Vol.2019-OS-147 No.5
2019/7/24

長期時系列データの蓄積と分析に適したストレージシステム

大辻 弘貴[1]　吉田 英司[1]

概要：映像やセンサデータの活用技術の進歩により，コンピュータシステムにより取り扱われるデータ量は増加の一途をたどっている．このようなデータは，時系列に長期間蓄積され，必要に応じてストレージ装置を増設することにより容量の拡張が行われる．しかしながら，容量が増えるほど設置面積や故障装置の保守頻度が増加することにより，運用コストも増大してしまう．このような状況に対しては，全領域に対して均一なアクセス性能を提供するのではなく，アクセス特性に合わせて最適化した装置を用いることによりコストを削減することが検討可能である．この点に着目し，保守交換を前提としない上，装置の同時稼働数を制限することにより，通常よりも高密度に実装した装置を用いてストレージシステムを構成する方法を検討した．このような装置はアクセス方法に制約があるため，データのレイアウトによっては要求された性能を満たすことができない可能性がある．そこで，本稿においては，取り扱われるデータのアクセス性能要件を満たしつつ，利用方法に制約のあるストレージ装置を活用するためのデータレイアウト手法について述べる．本提案手法について，シミュレーションによる評価を行い，長期時系列データに求められるアクセス性能を満たしつつ，保守コストの大部分を削減した上であっても従来システムと同等のアクセス性能を提供することが可能であることを示した．

1. はじめに

ストレージシステムは，一様かつ頻繁にアクセスされるデータを扱う通常のシステムと，主に時系列のデータ蓄積を目的としたアクセス頻度に偏りのあるコールドストレージに分類される．前者のケースでは，並列ファイルシステム [1] や高性能なブロックストレージシステムが用いられており，全てのデータに対して常に最大のアクセス性能と信頼性を提供することが目的となっている．一方で，後者のケースにおいては，データは時間を追うごとに増加し，必要に応じて装置の増設が行われる．また，アクセス特性についても，保存されて間もないデータに対してはアクセスが集中する一方で，長期間保管されているデータに対してはアクセス頻度が低い傾向が見られる．時系列データを保存する場合，必然的に後者のアクセス頻度の低いデータが多数を占めることになるため，通常のアクセス性能をストレージ領域全体に対して一様に提供することは無駄が多い [2–5]．加えて，大容量のデータを扱う場合，構成される記憶装置の数も増加するため，故障発生数もあわせて増大してしまう．この結果，通常のストレージ装置やシステムを用いてコールドストレージを構成した場合，過剰な性能を提供することによるオーバーヘッドと，頻発する故障に対応するための保守とコストの双方により，維持が困難になる．現時点においてコールドストレージに特化したサービスとして，内部構造の開示は行われていないが [6] や [7] が挙げられる．いずれも，通常のストレージサービスと比較すると保管コストが安価である一方，データの取り出しコストは高価に設定されており，取り出しまでの待機時間が存在するなど，コールドデータを保存する上での最適化が行われていると推察される．

コールドストレージを実現する上では，電力や放熱の制約を設けて同時稼働する記憶装置の数を制限することや，保守交換を前提としない構造を取ることにより，高密度化や低コスト化を図ることが考えられる．しかしながら，このような装置を用いる場合，全ての装置に対して常に最大性能でアクセス可能な状態を維持することができないため，アクセス方法に制約が生じる．また，保守交換を行わないことを前提とする場合，従来の RAID [8] だけでは故障時のリビルドが困難になるため，継続運用が難しくなる．

本研究は，保守交換を行わない運用を続けると，ある装置内において利用可能な記憶装置数が減少することに着目して最適化を行った．同時に稼働させる記憶装置数に制限のある装置においては，記憶装置の故障が発生すると，残存した記憶装置を基準に計算した見かけ上の稼働率（実効稼働率）が上昇する．この効果を活用することにより，保存されて間もないアクセス頻度の高いデータの配置先を確

[1] 株式会社富士通研究所
ICT システム研究所
データシステムプロジェクト

保することが可能である．

本稿では，以上に示した制約を持つ構成の装置を前提として，故障による実質的な装置稼働率の変動を活用しつつ，時系列に増加を続ける大容量のデータの保存とアクセスを実現するためのデータレイアウト手法を提案する．

