

Internet Week 2009

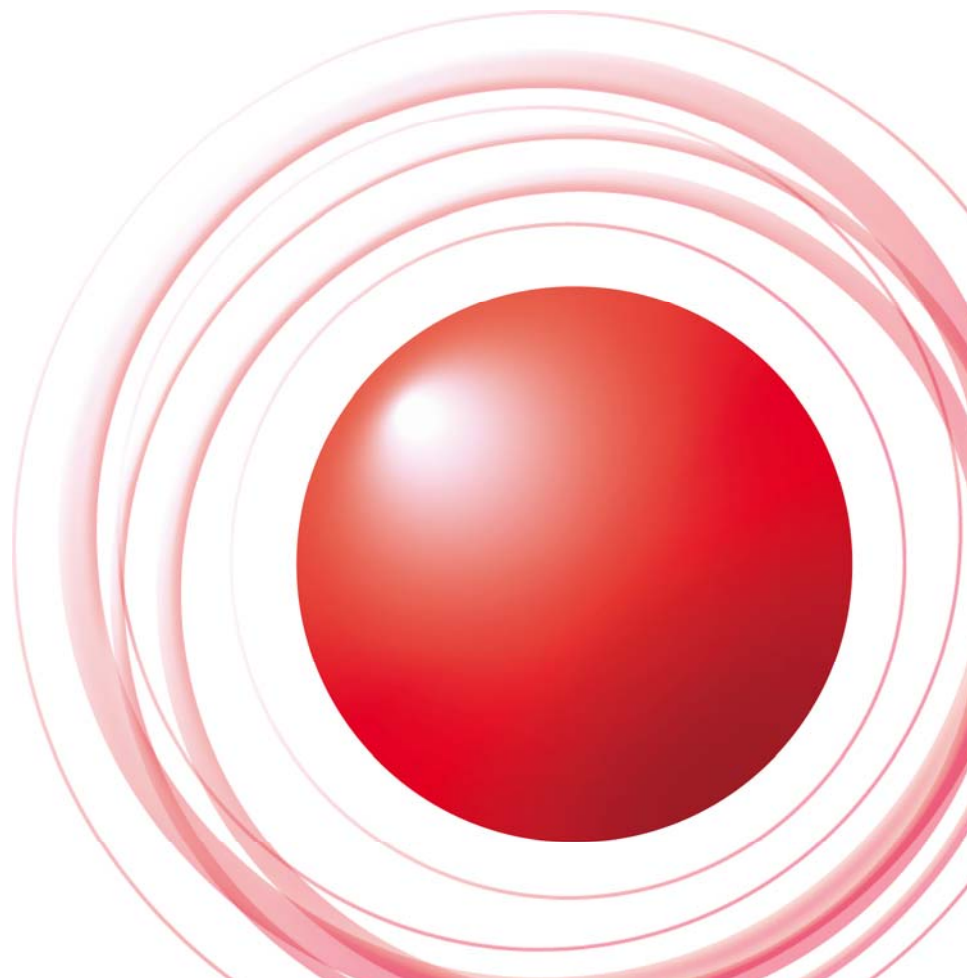
分散システムの実装とトラフィック解析への応用



2009年11月26日

株式会社インターネットイニシアティブ
前橋 孝広

Ongoing Innovation



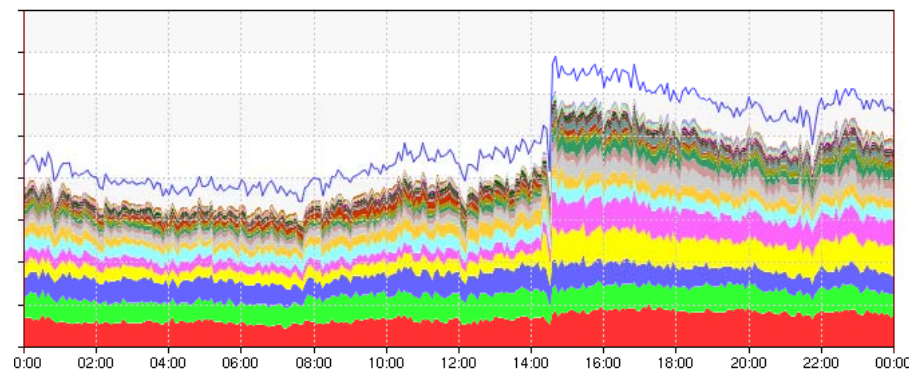
3行でまとめると

- 大量のデータを扱いたい
- 良いデータベースがないから自分で作った
- 日々、活用しています

