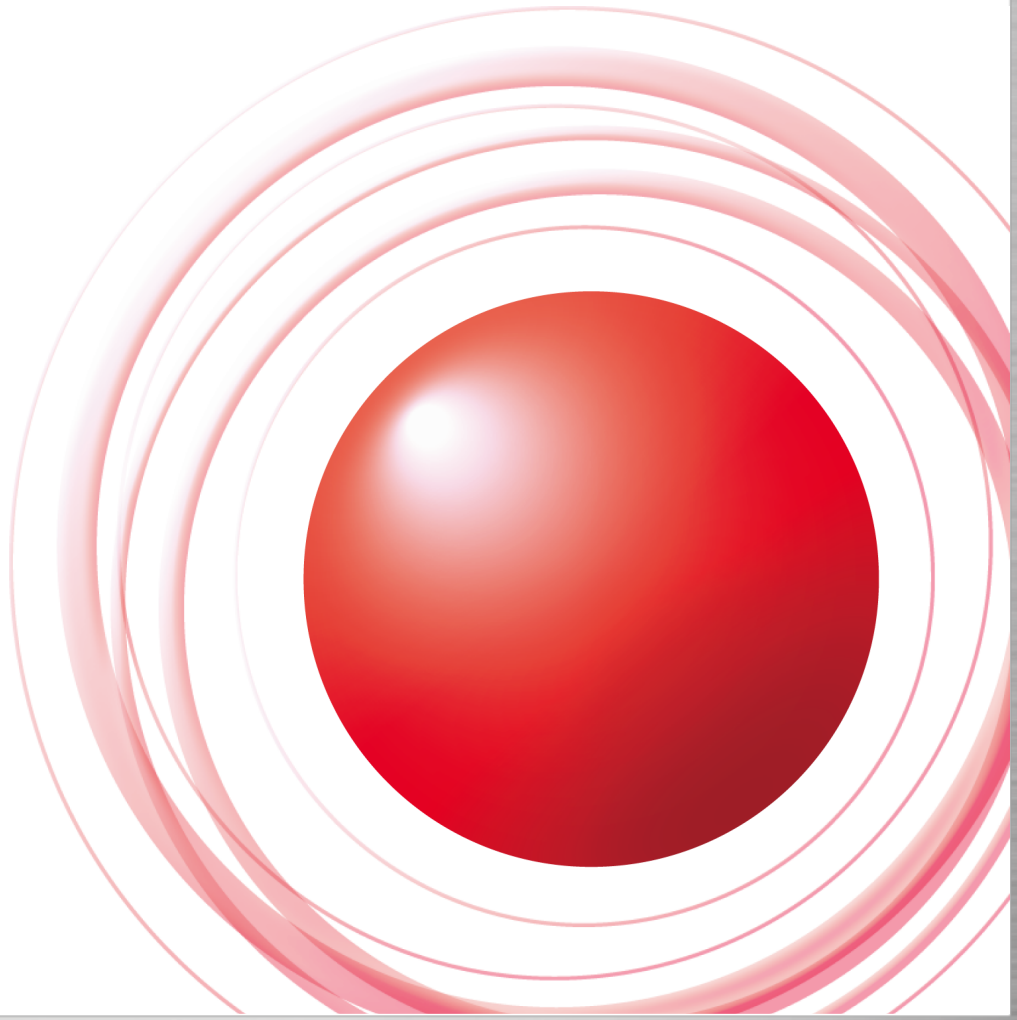
IIJ Technical WEEK 2011

IIJの分散処理プラットフォーム「dplat」について

IIJ Internet Initiative Japan



2011/11/09

株式会社インターネットイニシアティブ
前橋 孝広

Ongoing Innovation

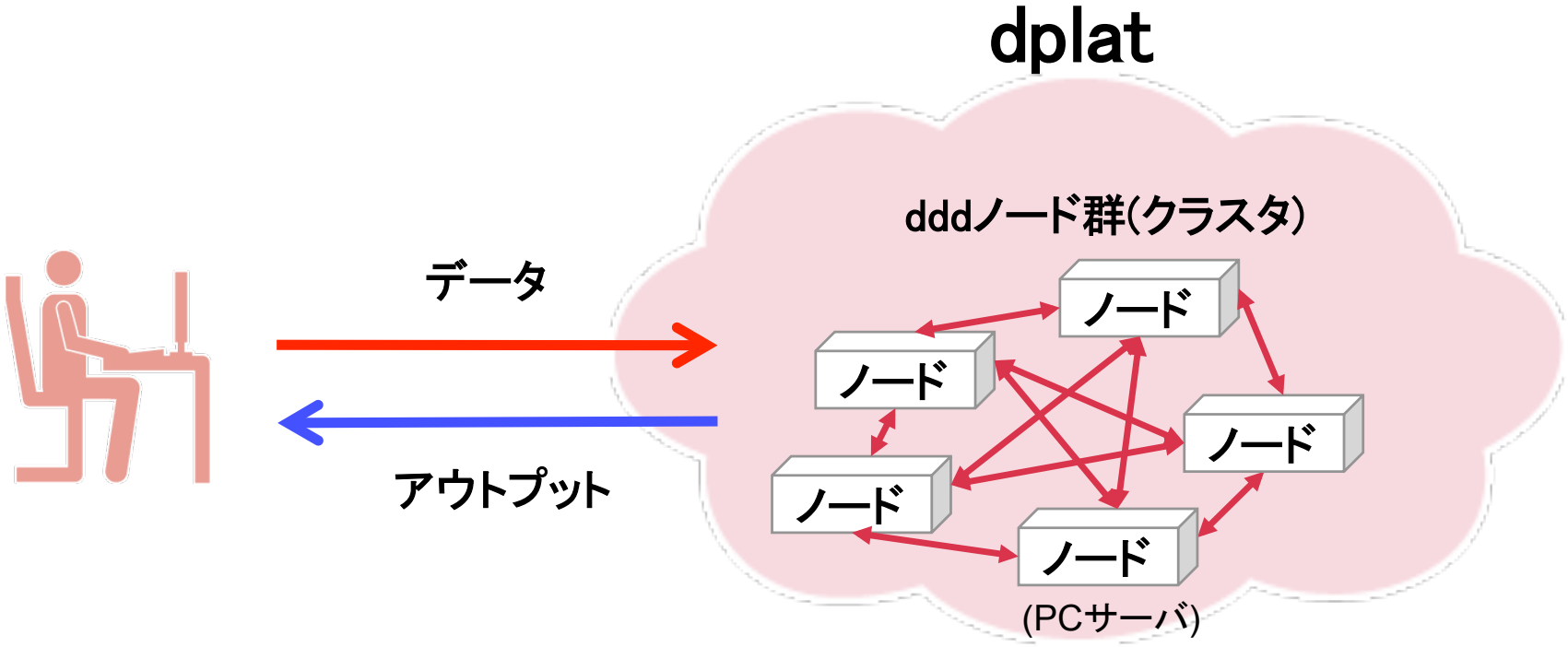
本日の内容

- dplat とは何か?
  - 概要
  - 技術要素
- 広域分散とディザスタリカバリ
- 松江データセンターパークについて

dplatの概要

dplat とは?