2. 関連研究

本稿が対象とするシステムに関連した研究としては [9,10] が挙げられる．[9] では，記憶装置（HDD）を高密度にラック内に実装し，電源制約と冷却制約を異なる軸に設定している．制約を設けることにより通常のストレージ装置と比較して高密度に実装することが可能となっているが，要求性能を満たすためにはデータの書き込みレイアウトやリビルドスケジューリングに工夫が必要であるとしており，試作した装置構成に特化したデータレイアウトや IO スケジューラを提案している．本研究との差異は，[9] が故障した記憶装置の交換やリビルドを前提としており，故障が発生した状態での継続運用は行っていないことと，装置全体がコールドストレージ用途となっている点が挙げられる．本研究では，記憶装置の故障が発生した装置をアクセス密度の高いデータを配置するための場所として積極的に活用しており，この効果によって時系列データに対して発生するアクセス要件を単一装置で満たすことが可能となる．

3. 前提とする装置の構成

3.1 装置構成

本章では，本研究が対象とするストレージシステムの構成について述べる．**図 1** にシステム全体の構成を示す．複数の記憶装置（HDD) を有するノードを最小単位とし，複数のノードがネットワークで接続されている．同ネットワーク上には制御ノードが接続されており，データの配置に関する情報の管理や，ストレージ装置の状況監視，データ再配置の指示を行なう．

それぞれのノードには記憶装置が多数搭載（ N 台) されているが，本研究が前提とする構成では，電源供給能力や冷却能力は全ての記憶装置に必要な分を確保しておらず，以後一つの装置で同時稼働可能な記憶装置の数を P と記し，正常動作する記憶装置に対する同時稼働可能台数の割合 C とする．

3.2 論理構成

本節では，本論文が対象とするシステムの論理的な構成について述べる．ノード内では，複数のディスクを束ねてボリュームが構成される．この様子を**図 2** に示す．ボリュームは故障に備えて冗長化が行われており，RAID [8] 構成を取ることが可能である．ここでは，冗長度が 2 である RAID-6 を前提としてボリュームを取り扱うこととする．

それぞれのボリュームにはアクセス要求に関する 2 つの

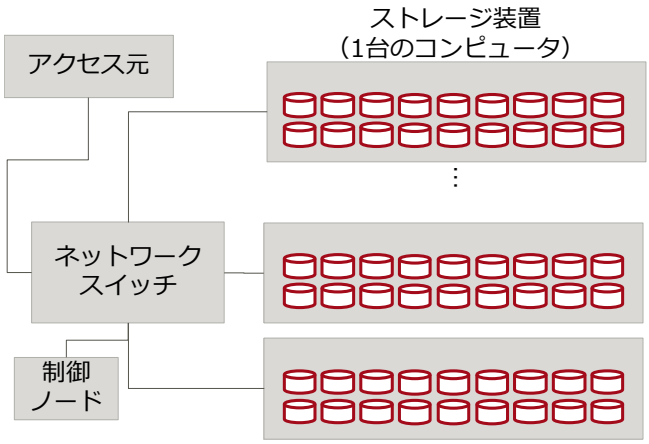

図 1 システム全体の構成を示す．システムはストレージノードと制御ノード，アクセス元から構成され，それらがネットワークで接続されている．ノード内においては複数の記憶装置によりボリュームが構成されており，それらの構成情報は制御ノードによって管理されている．

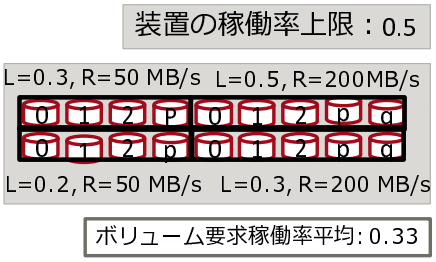

図 2 ボリュームの論理構成を示す．装置内の複数の記憶装置を束ねてボリュームが構成される．それぞれのボリュームには，要求稼働率とスループットが設定されている．

属性が設定されており，それぞれ利用率 L と，最大アクセススループット R がある．利用率 L は，ボリュームに対してアクセス可能な時間の割合を表しており，ある装置内のボリュームの利用率平均は装置設計の上限を超えることは出来ない，最大アクセススループット R は，あるボリュームについて要求される読み書き性能を設定し，ボリュームを構成する際には構成される記憶装置の合計性能が要求を満たすようにしなければならない．

