サイエンティフィック・システム研究会 2011 年 10 月 19 日
2011 年度科学技術計算分科会 会合

—SS 研大規模ストレージ WG 成果報告—

大規模ストレージシステムの課題と今後の展望

藤田 直行

宇宙航空研究開発機構

[アブストラクト]

今年３月に活動を終了した大規模ストレージワーキンググループ(WG)の活動報告を行う。スーパーコンピュータの演算性能の向上や、観測機器・測定機器の高精度化・大型化により、近年、入出力データの大規模化が急速に進んだ。この急速な変化の中で、大規模ストレージの抱える課題の存在がぼんやりとではあるが一般に認識され始めている。一方、SS 研では、ストレージシステムの重要性を古くから認識し、精力的に活動を展開してきた。本 WG は、過去 3 回の SS 研でのストレージに関する WG 活動での広範な技術動向と課題の検討を受け、大規模ストレージに特化した技術課題と解決策について検討した。ぼんやりと認識され始めた課題を、古くから問題意識を持つメンバの議論を通して明確化された６つの課題：「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」として整理した。本稿では、大規模ストレージの技術動向（２章）、上述した６つの課題とその対策（３章）、活動報告書[(1)]で提案している”ファイルシステム健康診断”（４章）について報告する。

[キーワード]

大規模ストレージ、HPC、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

1. はじめに

これまで SS 研では、「ネットワーク時代の統合ストレージマネージメント WG」、「ストレージを中心としたシステムマネージメント WG」、「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」を発足させ、その時点でのストレージに関する広範な技術動向と課題を様々な側面から検討してきた。現在 HPC の世界では、データの大規模化や計算機の高性能化が一層促進されてきており、今後のシステム検討において、大規模ストレージがますます重要な構成要素となってきている。この様な状況を踏まえ、大規模ストレージに特化し、大規模ストレージを有する各サイト関係者の参加のもと、大規模ストレージが抱える技術課題と解決策について検討する WG を発足することとなった。

WG では 7 つの視点（①ストレージ製品の現状と動向の調査、②参加機関毎の特徴的なストレージ/IO 例の整理、③ストレージ/IO に対するニーズの明確化、④ストレージ/IO の使われ方や考え方の現状把握、⑤効果的なストレージ/IO システムの検討、⑥ストレージ運用管理機能の検討と製品への反映検討、⑦ストレージ性能測定ツールのモデル化、測定・結果評価および性能評価指針案の策定）で、2009 年 1 月から 2011 年 3 月まで、合計 10 回の会合等を通じて、調査・検討を行った。その結果、大規模ストレージの課題を、3章に示す6つに整理すると共に、それらに対して現時点で考えられる対策をまとめた。

2. 大規模ストレージシステムの動向

大規模ストレージの課題とその対策を述べるに当たり、今後の技術の方向性に関する意識の共有を行うために、大規模ストレージシステムの主要技術の動向について WG で紹介された議論を記しておく。

2.1. ホストインターフェース

ブロックアクセス型のストレージとして現在はファイバチャネルインターフェースが主流であるが、今後のロードマップを考えると中期的には FCoE(Fibre Channel over Ethernet)が有望なインターフェースであると考えられる。ただし、現段階では CNA(Converged Network Adapter)および対応スイッチ間での接続性などに制限が多いが、2012 年頃から第二世代のスイッチや CNA がベンダーから供給され、相互接続性の問題も解消すると考えられている。

表1 ホストインターフェースの動向

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

2.2. ディスクドライブ

高性能を目指すオンラインディスクは現在 3.5″ 15,000rpm が主流であるが将来的(2012 年末頃)には供給停止になるため、今後は 2.5″ 10,000rpm のドライブが主流となる。あわせてドライブインターフェースも FC(4Gbit/s)および SATA(3Gbit/s)中心から SAS 2.0(6Gbit/s)に移行していく。さらに将来は NAND Flash を使用した SSD(Solid State Drive)が容量単価でオンラインディスクに近づくことが予想されており、オンラインディスクにとってかわることも考えられる。

これに対し容量対価格重視のニアラインディスクは容量単価で優位な 3.5″ 7,200rpm が今後も主流となる。ディスクドライブは 1.5～2 年で容量が 1.5～2 倍となるが、性能は 15～30%程度しか向上しない。これに対し HPC システムとして要求される容量・性能はシステム更新の度に 10～100 倍となっており、ドライブの改善だけでは吸収できない。そのため、ディスクドライブだけでなくストレージ装置自体も増加しシステム全体としての消費電力や設置スペース、コストなど新たな問題が起きている。

表2 ディスクドライブの動向

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	
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。

