# 自律ディスク上の分散ディレクトリの負荷均衡機構を用いたクラスタ再構築

東京工業大学大学院
情報理工学研究科
計算工学専攻 横田研究室
伊藤 大輔

東京工業大学
学術国際情報センター
横田 治夫

# 発表内容

- 研究の背景
  - 既存の手法の問題点
  - 自律ディスクのコンセプト
- 自律ディスククラスタのオンライン再構築
  - クラスタ再構築プロトコル
  - オンラインクラスタ再構築と排他制御
- 評価実験
- まとめ

# 研究の背景

- コンピュータ上で扱うデータが増加
  - 管理コストも増加
- 解決案：ストレージセントリック・コンフィグレーション
  - システムの中心にアーキテクチャ非依存のストレージを配置
    - ネットワーク接続ディスク
  - クライアントは全てのディスクにアクセス可能
  - ストレージの一元管理可能
- しかし、全ての問題を解決したわけではない

# 既存のストレージ技術の問題点

1. パッシブデバイスであること
   - ストレージはネットワーク越しのクライアントの命令によって動作
   - 管理する際の通信コストが高い
     - ミラーバックアップ
2. 特定の中央サーバへ依存していること
   - 通常はデータ配置を専用サーバで行う
   - ボトルネック (限られた拡張性) / SPOF

# 目的

- 既存のストレージセントリック・システムを改善:
  - 高い拡張性：ボトルネックの要因をなくす
  - 高い信頼性：複雑な中央制御を止める
- 自律分散制御が適している

**自律ディスク**

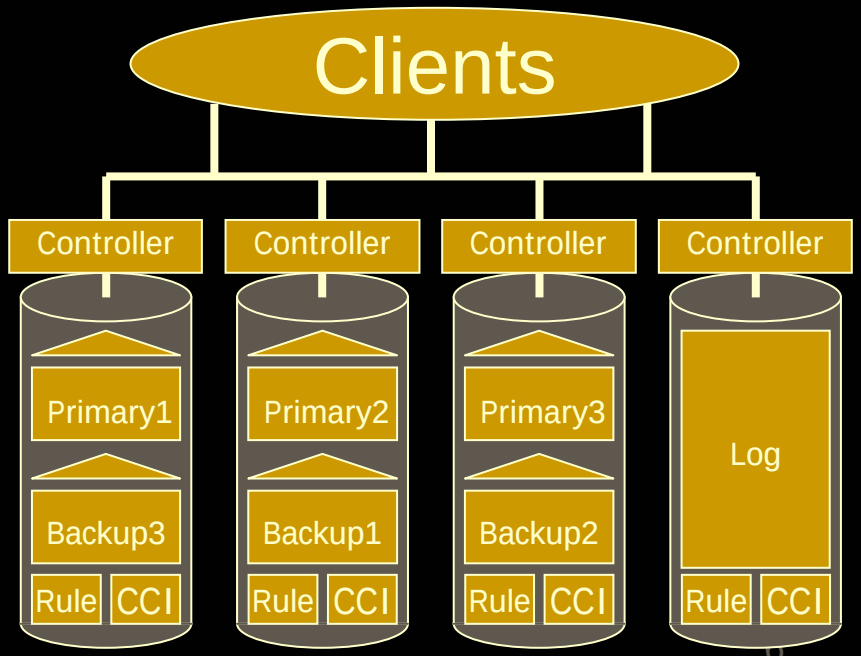
**ディスク上の演算能力を用いて**

**ディスク自体を自律分散制御**

# 自律ディスク

- 本研究と関連の強い六つの特性

1. ネットワーク上でストレージクラスタを構成
2. 高機能な分散ディレクトリを採用
3. ルールを用いてカスタマイズ可能なコマンド階層
4. ログを用いた非同期プライマリバックアップ
5. 全てのディスクがクラスタ構成情報 (CCI) を保持
6. ストリームインターフェース



# 発表内容

- 研究の背景
  - 既存の手法の問題点
  - 自律ディスクのコンセプト
- 自律ディスククラスタのオンライン再構築
  - クラスタ再構築プロトコル
  - オンラインクラスタ再構築と排他制御
- 評価実験
- まとめ