これらの論理構成は，制御ノードによって管理が行われる．**表 1** は，制御ノードが管理するデータの構造を表しており，データの読み書きや故障発生，ボリューム確保の際に変更や参照が行われる．

表 1 管理データの構造

volume_id	ボリューム ID
volume_name	ボリューム名
node_id	ノード ID
disk_list	構成ディスクリスト
req_load	要求稼働率
req_throughput	要求スループット

データアクセスが行われる際の各装置におけるやりとり

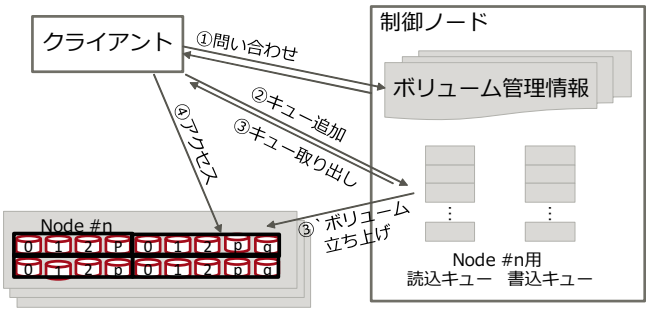

図 3　データに対するアクセス手順

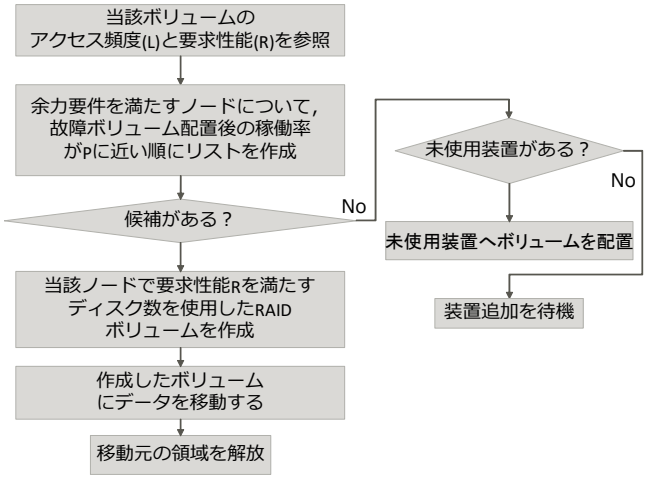

図 4　記憶装置故障時の動作手順

を図 3 に示す．

最初にクライアントからのアクセスが行われる（図中 1）．クライアントはボリューム名またはボリューム ID をもとに，管理データから保管先の問い合わせを行い，データが保管されているノード ID と構成ディスクのリストを得る．書き込みアクセスは他の読み込み処理よりも優先して処理を行なうため，既に稼働率が上限に達していた場合は，読み込みアクセスを中断して書き込み処理を行なう．

読み込み処理については，後に 4.2.2 で示すように，一旦リクエストをキューに蓄積した上で装置の稼働率上限の範囲内で処理を行う．

4. 保守交換を不要とするデータレイアウトとアクセス手順

4.1 データレイアウト

本章では保守交換を前提としない装置の利用方法について述べる．論理構成については 3.2 に示したように，装置内の複数のドライブを用いてボリュームを構成する．このため，ボリューム作成時や記憶装置故障時のデータ移動の際に，3.1 に示した制約を満たした記憶装置の選択を行う必要がある．以降の節では，ボリュームの作成や装置故障時の動作手順について述べる．

4.2 データアクセスの手順

4.2.1 ボリューム作成

3.2 に記したように，ボリュームは利用率（アクセス頻度）L と最大スループット R が設定されているため，これらの要件を満たす記憶装置を選択してボリュームの作成を行なう必要がある．最初に，ボリュームを配置するノードを選択する必要がある．1 章で述べたように，時系列データでは直近に書き込まれたデータほどアクセスされる可能性が高いことから，L が大きな値を持つボリュームが作成される．そのため，本手法では，新規ボリューム配置後の平均稼働率（ボリュームごとの稼働率の平均）が，装置の実質的な稼働率である C と最も近くなり，かつ最大スループット R を満たすだけのドライブ数を有するノードを選択する．C を考慮せずにボリュームの配置を行うと，高頻度にアクセスされるボリュームが健全な記憶装置を多く有するノードに配置され，稼働率の制約のためそれ以上のボリュームを配置できない状態に陥る可能性がある．尚，要件を満たす記憶装置を選定できない場合は，ボリュームの作成は失敗する．