性能・消費電力の観点で期待の大きい SSD は、アクセス形態による性能差が大きい特徴がある。小 IO サイズでのリード処理ではオンラインディスクに比べ 50 倍以上の性能が出るが、大きな IO サイズでの連続したライト処理を 1 多重で実行した場合はオンラインディスクとほぼ同等という性能特性を持っている。また、一般には Flash 素子への書き込み上限があり、この問題への対応も必要である。ETERNUS で採用している SSD にはウェアレベリング技術により Flash 素子全体を平均に使うことにより特定の素子が書き込み上限に達して使えなくなることを防止している。

現在、SSD(Flash ドライブ)はストレージの内部的には、磁気ディスクドライブと同一のインターフェースとフォームファクター(3.5″や 2.5″など共通の外部形状)を採用しているが将来的には高速化のため、SAS などのデ

ィスク用インターフェースではなく、よりオーバーヘッドの少ない PCI-Express などのようなインターフェースを使用した Flash ドライブによるストレージを構成することも検討されている。

2.3. ストレージシステムとファイルシステムの融合(将来技術)

ストレージに関する業界団体である SNIA(Storage Networking Industry Association)では OSD(Object Storage Device)に関する検討が行われ、標準化向けた取組が進められている。

OSD は高速・大容量の共有ファイルシステムとして研究が進められているもので、従来のファイルシステムはサーバ側でデータが配置されているブロック情報も管理されているが、OSD ではサーバ側でのブロック管理は不要となり、ファイル名とオブジェクト ID のみを管理するようになる。一方、ストレージ側はオブジェクト ID とデータおよびアトリビュート情報を管理・制御するようになる。

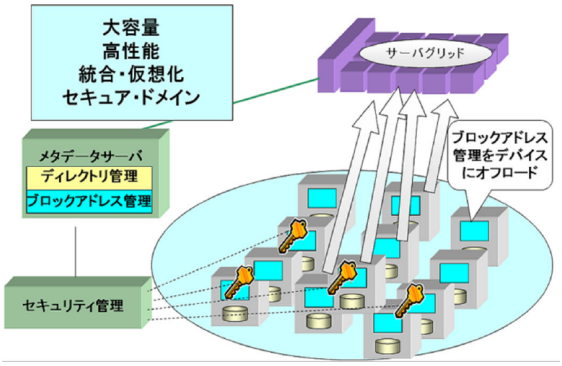

図1 Object Storage Device

OSD には以下のような特徴がある。

- 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、リテンションなど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

OSD の機能は Ceph として Linux のカーネルにも盛り込まれており、今後の開発と普及が期待される。

2.4. 高性能ファイルシステム

表3に各社の提供している高性能ファイルシステムを示す。堅牢で大規模な SMP サーバを用いた高性能ファイルシステム SRFS,HFS、ストレージネットワークをベースにした GFS,クラスタファイルシステムとしては、Lustre ファイルシステムベースの LXFS, CFS のほか、GPFS, CXFS, PanFS が製品として提供されている。

表3 各社の HPC 向けファイルシステム

ベンダー	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 の携帯で利用 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

クラスタファイルシステムは、複数のサーバを束ねて高いファイル IO 性能を実現するファイルシステムである。例えば Lustre ファイルシステムを構成するサーバはディレクトリ構造などを司るメタデータサーバ(MDS)とファイルのデータを扱うオブジェクトストレージサーバ(OSS)に分類される。論理的に MDS, OSS 共にクラスタ構成を組むことができるが、Lustre の現状のバージョン(2.0)では、MDS のクラスタ構成はサポートされていない。こ

のため、Lustre ファイルシステムは大規模なデータ転送には適するが、ディレクトリアクセスや小規模ファイルの大量アクセスなど、MDS に負荷がかかる処理は得意ではない。Lustre のロードマップには MDS のクラスタ構成も提供スケジュールに入っているが、現状ではサポートされていないためファイルシステムを分割するなどの工夫が必要である。一方、GPFS は MDS のクラスタ構成もサポートしているほか、ストレージの動的削除、スナップショットなど Lustre が備えていない機能をもっている。

今後の動向として、HPC システムの巨大化と共にファイル IO 性能に対する要求はますます高まっていくため、Lustre をはじめとするクラスタファイルシステムへの移行が進むと考えられる。しかし、ファイル IO 性能の向上への要求は、すなわち、所要ストレージデバイス(HDD)数と OSS を構成するサーバ数の増大となって現れてくる。このため設置面積の増大、ストレージデバイスとサーバ数増加によるコストの増大とその消費電力を如何に小さくするかが課題になってくる。コストパフォーマンス、省スペース、低消費電力を実現するストレージと高性能ファイルシステムが実現の課題となっている。

