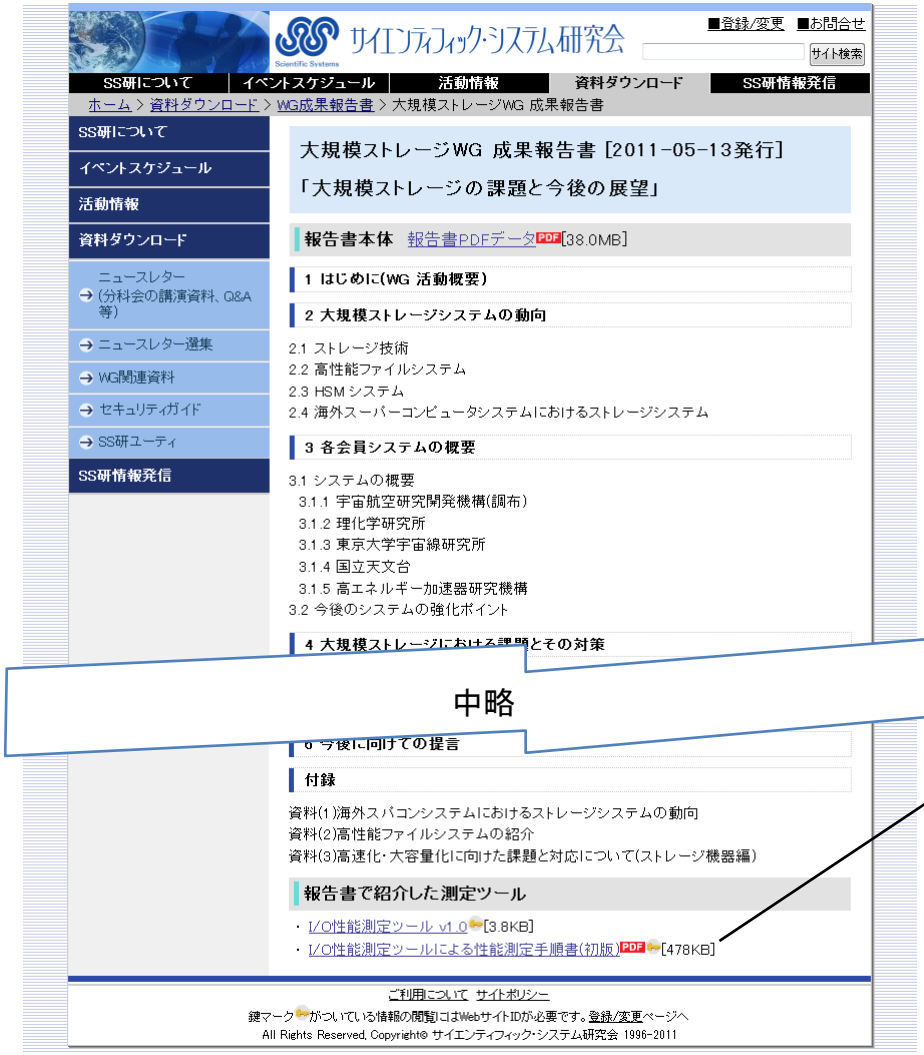
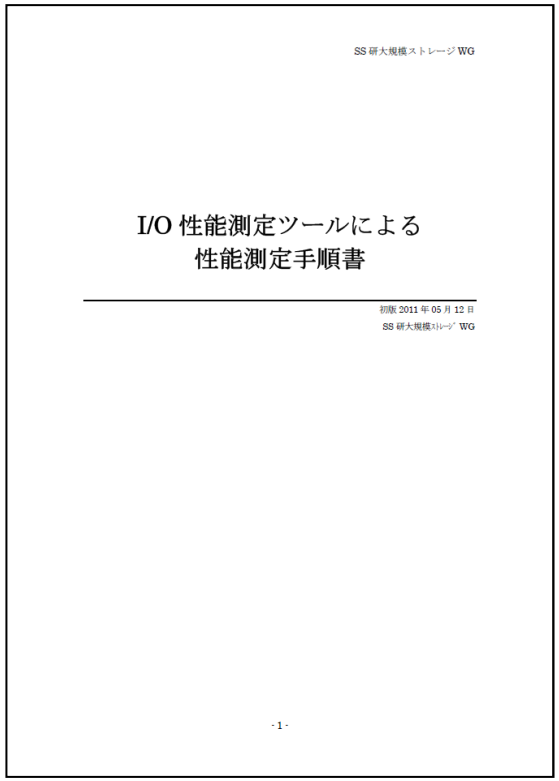


図 5.7.9 理化学研究所
メタデータアクセス(ファイル)

■入手方法、連絡先

http://www.ssken.gr.jp/MAINSITE/download/wg_report/lsstorage/index.html





【お問い合わせ】
本ツールに対するお問い合わせは、以下のメールアドレスにご連絡ください。
office@ssken.gr.jp

おわりに

- 大規模ストレージWGの活動報告を行った
  - ✓ 大規模ストレージシステムの動向
  - ✓ 大規模ストレージシステムの課題と対策
    大容量化、高速化、バックアップ、リビルド、データ移行、高可用性
  - ✓ ファイルシステム健康診断
- 今後のスパコン演算性能向上、観測/測定機器の高精度化・大型化により、大規模ストレージの役割はますます重要になっていくと予想される。
- 今後もストレージシステムの議論がSS研で継続的に行われることを期待している。

最後になりましたが、

WGに参加し活発な意見交換をしていただいた皆様、SS研事務局の皆様に感謝の意を表します。

ご清聴ありがとうございます。