大量のデータを扱いたい

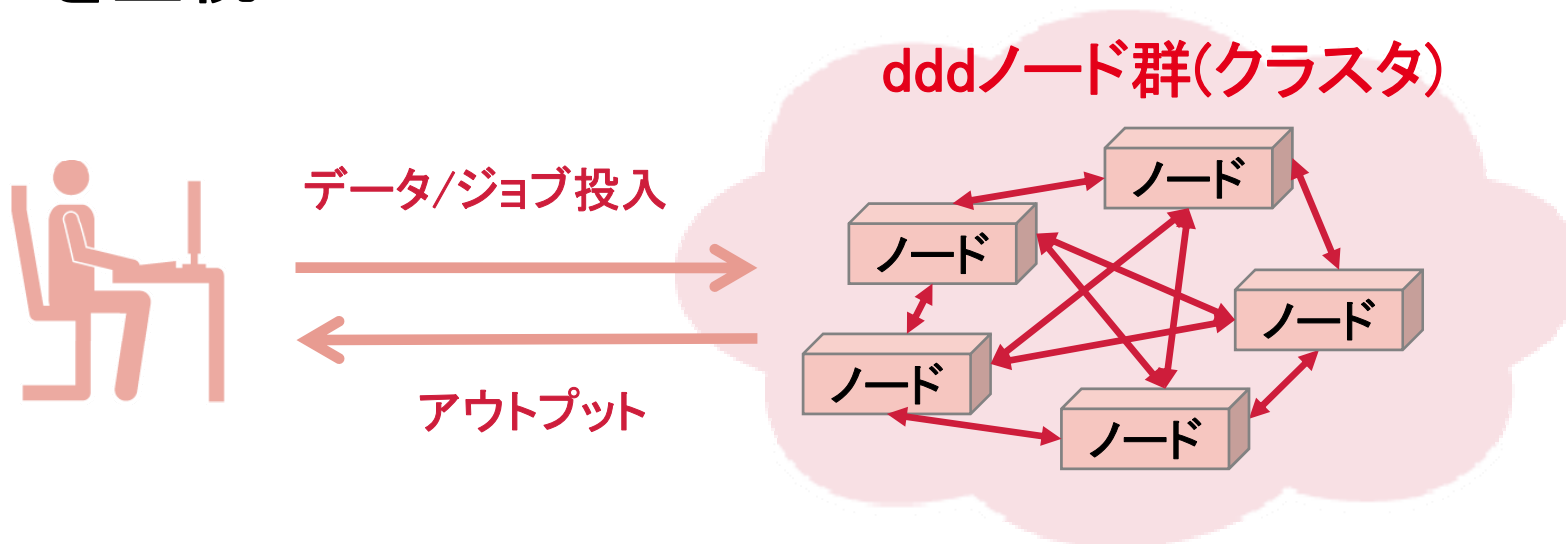
- **ISPにおいて、トラフィックの状態把握は必須**
 - 運用や設計を最適な状態に維持するため
 - 異常の検知・障害への対応のため
- **状態把握のための計測手法**
 - SNMP
 - インタフェース単位、比較的小さいデータ量
 - NetFlow (フロー統計情報)
 - 詳細だが**膨大なデータ量**

NetFlowによるトラフィック解析グラフ例



dddとは?

