読み出し性能と書き込み性能を選択可能なクラウドストレージ

中村　俊介†　　首藤　一幸†

† 東京工業大学情報理工学研究科数理・計算科学専攻　〒 152-8552 東京都目黒区大岡山 2-12-1
E-mail: †{nakamur6,shudo}@is.titech.ac.jp

あらまし　読み出し性能重視となるか書き込み性能重視となるかは, 分散データストアにおいてもストレージエンジンが決めると予測した. もしそうであれば、読み出し性能と書き込み性能を, データストア自体の使い分けではなく, ストレージエンジンを差し替えることで調整ができる. そこで実証のために, ストレージエンジンの差し替えが可能な分散データストア MyCassandra を開発し, クラスタ上で元の Cassandra とストレージエンジンを MySQL とした場合を比較した結果, 書き込み比率の高いワークロードでは前者の書き込み遅延が後者より 41.4%小さく, 読み出し比率の高いワークロードでは後者の読み出し遅延が前者より 49.4%小さくなることを確認した.

キーワード　分散データベース, クラウドストレージ, 性能トレードオフ

A Cloud Storage Adaptable to Read-Intensive and Write-Intensive Workload

Shunsuke NAKAMURA† and Kazuyuki SHUDO†

† Dept. of Mathematical and Computing Sciences, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro, Tokyo, 152–8552 Japan
E-mail: †{nakamur6,shudo}@is.titech.ac.jp

Abstract　We expect that a storage engine determines whether a cloud storage is read-optimized or write-optimized. It means that a single cloud storage can be both of them just by replacing its storage engine with another one. It is not necessary to use another cloud storage to adjust a balance of read and write performance. The expectation was confirmed by performance comparison of Bigtable-based storage engine of the original Cassandra and MySQL storage engine. Write latency of the former is 41.4% lower than the latter with a write-heavy workload, and read latency of the latter is 49.4% lower than the former with a read-heavy workload.

Key words　distributed database, cloud storage, performance trade-off

1. はじめに

RDBMS が満たしきれない要求に応える分散データストアとして NoSQL や Key-Value Store といったクラウドストレージが注目されている. これらは従来の RDBMS と比べてデータモデルやクライアント側の機能を制限し, また一貫性についての条件を緩めることで, 負荷の分散を比較的容易にし, ノード数をスケールアウトさせやすいという点が共通の特徴である.

しかし, それぞれのクラウドストレージは様々な点で異なっている. 例えば, データモデルには key-value 形式や多次元マップ (multi dimentional map) 形式があり, 分散の構成は master/worker 型やツリー型や非集中分散型があり, データを全てメモリに保持することで遅延を抑えたものや永続化が可能なもの, 複製の配置方法やそれを同期/非同期で行うかなどというように様々な設計方針がある. また, 従来のクラウドストレージには書き込み性能重視のものと, 読み出し性能重視のものがある. 利用者がこのように多くのクラウドストレージの中から利用用途に忠実に沿うものを選択するのは困難である. そこで我々は既存クラウドストレージを「分散のための機構」と「ストレージエンジン」に分離し, 後者の「ストレージエンジン」を同一データストア内で差し替えることで, 全く異なる読み出し/書き込み性能の性質を得られるということを提案する. 我々はこの提案を示すために Apache Cassandra をベースに, MyCassandra というモジュラーな分散データストアを開発した.

表 1　既存クラウドストレージの特徴

	Cassandra, HBase	Sherpa, sharded MySQL
書き込み	Diff Sequential	Key lookup Update
読み出し	Diff Merge	Single Read
性能	書き込み性能重視	読み出し性能重視
ストレージエンジン	Bigtable 由来	MySQL

2. 提案手法・設計

クラウドストレージを用いる大規模なシステムでは, 次にどのデータが参照・更新されるかを予測することは難しく, 全てのデータがメモリ上に収まらない限り, 読み出し時にディスクへのランダム I/O が発生してしまう. 書き込み時にランダム I/O を行うよりも, ディスクへのシーケンシャル I/O のみでログを記録する方が, 高いスループットを実現できるが, 一方このようなログ記録方式では, 読み出し時はログから一連の更新結果を拾い集めなければならないので効率が悪い. つまり, 読み出し性能と書き込み性能は, 常にトレードオフの関係がある.