2.5. HSM システム

まず議論の明確化のために HSM(Hierarchical Storage Management)を以下のように定義する。

「Disk,Tape など速度・容量の異なるメディア間でデータのステージング・デステージングを透過的に行う」

現在のHSMの利用方法は以下のように3つに大別される。

a)階層ストレージ…大規模ストレージの低コスト化：小容量 Disk、大容量 Tape、頻繁なステージング

b)アーカイバ…データ蓄積：ほとんどが Disk→Tape のみ。Disk に戻す頻度は少ない

c)バックアップ…ファイル/ファイルシステム単位、ほとんどが Disk→Tape のみ。Disk に戻すのは緊急時

現在は昨今の Disk 高性能・大容量化・低価格化により元来の HSM 使用方法である a)よりも b),c)の用途が多くなっている。HSM ソフトは大規模 HPC ユーザでは SAM-FS、HPSS や PetaServe が殆どであるが、ここ数年は HSM 機能としては画期的な進歩は無い。一方ストレージ(NAS,DAS)はインテリジェント化が進み、容量仮想化、ストレージによる自律 snapshot や自動マイグレーションを実現する方向であり、HSM ソフトとストレージの進化は独立して進んでいる。

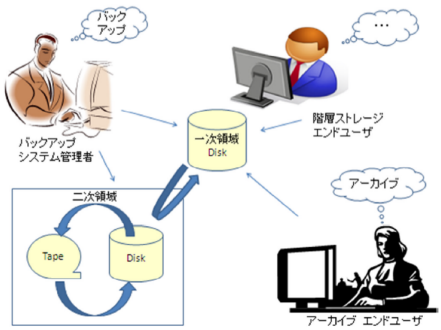

図2 HSM の利用用途

今後の動向として、HSM の利用は、単純バックアップ・アーカイブとハイエンド HSM へ二極化して行き、HSM ユーザ数は減少方向へ向かうと考えらえる。

3. 大規模ストレージシステムの課題と対策

WG では、各参加機関の大規模ストレージシステムの概要、特徴、認識している課題、将来構想を紹介・議論し、大規模ストレージの課題を、「高速化」、「大容量化」、「バックアップ」、「リビルド」、「データ移行」、「高可用性」の６つに整理した。整理の検討過程は参考文献(1)の３章をご覧いただきたい。本章ではこれらの背景と対策についての結論を記載しておく。

3.1. 大容量化

計算ノードの総数、総コア数や搭載されるメモリ量は益々増加する傾向にある。その結果、必要な総 IO スループットも増大し、ファイルシステムの主流が「SAN 共用ファイルシステム」から「ネットワーク経由のファイルシステム」に変わり、ファイルサーバも単一サーバ型からクラスタ化へと変わりつつある(図1)。

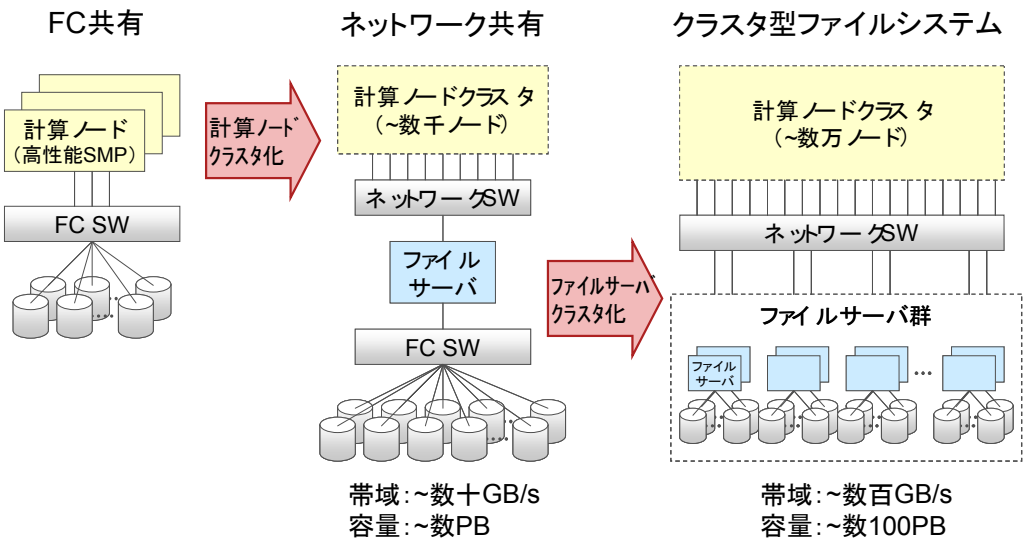

図1　大規模ストレージ向けファイルシステムの方向性

(a)ハードウェアでの対策

一次ストレージの大容量化への対応はディスクの容量と数によって決まる。前述のとおり、ディスクドライブの容量は 1.5～2 年で 1.5～2 倍となっているが、システム更新時の要求容量は性能と同様に 10～100 倍となっていて両者には大幅な乖離がある。従ってディスクドライブの数が大幅に増え、設置面積・消費電力・空調能力・コストといった問題を引き起こしている。

