

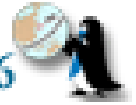
MySQL Clusterの最適構成

住商情報システム
IT基盤ソリューション事業部
オープンソースシステム部

廣濱 顕司

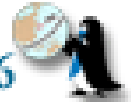
目次

1. DBMSのクラスタリング
2. MySQLとMySQL Clusterのアーキテクチャ
3. IPAプロジェクトでの検証結果
4. まとめ



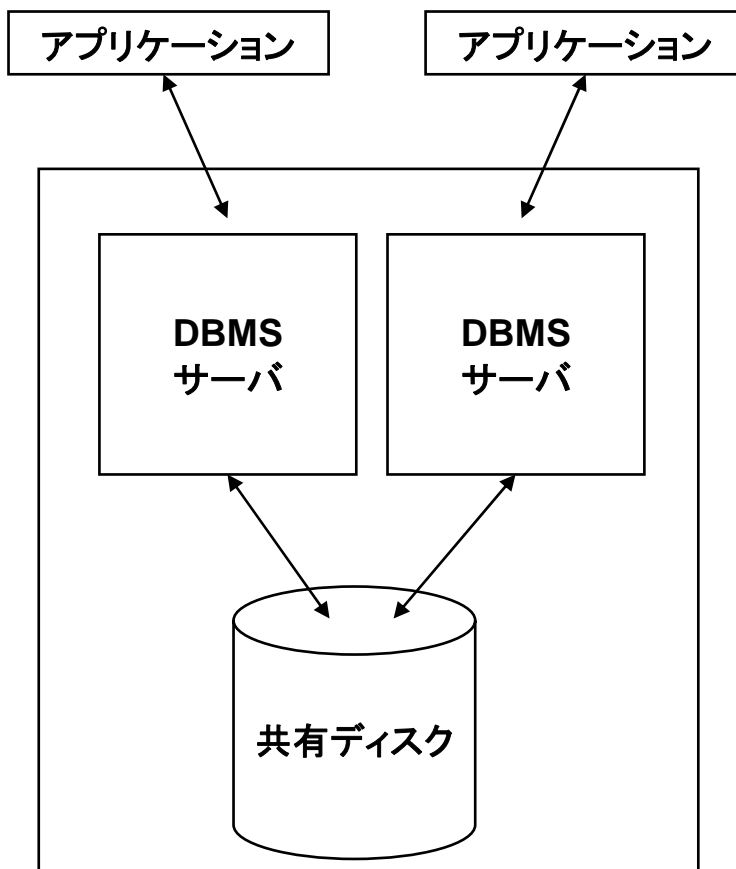
1. DBMSのクラスタリング

- DBMSクラスタリングのアーキテクチャ
 - 共有型
 - 非共有型



DBMSのクラスタリング:共有型

□ 共有型



長所

- ・データを一元管理できる

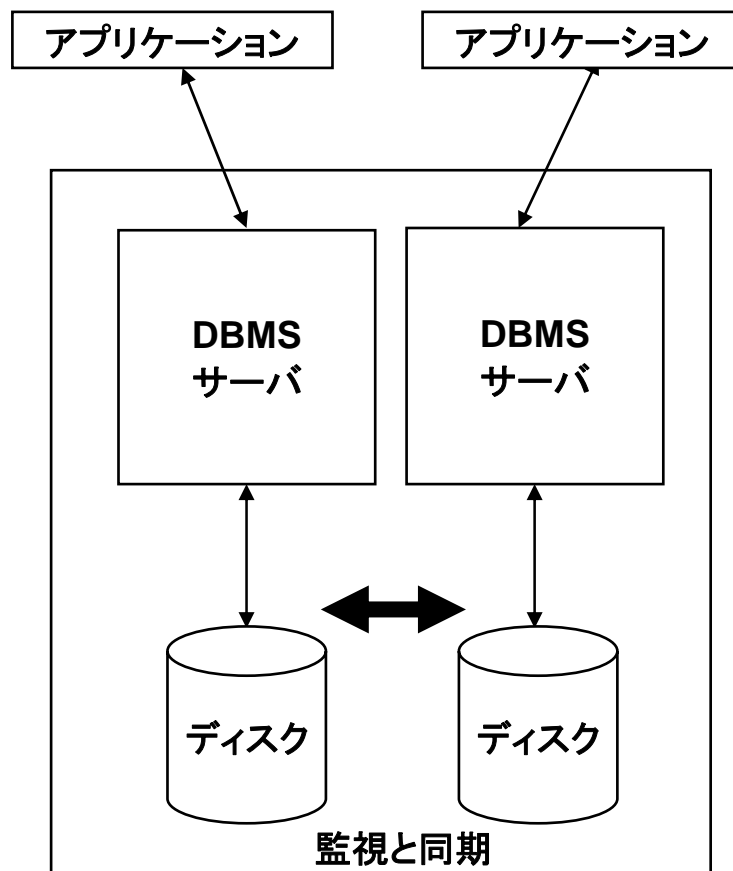
短所

- ・共有ディスクがSPoFとなり得る
- ・共有ディスクの排他制御が必要



DBMSのクラスタリング:非共有型

□ 共有型



長所

- ・H/WにSPoFが無い
- ・共有ストレージなど高価なH/Wを必要としない

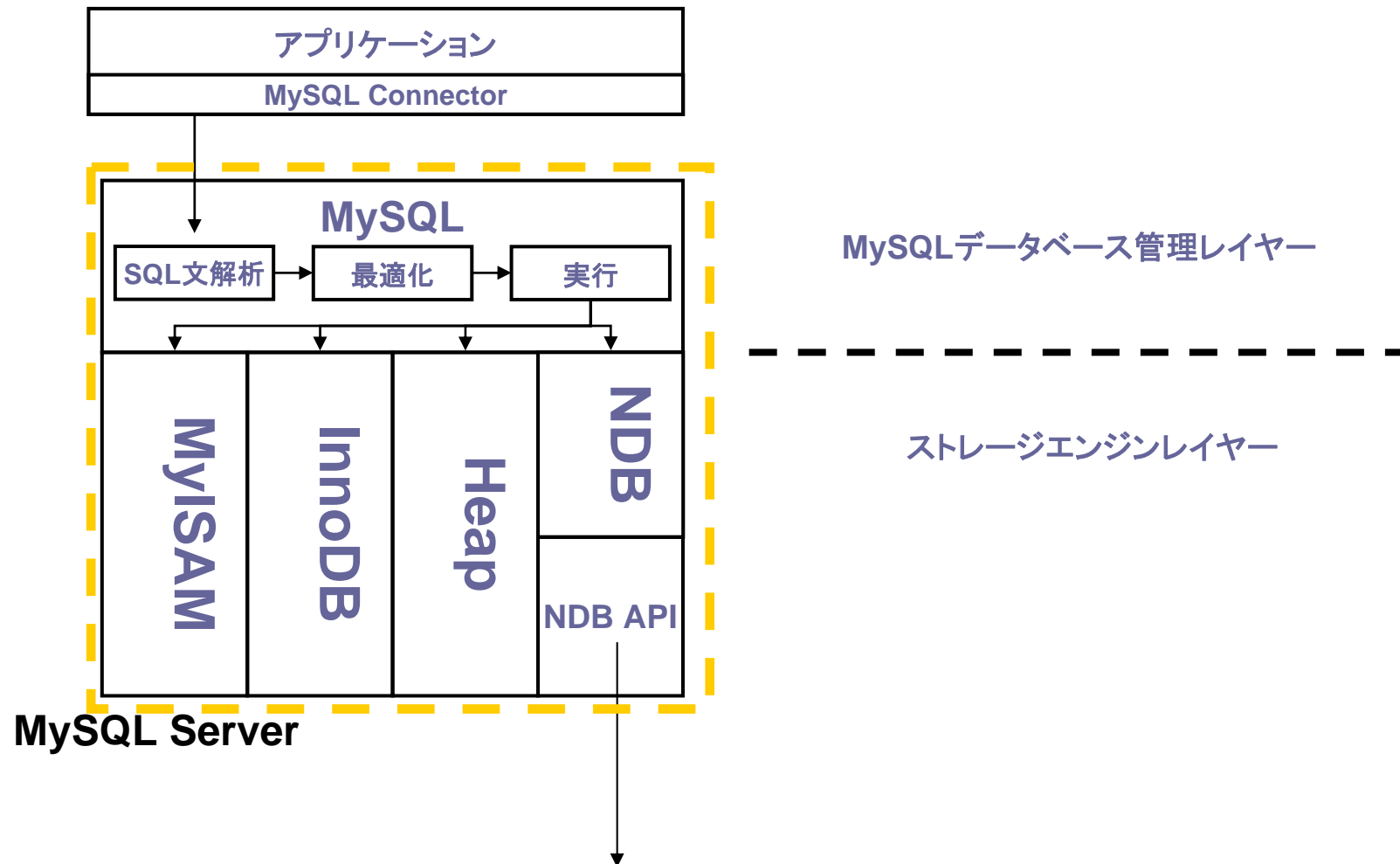
短所

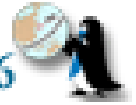
- ・複数のディスクで同期をとる必要がある
- ・インターコネクト上に大量データが流れる可能性がある

2. MySQLとMySQL Clusterのアーキテクチャ

- MySQL
- MySQL Cluster
 - Data Node内のデータ配置
 - レプリカとノードグループ
 - 同期レプリケーションと二相コミットメント