- 大量のデータを保持・処理するための基盤
- IIJ社内に対し、分散システムのプラットフォームを提供する



dplatの特徴

- IIJ独自開発の分散システム ddd を使用
- スケーラブル
  - 大量のデータを保持し、高速に処理できる
  - ノードを動的に追加することで容量や処理能力をUP
- 高可用性
  - ノードの一部が故障しても全体としては動き続ける
  - 地理的に離れた場所に分散配置されている
- 自分たちのニーズを満たすために開発
  - 直接社外に対してサービスしているものではない
  - 社外向けサービスのバックエンド等として使っている

dddとdplatの関係

- ddd
  - 分散システムソフトウェア
- dplat
  - dddを利用したプラットフォーム

dplatが提供するサービス

- 分散ファイルシステム(ストレージ)
- 分散処理機能 MapReduce
  - 検索
  - 並べ替え
  - 集計
  - etc

用途 1: トラフィック解析システム

- 膨大な量のトラフィック情報を保持
- ほぼリアルタイムで解析・視覚化
- 用途
  - 障害対応
  - 攻撃対応
  - 設備増強の参考

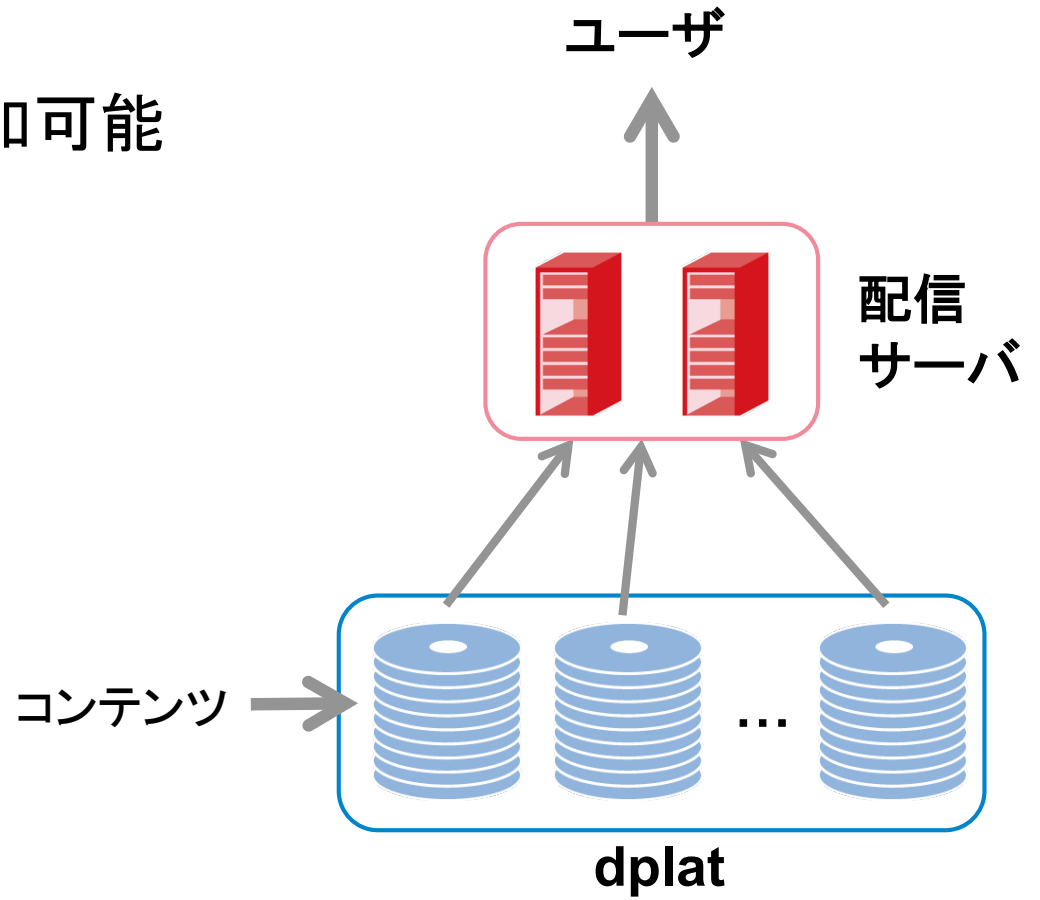
用途 2: レポートサービスバックエンド

- 顧客のログを保存、解析してレポート表示

例: セキュリティ系レポートサービス

用途 3: コンテンツ配信サービスのストレージ(予定)