設置面積の改善としては高密度実装による対応が考えられるが、データセンターの床耐荷重・電力供給・空調を合わせて対策する必要がある。消費電力の削減については高効率電源の採用やコントローラ部分の小型化、冷却用ファンの削減や回転数を細かく制御することによる電力削減がある。ただ、ディスクドライブについては、その種類によって消費電力は決まっており、ディスク数に比例して消費電力が増える。将来的には SSD(Flash ドライブ)に移行することにより大幅な消費電力の改善が期待される。現時点ではドライブ単体での消費電力の削減効果は約 50%であるが、容量あたりでは約 150%となりかえって増えてしまうことになる。これは SSD の容量の少なさが要因であり、今後 SSD の大容量化により、容量あたりでも消費電力の改善が期待されている。

バックアップやアーカイブ用途のストレージでは、一次ストレージとは異なり大容量・低価格・省スペース・省電力などが求められる。それらのニーズに対応する新しい技術として Thin Provisioning や De-duplication といった機能が提供されており、従来のテープメディアに代わるものとして注目を集めている。これらの技術はいずれもディスクメディアを使用するが利用効率の向上によりコストを削減しながら、オンラインアクセスを確保することが出来、ユーザ利便性を向上させることを可能にしたものである。また一部のファイルシステムではストレージのエコモード(MAID 技術)を利用できるものもあり、アクセスのない時間帯でディスクドライブの回転を止めることによる消費電力の削減という方法も考えられる。

一方でアクセス頻度の極端に低いデータの格納領域として保管コストの観点でテープメディアを使った HSM などのニーズもある。多くのユーザでは利便性の観点でディスクメディアを使いたいという意向があるものの、コスト観点でテープメディアを選択せざる得ない状況となっている。テープメディアとしては LTO(Linear Tape Open)が実質的な業界標準となっている。現在は第 5 世代が出荷されているが、第 8 世代までロードマップが示されており長期にわたって使用できるものとなっている。

(b)ソフトウェアでの対策

ここでは、今後ソフトウェア面で改善しなければならない課題を列挙する。これら課題の解決は、今後、例えば「京」のストレージシステム開発時に技術の発展を期待するものである。

(1)大規模化

演算速度の高速化が進むと単位時間あたりのデータ量が増加してより高い IO 性能が必要になり、また搭載メモリ量の増大が進むとトータルデータ量が増加してより大容量のストレージが必要になる。数年後のファイルシステムは現状の 100 倍規模の性能・容量が必要になると予想でき、TB/s 級のファイル性能と EB 級のファイルシステム規模やファイルサイズへの対応が必須である。

(2)スケーラビリティ

多数のファイルサーバ群を束ねるクラスタファイルシステムでは、並列化によりスループット性能をスケールアップ、台数増により容量をスケールアップでき、原理的にはサーバ・ストレージ台数に比例して性能・容量をスケールアップすることが可能である。

(3)使い易さ

・動的なハードウェア増設

大規模ファイルシステムでは、運用を止めることなくファイルサーバやディスクを増設できることが必要である。同じファイルシステム内に異なる性能のファイルサーバや異なる容量のディスク装置が混在していると、性能のばらつきや空き容量の偏りが生じる。システムとしての性能と容量のバランスを保つためには、増設元と同じ構成・容量のファイルサーバ・ディスク装置を増設する必要がある。

・運用状態の監視

システム管理者が想定した運用が行えているかどうかを、簡単に監視できる手段が必要である。全体の稼働状況が一目で分かるように見える化できること、その結果を元にチューニングが容易にできること、異常な IO を出しているユーザを特定できることが求められる。

3.2. 高速化

(a)ハードウェアでの対策

ストレージの性能はストレージに使用されるドライブ種とその数によってほぼ決まる。そのため、高速ドライブの採用と分散処理が高速化の基本となる。

ドライブの動向としては2章でも触れたように SSD(Flash ドライブ)が高速化のキーとなることが予想されている。小 IO サイズでのリード処理ではオンラインディスクに比べ 50 倍以上の性能が出るが、大きな IO サイズでの連続したライト処理を 1 多重で実効した場合はオンラインディスクとほぼ同等という性能特性を持っている。ただし、同一 SSD へのアクセスが増えると 4 倍程度の性能となるためより多くの IO 処理が行われる環境では効果が出やすいと考えられる。また、性能が出しにくい数 KB 程度の小サイズの IO については SSD による性能向上の効果はより大きくなる。分散処理についてはハードウェアでのストライピングだけでなくファイルシステムでの対応も解決方法の 1 つとなる。