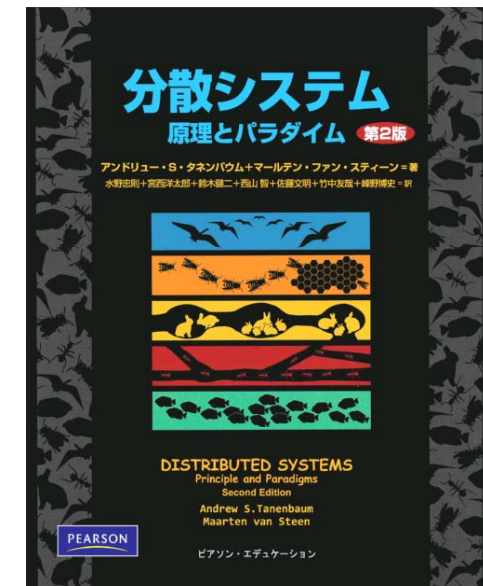
- IIJが独自開発した分散システムソフトウェア
 - 分散ストレージ + 分散データ処理機能(MapReduce)
 - とりあえずデータベースと置いていただければ...
- 大量のデータを扱うため、スケーラビリティと可用性を重視



イメージ図

分散システムの定義と目的

- 分散システムはそのユーザに対して単一のコヒーレントシステムとして見える独立したコンピュータの集合である(Tanenbaum)



- 分散システムを使う目的
 - 1台でできないことを複数台で分担することで**能力向上**
 - 一部のマシンが壊れても全体としては動き続ける**可用性**

ddd開発の経緯

- **ddd以前**

- リレーショナルデータベースを使ってデータを格納
- 数千万レコード程度が限界、スケールしない

- **要求事項**

- 大量のデータを保存できること(数百億レコード超)
- 任意の条件で高速にデータを抽出・集計できること
- 一部のノードが故障してもシステム全体は動き続けること
- 後から動的に拡張できること