- dplat を巨大コンテンツストレージとして利用
- dplatの以下の特徴を活用
  - 高い冗長性
  - どうてきに容量増加可能



プラットフォームの意味

- インタフェースを通じて(社内に向けて)データ処理サービスを提供する

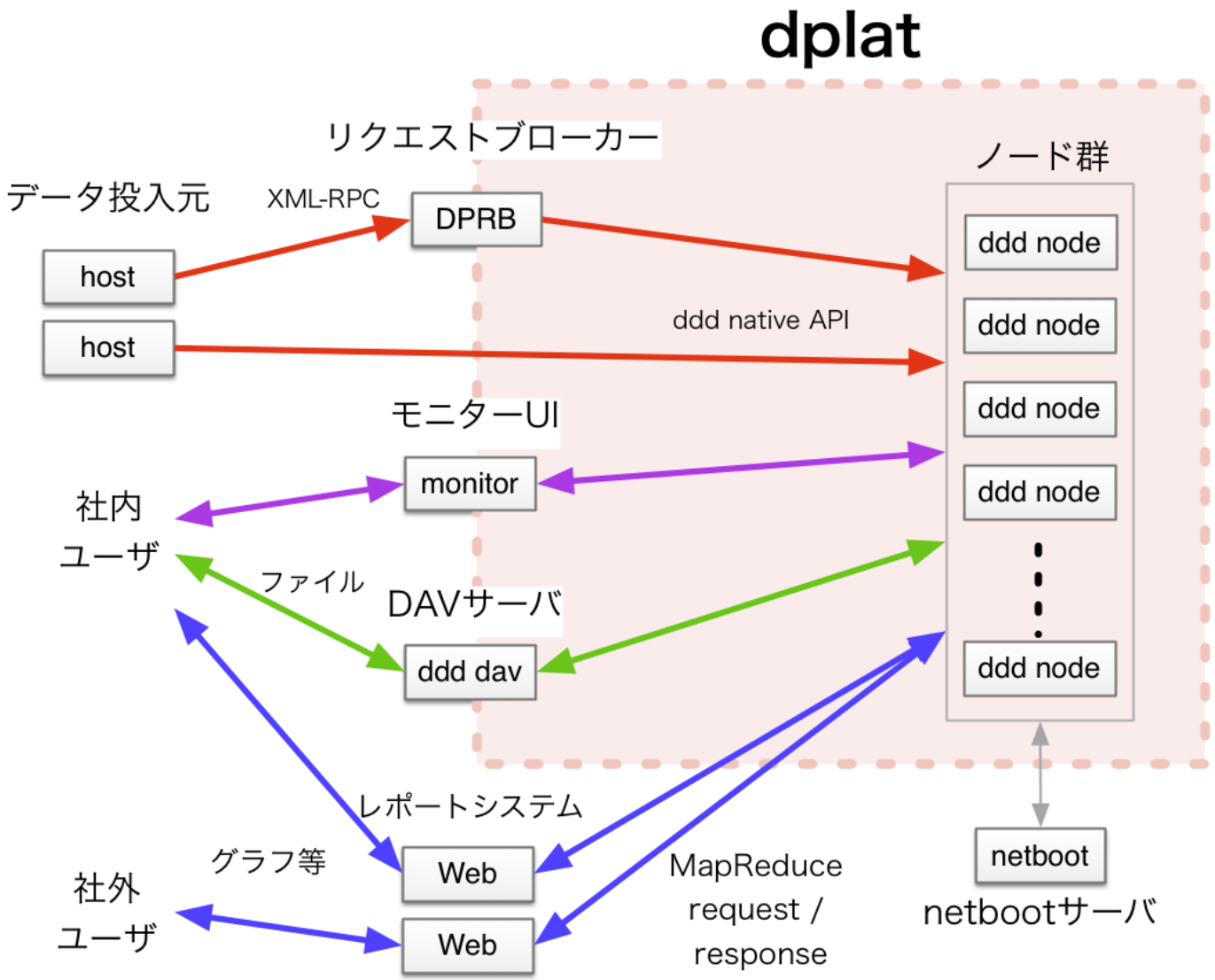
メリット

- プロジェクト毎に分散システムを構築しなくてよい
- とりまとめたほうがスケールメリットがある

デメリット

- プロジェクト毎に事情が異なるのでメンテナンスや機能追加がやりにくい

構成要素



用途により複数のクラスタに分割して使い分け

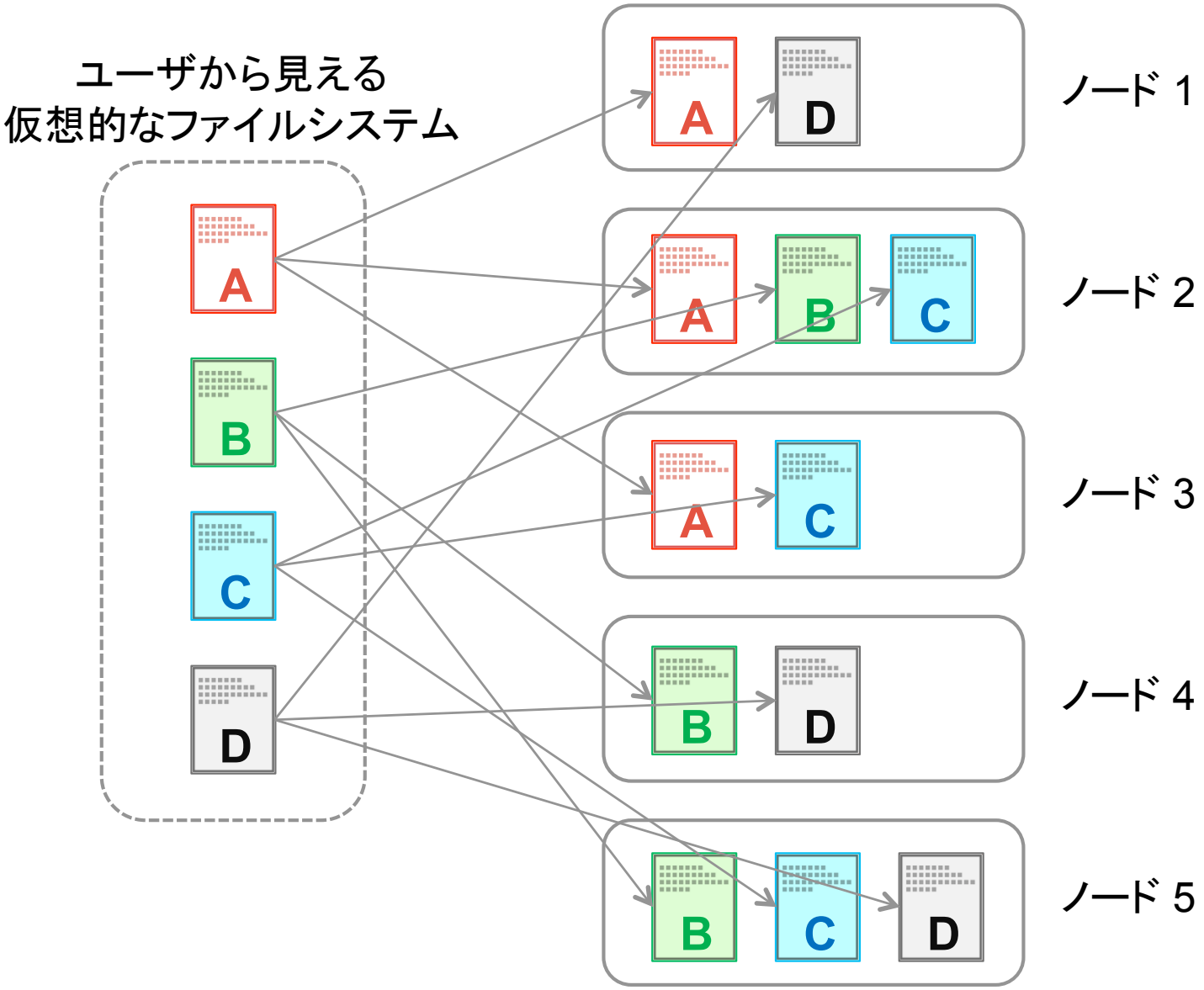
全体の実際の台数は非公表

技術要素

分散ファイルシステム

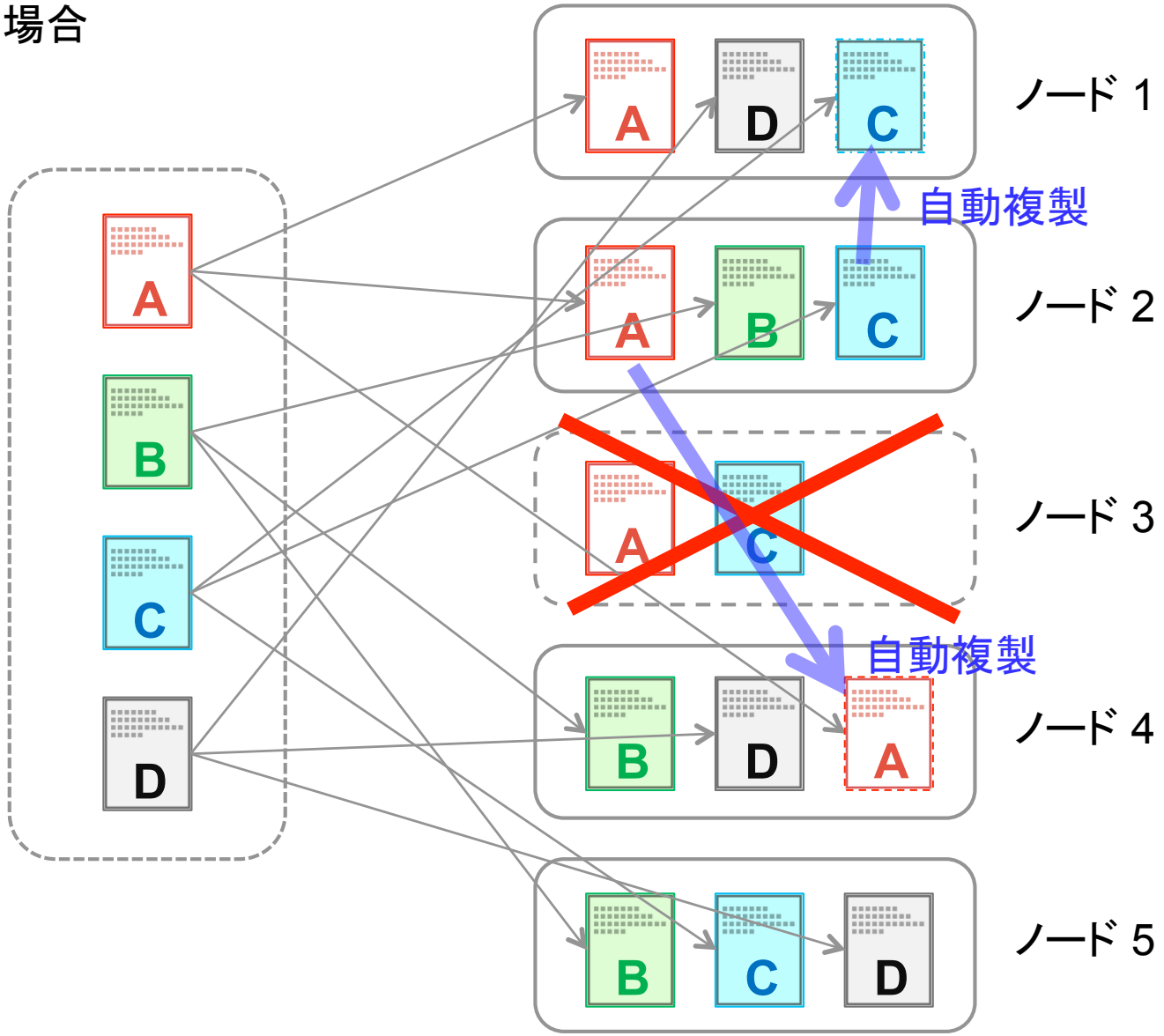
- 分散キーバリューストアをファイルシステムっぽく見せかけている
  - キー: ファイル名
  - バリュー: ファイル本体
- すべてのデータは異なる3つのノードに複製

