

分散 database Jungle に関する研究

A Study of distributed database  
Jungle

平成25年度 学位論文(修士)



琉球大学大学院 理工学研究科  
情報工学専攻

大城 信康

本論文は、修士(工学)の学位論文として適切であると認める。

論文審査会

印  
\_\_\_\_\_  
(主査) Mohammad Reza Asharif 氏

印  
\_\_\_\_\_  
(副査) 宮城 隼夫 氏

印  
\_\_\_\_\_  
(副査) 長田 智和 氏

# 要 旨

スマートフォンやタブレット端末の普及により、大量の通信を扱うウェブサービスが現れてきている。しかしそれに伴い、サーバサイド側への負荷も増大しウェブサービスがダウンする事態が出てきている。そのため、スケーラビリティはウェブサービスにおいて重要な性質の1つとなっている。スケーラビリティとは、ある複数のノードから構成される分散ソフトウェアがあるとき、その分散ソフトウェアに対して単純にノードを追加するだけで性能を線形に上昇させることができる性質である。そこで、スケーラビリティを持たせるためにアーキテクチャの設計から考えることにした。当研究室では非破壊的木構造を用いたデータベースである Jungle を開発している。非破壊的木構造とは、データの編集の際に一度木構造として保存したデータには触れず、新しく木構造を作成してデータの編集を行うことを言う。Jungle は分散データベースとして設計されていたが、データ分散や永続性といった部分の実装がまだ行われていない。

本研究では、Jungle に分散データベースの実装を行い、スケーラビリティをもつアーキテクチャの追求を行う。データ分散部分には当研究室で開発中である並列分散フレームワークである Alice を使用した。結果、複数のサーバノード間でデータの分散を行うことを確認した。

# Abstract

Smartphone and tablet pc are widely used, thereby Web services that handle large amounts of data are emerging. It has caused the webservice is down. Therefore, scalability is important software factor today. Scalability in distributed system is able to increase performance linearly when just added new node to system. In order to make provide scalability, we considered design of architecture.

We are developing database Jungle. It is use non-destructive tree structure. Non-destructive tree structure is not the destruction of data. Editing of data is done creating by new tree. Jungle was designed as a distributed database. But data distribution and persistent has not yet been implemented in the Jungle.

In this paper, we develop distributed database on jungle for pursuit architecture with scalability. Distributed data on Jungle is developing using parallel distributed framework Alice. As a result, we confirmed that data is distributed between the server node.

# 目次

<b>第 1 章</b>	<b>序論</b>	<b>1</b>
1.1	序論	1
1.1.1	研究背景と目的	1
1.1.2	本論文の構成	2
<b>第 2 章</b>	<b>既存の分散データベース</b>	<b>3</b>
2.1	Cassandra	3
2.2	MongoDB	3
2.3	Neo4j	4
<b>第 3 章</b>	<b>木構造データベース Jungle の実装と分散設計</b>	<b>5</b>
3.1	木構造データベース Jungle	5
3.1.1	破壊的木構造	5
3.1.2	非破壊的木構造	6
3.2	Jungle におけるデータ編集	8
3.2.1	TreeOperationLog	9
3.3	データ衝突時におけるマージによる解決	10
<b>第 4 章</b>	<b>Jungle の分散実装</b>	<b>11</b>
4.1	TreeOperationLog を用いての分散データベースの実装	11
4.2	並列分散フレームワーク Alice	11
4.3	Alice のトポロジーマネージャーの利用	12
4.4	Alice を用いての分散実装	13
4.5	ログのシリアライズ	13
4.6	Jungle の永続性の実装	14
4.7	掲示板プログラムにおけるマージの実装	14
<b>第 5 章</b>	<b>分散木構造データベース Jungle の評価</b>	<b>17</b>
5.1	実験方法	17
5.2	実験環境	17
5.3	実験	17
5.4	実験結果	17

第 6 章 結論	18
6.1 まとめ	18
6.2 今後の課題	18
6.2.1 データ分割の実装	18
6.2.2 Merger アルゴリズムの設計	18
6.2.3 Compaction の実装・分代替性の実装	18
謝辞	19
参考文献	20
発表文献	21

# 目 次

3.1	破壊的木構造の編集	5
3.2	非破壊的木構造の編集	6
3.3	非破壊的木構造の編集 1	7
3.4	非破壊的木構造の編集 2	7
3.5	非破壊的木構造の編集 3	7
3.6	非破壊的木構造の編集 4	8
3.7	非破壊的木構造による利点	8
3.8	Node の attribute と NodePath	9
3.9	TreeOperationLog の具体例	10
4.1	リング型の Network Topology	11
4.2	ツリー型の Network Topology	11
4.3	Alice によるネットワークトポロジー形成	13
4.4	Jungle による掲示板プログラムのデータ保持方法	14
4.5	他サーバノードの編集データ反映による整合性の崩れ 1	15
4.6	他サーバノードの編集データ反映による整合性の崩れ 2	15