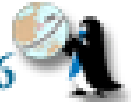
まずは、MySQLのストレージエンジンアーキテクチャ



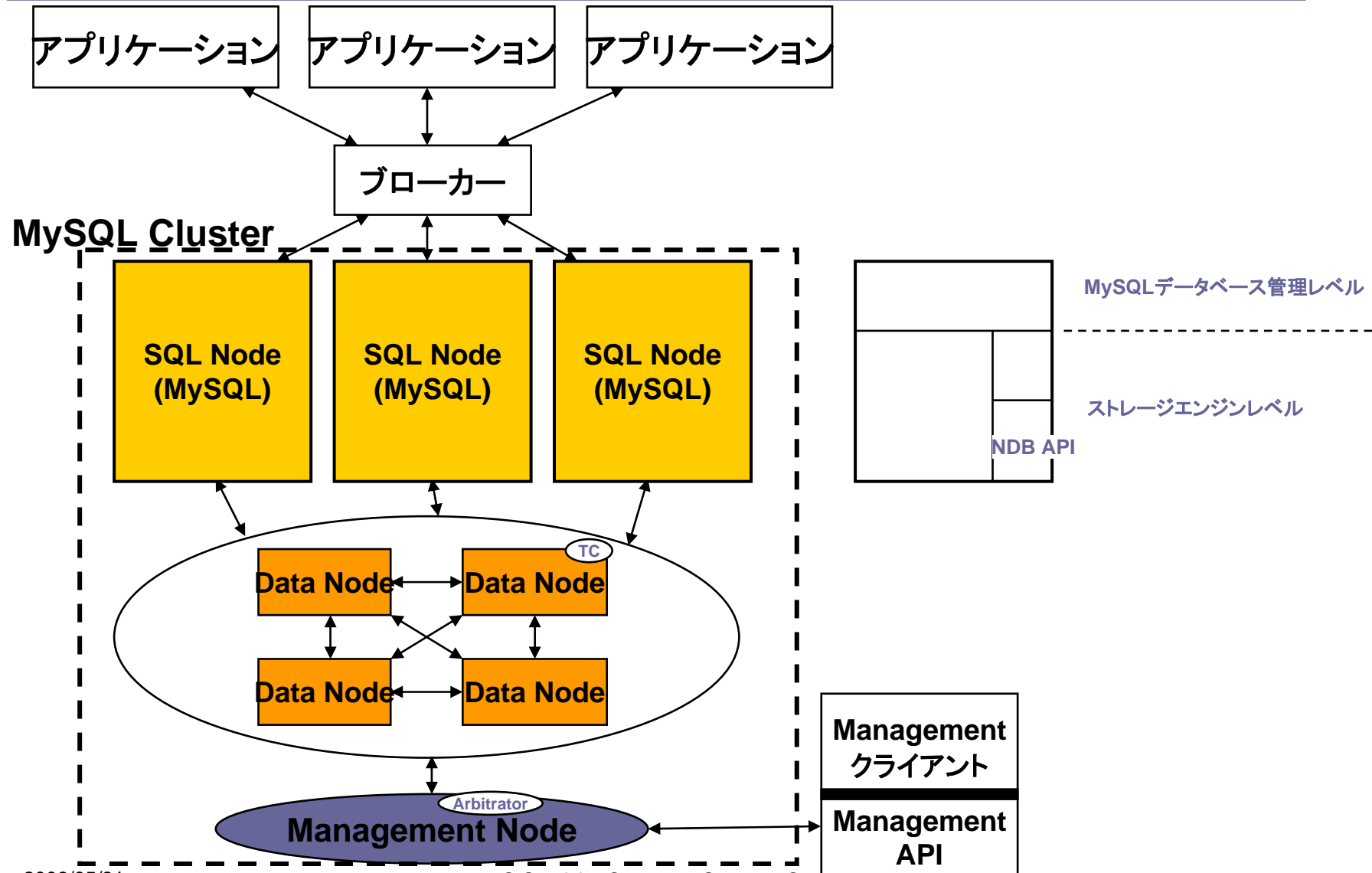


MySQL Clusterの特徴

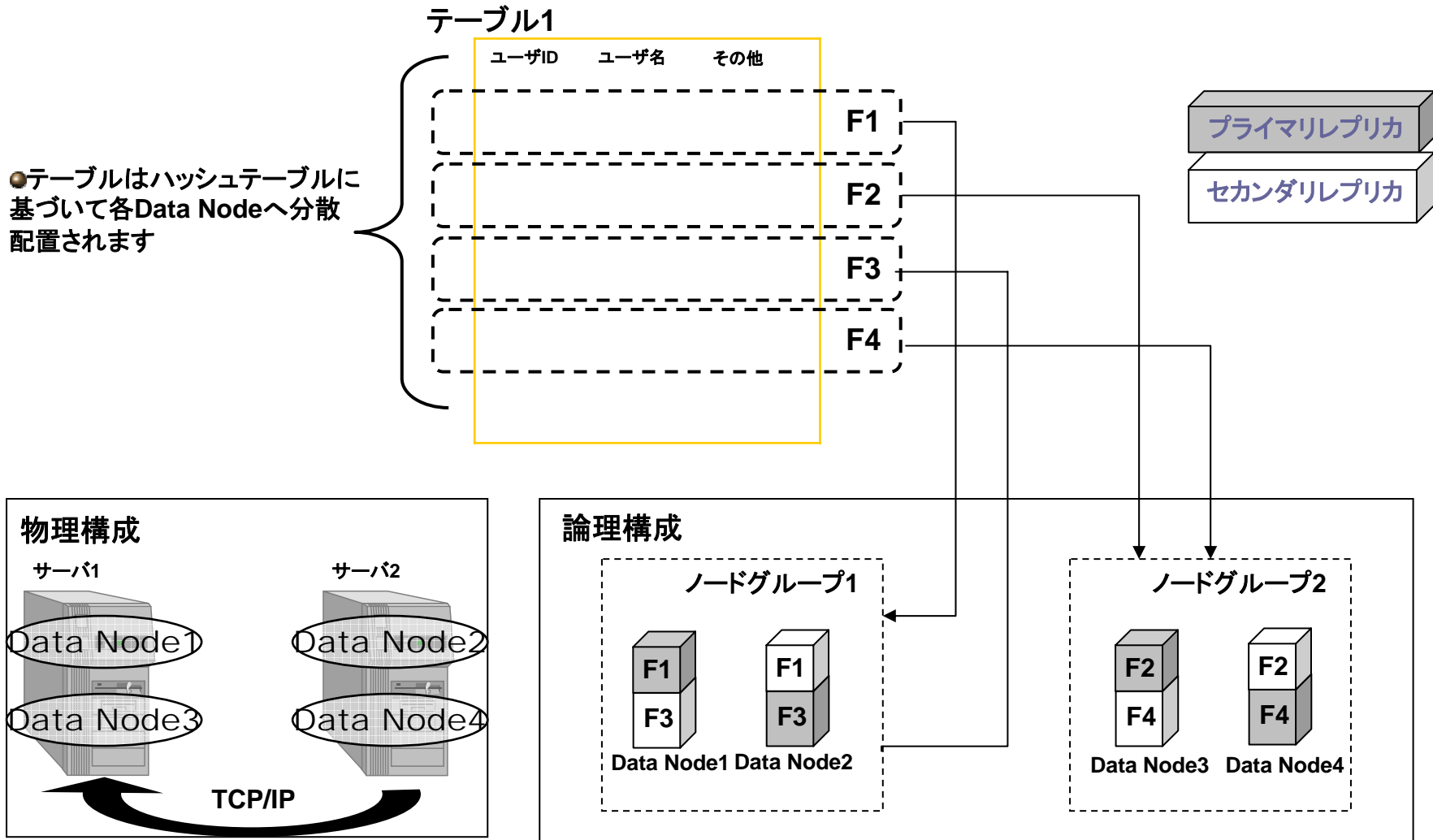
- メモリデータベース(5.1からディスクも選択可)
- 非共有型のクラスタリング
- アクティブ・アクティブ構成
- 自動リカバリ



MySQL Clusterのアーキテクチャ



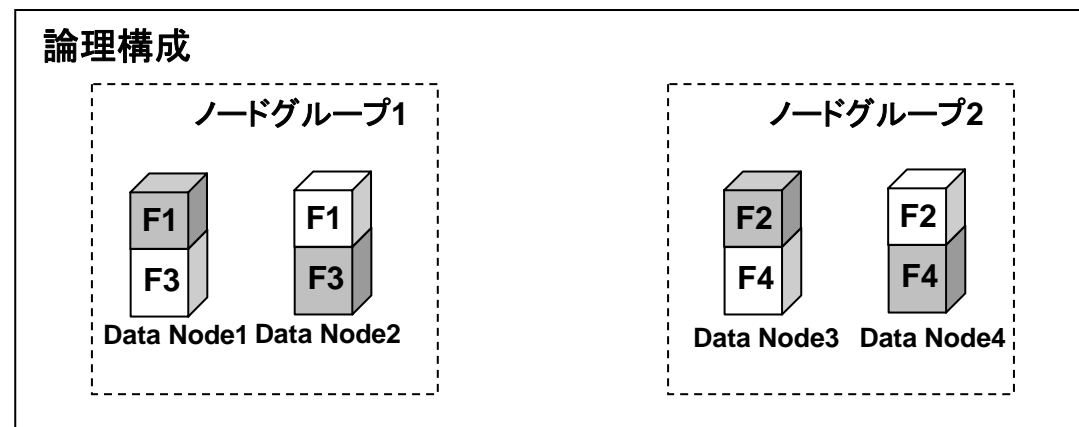
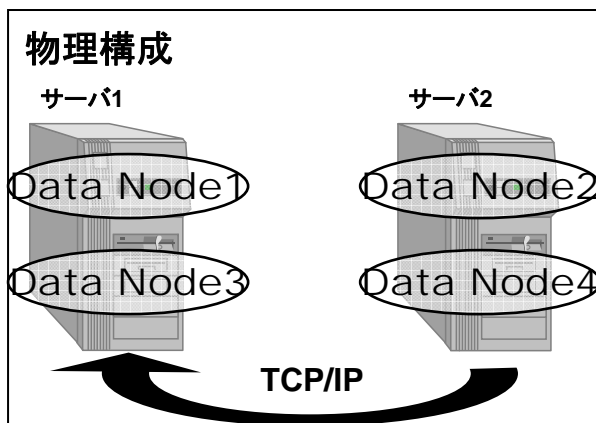
Data Node内のデータ配置