すべてのファイルは異なる3つのノードに複製



一部のノードが故障しても他ノードから自動複製

ノード3が壊れた場合

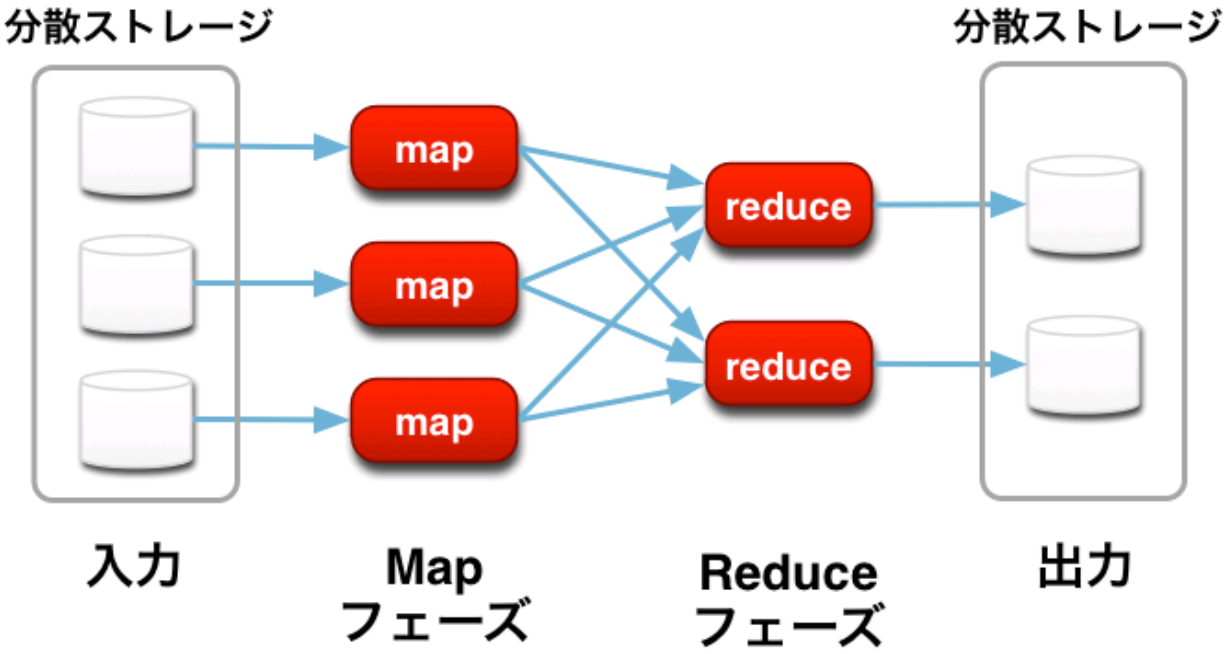


整合性の問題

- 分散ストレージには整合性の問題がつきまとう
- dddでは…
  - ファイルを一度書き込んだら変更不可(write once)
  - ファイルを消して書きなおすことは可能
- 結果整合性
  - ファイルを消して書き直した直後等、ノード間のデータの整合性が崩れることがある
  - 十分に時間が経てばすべての複製を含めたデータの一貫性が保たれる

並列分散処理フレームワーク MapReduce

- ひとつのジョブを多数のタスクに分割して並列実行
- 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・集計



広域分散とディザスタリカバリ

大災害のリスク

- dplatもデータを預かるサービスである以上、データの喪失はなんとしても避ける必要がある

- 東日本大震災以降、情報システムのディザスタリカバリに再考が求められる

南三陸町の戸籍データ消失、法務局保存分も水没

東日本巨大地震で被災した宮城県南三陸町で、戸籍の全データが津波で消失した可能性が高いことが19日、明らかになった。
（2011年3月20日　読売新聞）

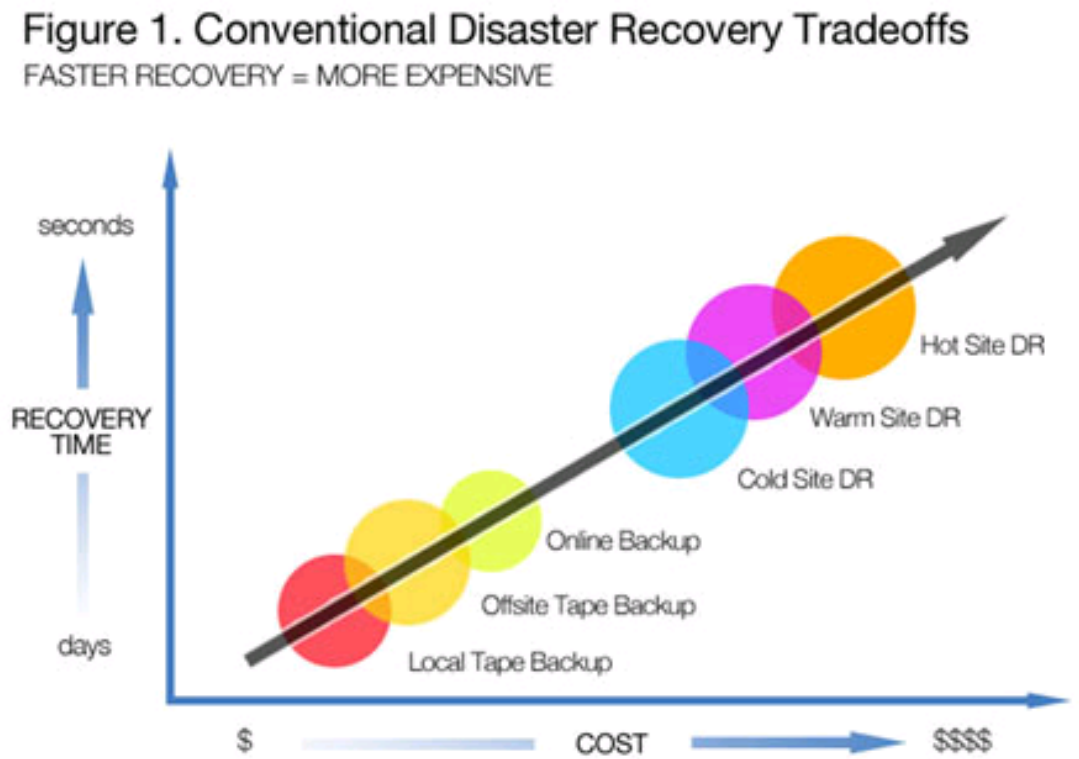
– (後に約1年前の副本データが残っていることが判明)