このような理由から, 既存の永続型クラウドストレージはあらゆるワークロードで優れた性能を提供しているのではなく, その設計は書き込み性能・読み出し性能どちらか一方に偏る. 幾つかのクラウドストレージの性質を表 1 に示す. Apache Cassandra [1] や Apache HBase [2] は書き込み性能を重視した設計がされているために, 書き込み比率の高いワークロードに向いており, 一方, Sherpa や sharded MySQL(MySQL の sharding 構成) は読み出し性能を重視した設計がされているために, 読み出し比率の高いワークロードに向いている.

我々はこの読み出し/書き込み性能の性質は主にストレージエンジンの違いによるものであると予測した. なぜならば, 従来のデータストアのボトルネックになりがちなのは, 各ノードでのディスク入出力部分であり, それは分散のための機構に依存せず分散データストアにも当てはまると考えられる. 実際に Cassandra や HBase は Bigtable のストレージエンジン (2.2.2 節参照) を採用することでディスクへの書き込みを差分のみシーケンシャルに書くことでランダム I/O が発生せず, 書き込みを高速に行うことができるが, 読み出し時には差分のマージ処理が必要となり, 複数回のランダム I/O が発生するため読み出し性能が犠牲になる. 一方, MySQL や Sherpa は従来のバッファプール方式により読み出しは 1 回の I/O で最新のレコードを引き出すことができるが, 書き込みには古いレコードの読み出しが必要なため, ランダム I/O が発生するため書き込み性能が犠牲になる.

ところで RDBMS の 1 つである MySQL の設計はコネクションやデータの分散アルゴリズムとストレージエンジンは独立したコンポーネントとして構成しているため, 利用者は欲しい性能や機能に応じたストレージエンジンを選択することでデータモデルやノードの分散構成などを変更することなく性能の調整を行うことができる. しかし, 既存のクラウドストレージはそのような設計はなされておらず, 利用者は足りない機能を他のソフトウェアと併用したり, もしくは自身で新しいクラウドストレージを実装しているのが現状である.

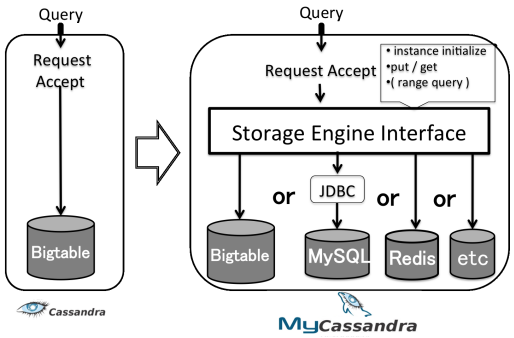

図 1　MyCassandra の Storage Engine Interface

そこで本研究では Apache Cassandra をベースに, 同じシステム内で MySQL をはじめ, 様々なストレージエンジンから選択が行えるクラウドストレージ MyCassandra を開発した.

以下ではまず MyCassandra のアーキテクチャを説明した後に, Apache Cassandra のアーキテクチャと, 評価の為に MyCassandra に追加した各ストレージエンジンについて述べる.

2.1 MyCassandra

MyCassandra は, Cassandra をベースとし, 「分散のための機構」と「ストレージエンジン」に分離した分散データストアである. これにより, 後者のストレージエンジンを他の部分に影響を与えることなく, 差し替えることができる.

図 1 は Cassandra と MyCassandra での各ノード上での読み書きに関係する部分を示している. MyCassandra には Cassandra のリクエスト受理部分と各ストレージエンジンの間に Storage Engine Interface を設けた. このインタフェースが規定する以下の関数を実装することで, MyCassandra に新たなストレージエンジンを追加できる.