MySQL Cluster : 4 Data Node構成 (2 dual processor サーバ) レプリカ数=2

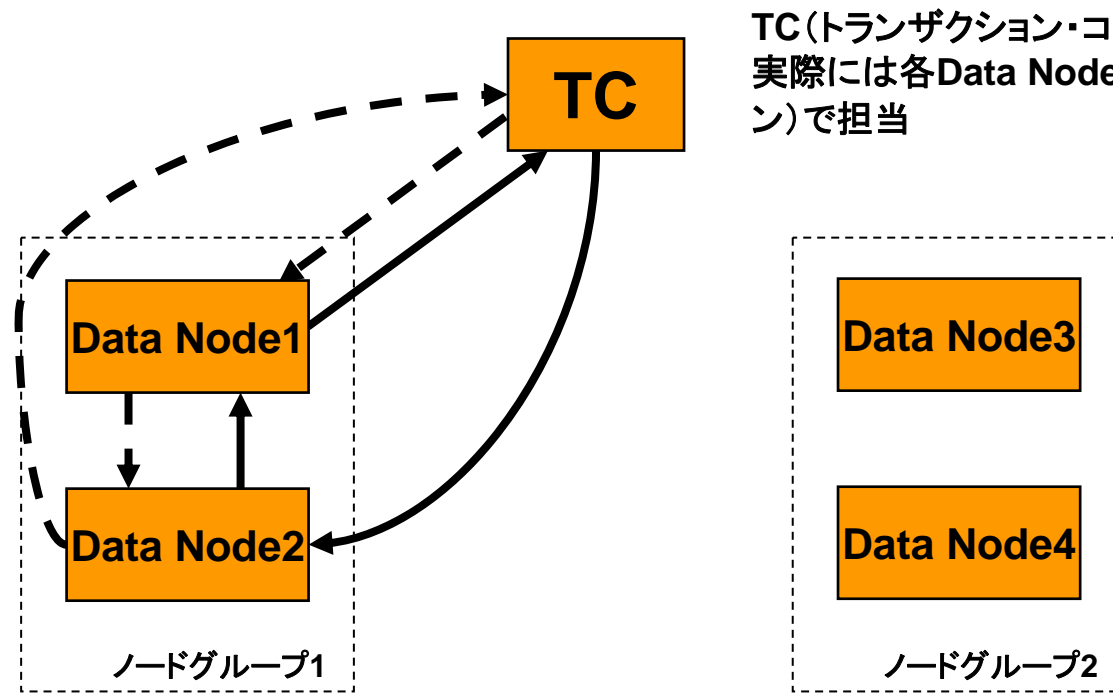
レプリカとノードグループ

- レプリカ
 - 複製
- ノードグループ
 - 同じデータを持つData Nodeの集合
- レプリカ数とノードグループ数の関係
 - 総ノード数 = レプリカ数 × ノードグループ数



MySQL Cluster : 4 Data Node構成 (2 dual processor サーバ) レプリカ数=2

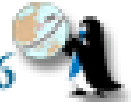
同期レプリケーションと二相コミットメント



TC(トランザクション・コーディネータ)の役割は、
実際には各Data Nodeが交代(ラウンドロビン)で担当

二層コミットメント

- - - 準備フェーズ: 状態確認と変更するデータの受け取り
- コミットフェーズ: 変更が実行される



3. IPAプロジェクトでの検証結果

- 2005年度 OSS推進フォーラム-開発基盤WGプロジェクト
 - 目的
 - サーバLinux、OSSの更なる普及・拡大のためのベンダサイドの課題解決
 - SCS担当
 - DBT-1によるMySQL評価
 - MySQL Cluster評価
- 検証システム構成
- 検証の目的と実施項目
- 検証結果

検証システム構成

□ 検証システム環境

- HW: 日立BladeSymphony 7枚のサーバモジュール
 - Xeon 3.0GHz / 2GB Memory / 160GB HDD
- OS: MIRACLE LINUX V3