# 表 目 次



# 第1章 序論

## 1.1 序論

### 1.1.1 研究背景と目的

スマートフォンやタブレット端末の普及により、大量の通信を扱うウェブサービスが現れてきている。しかしそれに伴い、サーバサイド側への負荷も増大しウェブサービスがダウンする事態が出てきている。そのため、スケーラビリティはウェブサービスにおいて重要な性質の1つとなっている。スケーラビリティとは、ある複数のノードから構成される分散ソフトウェアがあるとき、その分散ソフトウェアに対して単純にノードを追加するだけで性能を線形に上昇させることができる性質である。そこで、スケーラビリティを持たせるためにアーキテクチャの設計から考えることにした。当研究室では非破壊的木構造を用いたデータベースである Jungle を開発している。非破壊的木構造とは、データの編集の際に一度木構造として保存したデータには触れず、新しく木構造を作成してデータの編集を行うことを言う。Jungle は分散データベースとして設計されていたが、データ分散や永続性といった部分の実装がまだ行われていない。

本研究では、Jungle に分散データベースの実装を行い、スケーラビリティをもつアーキテクチャの追求を行う。データ分散部分には当研究室で開発中である並列分散フレームワークである Alice を使用した。結果、複数のサーバノード間でデータの分散を行うことを確認した。

## 1.1.2 本論文の構成

## 第2章 既存の分散データベース

### 2.1 Cassandra

Cassandra は2008年7月に Facebook によってオープンソースとして公開された Key-Value なデータベースである。Amazon の Dynamo という分散キーバリューデータベースの影響を受けて作られている。スキーマレスな NoSQL データベースになる。

Cassandra はサーバノードの配置にコンシステント・ハッシングアルゴリズムを用いる。コンシステント・ハッシングによりノードは論理的にリング上に配置される。リングには数値で表される位置がある。データを書き込む際には、キーとなるハッシュ値に従いそのリングの位置から時計回りに近いサーバノードへと書き込まれる。コンシステント・ハッシングを用いることで、ノードの数が増減した場合に、再配置をしなくてもよいという利点がある。データの偏りにより少数のサーバへの負荷が大きき場合に、負荷が高いハッシュ値が指すリング上に新たなノードを追加することで負荷を下げるといった手段もとれる。

また Cassandra はデータの読み込み、書き込みの際に Consistency Level を指定することができるのも Cassandra の特徴である。データ取得の際に、いくつかのノードに問い合わせを行うのか設定するのが Consistency Level である。Consistency Level には ONE, QUORAM, ALL があり、それぞれ1つのノード、全体のノードの過半数+1、全体のノードへの処理を設定することができる。Cassandra は高い可用性と分断耐性をもつ。

### 2.2 MongoDB

MongoDB は2009年に公開された NoSQL のデータベースである。JSON フォーマットのドキュメントデータベースであり、これはスキーマが無いリレーショナルテーブルに例えられる。スキーマが無いため、事前にデータの定義を行う必要がない。そのためリレーショナルデータベースに比べてデータの追加・削除が行いやすい。

MongoDB は保存したデータを複数のサーバに複製をとる (replicate)。これはレプリケーション (replication) と呼ばれる。また、1つのサーバが全てのデータを持つのではなく、ある範囲の値を別々のサーバに分割 (シャード) させて保持する。これをシャーディング (sharding) という。MongoDB はレプリケーションとシャーディングにより分散データベースとして機能する。

## 2.3 Neo4j

Neo4j は、グラフデータベースと呼ばれる NoSQL のデータベースである。データをグラフとして保存する。グラフはノードとリレーションシップにより表され、それぞれがプロパティを持つことができる。リレーションシップはグラフでいうところのエッジにあたる。ノードからリレーションシップを辿り、各プロパティをみることでデータの取得を行うことができる。通常データベースでは、データの取り出しに値の結合や条件の判定を行う。だが、グラフデータベースグラフはどれだけデータが大きくなろうとノードからノードへの移動は 1 ステップですむ。そのため、どれだけデータが大きくなろうと、データが小さい時と同じ計算量でデータの取得が行える。

Neo4j はマスターとスレーブの関係になるクラスタを構成することで分散データベースとして機能する。マスターに書かれたデータはスレーブに書き込まれるが、すぐに全てのスレーブに書き込まれるわけではない。したがってデータの整合性が失われる危険がある。スレーブサーバは現在保持しているデータを返すことができる。それにより高い読み取り性能の要求に答えることができる高可用性を実現している。

# 第3章 木構造データベースJungleの実装と分散設計

## 3.1 木構造データベースJungle

Jungle はスケーラビリティのある CMS の開発を目指して当研究室で開発されている非破壊的木構造データベースである。一般的なコンテンツマネジメントシステムではブログツールや Wiki・SNS が多く、これらのウェブサイトの構造は大体が木構造であるため、データ構造として木構造を採用している。