そのようなデータベースは存在しない

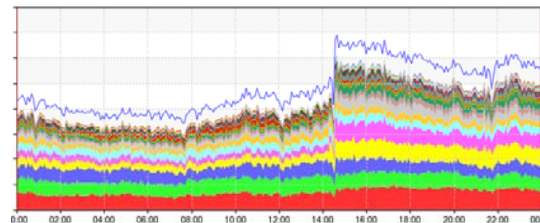


自分で作った

dddの使用用途例

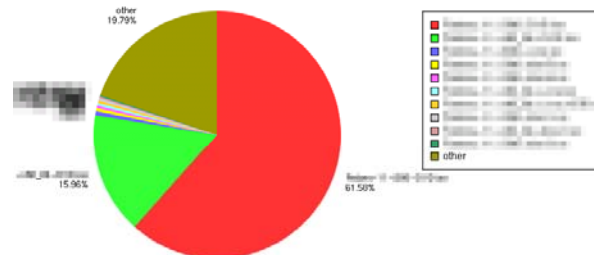
- **トラフィック解析**

- NetFlowデータを蓄積・解析
- 障害対応、DDoS攻撃の分析、設備増強計画の参考情報として利用



- **ログ解析**

- 各種サーバのログ
- ファイアウォール、接続認証サーバ、Apache など



- **コンテンツ配信**

- dddを巨大な分散ストレージとして利用

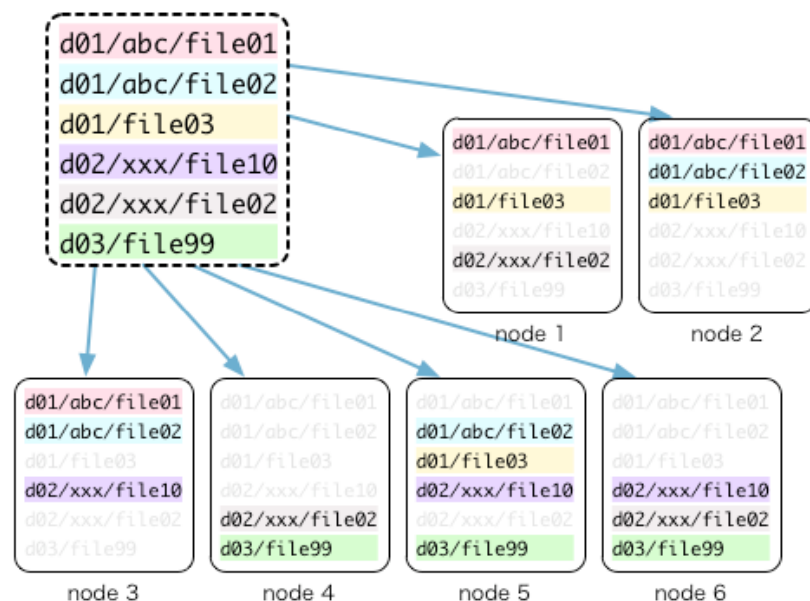
いずれも「大量」「消えたら困る」データ

dddを構成する技術要素

- **P2P**
 - 各ノードは対等、単一障害点なし
- **分散キーバリューストア(KVS)**
 - データをキーとバリューストアの組み合わせで格納
 - コンシステントハッシュ法、バーチャルノード
- **自動データ複製**
 - 同じデータを異なる3つのノードに複製
- **MapReduce**
 - 並列分散処理フレームワーク
 - バッチ処理&インタラクティブ処理両用

分散キーバリューストア(KVS)によるストレージ

- **KVS をファイルシステムっぽく見せかけている**
 - キー：ファイル名(パス名)
 - バリュー：ファイル本体
- 各ノードで、キーで示されるパスにファイルを保存
- ls は、全ノードで実行した結果をマージ
- 異なる3つのノードに複製
- アクセス方法
 - 独自API
 - WebDAV
 - FUSE



分散ストレージへのトラフィック情報の保存

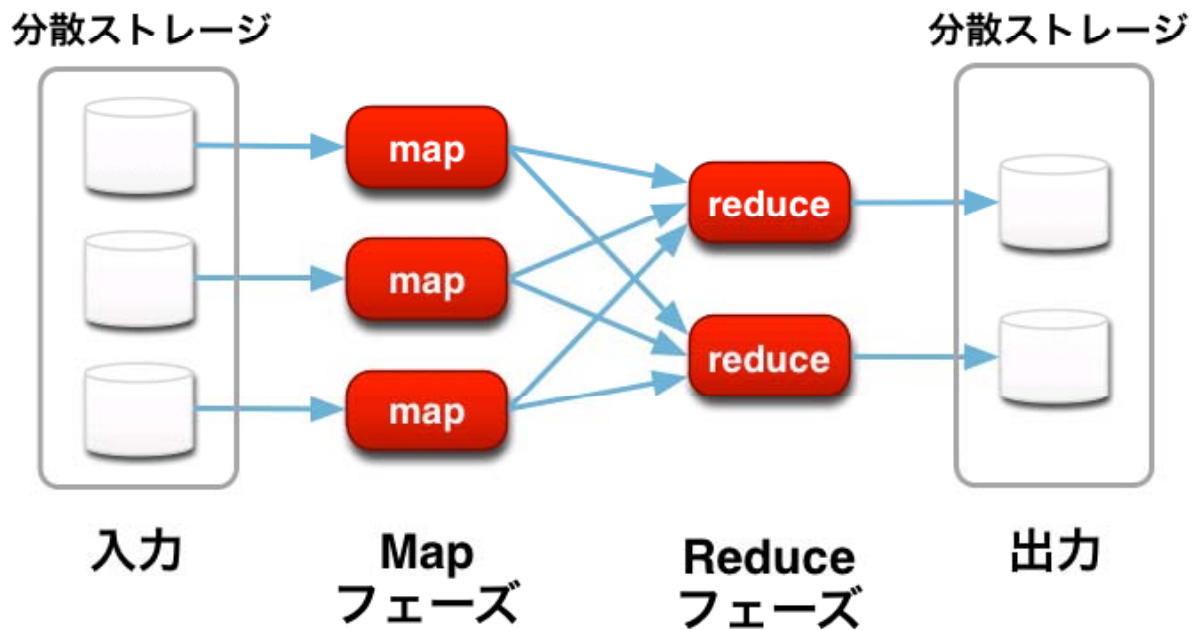
- ルータ名(ホスト名)+時刻をファイル名に
 - 時刻はある程度まとめて1ファイルに複数レコード保存
- ファイル名をキーにしてKVSにストア
 - コンシステントハッシュ法により適度に別々のノードにばらける

例:

```
router01.iij.ad.jp/20091101/00  
router01.iij.ad.jp/20091101/01  
router01.iij.ad.jp/20091101/02  
router01.iij.ad.jp/20091101/03  
router01.iij.ad.jp/20091101/04  
...
```