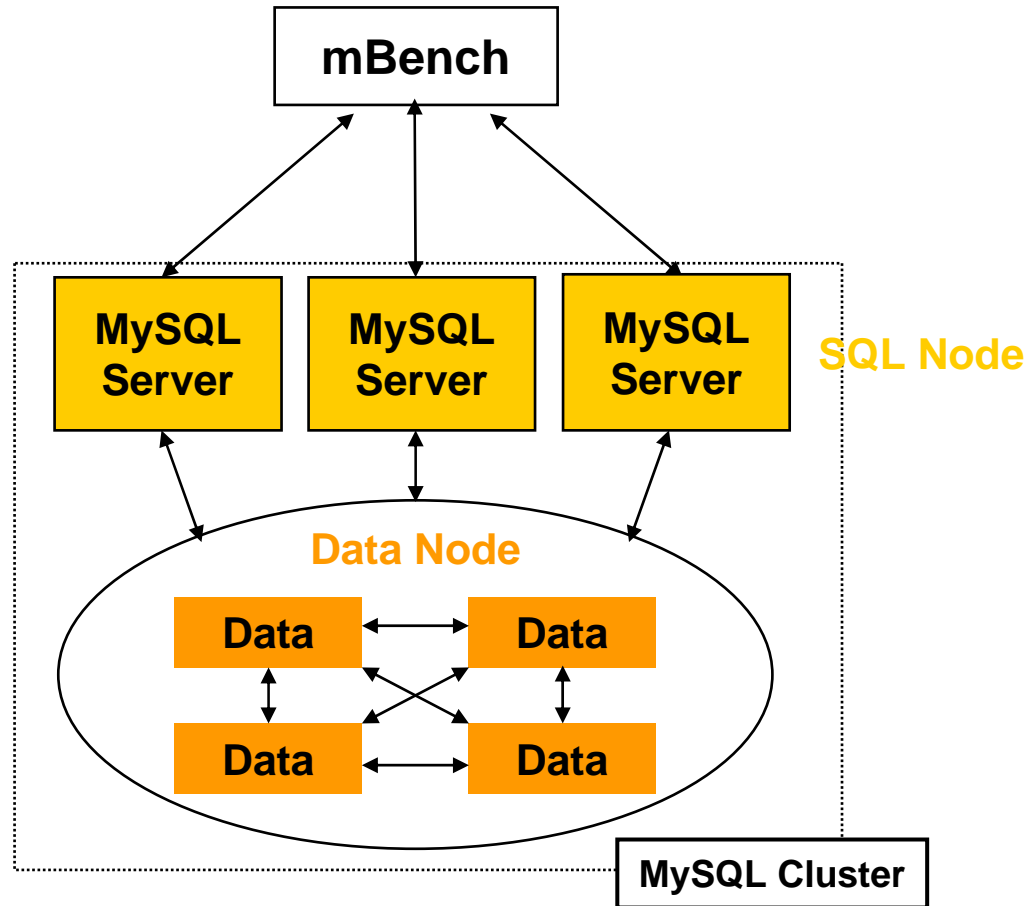
□ 検証内容

- mBenchより基本的なSQLクエリを発行
- トータルで60,000件のトランザクションを完了するまでの時間を計測してスループットを算出

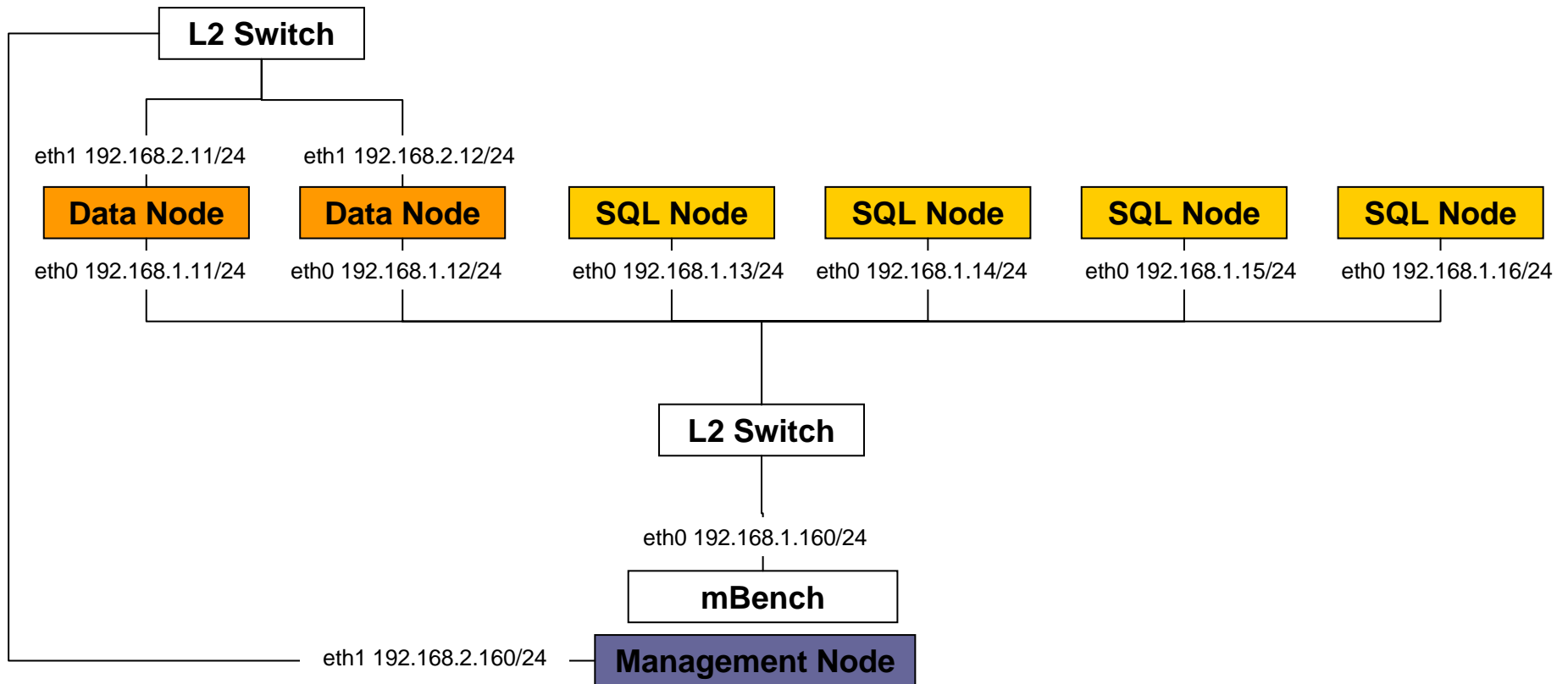
□ ベンチマークツール: mBench

- IPAプロジェクトの一環でSCSが作成したJavaベース ベンチマークFW
- <http://mbench.sourceforge.net/>にてGPLでソース公開中