ディザスタリカバリのポイント

- データの保全
  - バックアップ先との地理的距離
  - バックアップ間隔
- システムの継続
  - データ保全を前提とし別サイトでシステムを再稼働
  - システム復元の容易さ
  - 再稼働までの所要時間

ディザスタリカバリのトレードオフ

- 素早い回復を求めるとコストがかかる

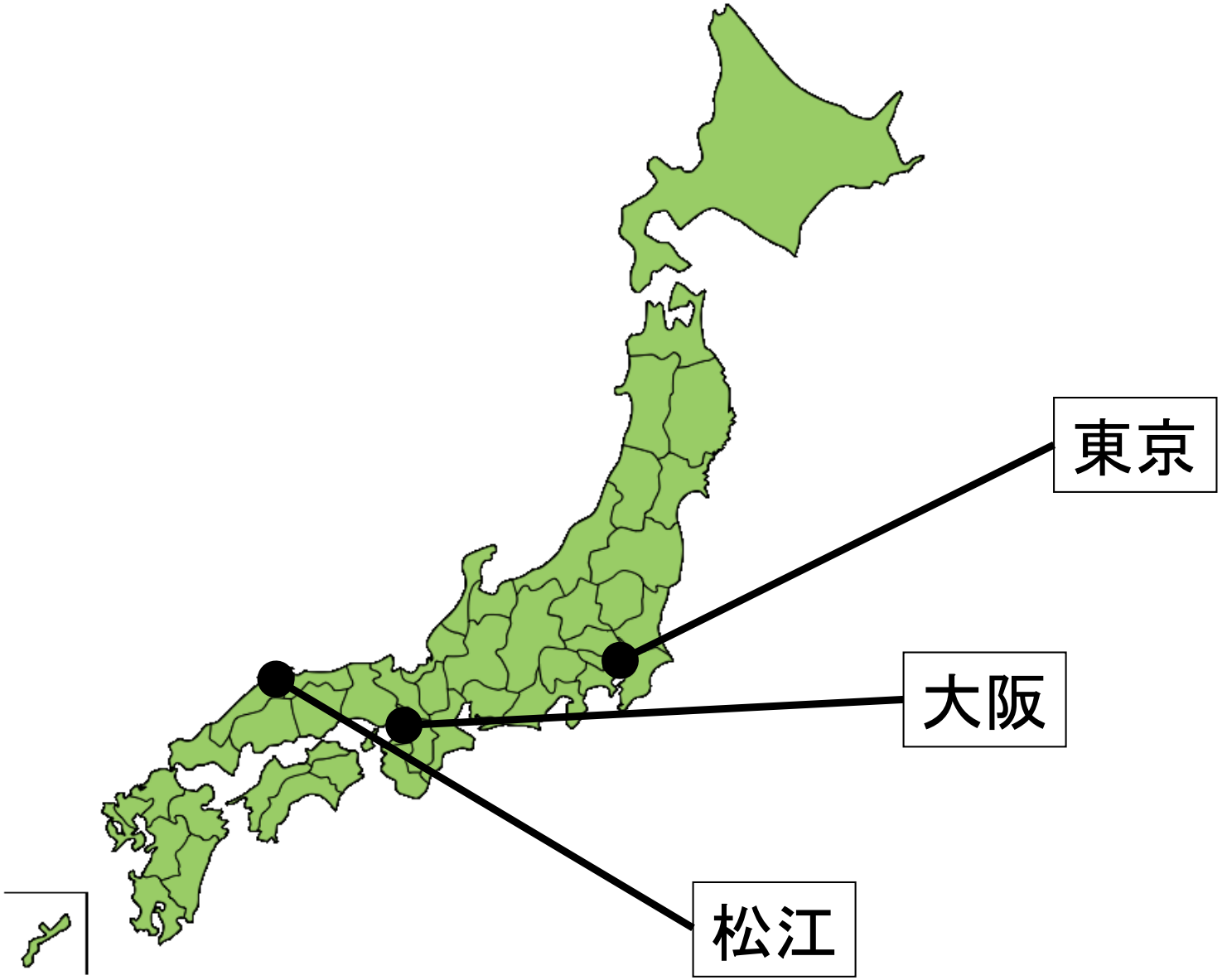


Data Center Knowledge: How The Cloud Changes Disaster Recovery
http://www.datacenterknowledge.com/archives/2011/07/26/how-the-cloud-changes-disaster-recovery/

dplatの広域分散(1)

- すべてのファイルは異なる3つのノードに複製
- さらに、3つのノードがすべて同一データセンターに属することを避けるように分散する
  - 各ノードは設置場所を属性として持つ
- 3つのノードはすべて対等であり、プライマリ／セカンダリといった区別はない