以上の手順に従うことにより，ボリュームのアクセス性能要件と記憶装置稼働率要件を満たしつつ，出来る限り多くの記憶装置を用いてボリュームを作成することが可能となる．

4.2.2 データ読み出し

3.1 で述べたように，装置には同時稼働可能な記憶装置の台数に制約がある．したがって，データの読み出しを行なうためには，記憶装置の立ち上げが必要である．前節でアクセス頻度を考慮したボリューム配置が行われているが，同時にアクセスされる可能性を排除するものではないことから，一旦非同期のアクセスを行わなければならない．そのため，読み出し要求は一度制御ノードで待ち行列に追加され，読み出しアクセスに伴って上昇した記憶装置の稼働率が装置の許容上限を超えないよう待機させる．

4.2.3 ボリューム要件の見直し

1 章で述べたように，長期に渡り保存されたデータはアクセス頻度が低下する傾向にあるため，定期的にボリュームの設定値の見直しを行い，L や R を逓減させる必要がある．これらの値を再設定することにより，ノードの稼働余力が改善する可能性がある．

稼働率が上限に達したことで記憶装置を使い切れないノードが存在する場合，そのノード上に配置されたボリュームを余力があるノードに再配置することにより，記憶装置の利用率を向上させることが可能になる．

この場合にはマイグレーション処理を行い，ボリュームの移動を行なうことが可能となる．ボリュームの移動先のノードは，4.2.1 に示した，ボリューム作成時の手順と同等である．

4.3 記憶装置故障時の動作

記憶装置の故障が発生した場合，3.2 に記したようにボリュームはパリティを持つことから，継続してアクセスすることが可能である．しかしながら，本稿は保守交換を前提としない構成で装置を構成しているため，交換を待ってリビルドを行なうことはできない．そこで，ノード内の残存記憶装置を利用したリビルドか，他ノードへのマイグレーションを行なう．この手順を**図 4** に示す．

最初に，故障した記憶装置が含まれるボリュームを存在しないものとみなし，4.2.1 における新規ボリューム配置先の選定手順と同様の方法で，配置すべきノードを決定する．この結果が現在のノードと同一であれば，当該ノードの残存記憶装置を用いてリビルドと同等の操作を行ない，異なる場合には，そのノードに対してデータのマイグレーションを行なう．これにより，ノード間における稼働余力をバランスしつつ，故障した記憶装置によるデータロスを防ぐことが可能となる．

尚，マイグレーションに伴うアクセス負荷については，事前にシステムに対して許容量を設定することにより，装置の許容上限を超えないように処理を行う．

5. 評価

5.1 評価環境

本稿で提案した手法の特性を，シミュレーションにより評価を行った．評価にあたっては，保守交換を行なう装置と，本稿が対象とする装置の間において記憶装置の利用効率や可用性の比較を行った．

本評価に用いたパラメータや装置構成を**表 2** に示す．

表 2 評価パラメータ

ノードあたりのドライブ数	24
ドライブあたりの容量	10 TB
装置の稼働可能記憶装置の割合	0.5
記憶装置の年間平均故障率	バスタブ曲線 (0.02-0.085)
ノードの運用年数	10 年
ノードあたりの最小記憶装置数	5
マイグレーション負荷率	0.1

5.2 評価シナリオ

評価にあたっては，時系列データの運用ワークロードを想定してシミュレーションを行った．ワークロードは，1 日あたり 5 TB の書き込みを行い，読み出しアクセス頻度は**図 5** に示すように時間の経過とともに減少するものとした．以上の条件のもと，以下に示す運用状況の推移やコスト比較，記憶装置の利用効率や可用性を評価した．