検証システム構成図：論理構成



検証システム構成図：物理構成



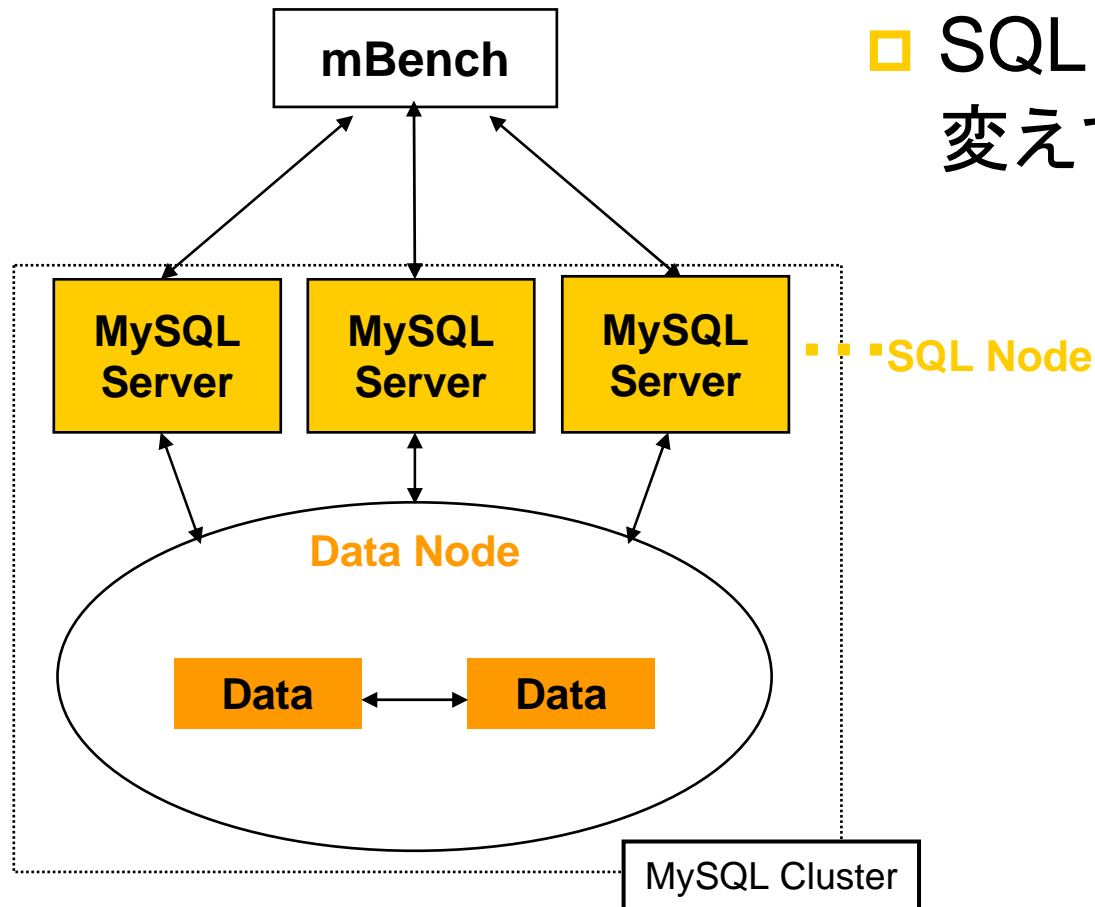
検証の目的と実施項目

- 目的
 - MySQL Cluster最適構成の指標を得る
 - クラスタシステムとしての基本機能の検証
- 実施項目
 - a. ノード構成
 - SQL Node数
 - Data Node数
 - b. インターコネクト
 - c. 32ビット vs. 64ビット
 - d. HA性能