まず破壊的木構造と、非破壊的木構造の説明をし、Jungle におけるデータ編集の実装について述べる。

### 3.1.1 破壊的木構造

破壊的木構造の編集は、木構造で保持しているデータを直接書き換えることで行う。図 3.1 は破壊的木構造の編集を表している。

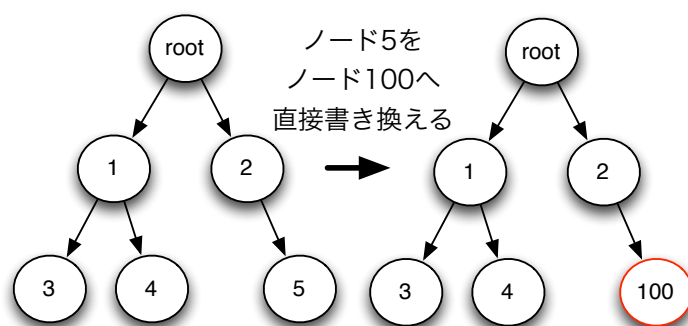


図 3.1: 破壊的木構造の編集

破壊的木構造は、編集を行う際に木のロックを掛ける必要がある。この時、データを受け取ろうと木を走査するスレッドは書き換えの終了を待つ必要があり、閲覧者がいる場合は木の走査が終わるまで書き換えをまたなければならない。これではロックによりスケーラビリティが損なわれてしまう。

### 3.1.2 非破壊的木構造

非破壊的木構造は破壊的木構造とは違い、一度作成した木を破壊することはない。非破壊的木構造においてデータの編集は、ルートから編集を行うノードまでコピーを行い新しく木構造を作成することで行われる。図 3.2 は非破壊的木構造のデータ編集を示している。

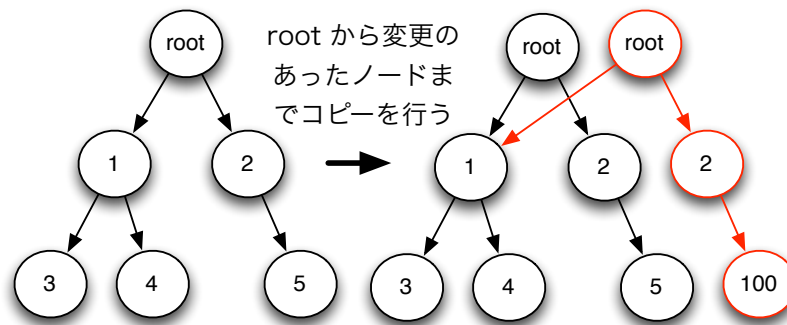


図 3.2: 非破壊的木構造の編集

非破壊的木構造におけるデータ編集の手順を以下に示す。

1. ルートから編集を行うノードまでのパスを調べる (図 3.3).
2. 編集を行うノードのコピーをとる。コピーをとったノードへデータの編集を行う (図 3.4).
3. 調べたパスに従いルートからコピーしたノードまでの間のノードのコピーをとり繋げる (図 3.5).
4. コピーしたルートノードは編集を行っていないノードへの参照を貼り新しい木構造を作る (図 3.6).