# オンラインクラスタ再構築の必要性

- 大規模ストレージの運用に伴うイベント
  - 仮想化されたクラスタ同士のバンド幅 / 容量の変更
  - ストレージプールの拡張
  - ディスク故障時の交換
- これらのイベントはクラスタ構成の動的な変更を必要とする
- メンテナンスに伴うダウンさえ許されないシステム

# 自律ディスクのクラスタ再構築

- クラスタの再構築を以下の2パートに分類
  1. クラスタ構成情報 (CCI) の更新
  2. クラスタメンバ間のデータ移動
- データ移動は分散ディレクトリを用いて対応
  - 負荷均衡化機構を利用
    - 再構築に伴うデータ移動 → 負荷の極端な偏りの解消
  - ディレクトリは細粒度のロックを取れる
- CCIの更新を全メンバで同期させる必要
  - 同期を保証するコーディネータが必要
  - 自律ディスクはサーバレスクラスタ

# クラスタ再構築プロトコル

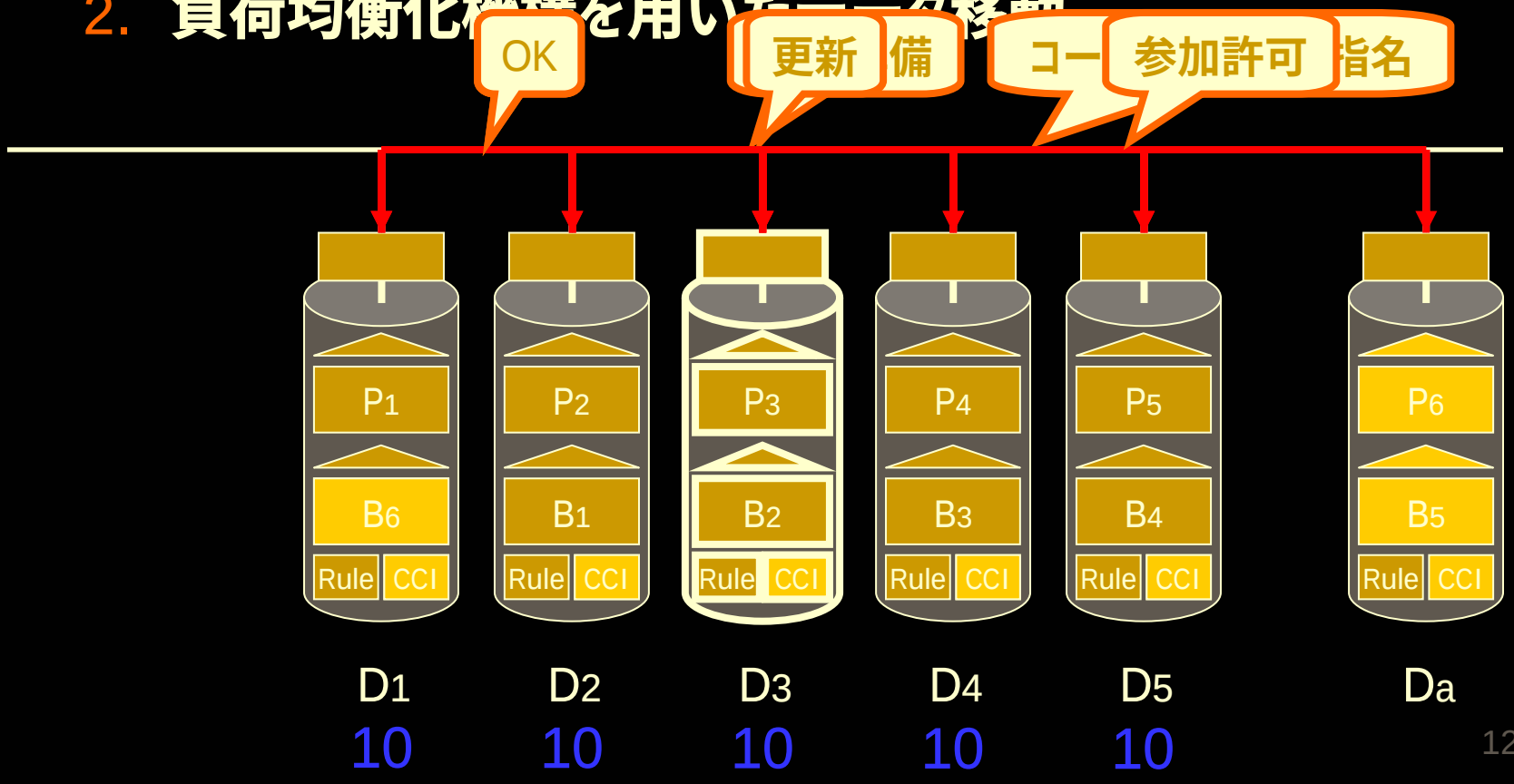
- **クラスタ再構築プロトコル** [PRDC2001, DEWS2001]
  - CCIの更新を同期させるプロトコル
  - 想定したクラスタ再構築を8つに分類し、それぞれ専用のプロトコルを定義
  - 同期手法：ダイナミックコーディネータを用いた多重フェーズ構成同期 (cf. 2PC)
    1. 動的にコーディネータを立てる
       - 再構築の要因となるディスクが任意のディスクを指名
    2. コーディネータがメンバーの合意を取る (prepare-commit)
    3. コーディネータが更新を指示する (commit)
- **サーバレスクラスタで保証付きの同期更新を実現**