a. ノード構成

- SQL Node数
- Data Node数
 - レプリカ数
 - ノードグループ数

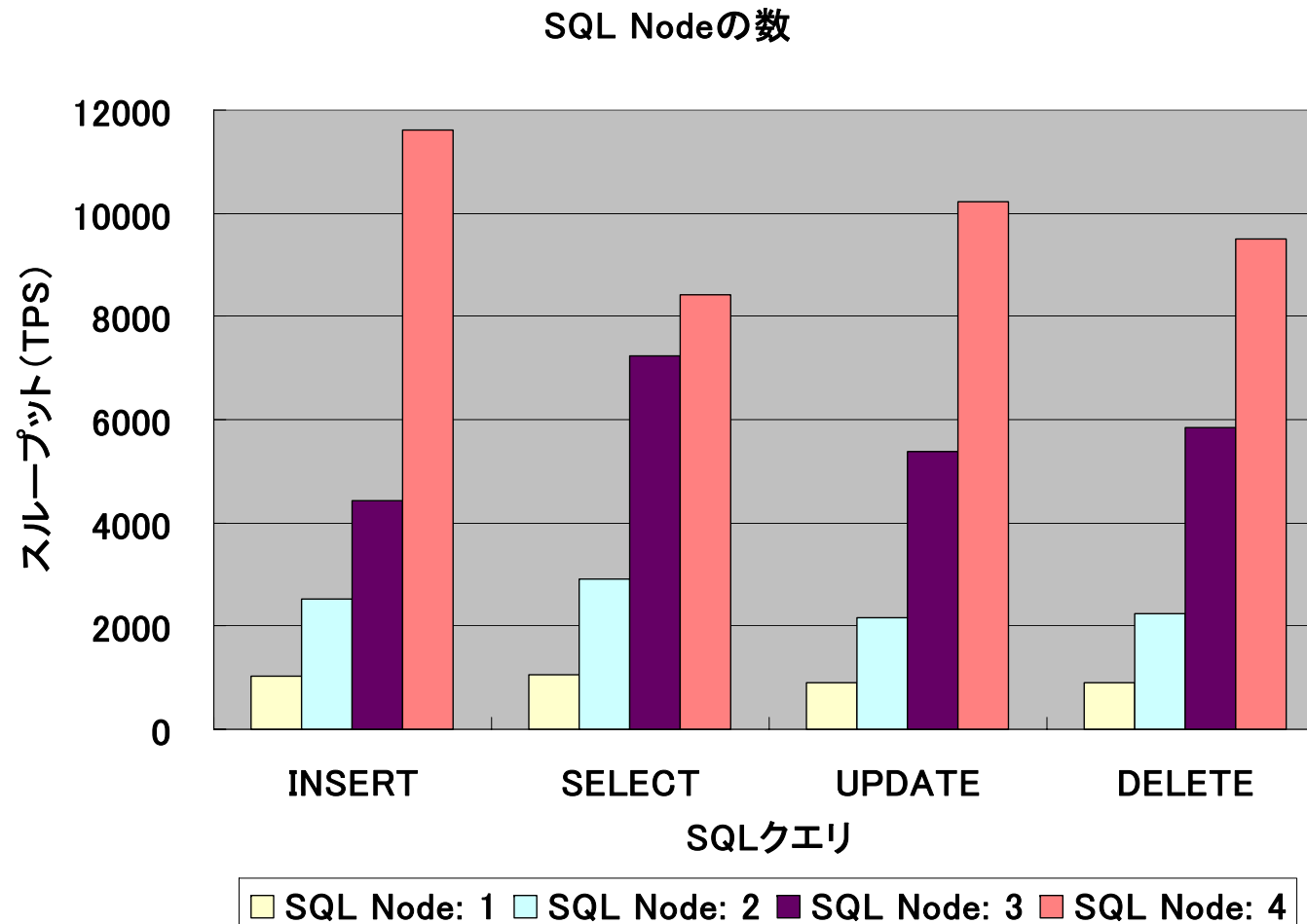
SQL Node数: 論理構成



- SQL Nodeの数を1~4に変えて計測

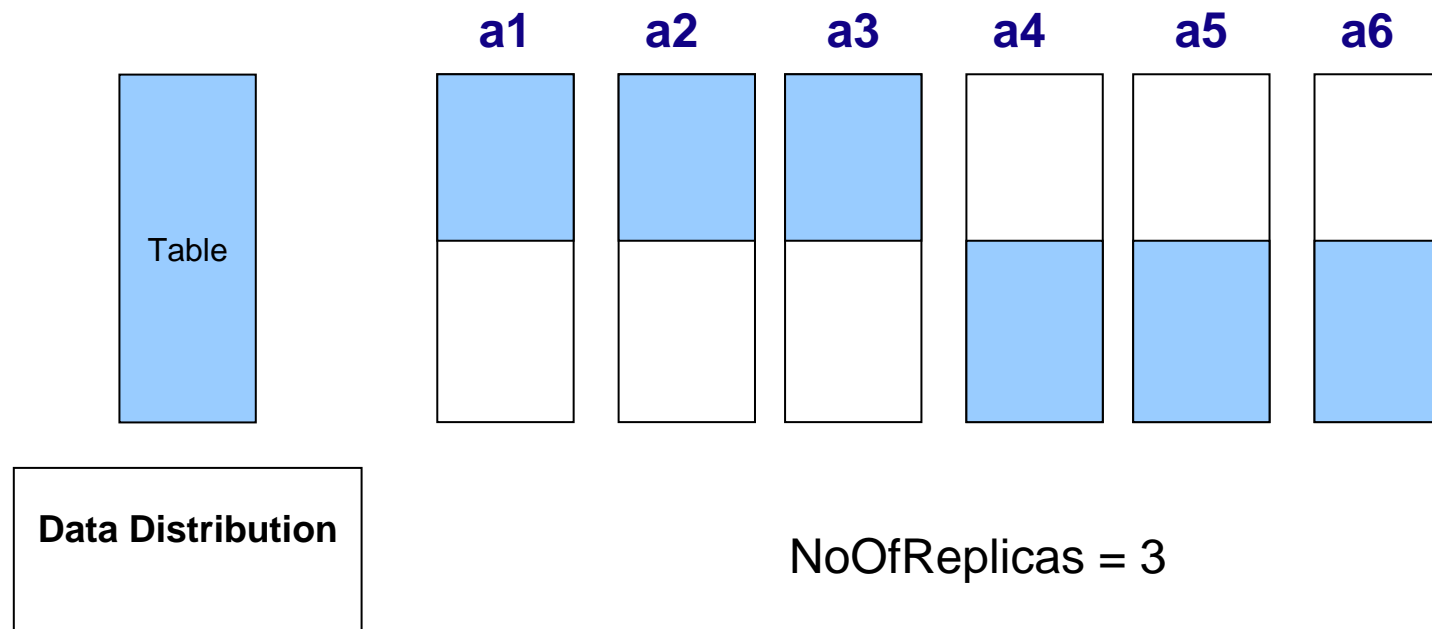
SQL Node数: 結果

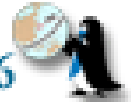
- SQL Nodeの数を増やすことでスループットが向上した



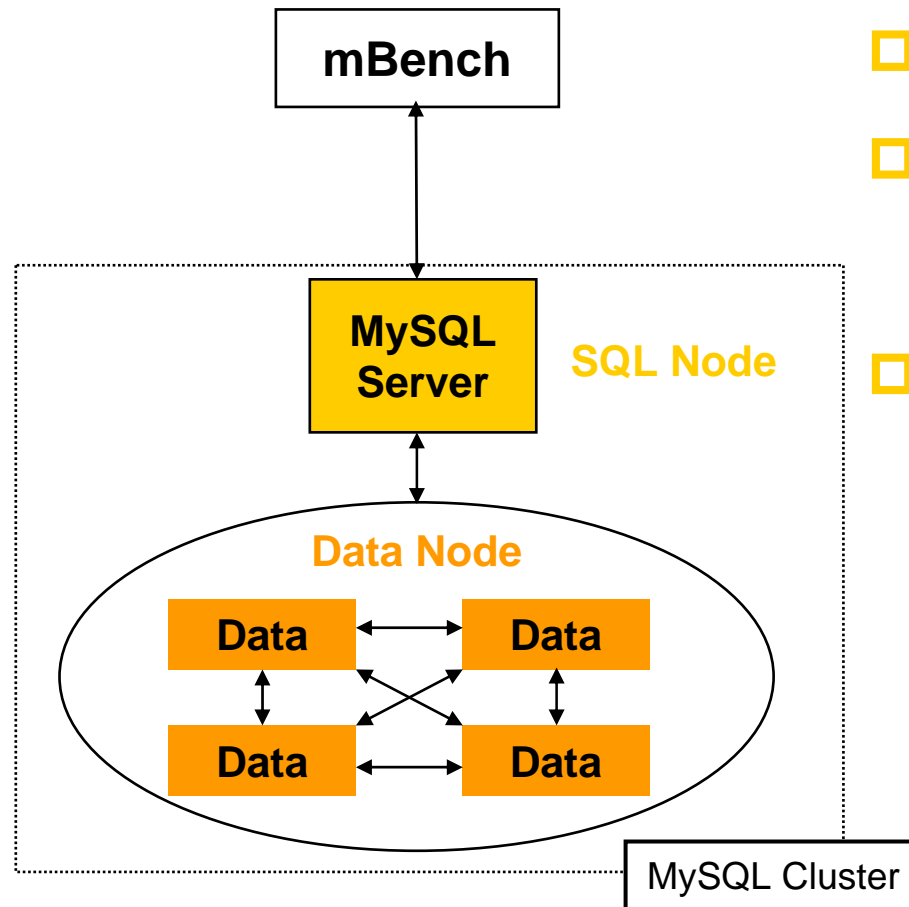