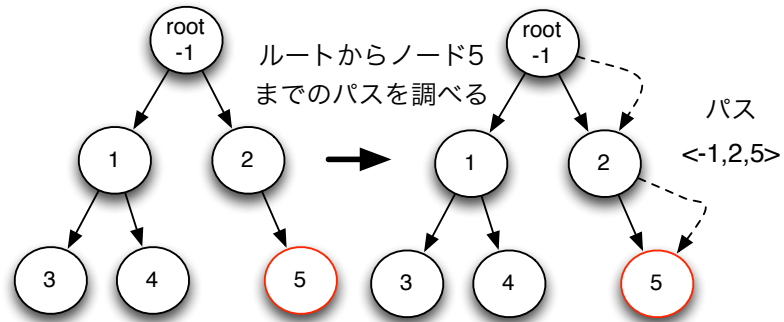


図 3.3: 非破壊的木構造の編集 1

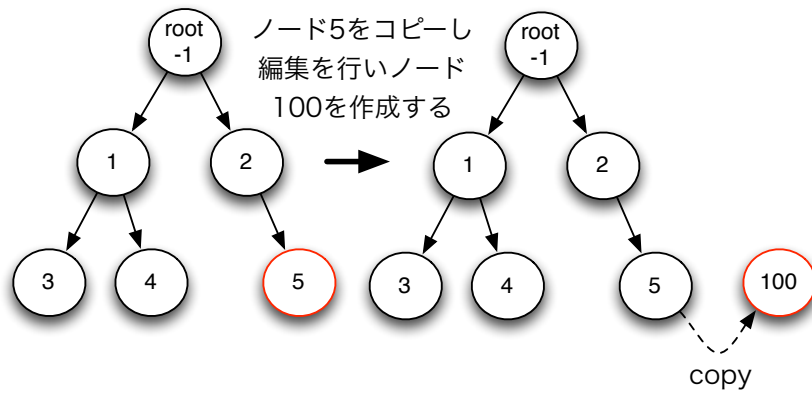


図 3.4: 非破壊的木構造の編集 2

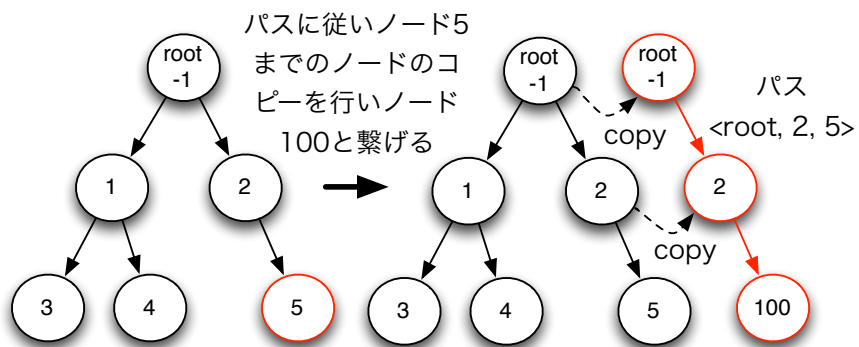


図 3.5: 非破壊的木構造の編集 3

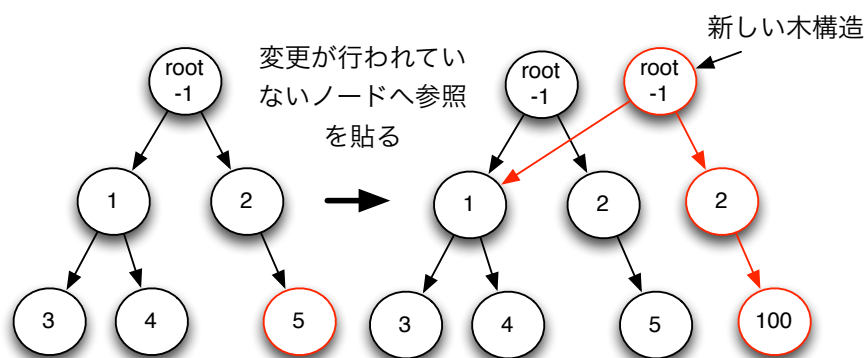


図 3.6: 非破壊的木構造の編集 4

非破壊的木構造により, データの読み込みと編集を同時に行うことが可能になる.

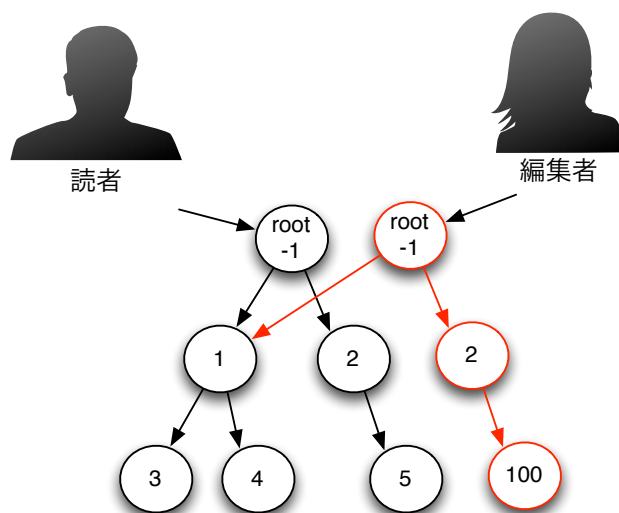


図 3.7: 非破壊的木構造による利点

## 3.2 Jungle におけるデータ編集

Jungle ではデータをそれぞれの Node が attribute として保持する. attribute は String 型の Key と ByteBuffer の value のペアにより表される. Jungle でデータ編集を行う場合, この Node に対して削除や attribute の追加等を行うことを指す. どの Node へデータの編集を行うかはパスで示す. このパスは NodePath と呼ばれる (図 3.8).

Node の編集は Node の追加・削除, それと attribute の追加・削除を行うことを指す. Node の編集のためには次の 4 つの API が用意されている.