一次ストレージ装置に求められる性能を満たすためには、最先端のインターフェースやドライブの採用と共にコントローラ部分のスループット性能向上が不可欠である。そのために最新のプロセッサ、メモリ、バスといったコモディティー素材をいち早く取り込むことと、チェックコード生成機能付きのプロトコルチップなどのストレージ装置特有の専用部品については関連ベンダーと並行開発による早期の製品化と安定化を図る取り組みが必要となっている。

(b)ソフトウェアでの対策

ここでも、今後ソフトウェア面で改善しなければならない課題を列挙する。これら課題の解決には今後の技術発展を期待するものである。

(1)性能向上

・メタデータアクセス

Lustre などクラスタ型ファイルシステムでは、多数サーバを並べることで台数に比例して高スループットを実

現できるが、メタデータサーバ(MDS)は 1 台のサーバで管理するため負荷が集中する。このため、大規模システムでは特に MDS の性能向上が重要である。本 WG ではこれらについて、いくつかの検討を行っている。詳細は参考文献(1)の 4.2.2 小節をご覧いただきたい。

・TSS レスポンス保障

共有ファイルシステムは、計算ノード側からのバッチジョブによるファイル IO に加えて、利用者が TSS ノード(ログインノード)からファイルにアクセスする運用が考えられる。バッチジョブからの共有ファイルシステムへのファイルアクセスにより、ログインノードからの TSS レスポンスが低下するという課題がある。

対策として、TSS ノードに対する IO 帯域とジョブ用の IO 帯域を分離し、TSS ノードへの IO 帯域を常時確保しておくことで、利用者に安定したレスポンスで共有ファイルを利用させることができる。

・小サイズ IO・多数ファイル

サイズが小さい多数ファイルへのアクセスでは、IO 時間に占めるファイルサーバへのアクセスコストの比率が高くなり、これがオーバーヘッドとなって IO 性能が低下する課題がある。

対策として、データの実体に加えてメタデータをクライアントでキャッシュすることで、クライアントからサーバへの IO 発生を抑えて性能を向上させることができる。本 WG では、実現例を示し効果の確認を行っている(参考文献(1) 4.2.2)。

・ディスク性能超え(サーバキャッシュ)

ファイルシステムのボトルネックはディスク装置であるが、サーバキャッシュを使うことでディスク性能を超える IO 性能を実現できる。サーバキャッシュにはディスクへの書き込みを保障しない「ライトバック型」と、ディスクへの書き込みまで行う「ライトスルー型」がある。前者の例には SRFS、後者の例には Lustre がある。

(2)運用ポリシー選択

多数の利用者が共同で利用するセンター運用では、1 ユーザによる大量のファイルアクセスにより他の利用者のレスポンスが低下したり、特定ジョブによる大量のファイル IO により他ジョブの実行時間がばらつくといった問題が発生している。このため、センターの運用方針に応じて、特定ユーザの IO が他の利用者に影響を与えないようにするなど、QoS の運用ポリシーを選択できる仕組みが求められている。

(3)輻輳防止(動的フロー制御)

クライアントからファイルサーバの性能を超える IO 要求が発生すると、再送による輻輳が発生してファイルシステムの性能が大幅に低下する。全クライアントからのメタデータ処理を行うメタデータサーバでは特に発生する危険性が大きく、システム全体に与える影響も大きい。このような輻輳を防止するため、動的なフロー制御により常にサーバが最大性能を発揮できる状態を維持することが重要である。

対策としては、サーバ側で受信可能な要求数と、クライアントが送出可能な要求数を相互にカウントし、クライアント・サーバ各ノードがそれぞれ与えられた要求数の範囲内で送受信を行うクレジット方式がある。

3.3. バックアップ

大規模ストレージにおいては、ストレージ容量の検討とあわせてデータの保管をどのように行うかということについて検討する必要がある。本節では大規模ストレージにおけるバックアップの課題および今後の対策について示す。

(1)課題

データの保管方法として、以下の 3 つのパターンが挙げられる。

・物理バックアップ:二次メディアに対する完全なバックアップを取得。物理障害からデータを保全。

・論理バックアップ:ある時点のファイルシステムのスナップショットを取得。論理障害からデータを保全。

・データ移動:保存すべきデータを一次領域から保存領域へ移動。データ保全ではない。

各サイトでは、データ種別(観測データ／解析データ／ソースファイル等)やデータ容量、運用ポリシー等を基にデータの保管方法を決定しており、現状では物理バックアップが一般的に行われている。そのため、今