# 例：データディスク追加

- 負荷均衡機構 (skew handler) の動作
  - トークンバス方式
    - より効率的な方式も存在するが、1stステップとして実装
  - クラスタ性能ベースの重みを考慮し、インターバルを用いる

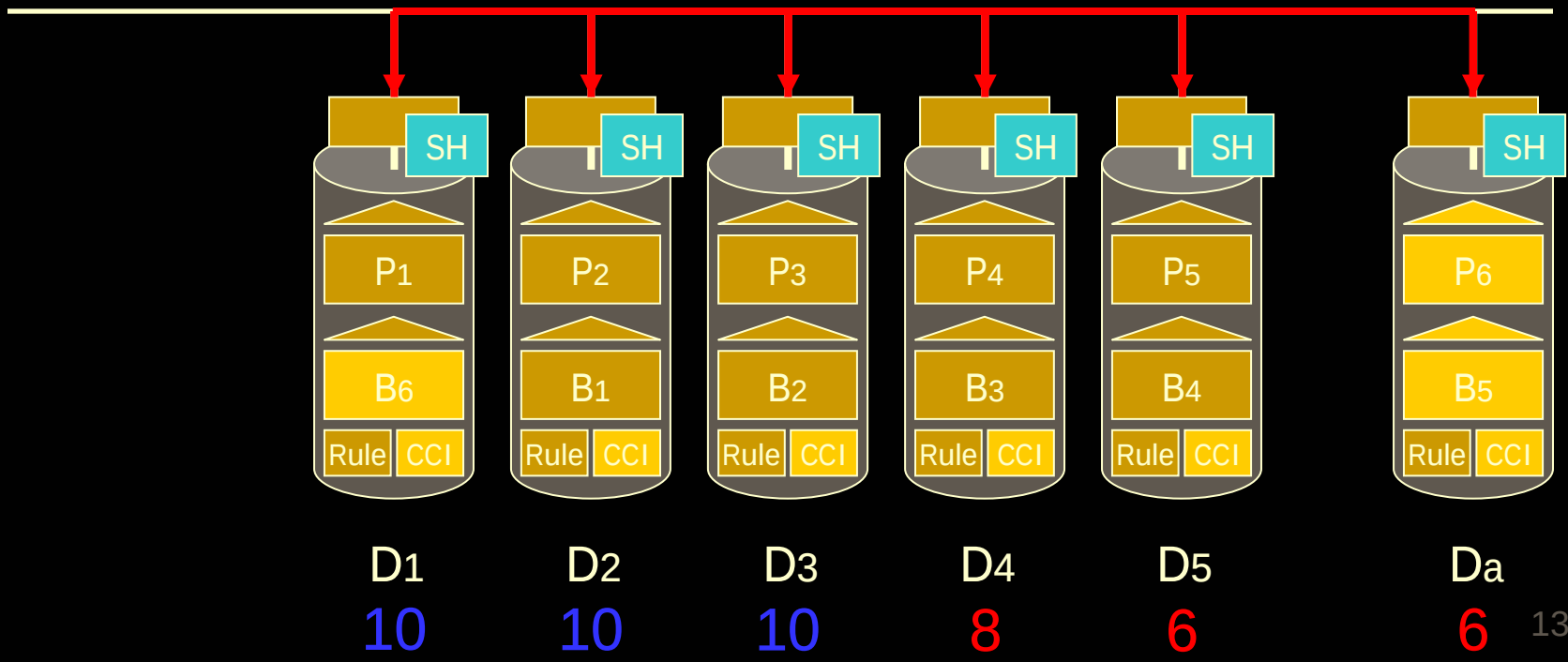
# 例 :データディスク追加

- ディスクDaをクラスタに追加
  1. ダイナミックコーディネータを用いた多重フェーズ構成同期
  2. 負荷均衡化機構を用いたデータ移動



# 例 :データディスク追加

- ディスクDaをクラスタに追加
  1. ダイナミックコーディネータを用いた多重フェーズ構成同期
  2. 負荷均衡化機構を用いたデータ移動



# クラスタ再構築所要時間の見積もり

- 一台のデータディスクを追加

$$
\begin{aligned}
T_{\mathrm{ADisk}} &= T_{\mathrm{ADisk,\ sync}} + T_{\mathrm{ADisk,\ migrate}} \times A_{\mathrm{leaf}} \ / \ 2 \\
&\cong T_{\mathrm{ADisk,\ sync}} + (T_{\mathrm{ADisk,\ insert}} + T_{\mathrm{ADisk,\ delete}} + T_{\mathrm{interval}}) \times A_{\mathrm{leaf}} \ / \ 2
\end{aligned}
$$

# 発表内容

- 研究の背景
  - 既存の手法の問題点
  - 自律ディスクのコンセプト
- 自律ディスククラスタのオンライン再構築
