非集中型クラウドストレージのスケーラビリティ評価

奥寺 昇平†　 中村 俊介‡ 首藤 一幸‡

東京工業大学†　　東京工業大学‡ 東京工業大学‡

■1.はじめに

Amazon Dynamo、Cassandraをはじめとした単一故障点がなく、負荷が自動的に分散される非集中型のクラウドストレージが普及しつつある。特にこれらのストレージに共通する傾向は、強い一貫性が求められない、つまりイベンチュアル・コンシステンシーな一貫性を保証していることである。

このような非集中な分散ストレージでは、各ノードが適切なノードにホップするために各自が持つ経路表情報を参照することが多い。特に１ホップで対象ノードに転送するシステムでは、最新のシステム全体のノード情報を如何に最新の状態に反映するかが不可欠になる。そこで、gossipをベースとした信頼性の高く、スケーラブルな情報分散・共有を行うプロトコルが開発され、効率よく通信を行うことが可能になった。

しかし、このような情報分散・共有を行う処理がシステム全体に、データの読み書きなどのフロントエンドのプロセスにどれくらいの負荷をかけているのかといったことは明らかではない。また、当然このようなバックエンドの処理は、フロントエンドに負荷をかけるべきでない。

そこで、システム全体の負荷を通信量という切り口で計測していく。本研究では、特にgossipベースのプロトコルを対象にノード台数に応じてシステム全体の通信量がどのようにスケールしていくかを考察する。

* 2.gossipベースの情報共有Protocolについて

一般的なgossipは以下のような手順で行われる。

0. ノードPのデータXが更新されたとき、自らが更新されたことを伝える。

1. ノードPがランダムに他のノードQを選択して更新情報をQに反映させる。

The Evaluation of the scalability in the Non-Decentrized type Cloud Storage

†Shohei okudera: Tokyo Instriute of Technology Information Science

2.ノードQがすでに更新済みであったときは、ノードPは他のノードに更新情報を伝えるのはやめる。

Gossipに代表されるエピデミックなアルゴリズムを利用したProtocolの長所は、プロセス間で同期する数が他の伝搬手法と比べて比較的少ないために、スケールしやすいところにある。

しかし、gossipのために使われる通信量がどれくらいになっているのか、ノードが増えたときに通信量がどのようにスケールしていくのかなどは明らかではない。そこで我々はgossipベースのProtocolの通信量を測定する実験を行った。

■3.実験

実験には、米Facebook社が開発したCassandraを利用した。Cassandraは、大量の構造化データを扱うことができる分散ストレージシステムである。

Cassandraは、オーバーレイネットワーク上に連想配列 (key-value) を構築する技術であるDHTを用いて、テータを複数ノードに分散して配置する。各CassandraノードとデータにはToken(ID番号)が付けられており、そのIDによって、どのノードにデータが保存されるかが決定される。また、各Cassandraノードは機能的に対等かつ独立で、自身が持つ経路表だけを参照して、保管されるノードを決定する。経路表情報はGossiipプロトコルで交換されることで、最新のノード全体の状況と反映できるようにしている。

次に、CassandraにおけるGossip-baseの情報共有Protocolについて説明する。Cassandraには、Seedと言われるノードが各ノードに独立で決められていて、新規にシステムに参加する際に最初にコンタクトする。Cassandraでは、自らのノードからみて、生存しているノードの数や、生存していないノードの数、Seedの数に応じて、1-３回のgossipが行われる。

・実験では、複数のCassandraノードを一台のマシン上で起動させ、マシン間で発生するトラフィックを解析した。

起動するスクリプトは、シェルスクリプトによって作成し、トラフィック計測はtcpdumpコマンドを利用して、tcpパケットのサイズを計測した。

以下の性能のマシンを10台使用した。

・Cassandra 0.6

・OS:Fedora release 14 (Laughlin)

・JRE: build 1.6.0\_21

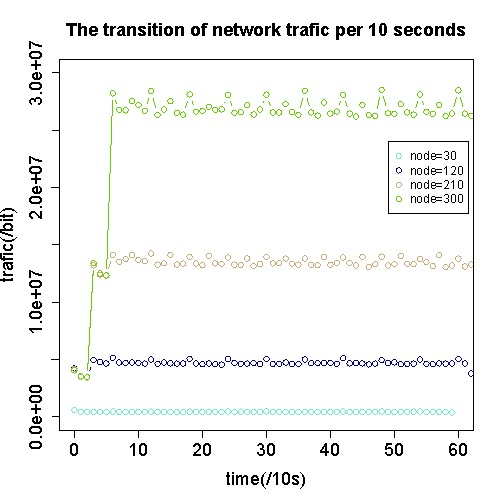
・CPU: Intel(R) Xeon(R) CPU \* 16

・メモリ:33.00G

実験シナリオは説明する。指定した数のCassandraノードを１マシンあたり10台を30秒ずつずらして起動し、10分間トラフィックを計測した。ここで指すトラフィックとは、Cassandraノードがgossipの為に利用するTCPパケットの長さのことである。

■4.評価

10秒あたりの通信量の変化を表したグラフが下図である。(X軸の単位は/10s、Y軸はbit)



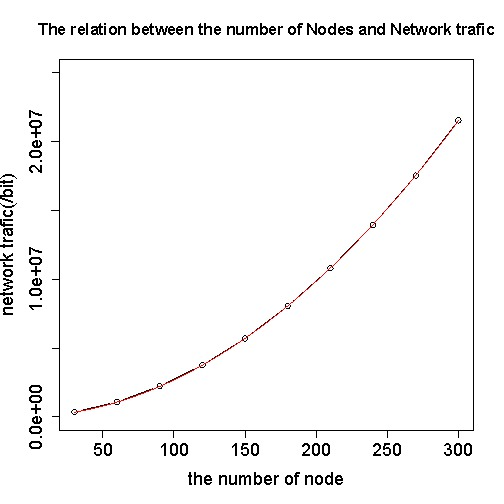
一定時間経つと、通信量は安定することがわかる。

次にノード数と通信量が安定した時の1秒あたりの通信量の平均をプロットしたのがグラフが左上図である。(X軸の単位はノード、Y軸はbit)

図中の曲線は、プロットした点から、二次関数にフィッテイングしたものである。得られた関数は、

y = 224.6 \* x^2 + 4314.8 \* x

である。



この関数から、ノード台数をパラメータとしてCassandraもといgossipで発生しうる全体の通信量を推測することができる。例えば、x = 1000 のとき、全体の通信量y = 229Mbpsとなる。逆に、一般的なスイッチの転送速度を1Gとすれば、そのスイッチの限界値を超えるには、x = 2100台の時であると推測できる。

この実験とその評価により、二つのことがわかった。一つ目は、gossip-baseの分散共有プロトコルに必要な通信量の値が把握できた

。二つ目は、ノード台数に対してgossipプロトコルの通信量が２次関数にしたがって変化することが確認できた。

* 5.Future Work

フューチャーワークとしては、二つを考えている。

一つ目は、gossip以外の分散協調プロトコルの通信量を調べることである。Gossip以外のプロトコルの通信量を知ることで、プロトコルの比較、良し悪しを通信量という観点から語ることができるようになる。二つ目としては、チューン状態の時の通信量を計測することである。チューン状態とは、クラスタを構成するマシンのノードの出入りが激しいときである。チューン状態の時の通信量の変化に加え、各ノードがもつ経路情報と現実のノード状態を比較することで、分散協調プロトコルの情報伝達度合いの性能を計測することができる。