5.3 運用状況の推移

表 2 および 5.2 に示す条件において，本稿が提案する構成の装置を運用した場合の容量や稼働率の推移のシミュレーションを行った．

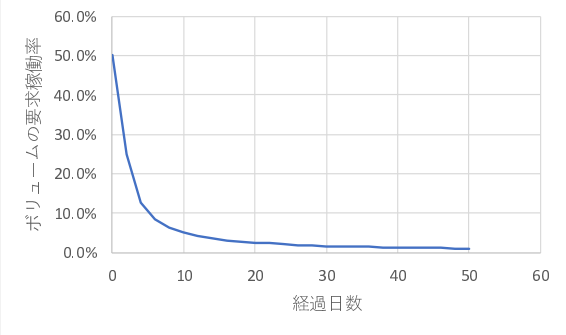

図 5 経過時間に対する要求稼働率の推移設定

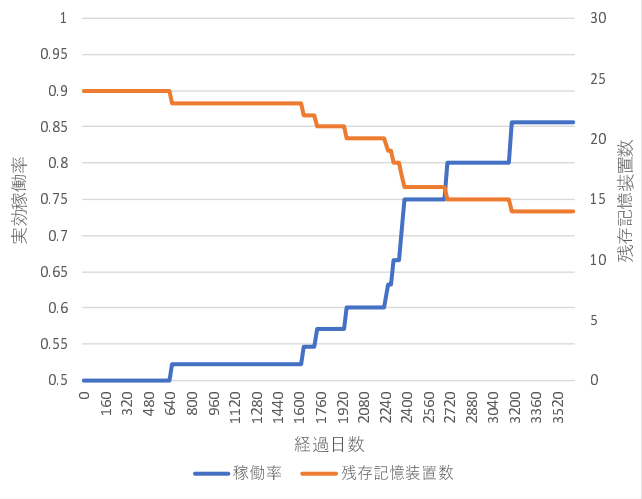

図 6 1 ノードを 10 年間運用した場合の稼働ディスク数と装置の実効稼働率上限の推移を示したグラフ．

図 6 は，1 ノードの装置について，10 年間稼働させた場合の稼働（残存）ディスク数と，実効稼働率上限を示したグラフである．運用期間が長くなるほど故障ドライブが発生することから，装置内の健全な記憶装置数に対する稼働可能記憶装置数の割合は高まり，結果として実質的な稼働上限率が上がる．これらの領域は，負荷率の高いボリュームの配置先として適した状態となる．

一方で，**図 7** は，5 TB/日の新規データを収容することを前提として，18 日ごとに装置追加を行った場合の実質稼働率分布を示している．導入初期のノードほど記憶装置の故障発生数が増えることから，高負荷領域のボリュームの配置先を確保しつつ，システム全体の記憶容量を拡大することが可能であることが分かる．

5.4 ボリュームマイグレーションの影響と保守コスト

記憶装置の故障に伴い，ボリュームのマイグレーションが発生する．本節においては，マイグレーションが通常のデータアクセス（読込）に対して与える影響について述べる．

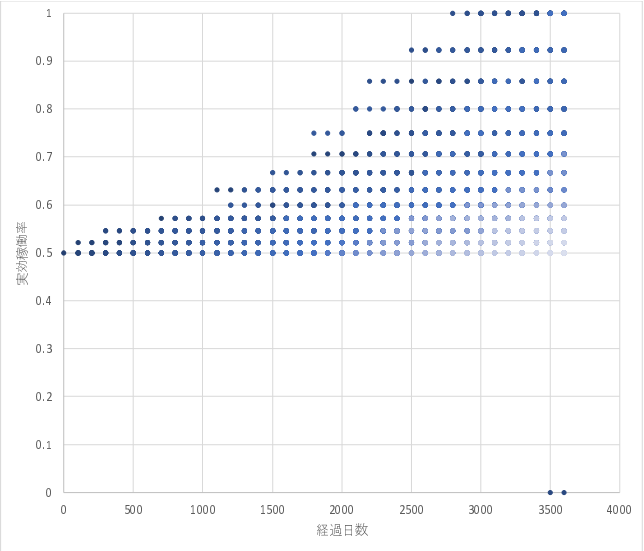

図 7 経過日数ごとのノードの実効稼働率の分布．
図中の点はノードに対応しており，初期に追加されたノードほど濃い色である．縦軸は実効稼働率であり，初期に導入されたノードは時間の経過ともに記憶装置の故障が発生し，実効稼働率が上昇する．このケースにおいては，10 年分のシミュレーションを行っている．