  - クラスタ再構築プロトコル
  - オンラインクラスタ再構築と排他制御
- 評価実験
- まとめ

# 評価実験

- 自律ディスクのクラスタ再構築の優位性の検証
  - 見積もり通りの性能が出ているか?
  - 再構築中にクライアントに与える影響 (性能劣化) がどの程度生じるか?
- 実験用模擬システムを用いて行う

# 実験システム

- PC上にJavaを用いて模擬自律ディスクを実装
- 実験システムのスペック
  - Intel Pentium3 933MHz
  - Linux 2.2.17, IBM JRE1.3.0
  - 100base-TX NIC with switched network
- 諸条件
  - ページサイズ 8KB, 更新に用いるタプルサイズ 4KB
  - データディスク1台あたりの最大葉ページ数 500ページ
  - ログディスク1台、単一バックアップ
  - バックアップの同期は20秒毎 (ただし負荷に応じて動的に調整)

# クラスタ再構築所要時間

- クラスタ再構築にかかわる諸条件
  - クラスタにあらかじめ400タプルのデータを均等に格納
  - 負荷均衡化機構のインターバルは50msまたは500ms
  - 偏り検出の閾値は5 (ディスク1台あたりの最大葉ページ数の1%) → 約200ページ移動

# 基本性能と見積もり

- 更新 (挿入) 操作のレスポンスタイム (データディスク6台)

$$T_{\mathrm{ADisk,\ insert}} = 6.3\,\mathrm{ms}$$
$$T_{\mathrm{ADisk,\ delete}} = 7.4\,\mathrm{ms}$$

- CCIの同期更新の所要時間 (データディスク6台)