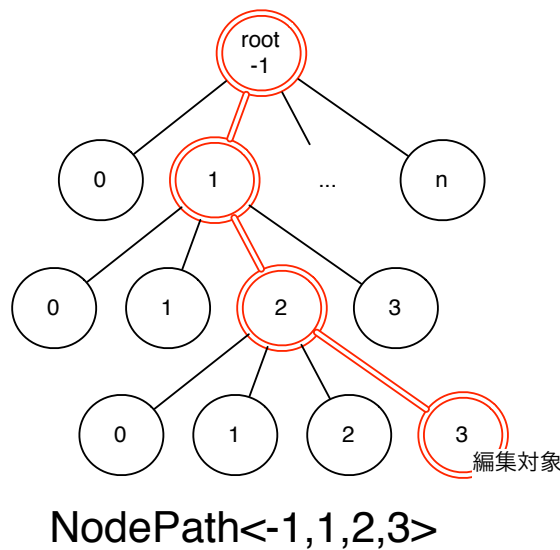


図 3.8: Node の attribute と NodePath

- `addNewChild(NodePath _path, int _pos)` NodePath で指定された Node に子供となる Node を追加する API である。pos で指定された番号に子供として追加を行う。
- `deleteChildAt(NodePath _path, int _pos)` NodePath と pos により指定される Node を削除する API である。
- `putAttribute(NodePath _path, String _key, ByteBuffer _value)` Node に attribute を追加する API である。NodePath は attribute を追加する Node を指す。
- `deleteAttribute(NodePath _path, String _key)` \_key が示す attribute の削除を行う API である。NodePath は Node を示す。

この Node 編集の為の API は NodeOperation と呼ばれる。

### 3.2.1 TreeOperationLog

API を使用すると, Jungle 内部では NodeOperation として順次ログに積まれていき, 最終的に commit されることで編集が行われる。この時ログに積まれる複数の NodeOperation を TreeOperationLog という。Jungle ではこの TreeOperationLog 単位でデータの編集が行われる。以下に TreeOperationLog の具体的な例を示す (3.1)。

Listing 3.1: トポロジーマネージャーの利用

```
1 [APPEND_CHILD:<-1>:pos:0]
2 [PUT_ATTRIBUTE:<-1,1>:key:author,value:oshiro]
```

```

3 [PUT_ATTRIBUTE:<-1,1>:key:mes,value:hello]
4 [PUT_ATTRIBUTE:<-1,1>:key:key,value:hoge]
5 [PUT_ATTRIBUTE:<-1,1>:key:timestamp,value:0]
    
```

このログは今回の研究で使用したベンチマーク用掲示板プログラムにおける書き込みにより行われるログである (図 3.9).

大文字の英字は実行した NodeOperation を表す.  $i, j$  により囲まれている数字は NodePath を示す. NodePath の表記以降は Node の position や attribute の情報を表している.

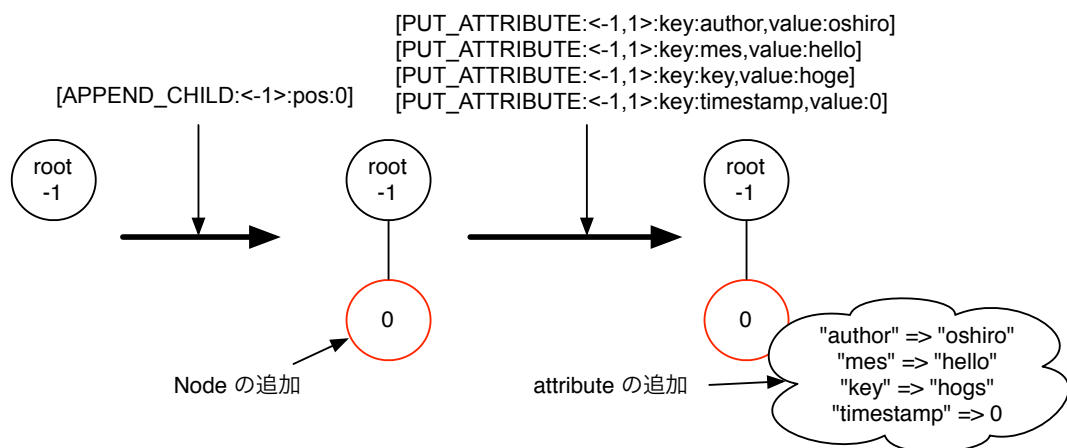


図 3.9: TreeOperationLog の具体例

### 3.3 データ衝突時におけるマージによる解決

Jungle への分散実装を行うことである問題が発生する. それは更新されたデータ同士が衝突を起こした時の処理である. Jungle はリクエストがきた場合, 現在もっているデータを返す. しかしそのデータは最新のものであるかは保証されない. よって, 別のサーバノードで動いている Jungle からの更新データと衝突する可能性がある.

この問題に対して Jungle はアプリケーションレベルでのマージを実装して貰うことで解決をはかる.