マイグレーションに伴い発生するアクセスについては，装置性能の 1 割を予め割り当てているため，マイグレーションによるトラフィックが事前の割当を超えない限り，アクセス性能に対しては影響を与えることがない．

マイグレーションの発生回数は，運用期間中のドライブ故障の発生数から求めることが可能である．また，保守コストに関しては，本提案手法による運用を行う限りは，記憶装置単体の故障交換が不要となる．

6. まとめと今後の課題

6.1 まとめ

本稿は，時間と共にアクセス頻度が逓減するコールドデータの保管および活用を目的として，ハードウェアの特性を最大限活用するデータレイアウト方式を提案した．

筐体内で同時に稼働する記憶装置の数および保守交換を制限することにより，従来よりも高密度に記憶装置を実装することができる．このような装置は，従来の方法で運用するとアクセス性能の低下やデータロスの可能性があるため，新しいデータレイアウト手法が必要であった．本稿では，コールドデータのアクセス特性を活用した上で，ドライブ故障を契機としたマイグレーションおよびデータレイアウト設定を行なうことにより，制約を持つ装置の性能を最大限に活用する手法を提案した．

本手法をシミュレータにより評価し，保守コストを大幅に削減しつつも，コールドデータに対するアクセス需要を十分に満たすことができることを示した．

6.2 今後の課題

本稿における実装では，ボリュームはノード内に閉じてレイアウトが行われるため，ノード内に内蔵されたコントローラが故障すると可用性が低下する課題がある．この課題は，複数のノードにまたがるボリュームを構成することで解決可能であるが，ネットワークを介した冗長化が必要となるため，モデルの複雑さが増す．より実用性を高めるためには，ノードレベルの冗長度が不可欠であることから，このようなケースにも対応できるよう，シミュレータの拡張を行なう．

参考文献

[1] Braam, P. J.: Lustre, `http://www.lustre.org/`.
[2] Zhang, G., Chiu, L. and Liu, L.: Adaptive Data Migration in Multi-tiered Storage Based Cloud Environment, *2010 IEEE 3rd International Conference on Cloud Computing*, pp. 148–155 (online), DOI: 10.1109/CLOUD.2010.60 (2010).
[3] Eldawy, A., Levandoski, J. and Larson, P.-A.: Trekking Through Siberia: Managing Cold Data in a Memory-optimized Database, *Proc. VLDB Endow.*, Vol. 7, No. 11, pp. 931–942 (online), DOI: 10.14778/2732967.2732968 (2014).
[4] Islam, N. S., Lu, X., Wasi-ur-Rahman, M., Shankar, D. and Panda, D. K.: Triple-H: A Hybrid Approach to Accelerate HDFS on HPC Clusters with Heterogeneous Storage Architecture, *2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, pp. 101–110 (online), DOI: 10.1109/CCGrid.2015.161 (2015).
[5] Xuan, P., Ligon, W. B., Srimani, P. K., Ge, R. and Luo, F.: Accelerating big data analytics on HPC clusters using two-level storage, Vol. 61, pp. 18 – 34 (online), DOI: https://doi.org/10.1016/j.parco.2016.08.001 (2017).
[6] Amazon S3 Glacier: `https://aws.amazon.com/glacier/?nc1=h_ls`.
[7] Google Archival Cloud Storage: `https://cloud.google.com/storage/archival/`.
[8] Patterson, D. A., Gibson, G. and Katz, R. H.: A Case for Redundant Arrays of Inexpensive Disks (RAID), *Proceedings of the 1988 ACM SIGMOD International Conference on Management of Data*, SIGMOD '88, New York, NY, USA, ACM, pp. 109–116 (online), DOI: 10.1145/50202.50214 (1988).
[9] Balakrishnan, S., Black, R., Donnelly, A., England, P., Glass, A., Harper, D., Legtchenko, S., Ogus, A., Peterson, E. and Rowstron, A.: Pelican: A Building Block for Exascale Cold Data Storage, *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*, Broomfield, CO, USENIX Association, pp. 351–365 (2014).
[10] Black, R., Donnelly, A., Harper, D., Ogus, A. and Rowstron, A.: Feeding the Pelican: Using Archival Hard Drives for Cold Storage Racks, *8th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage 16)*, Denver, CO, USENIX Association (2016).