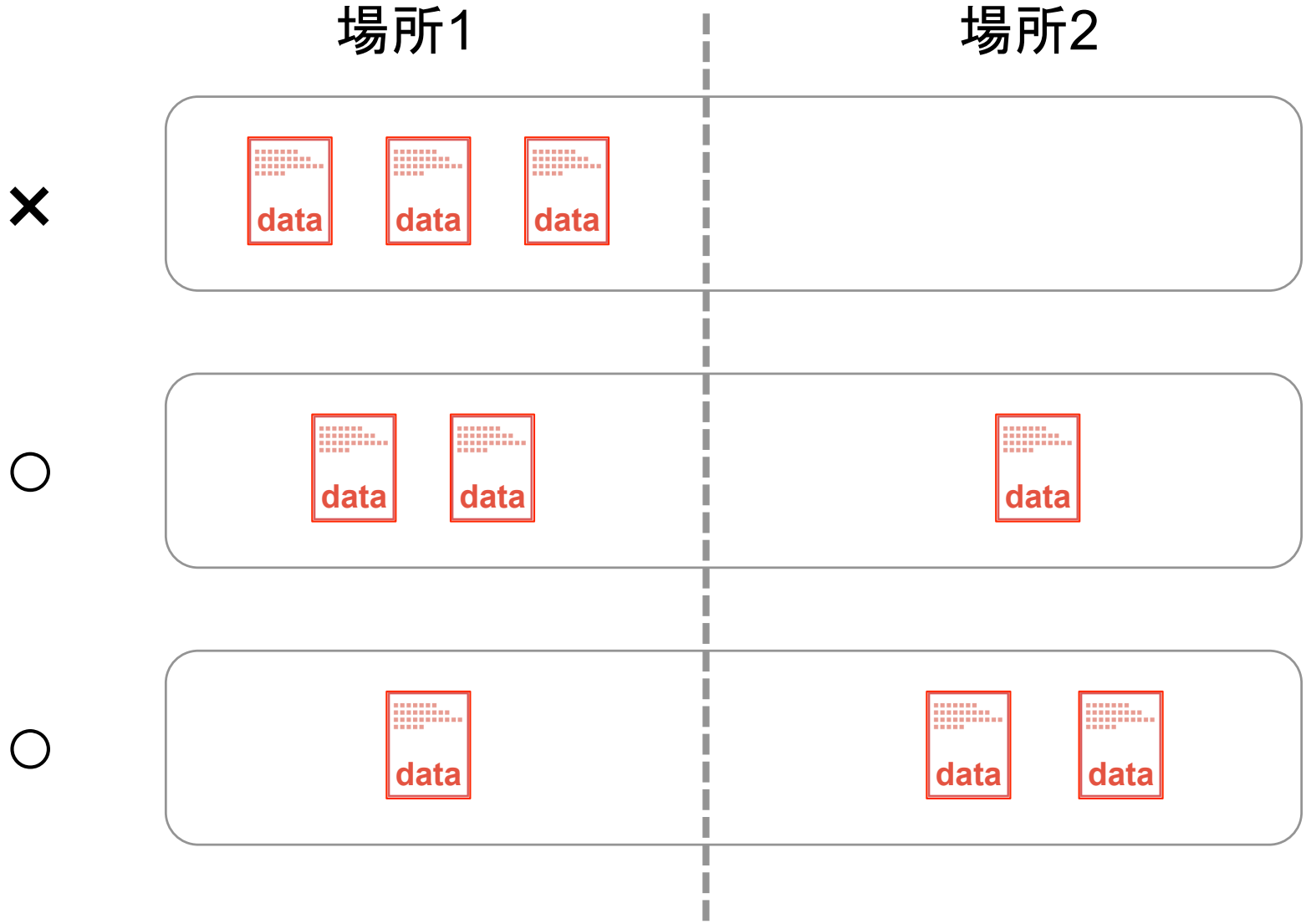
dplatノード群設置場所



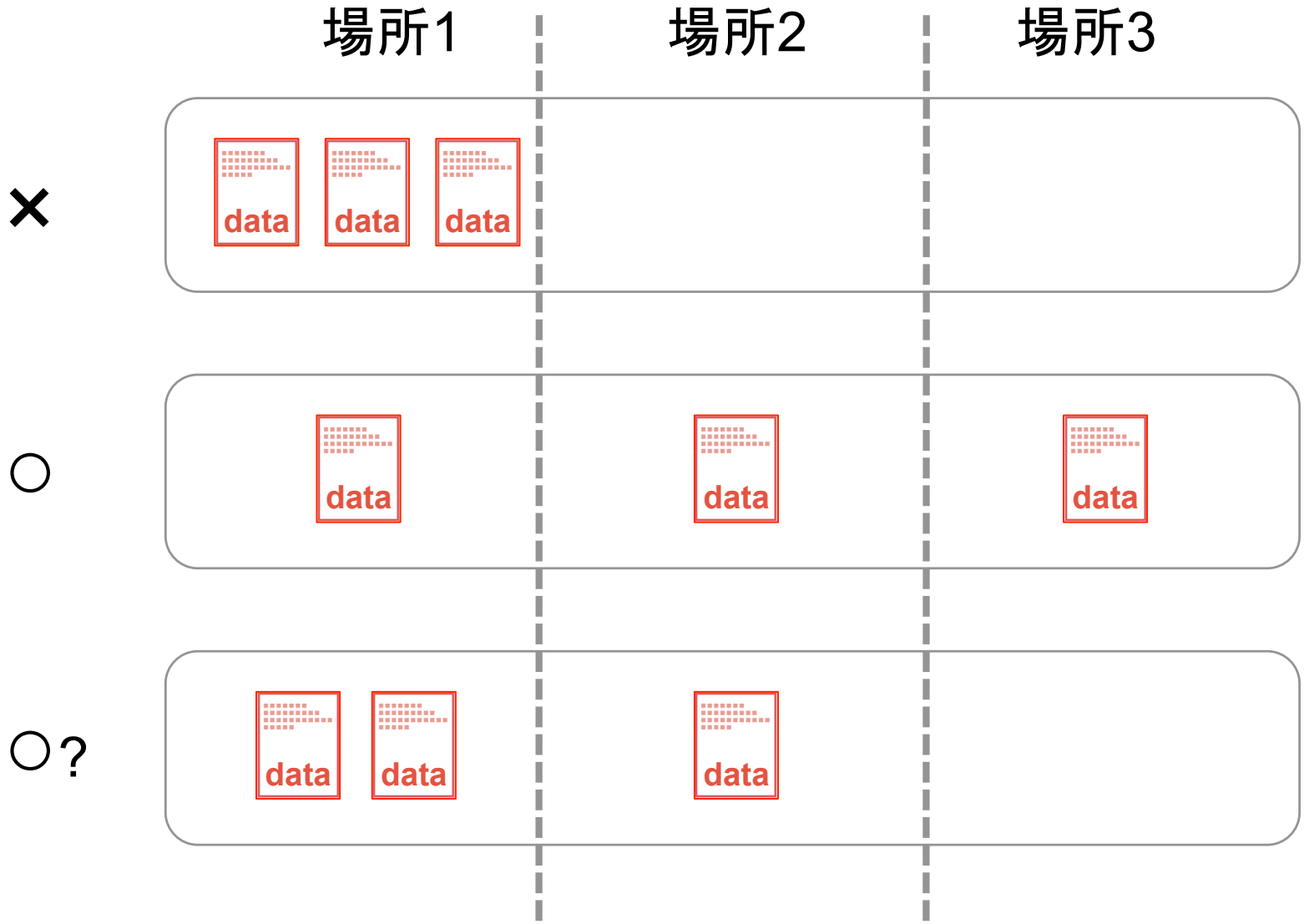
dplatの広域分散(2)

- 設置場所
  - 東京、大阪、松江
  - すべて電力会社が異なる
- コスト問題
  - サーバやハードディスクは安価な一般品を使用
  - コンテナと外気空調によりファシリティとランニングコスト低減
  - MapReduceの並列分散処理は全ノードを使用(スタンバイ機はない)

異なる場所に配置(場所が2箇所の場合)



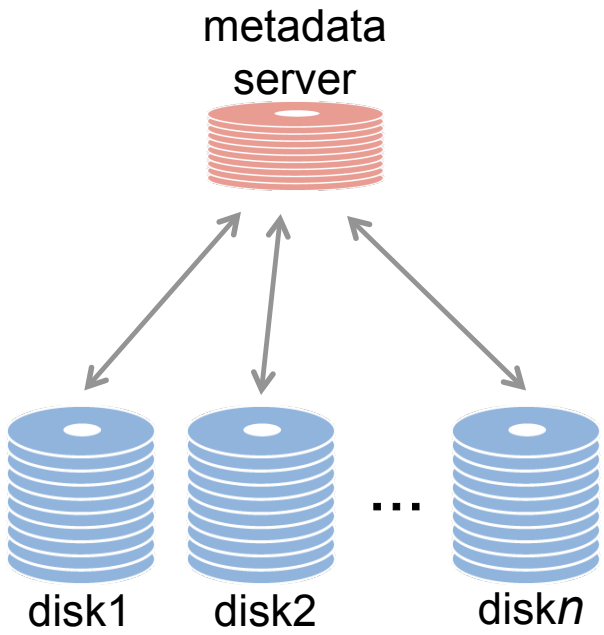
異なる場所に配置(場所が3箇所の場合)



分散配置時のノードの選択方法(1)