## 第4章 Jungleの分散実装

### 4.1 TreeOperationLogを用いての分散データベースの実装

Jungle でデータ扱うと TreeOperationLog として残ることは述べた。この TreeOperationLog を他のサーバへと送り、Jungle の編集を行って貰うことでデータの分散を行うことができる。ここで問題になることはネットワークポロジの形成方法であった。

Jungle で使用するネットワークポロジはツリー型を考えている。しかし、リング型といった他のネットワークポロジによる実装についても試す余地はある。自由にネットワークポロジの形成を行うことができる必要があった。

そこで当研究室で開発を行っている並列分散フレームワークである Alice を使用することにした。Alice により提供されるネットワークポロジ形成を用いて TreeOperationLog を各サーバノードへ配ることで並列分散フレームワークの実装を行う。

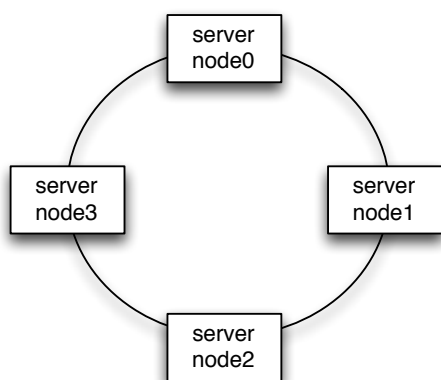


図 4.1: リング型の Network Topology

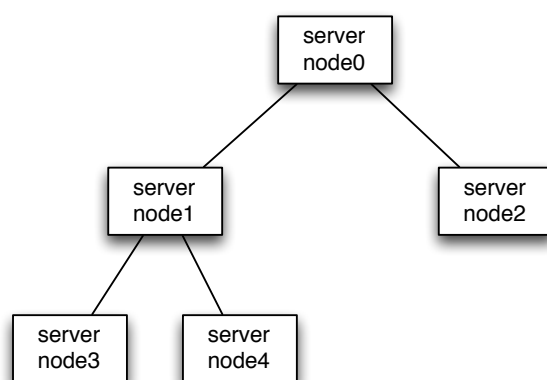


図 4.2: ツリー型の Network Topology

### 4.2 並列分散フレームワーク Alice

Alice は当研究室で開発している並列分散フレームワークである。Alice はデータを DataSegment, コードを CodeSegment という単位で扱うプログラミングを提供している。DataSegment として扱われるデータは

### 4.3 Alice のトポロジーマネージャの利用

Alice はサーバノード同士によるネットワークトポロジー形成の機能を持つ。トポロジーマネージャの起動は 4.2 の様にポート番号の指定と dot ファイルを引数として渡すことで行う。(4.1).

Listing 4.1: Alice によるネットワークトポロジーマネージャの起動

```
1 % java -cp Alice.jar alice.topology.manager.TopologyManager -p 10000 -conf ./topology/tree5.dot
```

ポート番号は Alice により記述された並列分散プログラムの起動時に渡す必要がある。dot ファイルには、トポロジーをどのように形成するかが書かれている。以下に、サーバノード数 5 で、2 分木ツリー構造を形成する dot ファイルの例を示す (4.2).

Listing 4.2: ネットワークトポロジー設定用 dot ファイル

```
1 % cat tree5.dot
2 digraph test {
3   node0 -> node1 [label="child1"]
4   node0 -> node2 [label="child2"]
5   node1 -> node0 [label="parent"]
6   node1 -> node3 [label="child1"]
7   node1 -> node4 [label="child2"]
8   node2 -> node0 [label="parent"]
9   node3 -> node1 [label="parent"]
10  node4 -> node1 [label="parent"]
11 }
```

node0 や node1 はサーバノードの名前を示す。サーバノードの間にはラベルがあり、Alice 上ではこのラベルに指定される文字列 (キー) を使うことで他のサーバノードのデータへアクセスすることができる。node0 -> node1 はサーバノード同士の繋がりを示している。次に続く label="child1" は、node0 が node1 のデータに "child1" という文字列を使うことでアクセスできることを示す。

dot ファイルを読み込んだ Alice のトポロジーマネージャに対して、サーバノードは誰に接続を行えばよいかを訪ねる。トポロジーマネージャは訪ねてきたサーバノードに対してノード番号を割り振り、dot ファイルに記述している通りにサーバノード同士が接続を行うよう指示をだす。

トポロジーマネージャは接続要求先を聞いてくるサーバノードに対して名前を割り振り、接続相手を伝える。dot ファイル 4.2 により形成されるトポロジーを図 4.3 に示す。