- 初期化 (インスタンス生成, ネットワーク接続)
- データの put 関数と get 関数

Cassandra はスキーマレスの多次元モデルをデータモデルとして採用しており, これはそのままでは RDBMS や key-value store には格納できない. そこで, データの形式を変換する必要がある. MySQL の場合はスキーマレスなデータモデルを構成するために key とその行の複数カラムをシリアライズした value という key-value のペアでストアする方法をとっている. 一方, Redis は同じく key-value のペア形式であるが, 複数の表を管理するモデルになっていないために key に ColumnFamily 名を prefix として付加している.

2.2 Apache Cassandra

Apache Cassandra は, Facebook 社が開発し, Apache Project としてオープンソース化したクラウドストレージである. 複数のデータセンターにまたがる数百台のノードで運用可能なスケーラビリティや, 非集中で単一故障点を持たないことによる高い可用性などを特徴とする.

2.2.1 Consistent Hashing

Cassandra はデータの分散アルゴリズムとして Amazon Dynamo [3] を参考にした Consistent Hashing という非集中な分

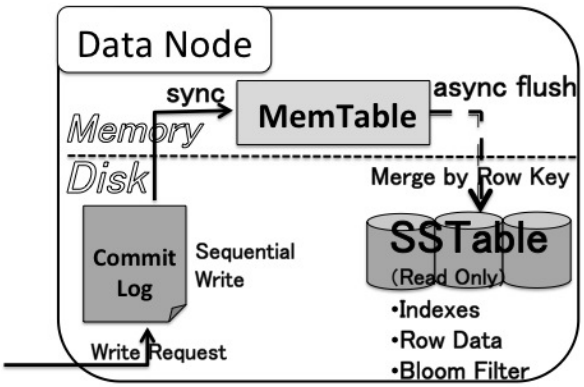

図 2　Cassandra の書き込みの流れ

表 2　YCSB ワークロード

Workload	Read	Update	Record Selection	App. Example
Update-Only	0%	100%	Zipfian	Log
Update-Heavy	50%	50%	Zipfian	Session Store
Read-Heavy	95%	5%	Zipfian	Photo tagging
Read-Only	100%	0%	Zipfian	Cache

表 3　実験パラメータ

データノード	実マシン× 6 台
クライアント	実マシン× 1 台
レコード	2,400 万件
単一レコード	10 カラム計 1KB

表 4　実 験 環 境

OS	2.6.35.6-48.fc14.x86_64
CPU	2.40 GHz Xeon E5620 × 2
Mem	32GB RAM
Disk	1TB SATA HDD × 2
JVM	Java SE 6 Update 21
MYSQL	6.0.10-alpha

散アルゴリズムを用いている.

Consistent Hashing の主な特徴は ハッシュ関数を用いることで各ノードが担当するデータ数が比較的均等になり易く, 負荷分散が容易であることと, データの担当範囲を集中管理するサーバーが必要ないので, 単一故障性の無い高い可用性を提供できることである. 各ノードは他ノードの位置を定期的に Gossip Protocol により情報交換している為, 任意ノードがプロキシとしてクライアントのリクエストに応じることができる.

データのレプリケーションは基本的に, リング上 ID 空間においてプライマリノードの右隣 N-1 個のノードに配置されるが, 異なるラックやデータセンター上に配置する設定も可能である.

2.2.2　Bigtable のストレージエンジン

Cassandra のストレージエンジンは Bigtable と同じアーキテクチャを採用しており, それは Commit Log, MemTable, SSTable の 3 つのコンポーネントから構成される. Cassandra の書き込み処理の流れを図 2 に示す. 書き込みはまず永続用にディスク上の Commit Log に差分をシーケンシャルに書き, 次に読み出しのパフォーマンスの為にメモリ上の MemTable に対してデータの更新を行い, クライアント側に書き込み成功のリプライを返す. MemTable のデータサイズが閾値を超えると, 古いデータから順に非同期でディスク上の SSTable に単位サイズごとにキーでソートされたファイルとして書き出される.