後のストレージ容量の増大に比例してバックアップ機器の規模(コスト、電力、設置場所)も増加するため、今後の大規模ストレージにおいては、いかにバックアップデータ量を抑えるかが課題となっている。

(2)今後の対策

バックアップデータ量の削減という課題に対して、現状、以下の対策が挙げられる。

・論理バックアップ(スナップショット)

・De-Duplication の導入

・運用(バックアップ世代数/頻度の削減、バックアップ対象の限定、ユーザ自身でのバックアップ)

3.4. リビルド

(1)課題

ディスク障害に対する運用継続性と信頼性確保の手段として RAID 技術が一般的となっているが、ディスク障害に伴う RAID のリビルド処理が長時間化してきている。リビルド処理時間は、ディスクの単体性能、ディスクの単体容量、RAID 種とストライプ量といった要素に依存する。参考例として、以下に富士通製ディスク装置 ETERNUS DX80 でのリビルド処理時間を示す。尚、リビルド処理中にはレスポンスが約 15%程度悪化する。

【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

今後、単体容量が 2TB や 3TB といったディスクが登場してきた場合、他の条件が同じであれば、リビルド処理時間はほぼディスク容量比で増加していく。リビルド時間の長時間化に伴い、リビルド中のレスポンス悪化の長時間化や別ディスクでの障害発生(ダブルフォルト)といったリスクもあわせて増加していく。

(2)今後の対策

リビルド処理においては、ディスク障害発生後にあらかじめ用意されたホットスペアディスクにデータを復元し(リビルド)、障害発生ディスクの交換後、ホットスペアディスクから交換後のディスクへのデータ書き込み(コピーバック)が実施される。リビルド処理時間の短縮化として、上述のコピーバックを行わず、交換したディスクを新規にホットスペアディスクとして認識させることが考えられる。しかし、この場合にはディスク交換に伴いホットスペアディスクのポジションが変わっていくため、保守が困難となるため、コピーバックをはずすことができないのが現実である。なお、主要なストレージベンダーの製品においても同様となっている。

そのため、運用における具体的な対策として、1RAID での性能低下の影響を局所化するようにファイルシステムを構築することが挙げられる。1 ファイルシステムを全 RAID でストライプした場合にはファイルシステム全体に影響が発生するため、1 ファイルシステムを複数の RAID グループで構築することにより、該当 RAID に存在するファイルだけが影響を受けるように局所化することが可能である。

現在のところ、リビルド処理の長時間化に対する根本解決策は見つかっていない状況であり、当面は運用において影響範囲を狭めることが現実解となっている。リビルド時間の長時間化は、実際のストレージ運用において必ず直面する問題であり、早々に根本的な対策が見出されることが望まれる。

3.5. データ移行

データ移行期間の見積もりは、運用を開始するための重要な要素である。各システムのデータの生成状態や仕組みにより、データ移行期間は大きく変化する。

(1)課題

データ移行の形態としては、ディスクからディスクへ、テープからテープへ、テープからディスクへ等様々なパターンが考えられる。

旧・現システムにおいて、ファイルシステムに互換性がある場合、媒体の物理的な移動やリパック等により短期間(数日間)に移行できるが、ファイルシステムに互換性がない場合、旧・現システムをネットワークにて接続

し、NFS+cp/tar、rsync 等によりデータ移行を実施することとなり、この場合、データ移行期間が長期化(数ヶ月間)する。更に、データ検証が必要な場合は、データ転送期間に加え sum コマンド等による転送元と転送先のデータの一致性確認のためデータ移行期間が長期化する。また、移行データ量が少なくてもファイル数が多い場合もデータ移行期間が長期化する。WG 参加機関の現行システムのストレージ容量より、次期システムへのデータ移行期間を算出した。算出条件として、現システムと次期システム間のデータ転送の経路はイーサネット、その実効転送能力は 50MB/s を仮定した。結果は、0.7 ヶ月～6.5 年間の移行期間が必要であることが解り、移行期間短縮化が大きな課題であることが分かった。

⑵今後の対策

データ移行期間短縮のための対策として、ハードウェア、ソフトウェア、運用、次世代テクノロジーの観点で対策案を表4にまとめた。これらの対策を各システムの課題に合わせ適用することにより、移行期間を 0.7 ヶ月～7.3 ヶ月へ短縮できるという見込みを得ることができた。ここで、対策前の最長移行期間を必要とした 6.5 年間のシステムの移行期間が 7.3ヶ月に短縮されたのではないことに注意する必要がある。各システムの短縮効果については、参考文献(1)の 4.3.3 小節をご覧いただきたい。

表4 データ移行期間短縮のための対策

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。
次世代テクノロジー	・OSD(Object Storage Device)を適用する。