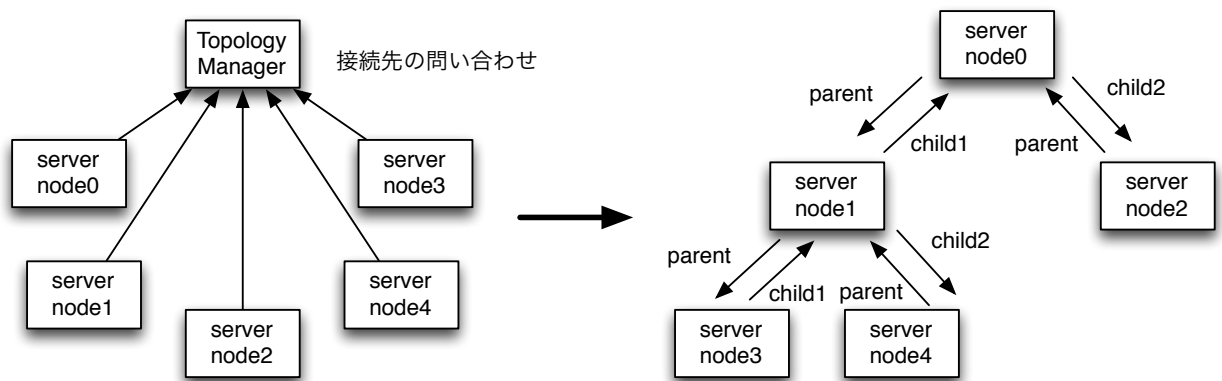


図 4.3: Alice によるネットワークトポロジー形成

矢印に書かれている文字列は、相手のデータにアクセスするキーを示す。 ”child1”, ”child2”, ”parent” というキーを使うことで別のサーバノードにあるデータを取得することができる。

トポロジーマネージャに最初に接続要求を行う並列分散プログラム側は、次のように記述する (4.3)

Listing 4.3: Alice を使用してのトポロジー形成

```

1 public static void main( String[] args ) throws Exception
2 {
3     RemoteConfig conf = new RemoteConfig(args);
4     new TopologyNode(conf, new StartBBSCodeSegment(args, conf.bbsPort));
5 }

```

そして、プログラムの起動時にはトポロジーマネージャが動いているサーバのドメインとポート番号を渡すことでトポロジーの形成が行われプログラムの処理がはしる。例えば、mass00.cs.ie.u-ryukyu.ac.jp というサーバ上でポート番号 10000 を指定してトポロジーマネージャを起動した場合は次のようになる (4.4)。

Listing 4.4: トポロジーマネージャの利用

```

1 % java Program -host mass00.cs.ie.u-ryukyu.ac.jp -port 10000

```

## 4.4 Alice を用いての分散実装

形成されたトポロジー上でのデータの送受信を行う部分について述べる。

## 4.5 ログのシリアルライズ

ここでログのシリアルライズについて述べる。

シリアルライズとは、データをネットワーク上に流しても良い形式に変換することである。

## 4.6 Jungle の永続性の実装

## 4.7 掲示板プログラムにおけるマージの実装

Jungle に分散実装を行った後の問題としてデータ衝突がある。他のサーバノードから送られてくるデータが既に手元で変更を加えた木構造を対象とした場合に発生する問題である。Jungle ではこれをアプリケーション毎にマージを実装することで解決させる。

今回分散実装を行い、例題として掲示板プログラムを用意した。掲示板プログラムに実装を行ったマージについて述べる。まず Jungle を用いた掲示板プログラムのデータ保持方法を図 4.4 に示す。

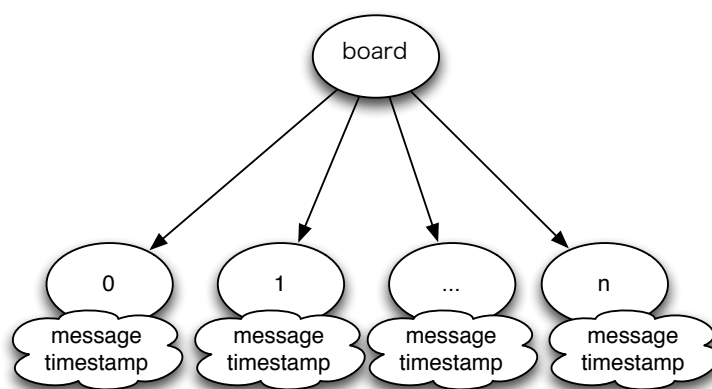


図 4.4: Jungle による掲示板プログラムのデータ保持方法

掲示板プログラムでは各掲示板毎に 1 つの木構造が作成される。掲示板への 1 つの書き込みは子ノードを 1 つ追加することに相当する。また、各子ノードは attributes として書き込みの内容である message と書き込まれた時間を表す timestamp を保持している。先に追加された順で子ノードには若い番号が割り振られる。

他サーバノードからの書き込みをそのまま子ノードの後ろに追加してしまうと、データの整合性が崩れてしまう。この時の状態を表しているのが図 4.5 と 4.6 になる。