もう一度、レプリカ数とは

- 同じデータの複製をいくつ持つか
- (例) Data Nodeが合計6台、レプリカ数3の場合





Data Node数(レプリカ数): 論理構成



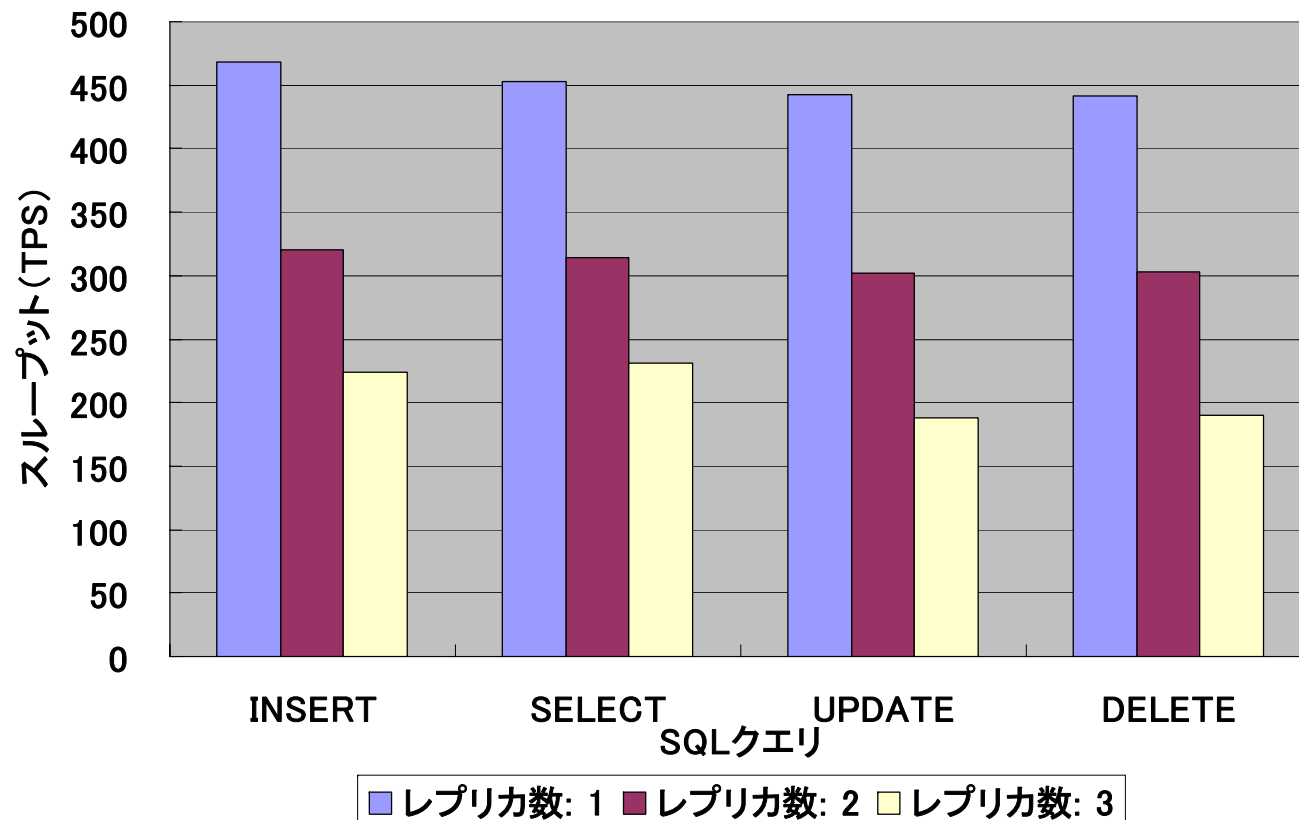
- SQL Node: 1
- レプリカ数: 3~1で計測

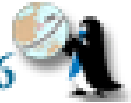
- Data Node数の変化
 - 合計6台、レプリカ数3
 - 合計4台、レプリカ数2
 - 合計2台、レプリカ数1



