分散版 Jungle データベースの性能測定方法

145762E 氏名 仲松琹 指導教員：河野 真治

Abstract

1 研究背景

スマートフォンやタブレット端末の普及にともない、年々 Web サービスの利用者は増加した一方で、データ量が増大し、サーバ側への負荷も増加している。これを解決するため、Web サービスには、よりシステムの処理能力を拡張する性質である、スケーラビリティが求められてきている。

スケーラビリティとは、高性能のマシンを用意したり、複数のマシンに処理を分散させたりすることで、システムの処理能力を向上させる性能を指す。本実験で指すスケーラビリティとは、後者の方である。Web サービスにスケーラビリティを付与する方法の 1 つに、データベースにスケーラビリティを持たせる事が考えられる。

そこで、当研究室ではスケーラビリティを持つデータベースとして木構造を持つ分散データベース Jungle を開発している。方法としては、分散環境上で複数のデータベース Jungle を起動することで、処理を分散させる。

研究の積み重ねにより、Jungle の性能は上がっている。しかし、分散環境上で Jungle の性能を測定する方法が確立されていなかった。これまでに行われた分散環境上での Jungle の性能を検証する実験 [2] では、使用するテストプログラムのフロントエンドに Web サーバー Jetty が使用されており、純粋な Jungle の性能は測定できていなかった。今回は、新たに改良された Jungle の性能を、Web サーバーを取り除いた Test プログラムを用いて測定する。

本研究では、最新版 Jungle の分散性能を測定する環境を構築することを目的とする。

2 分散データベース Jungle

Jungle は、当研究室で開発を行っている木構造の分散データベースで、Java を用いて実装されている。

Jungle は名前付きの複数の木の集合からなり、木は複数のノードの集合でできている。ノードは自身の子のリストと属性名、属性値を持ち、データベースのレコードに相応する。通常のレコードと異なるのは、ノードに子供となる複数のノードが付くところである。

通常の RDB と異なり、Jungle は木構造をそのまま読み込むことができる。例えば、XML や Json で記述された構造を、データベースを設計することなく読み込むことが可能である。また、この木を、そのままデータベースとして使用することも可能である。しかし、木の変更の手間は木の構造に依存する。特に非破壊木構造 [1] を採用している Jungle では、木構造の変更の手間は O(1) から O(n) となりえる。つまり、アプリケーションに合わせて木を設計しない限り、十分な性能を出すことはできない。逆に、正しい木の設計を行えば高速な処理が可能である。

Jungle はデータの変更を非破壊で行なっており、編集ごとのデータをバージョンとして TreeOperationLog[2] に残している。Jungle の分散ノード間の通信は木の変更の TreeOperationLog を交換することによって、分散データベースを構成するよう設計されている。

3 分散フレームワーク Alice による分散環境の構築

本研究では、分散環境上での Jungle の性能を確認する為、VM32 台分のサーバーノードを用意し、それぞれで Jungle を起動することで、Jungle 間で通信をする環境をつくる。Jungle を起動したサーバーノード間の通信部分を、当研究室で開発している並列分散フレームワーク Alice[1] にて再現する。

Alice には、ネットワークのトポロジーを構成する TopologyManager[2] という機能が備わっている。TopologyManager に参加表明をしたサーバーノードに順番に、接続先のサーバーノードの IP アドレス、ポート番号、接続名を送り、受け取ったサーバーノードはそれらに従って接続する。今回、TopologyManager は Jungle をのせた VM32 台分のサーバーノードを、木構造を形成するように采配する（図 1 ）。

トポロジー構成後、Jungle 間の通信でのデータ形式には TreeOperationLog を利用する。TreeOperationLog には、ノードの編集の履歴などの情報が入っている。TreeOperationLog を Alice によって他の Jungle へ送ることで、送信元の Jungle と同じ編集を行う。こうして、Jungle 間でのデータの同期を可能にしている。

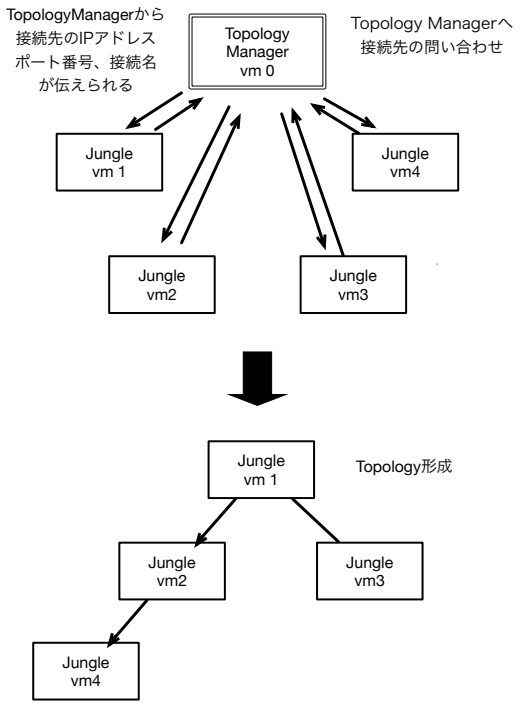

図 1: Alice による Jungle の木構造トポロジーの形成

4 TORQUE Resource Manager

分散環境上での Jungle の性能を測定するにあたり、VM32 台に Jungle を起動させた後、それぞれでデータを書き込むプログラムを動作させる。プログラムを起動する順番やタイミングは、TORQUE Resource Manager[3] というジョブスケジューラーによって管理する。