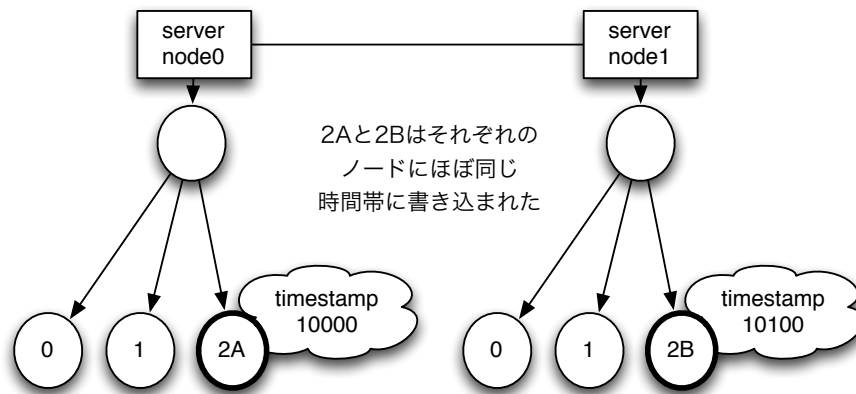


図 4.5: 他サーバノードの編集データ反映による整合性の崩れ 1

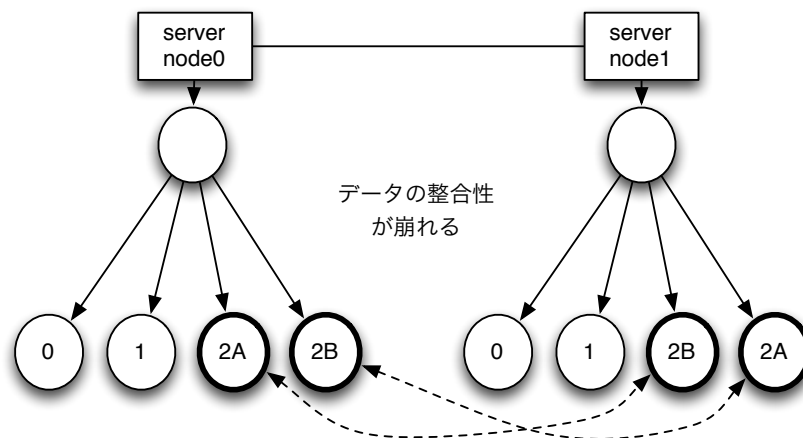


図 4.6: 他サーバノードの編集データ反映による整合性の崩れ 2

図 4.6 の server node0 の木の状態にするのが理想である。掲示板への書き込みの表示は、書き込みされた時間が早い順に表示されるようにしたい。これを timestamp を利用することで行う。他サーバノードから来たデータに関しては、timestamp を参照し、次に自分の保持している木の子ノードの timestamp と比べていくことでデータの追加する場所を決める。これが今回実装を行った掲示板システムにおけるマージになる。



# 第5章 分散木構造データベース Jungleの評価

- 5.1 実験方法
- 5.2 実験環境
- 5.3 実験
- 5.4 実験結果

## 第6章 結論

### 6.1 まとめ

### 6.2 今後の課題

#### 6.2.1 データ分割の実装

#### 6.2.2 Merger アルゴリズムの設計

#### 6.2.3 Compaction の実装・分代替性の実装

# 謝辞

本研究を行うにあたり、ご多忙にも関わらず日頃より多くの助言、ご指導をいただきました河野真治助教授に心より感謝いたします。

また、様々な研究や勉強の機会を与えてくださった、株式会社 Symphony の永山辰巳さん、同じく様々な助言を頂いた森田育宏さんに感謝いたします。様々な研究に関わることで自身の研究にも役立てることが出来ました。

研究を行うにあたり、並列計算環境の調整、意見、実装に協力いただいた谷成 雄さん、杉本優さん、並びに並列信頼研究室の全てのメンバーに感謝いたします。

最後に、大学の修士まで支えてくれた家族に深く感謝します。

## 参考文献

- [1] 玉城将士, 河野真治. Cassandra を使った cms の pc クラスタを使ったスケーラビリティの検証. 日本ソフトウェア科学会, August 2010.
- [2] 玉城将士, 河野真治. Cassandra を使ったスケーラビリティのある cms の設計. 情報処理学会, March 2011.
- [3] 玉城将士, 河野真治. Cassandra と非破壊的構造を用いた cms のスケーラビリティ検証環境の構築. 日本ソフトウェア科学会, August 2011.
- [4] Avinash Lakshman and Prashant Malik. Cassandra - a decentralized structured storage system. *LADIS*, Mar 2003.
- [5] Fay Chang and 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.
- [6] Nancy Lynch and Seth Gilbert. Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services. *ACM SIGACT News*, 2002.
- [7] 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.
- [8] Matt Welsh. The staged event-driven architecture for highly-concurrent server applications.
- [9] Eric Brewer Matt Welsh, David Culler. Seda : An architecture for well-conditioned , scalable internet services. *SOSP*.

# 発表履歴

- Continuation based C の GCC 4.6 上の実装について,  
大城信康, 河野真治 (琉球大学),  
第 53 回プログラミング・シンポジウム, Jan, 2012
- ディペンダブルシステムのための木構造を用いた合意形成データベースの提案と実装,  
大城信康, 河野真治 (琉球大学), 玉城将士 (琉球大学), 永山 辰巳 (株式会社 Symphony),  
情報処理学会システムソフトウェアとオペレーティング・システム研究会 (OS), May, 2013
- Data Segment の分散データベースへの応用,  
大城信康, 杉本優 (琉球大学), 河野真治 (琉球大学),  
日本ソフトウェア科学会 30 回大会 (2013 年度) 講演論文集, Sep, 2013