並列分散処理フレームワーク MapReduce

- **mapとreduceの2段階にわけてデータ処理**
 - ① map – 抽出・変換
 - ② reduce – 集約



Apacheログ解析のデータフロー例

元データ

HTTP ステータスコードの集計の場合

```
192.168.0.4 - - [22/Nov/2009:15:22:24 +0900] "GET /index.html HTTP/1.1" 200 449
192.168.0.4 - - [22/Nov/2009:15:22:24 +0900] "GET /favicon.ico HTTP/1.1" 404 209
172.16.100.1 - - [22/Nov/2009:15:22:28 +0900] "GET /info.html HTTP/1.1" 301 2352
...
```

map

```
1258870944 200 1
1258870944 404 1
1258870948 301 1
...
```

他のノードで map された
結果
...
...

他のノードで map された
結果
...
...

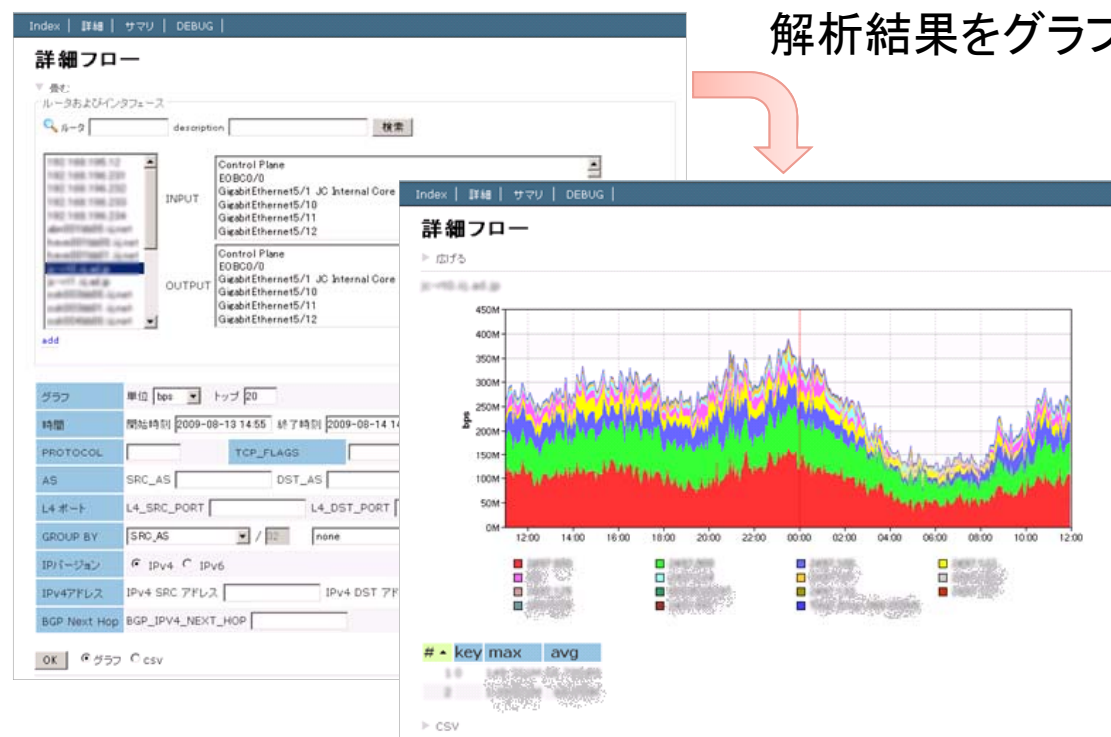
reduce

```
1258870800 200 1566409
1258870800 404 339071
1258870800 206 2209877
...
```