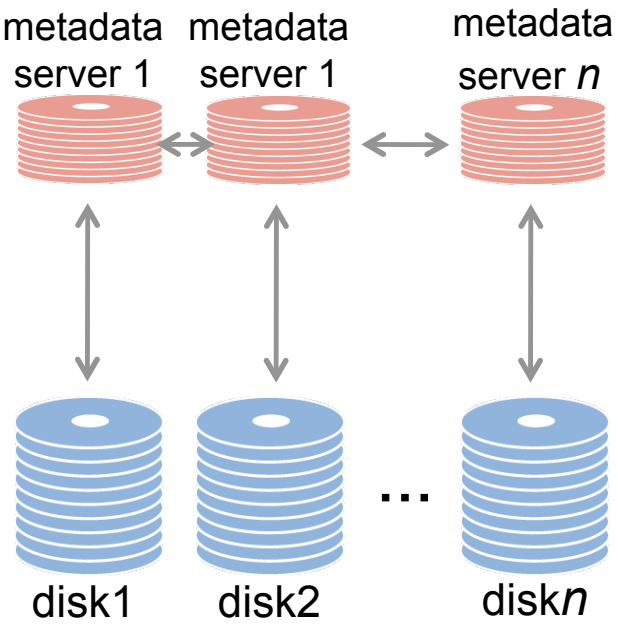
- メタデータ(metadata)を用いる方法

centralized metadata system



•単一障害点
•負荷集中

distributed metadata system

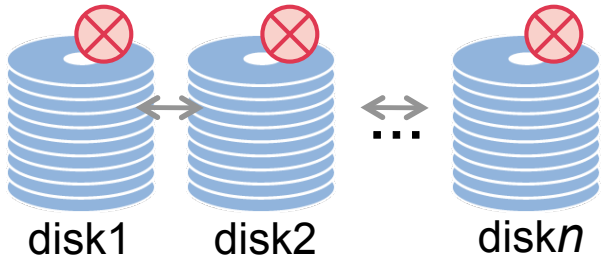


•metadataの整合性問題
•metadata間の通信量増大

分散配置時のノードの選択方法(2)

- アルゴリズムを用いる方法(dddの方法)
  - 配置場所はファイル名(パス名)によってアルゴリズムで決定
    - ノードの性能や空き容量無視(ただしノード毎にウェイトを設定可能)
  - 全ノードで同じアルゴリズムを使う