3.6. 高可用性

⑴課題

数百台規模のファイルサーバ・ストレージ装置で構成する大規模なファイルシステムでは、常にどこかのハードウェアが故障している状態が予想される。このような状況下でもファイルシステムには運用を止めることなくサービスを継続することが求められる。

⑵今後の対策

・ハードウェアの二重化

クライアントノードからストレージ装置までのデータ経路上にあるハードウェア各所において、単点故障のないハードウェア構成を取ることで、片系が故障した場合に他方に切り替えることで単点故障を迂回する。

・ソフトウェアによる交替・リカバリ制御

ソフトウェアでは、上述した二重化したハードウェアを利用し、ハードウェア故障の検出と運用系と待機系の切り替え、切り替え後のリカバリ処理を制御する。

Lustre では、キャッシュやバッファ上のデータが残っていても故障によるデータ消失を発生させないため、ジャーナリングに加えてプロトコルレベルでのリカバリの仕組みを取り入れている。ファイル IO 時にサーバ・クライアント間のトランザクション処理のログを両者で記録し、サーバ・クライアント間でアトミック性が必要なひとまとまりの処理が完了するまでログを保持する。ファイルサーバやネットワークがダウンして交替先に切り替えた後に、ログに基づいて未完了の処理を再開・リトライすることでクライアントからの IO 処理を正常に完了させる。

4. ファイルシステム健康診断

ファイルシステムの検討を行なう際に、ファイルシステム性能は非常に重要な要素である。しかし、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的なツールは存在せず、システム毎に個別で性能測定ツールを作成して性能評価を行なっている。更に、これらはファイルシステム固有の操作が必要となる場合があり、他ファイルシステムでのファイルシステム性能測定結果と比較することが難しい。そこで、本 WG において測定ツール、測定項目をモデル化し、そのツールを使った測定結果を元に、大規模ストレージシステムの特性を簡易に診断する方法を検討し、ツールの形でまとめた。

4.1. 目的

以下の二点を短時間でかつ汎用的に測定できる事を目的とする。

(1)導入システムの健康診断

システム管理者によるシステム導入時に狙った性能が出ているかを調べる。

(2)運用中の定期健康診断

運用中におけるユーザ視点における性能レンジの把握。運用性能を測定する。

このため診断ツールによる健康診断においては、以下の項目に着目して測定する。

・最大スループットバンド幅(キャッシュヒット性能)

・最小スループットバンド幅(ボトルネック部分の性能。ネットワークやディスク性能など)

・レスポンス性能

4.2. 診断モデル

(1)導入システムの健康診断

モデルは以下とする。

・最大スループットバンド幅性能

Write 直後の Read 性能。データはクライアントキャッシュ上にある事を想定。

・最小スループットバンド幅性能

Write 後に fsync を掛けキャッシュは全て実ストレージデバイスへ転送される事を想定。

・メタデータアクセス

キャッシュの有無は考慮しない。(キャッシュ管理をブラックボックステストでは制御不可能なため)

(2)運用中の定期健康診断

実運用中に本診断ツールを定期的に動作させ、状態把握を可能とする。この場合ユーザ視点の観点から最大性能(キャッシュヒット性能)を採取する事を想定している。しかし運用時の測定においてはシステム構成により過負荷を掛け運用に悪影響を与える事があり得る。

このため測定ツールを定期的に動作させるためにはシステム管理者による十分な事前確認と必要であれば測定ツールのカスタマイズ(測定規模縮小など)を行うこと。

(3)精密検査

ファイルシステム性能を測定する場合、ファイルシステムにはクライアントキャッシュ、サーバキャッシュが存在することが多く、単純に測定をおこなったのでは、キャッシュに対する IO なのか、ディスクに対する IO なのか区別が付かない。またファイルシステムごとにクライアントキャッシュ、サーバキャッシュのクリア方法は異なり、ファイルシステムによってはサービスの停止が必要なものもあり、共通の測定を汎用的に行なうことが困難である。

本診断ツールは、健康診断より詳細なテストが可能なように考慮されているが、キャッシュの影響を排除した上で Read におけるストレージデバイスの性能を測定するために、以下のいずれかの前処理を診断ツール内のスクリプトに各システムで追記して測定する。

・Write 後に一旦クライアント・サーバすべてからファイルシステムを umount する。

・クライアントやサーバのメモリ量の数倍のデータ量を読み込む。

・ファイルシステム固有のキャッシュ廃棄処理を行う。

4.3. 診断項目

前述の通りファイルシステム性能には大きく分けて二つの観点がある。一つはデータ転送の性能であるスループット性能であり、もう一つはファイル操作の性能であるレスポンス性能である。スループット性能に関しては、流体・分子力学計算等で見られるファイルサイズ、IO 長が大きい場合、遺伝子・たんぱく質解析等に見られるファイルサイズ、IO 長が小さい場合の 2 点に着目し、レスポンス性能と合わせて、計 3 点の観点で診断を行う。この 3 点について、IO 長、プロセス数(並列数)をシステム設計時に想定した範囲でパラメタとして測定を行う。