この方式の利点は, ディスクへの書き込みが常にシーケンシャルであることと, 一度ディスクに書かれた内容は変更されることが無いため, 書き込みロックが不要で常に書き込みが可能という点である. 一方, 欠点としては読み出し時に, 指定されたキーを持つ差分データを複数の SSTable から読み出してマージする処理が必要となる為に読み出し性能が犠牲になることである.

2.3　MySQL / InnoDB

今回, ストレージエンジンの 1 つとして MySQL 6.0 を MyCassandra に組み込んだ. MySQL へのアクセスには JDBC API を使用し, MySQL 内のストレージエンジンとしてはデフォルトである InnoDB を用いた.

2.4　Redis

Redis [4] は Key と Value のペアをメモリ上に保持する key-value store である.Redis の大きな特徴はメモリ上のデータをプライマリとしつつ, 非同期で定期的にディスクへ書き出すことができる点であり, 特定時点での永続性が保証される. データをオンメモリで扱うため, 読み出し・書き込み両方共に高速に行えることが利点ではあるが, 実メモリに乗りきらないデータ量は扱えないという問題があり, 利用できる用途が限定的である.

3.　ストレージエンジンの性能比較

MyCassandra で使うことのできるストレージエンジンのうち, 元の Cassandra が用いる Bigtable 由来のストレージエンジン, MySQL, Redis という 3 種の性能を比較する. 性能測定には Yahoo!'s Cloud Serving Benchmark(YCSB) [5] を用いる.

3.1　YCSB

YCSB は, 様々なクラウドストレージを公平に評価することを目的として Yahoo! Research が実装したオープンソースのベンチマークフレームワークである. 実アプリに近いコアワークロードが用意されている.

YCSB では, 読み出し処理と書き込み処理の回数の比率をユーザが指定できる. YCSB がデータストアに対して読み出しと書き込みを実行し, ワークロード全体のスループットと, 各処理に要した時間, つまり遅延を集計する.

表 2 に, 今回の測定で用いる 4 種類のワークロードを示す. 書き込み比率が高い Update-Only と Update-Heavy, 読み出し比率が高い Read-Only, Read-Only を用意した. 各ワークロードに対応する実アプリの例を右側に示している. 各ワークロードにおいて, アクセス対象データの分布として Zipfian 分布を用いる.Zipfian 分布とは, データ鮮度とは関係なく人気によってアクセス頻度が決まるようなアプリのデータアクセス分布を確率としてモデル化したものであり, 一部のデータがデータが常にヘッドであり, 大部分がテールとなるような分布である.

表 3 に実験パラメータを, 表 4 に実験環境を示す.

3.2　評価と考察

図 3 はクライアント数を調整して 5,000 回/秒のクエリを発行した際の読み出しと書き込みの遅延をワークロードごとに示している. Bigtable ストレージエンジンと MySQL の結果を比較すると, 書き込み遅延は Bigtable の方が小さく, 最大でも MySQL の 41.4%であり, 読み出し遅延は MySQL の方が小さ