TORQUE Resource Manager は、ジョブを管理･投下･実行する 3 つのデーモンで構成されており、ジョブの管理･投下を担うデーモンが稼働しているヘッダーノードから、ジョブの実行を担うデーモンが稼働している計算ノードへジョブが投下される（図 2 ）。

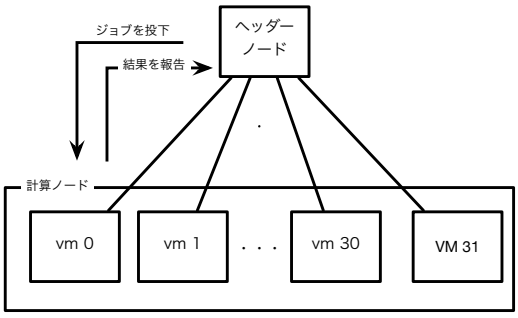

図 2: TORQUE の構成

ユーザーはジョブを記述したシェルスクリプトを用意し、スケジューラーに投入する。その際に、利用したいマシン数や CPU コア数を指定する。TORQUE は、ジョブに必要なマシンが揃い次第、受け取ったジョブを実行する。

5 性能測定用プログラム

これまでの分散環境上での Jungle の性能を測定する実験で使われたテストプログラムは、フロントエンドに Jetty という Web サーバーが使われていた。しかし、Web サーバーが仲介した測定結果となってしまい、純粋な Jungle の性能を測定できないという問題がある。そこで、Web サーバーを取り除き、これまでの研究により純粋に Jungle の性能を測定するプログラムを作成した。

まず、Jungle を起動する際に、木構造における子ノードに、データを複数書き込む機能である、-write オプション、書き込む回数を指定できる -count オプション を実装した。

複数の子ノードにデータをそれぞれ書き込み、最終的に root ノードへデータを merge していく。また、今回性能を測定するにあたり、root ノードに到達したデータが書き込まれた時間を計測し、出力結果に時間を表示するプログラムを、Alice に実装した。この機能は、TopologyManager を起動するコマンドに- showTime オプションをつけることで起動する。

6 評価実験

Jungle の分散性能を測定するにあたり、複数台の Jungle を通信させ、Jungle から Jungle に対する書き込みにかかる時間を計測する。複数台の Jungle を分散させる為に、学内共用の仮想マシンを 32 台使用した。分散した Jungle 同士の通信部分には、当研究室で開発している分散フレームワーク Alice の機能である TopologyManager を使用する。TopologyManager の起動には、仮想マシン 32 台のうちの 1 台を使用する。学科の仮想マシン 31 台上でそれぞれ 1 台ずつ Jungle を立ち上げ、ツリー型のトポロジーを構成する。そのうち 16 台の Jungle に対して 100 回ずつデータを書き込む。子ノードの Jungle は、次々と親ノードの Jungle へデータを書き込む。最終的にルートノードの Jungle へデータが到達し、書き込まれた時間を計測し、平均を取る。31 台中 16 台の Jungle から書き込まれたデータがルートノードの Jungle へ書き込まれる、一回あたりの時間を計測する実験である。(図 3)

7 まとめ

本研究では、Jungle の純粋な性能を測定するためのプログラムを Jungle,Alice に実装した。また、それらの機能を使用し、実際に Jungle の性能評価を行なった。

Jungle への書き込みを行う機能である -write オプションと、書き込みの回数を指定できる -count オプションの実装を行なった。

ツリートポロジーを構成した Jungle の分散環境上で、子ノードの Jungle に書き込まれたデータが、root ノードの

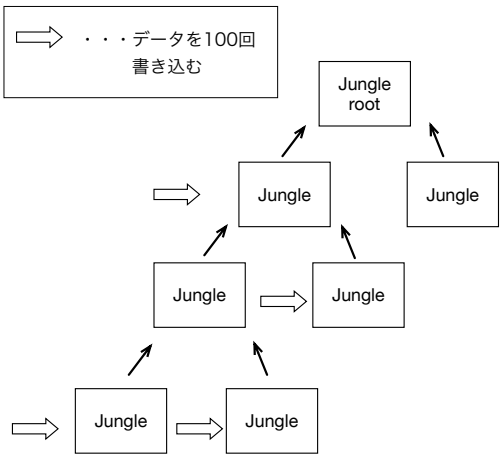

図 3: 複数の jungle に書き込まれたデータが root の jungle へ到達する時間を計測する

Jungle に到達し、書き込みが終了するまでの時間を表示する、 –showtime オプションの実装を行なった。

今回 Jungle の分散性能の評価を行い、() がわかった。

今後の課題として、今回確立した測定環境で、実際に現在の Jungle の性能をそくていする。また、Jungle は、トポロジー構成中に接続が切れてしまったとき、再接続すると node の中のデータが消えているため、データを再ロードする必要がある。その再ロードのプロトコルを定義したい。方法としては、接続した他のノード、もしくはデータを書き出すディスクを作り、そのディスクからデータを読み込みたい。その際に、ディスク上にしかないツリートポロジーを徐々に読み出すプロトコルを作りたい。

参考文献

[1] 金川 竜己 and 河野真治. 非破壊的木構造データベース jungle とその評価. 情報処理学会, 2015.

[2] 大城 信康. 分散 database jungle に関する研究. Master's thesis, 2013.

[3] 杉本　優. 分散フレームワーク alice 上の meta computation と応用. Master's thesis, 2014.

[4] 大城 信康 and 杉本 優 and 永山 辰己 and 河野真治. Data segment の分散データベースへの応用. 日本ソフトウェア科学会, apr 2013.