(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)

モデル:プロセス数とデータ量が比例＝解析モデルと計算量が比例

着眼点:多数プロセスからの多数ファイルへの一斉アクセスで性能を有効に使える範囲

期待値:IO 長、プロセス数増加に比例して性能(転送量／単位時間)は増加

性能劣化:IO 長、プロセス数が想定したシステムキャパシティを超える前に劣化

(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)

モデル:1 ジョブで解析対象は一定、並列数のみ増

着眼点:並列化を進めファイルが細分化されていくが、並列化効果が期待できる範囲

期待値:IO 長、並列度に比例して性能(転送量／単位時間)は増加

性能劣化:IO 長、並列数が想定したシステムキャパシティを超える前に劣化

(3)メタデータアクセス(レスポンス性能)

モデル:プロセス数とメタデータアクセス数が比例

着眼点:1 ジョブから数百～数千のファイルに対し stat、touch、rm を頻発させてもシステム性能を引き出せる範囲

4.4. 診断クラス

システムのスループット性能が数十 GB/s にも及ぶ大規模システムと、ワークステーション程度のシステムで同じ測定を実施すると、測定に何日も必要としてしまうことが想定されるため、システムの規模に応じて、プロセス並列数などのパラメータを変更することを検討した。システム規模の想定とモデルのクラス分けは表5の以下の通り。

表5 システム規模とクラス

クラス	規模	論理スループット(オーダー)	最大プロセス数	最大ディスク使用量	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

4.5. 診断ツール

本測定で使用する診断ツールには以下の条件が必要であると考えた。

・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

上記の条件を元に検討した結果、ファイルシステム性能測定(診断)の際に使用するツールとして、スループット性能は IOR、レスポンス性能は mdtest で行なうこととした。MPI は普及している OSS の MPI ソフトウェアである OpenMPI を使用することとした。各ソフトウェアの概略は表6のとおり。

表6 診断ツール概要

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。

4.6. 診断例と入手方法

参考文献(1)5.7 節に4つのシステムの診断例が記載されている。

SS 研公開 Web サイトには、「報告書で紹介した測定ツール」として、性能測定手順書等を記載してある。ぜひ一度お手に取って試していただきたい。

5. おわりに

スーパーコンピュータの演算性能の向上や、観測機器・測定機器の高精度化・大型化により、近年、入出力データの大規模化が急速に進んだ。この急速な変化の中で、大規模ストレージの抱える課題の存在がぼんやりとではあるが一般に認識され始めている。一方、SS 研では、ストレージシステムの重要性を古くから認識し、精力的に活動を展開してきた。本 WG は、過去3回の SS 研でのストレージに関する WG 活動での広範な技術動向と課題の検討を受け、大規模ストレージに特化した技術課題と解決策について検討した。SS 研会員からは古くから大規模なストレージシステムの運用管理を行ってきた日本有数の組織の方々と、賛助会員からは単体デバイスからソフトウェアまで広範な分野にわたる第一線で活躍する方々が検討メンバとして参加した。ぼんやりと認識され始めた課題を、古くから問題意識をもつこれらメンバの議論を通して、少しでも鮮明化することで、世間一般の大規模ストレージの議論を活発化させる一助となる成果を生み出すことが本 WG の大きな目的である。

検討の対象を大規模ストレージに絞ってはいたが、過去の WG の報告からも分かるように、ストレージのニーズや今後必要とされる技術は、組織のミッションやストレージシステムの位置付けによって各々異なるものとなる。従って、本 WG ではひとつの共通解を出すことには拘らず、ユーザ、管理者、メーカ等それぞれの視点からの議論をする中で、ニーズの明確化、技術課題の明確化を行ったつもりである。

WG の活動は報告書の形で完成したが、いくつかやり残したと考えている点もある。WG の議論を通して、あらためて現在のストレージシステムにおいて、「運用」がとても重要な要素であることを再認識し、より深い検討が必要であると感じた。また、「ファイルシステム健康診断」ツールを提案したが、ツールの評価は使われ始めてから行われるものであり、その意味で現時点は出発点に立っていると言える。今後、このような点を含めた、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。

[参考文献]

(1) 大規模ストレージ WG 成果報告書「大規模ストレージの課題と今後の展望」、サイエンティフィック・システム研究会大規模ストレージ WG、2011 年 5 月 13 日