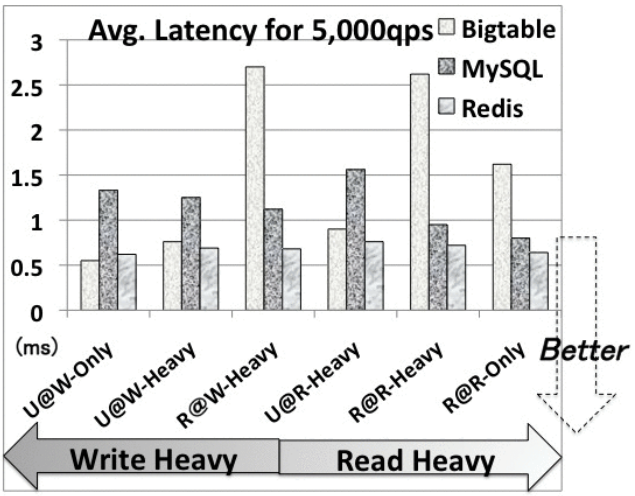

図 3　ワークロードごとの遅延

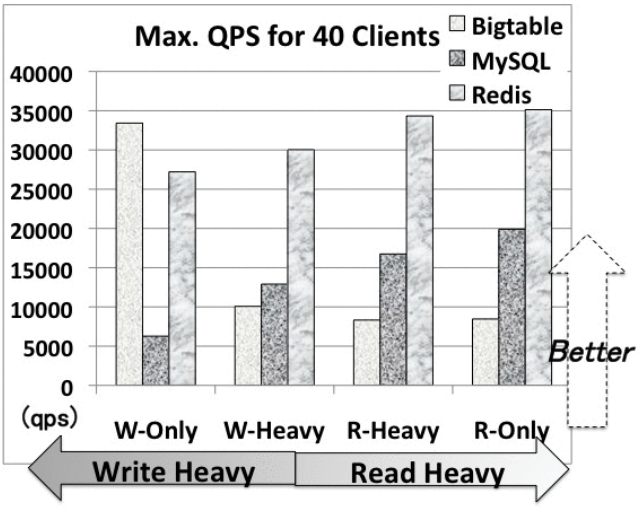

図 4　ワークロードごとのスループット

く最大でも Bigtable の 49.4%であるという結果が得られた.

図 4 はクライアント数を 40 として負荷をかけたときのスループットをワークロードごとに示している. 書き込みが多いワークロードでは, Bigtable の方が高く MySQL の最大 5.32 倍であり, 読み出しが多いワークロードで MySQL の方が高く Bigtable の 2.35 倍という結果が得られた. また, 同一ストレージエンジンについて各ワークロードでの結果を比較すると, Bigtable では書き込み比率が高いほどスループットが高く, MySQL では読み出し比率が高いほどスループットが高くなっていることが確認できる.

この通り, データストア自体を置き換えずともストレージエンジンを差し替えることで読み出しと書き込みの性能の傾向が異なるデータストアを得ることができた.

Redis は全データをオンメモリで扱うがゆえに, いずれのワークロードでも読み出しと書き込み両方の操作において高速に行うことができる.

4. 関 連 研 究

Anvil [6] は, 粒度の細かいコンポーネント dTable を組み合わせて構成されるデータストアである. アプリケーションのデータアクセスパターンに応じたデータストアを構成できる. データストアをモジュラーに構成しようという点が本研究と共通しているが, 分散データストアを対象としているわけではない.

Cloudy [7] は, クラウドストレージをモジュラーに構成しようという提案である. ストレージエンジン以外にも, ルーティング処理やロードバランスの方式もコンポーネント化し, 選択可能とすることを提案している. 性能は報告されていない.

Amazon Dynamo [3] は, Amazon 社がウェブ上で提供するサービス向けに開発した key-value store である. ストレージエンジンとして Berkeley DB や MySQL, メモリ上のバッファを用いることができる. ストレージエンジンを選択する際の指標の例として, 格納データのサイズが挙げられている. Dynamo で用いることのできるストレージエンジンでは, 読み出し性能と書き込み性能を調整することはできない.

本研究は読み出しと書き込みという対比について分散環境での定量的な評価を行えている.

5. まとめと今後

本研究では, 同一システム内のストレージエンジンを差し替えることで読み出し/書き込み性能を調整できるようなクラウドストレージの提案と実装を行い, ベンチマークにより実際にストレージエンジンごとに大きく性能が異なることを確認した.