結果

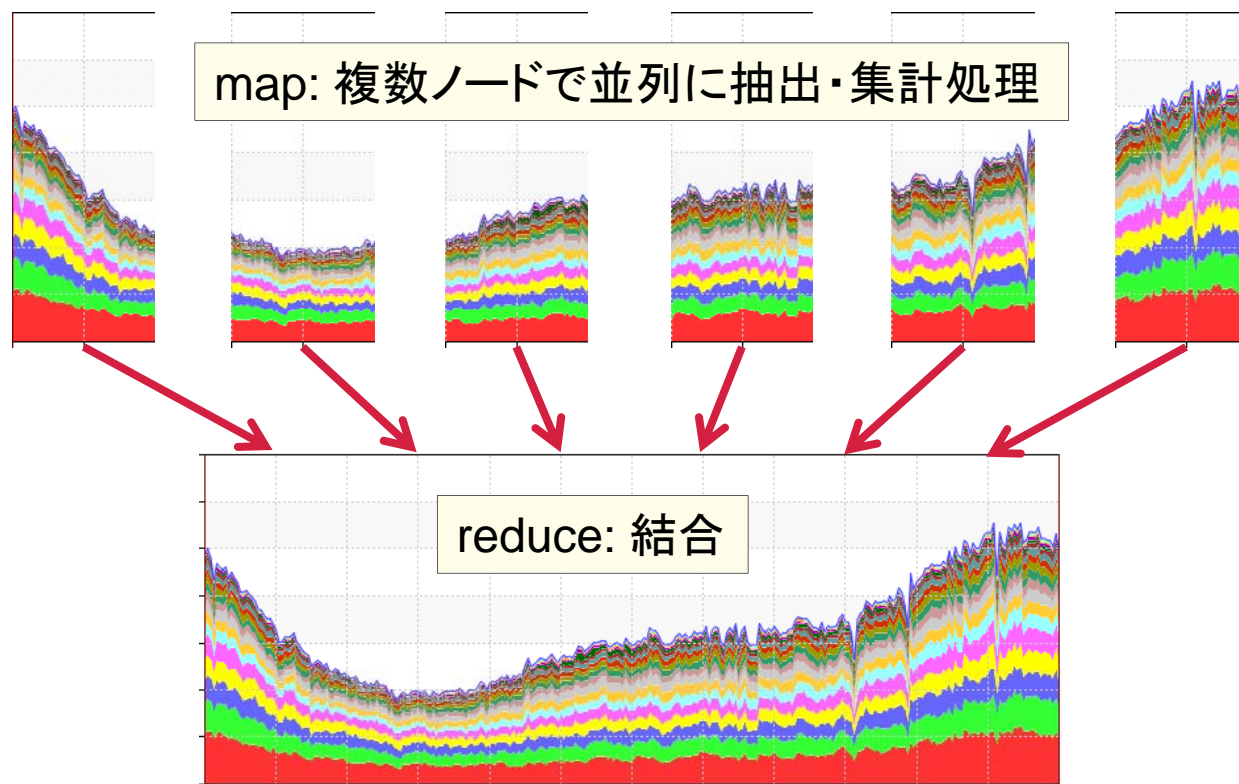
インタラクティブ型視覚化 UI

- MapReduce はバッチ処理向き(とされているが...)
- ddd ではインタラクティブな処理にも利用
 - MapReduceジョブ開始時のオーバヘッドが小さい



MapReduce のグラフ描画への応用

- ユーザが任意の抽出・集計条件を指定
- 時分割されたデータに対して複数ノードで並列処理
- それらを結合



まとめ

- **dddは多数のノードからなる分散システム**
 - 分散システム – 単一システムに見える多数のノード群
- **膨大なデータを処理するために ddd を独自開発**
 - スケーラビリティと可用性を重視
- **仕組み**
 - P2P, 分散キーバリューストア
 - MapReduce
- **トラフィック解析・ログ解析などに使われている**



ご清聴ありがとうございました

Ongoing Innovation

本書には、株式会社インターネットイニシアティブに権利の帰属する秘密情報が含まれています。本書の著作権は、当社に帰属し、日本の著作権法及び国際条約により保護されており、著作権者の事前の書面による許諾がなければ、複製・翻案・公衆送信等できません。IIJ、Internet Initiative Japanは、株式会社インターネットイニシアティブの商標または登録商標です。その他、本書に掲載されている商品名、会社名等は各会社の商号、商標または登録商標です。本文中では™、®マークは表示しておりません。©2009 Internet Initiative Japan Inc. All rights reserved. 本サービスの仕様、及び本書に記載されている事柄は、将来予告なしに変更することがあります。