アルゴリズム:
コンシステントハッシュ法をベースにした独自のもの



松江データセンターパークについて

松江データセンターパークのコンテナ

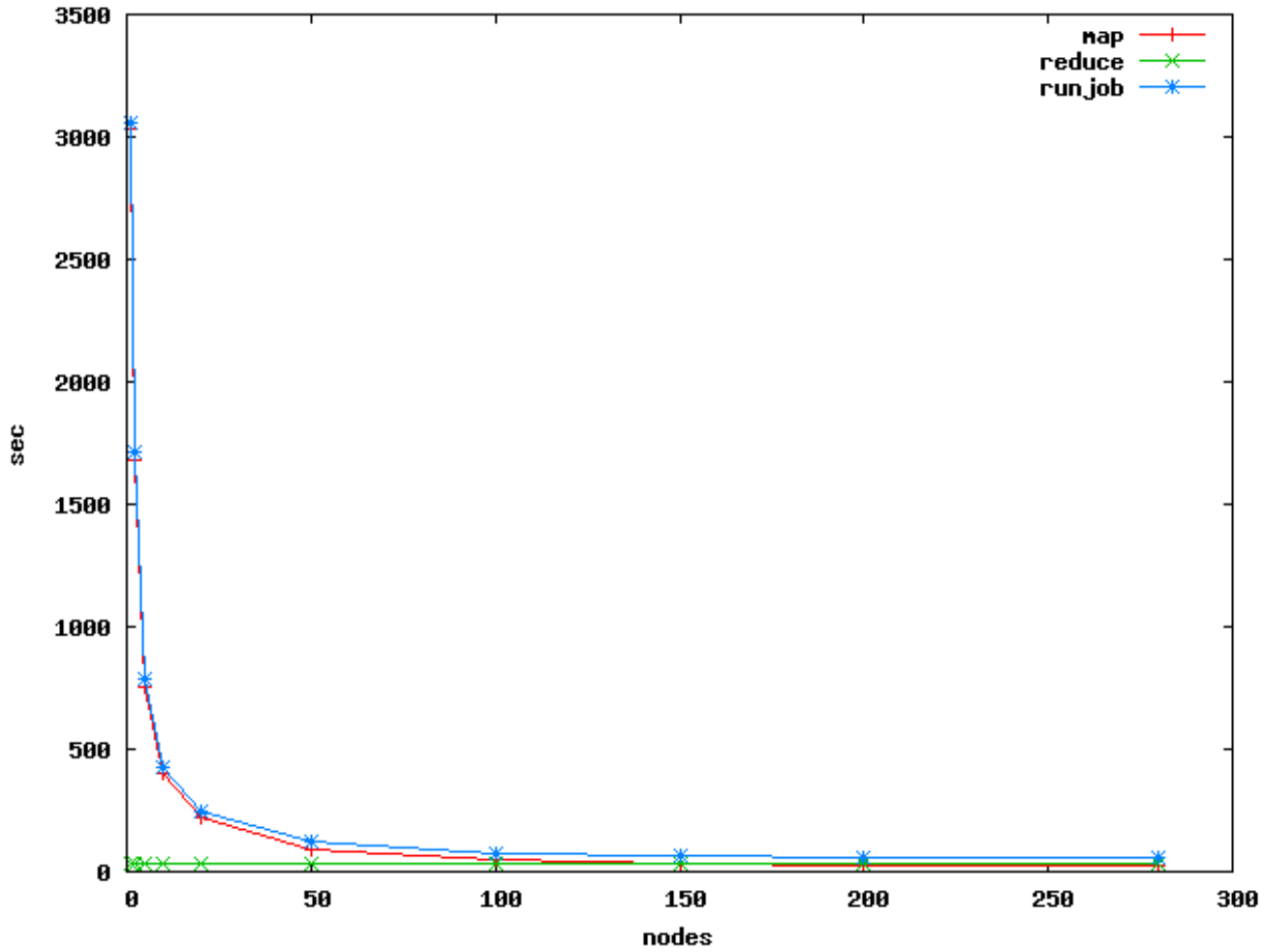
ITモジュール「IZmo(イズモ)」の中身

- IZmo S(スリム): ラックを傾斜配置
  - 他に IZmo W(ワイド)もあり

コンテナによるdddのスケーラビリティテストを実施

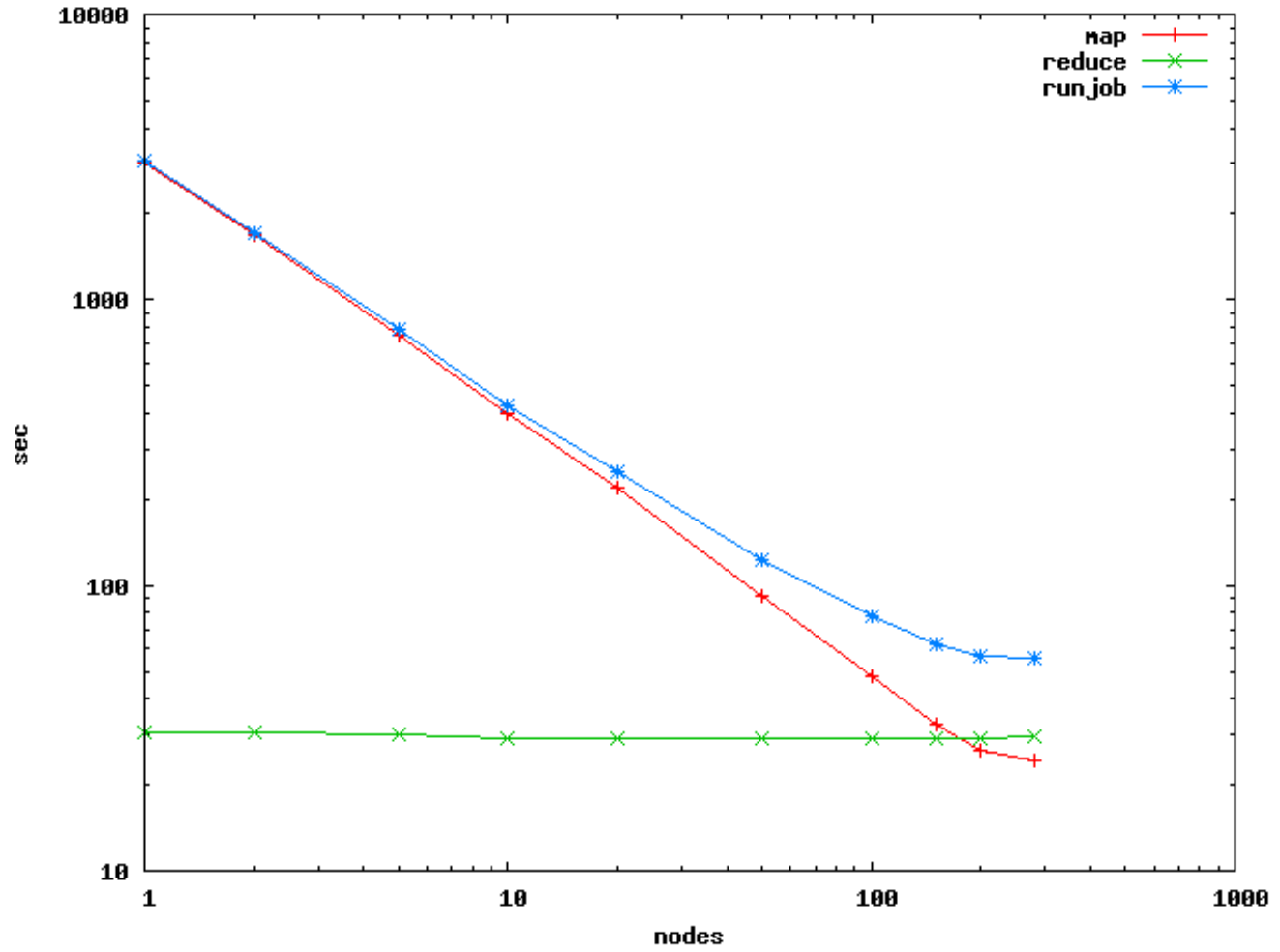
- ノード数を変化させつつ、一定のジョブを走らせ、処理時間がどのように変化するか
  - ノードが増えるほど処理時間は短くなるはず
- 処理内容
  - トラフィックデータを解析し、各ソースIPアドレスの出現数をカウント
- 2パターン
  - 2000ファイル、2000 map タスク
  - 100000ファイル、100000 map タスク

2000タスクの処理時間グラフ



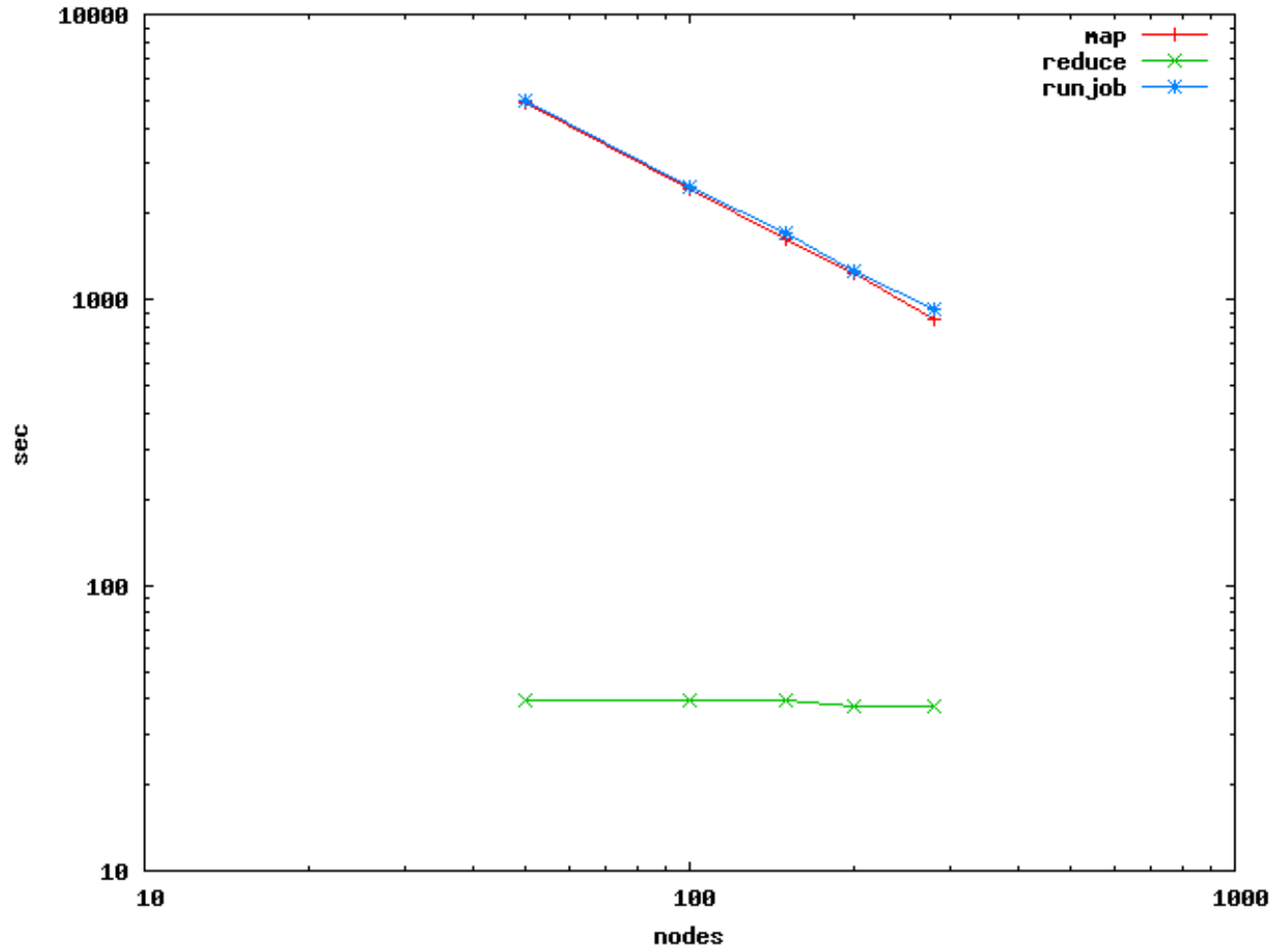
- 台数が増えると処理時間が減少

2000タスクの処理時間グラフ(対数)



- ノード数に対してタスク数が少ないと、暇なノードが出てくる

100000タスクの処理時間グラフ(対数)



- ほぼノード数に応じて処理時間が減少

温度や消費電力測定

- 大きく分けて2系統
- IPMI – サーバをモニタするための標準インタフェース仕様
  - 各部温度、消費電力、ファン回転数、電圧等
- IZmo管理システム
  - 温度、湿度、室外機運転台数、運転モード、ダンパ開度等

まとめ

- dplatは、IIJ社内向け分散システムプラットフォーム
- 広域分散によりデータ喪失を防ぎ、システムの継続稼働をめざしている
- 松江データセンターパークについて

IIJ Internet Initiative Japan

ご清聴ありがとうございました

お問い合わせ先　IIJインフォメーションセンター
TEL：03-5205-4466（9：30～17：30 土/日/祝日除く）
info@iij.ad.jp
http://www.iij.ad.jp/

Ongoing Innovation