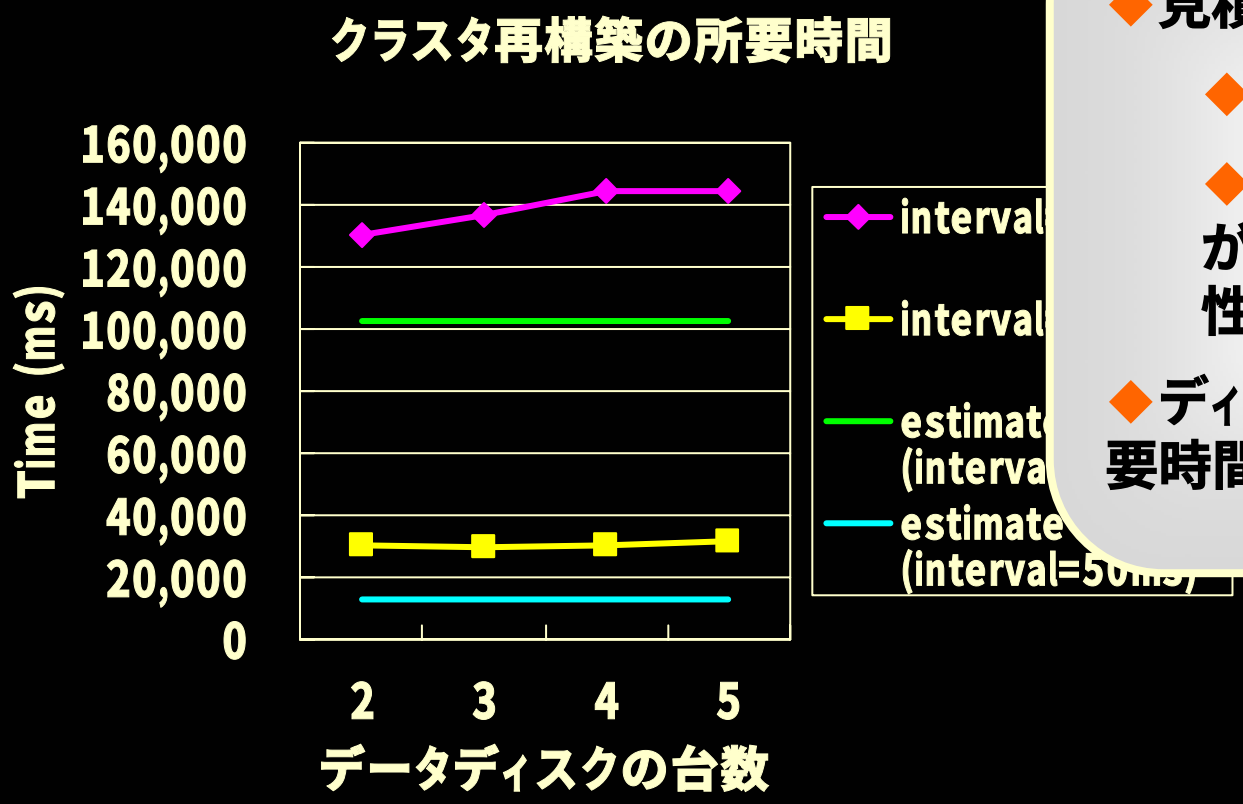
$$T_{\mathrm{ADisk,\ sync}} = 24\ \mathrm{ms}$$

- 見積もり
  - クライアントからの負荷を与えない場合

$$T_{\mathrm{ADisk}} \cong \begin{cases} 12,800\ [\mathrm{msec}] & (T_{\mathrm{interval}} = 50\mathrm{ms}) \\ 102,740\ [\mathrm{msec}] & (T_{\mathrm{interval}} = 500\mathrm{ms}) \end{cases}$$

# クラスタ再構築所要時間の実測

◆クライアントからの負荷を与えない場合




◆見積もりより悪化
- ◆ログリプライの影響
- ◆ディレクトリのロックが衝突している可能性

◆ディスク台数と再構築所要時間は無関係

# クラスタ再構築所要時間の実測

- クライアントから負荷を与えた場合 (オンラインクラスタ再構築)
  - クライアントは1台、毎秒20insert
  - スループット確保のため、最大構成のクラス
  - 負荷均衡化機構には50msのインターバ

再構築所要時間

| 条件 | 総移動葉ページ | 所要時間 (ms) |
|---|---|---|
| 負荷あり | 340 | 115,42 |
| 負荷なし | 200 | 31,37 |

再構築中の性能

| 条件 | レスポンスタイム (ms) | スループット (Mbps) |
|---|---|---|
| 再構築中 | 6.80 | 19.4 |
| 平常時 | 6.28 | 19.0 |

- 再構築所要時間はさらに悪化
  - コントローラ内のスレッド優先順位問題
- クライアントから見た性能低下は殆どなし
- 性能を劣化させないオンラインクラスタ再構築が実現可能

# まとめ / 今後の課題

- まとめ
  - クラスタ再構築機能を模擬自律ディスクへ実装
    - オンライン再構築機能を実現
    - 再構築に伴うパフォーマンスの低下を殆ど与えないことを実証
- 今後の課題
  - クラスタ再構築プロトコルの他への応用
    - サーバレスクラスタ内での保証付き同期更新
    - 設定ファイルに基づいて動作する無共有クラスタに有効
  - 実装上の問題の解決
    - 負荷均衡処理に伴うロックの問題

ご清聴を感謝します