5.1 SSD 上での評価

今回の実験環境は HDD であったが, 今後は SSD 上でも評価を行う. SSD は HDD と比較するとランダム I/O の性能が優れているため, 各ストレージエンジンが苦手とする処理の性能が向上し, HDD との結果と比べると Bigtable と MySQL の性能差は縮まるものと予測される.

5.2 読み出し性能と書き込み性能の両立

また次のステップとして, ストレージエンジンの異なるノードを組み合わせて用いることで, 単一のクラスタで, 読み出し性能と書き込み性能を両立させることを狙う. 現在, 性能評価を進めている.

具体的には, 図 5 の通り, 読み書き性能についての性質が異なるノード群に複製を配置する. そして, 書き込みは Bigtable と Redis に対して同期的に行い, 読み出しには MySQL と Redis に対して同期的に行うというように, それぞれのストレージエンジンが得意とする処理を同期的に行い, 得意でない処理についてはリクエストへの応答とは独立して非同期で行うよう, プロキシ側でリクエストの振り分けを行う. 非同期の読み出しはクライアントにデータを返すためではなく, 複製の間で整合性をとるために行われるので, 遅延が大きくても読み出し性能には影響しない. これらを行うためには, 各ノードが全ノードのストレージエンジンの種類を知っておく必要があるので, Cassandra のノード生存確認に利用される Gossip Protocol にストレージエンジンに関するメタ情報を加える.

このとき, レプリカ間の一貫性が課題となる. 例えば, 書き込んだデータをすぐ読み出す場合, 非同期で書き込みが行われたノードから読み出しを行うと古いデータが得られてしまう可能

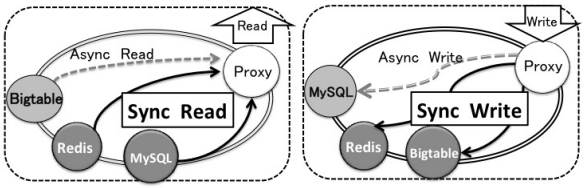

図 5　MyCassandra Cluster

性がある. しかし, Cassandra は, 直前に書き込まれたデータを読み出せる, ということを Quorum（多数決）で保証できる.

MyCassandra には, その他に次の課題がある.

- ネットワーク近接性 (proximity) とストレージエンジンに応じた書き込み先ノード選択の両立
- 負荷分散

ネットワーク近接性, つまりラックやデータセンターを考慮しつつも, ノードごとに読み出し性能と書き込み性能のどちらに優れるのかも考慮して, 複製を配置する必要がある. また, ノードごとに読み書きのどちらが得意なのかが異なるため, 同期書き込みを処理するノード, 同期読み出しを処理するノード, 両方を処理するノード, と分かれ, 負荷分散に問題が生じかねない. この課題に対しては, 単一のサーバに複数のノードを動作させることで対処できる. その際, 同一サーバに複数の複製が作られないように配慮する必要がある.

文　　献

[1] Avinash Lakshman and Prashant Malik. Cassandra - a decentralized structured storage system. In *Proc. LADIS '09*, 2009.

[2] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E. Gruber. Bigtable: A distributed storage system for structured data. *In Proc. OSDI '06*, Vol. 7, pp. 205–218, 2006.

[3] Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall, and Werner Vogels. Dynamo: Amazon's highly available key-value store. In *Proc. SOSP '07*, 2007.

[4] Redis. Redis. http://code.google.com/p/redis/, March 2010.

[5] B. F. Cooper, Adam Silberstein, Erwin Tam, Raghu Ramakrishnan, and Russell Sears. Benchmarking cloud serving systems with ycsb. In *Proc. SOCC '10*, 2010.

[6] Mike Mammarella, Shant Hovsepian, and Eddie Kohler. Modular data storage with anvil. In *Proc. SOSP '09*, 2009.

[7] Donald Kossmann, Tim Kraska, Simon Loesing, Stephan Merkli, Raman Mittal, and Flavio Pfaffhauser. Cloudy: A modular cloud storage system. In *Proc. VLDB '10*, 2010.