Data Node数 (レプリカ数) : 結果

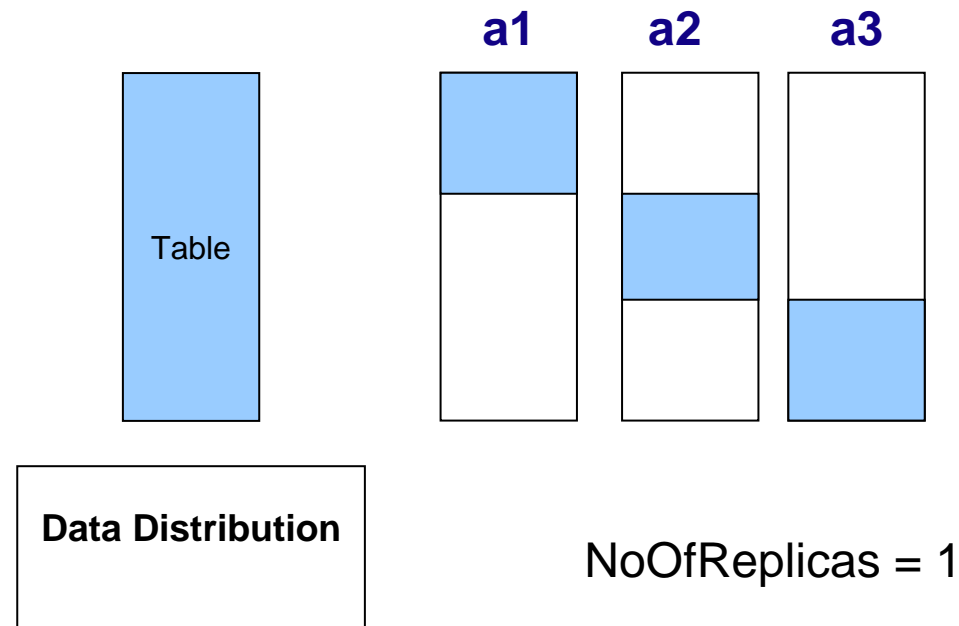
- Data Nodeの数及びレプリカ数を増やすとスループットは劣化する
- 但し、可用性は向上する



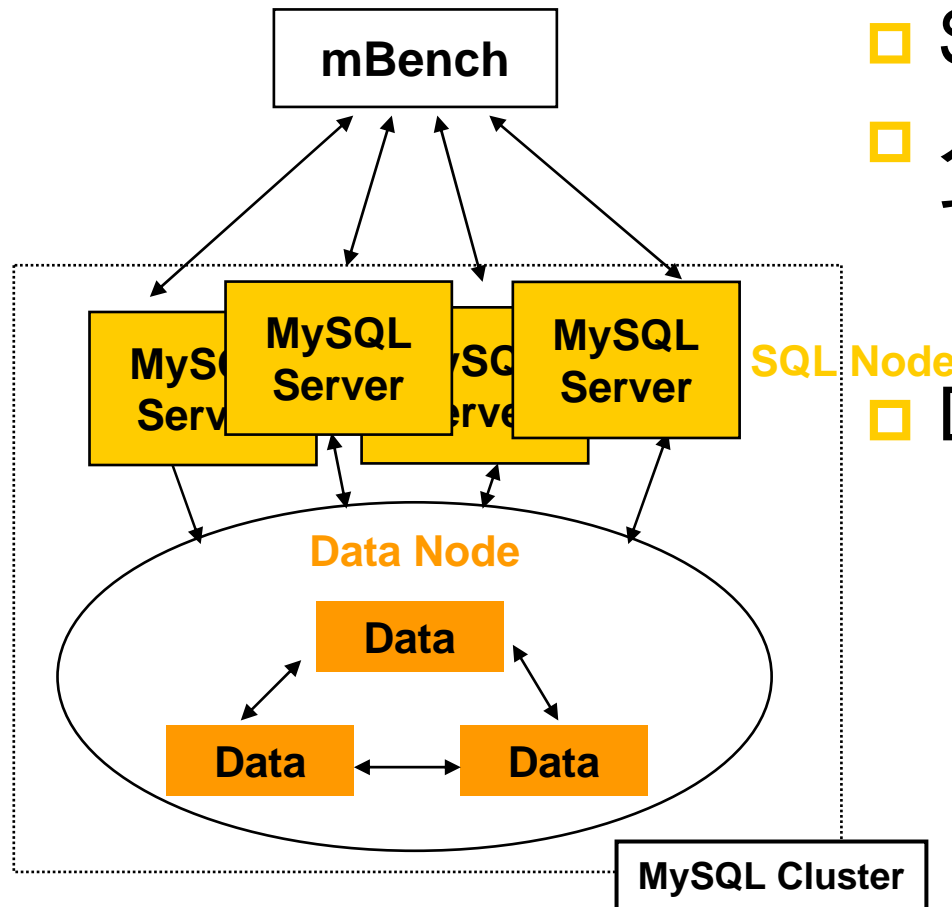


もう一度、ノードグループ数とは

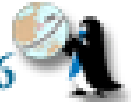
- 1つのテーブルを何分割するか
- (例) Data Nodeが合計3台、レプリカ数1、ノードグループ数3の場合



Data Node数(ノードグループ数): 論理構成

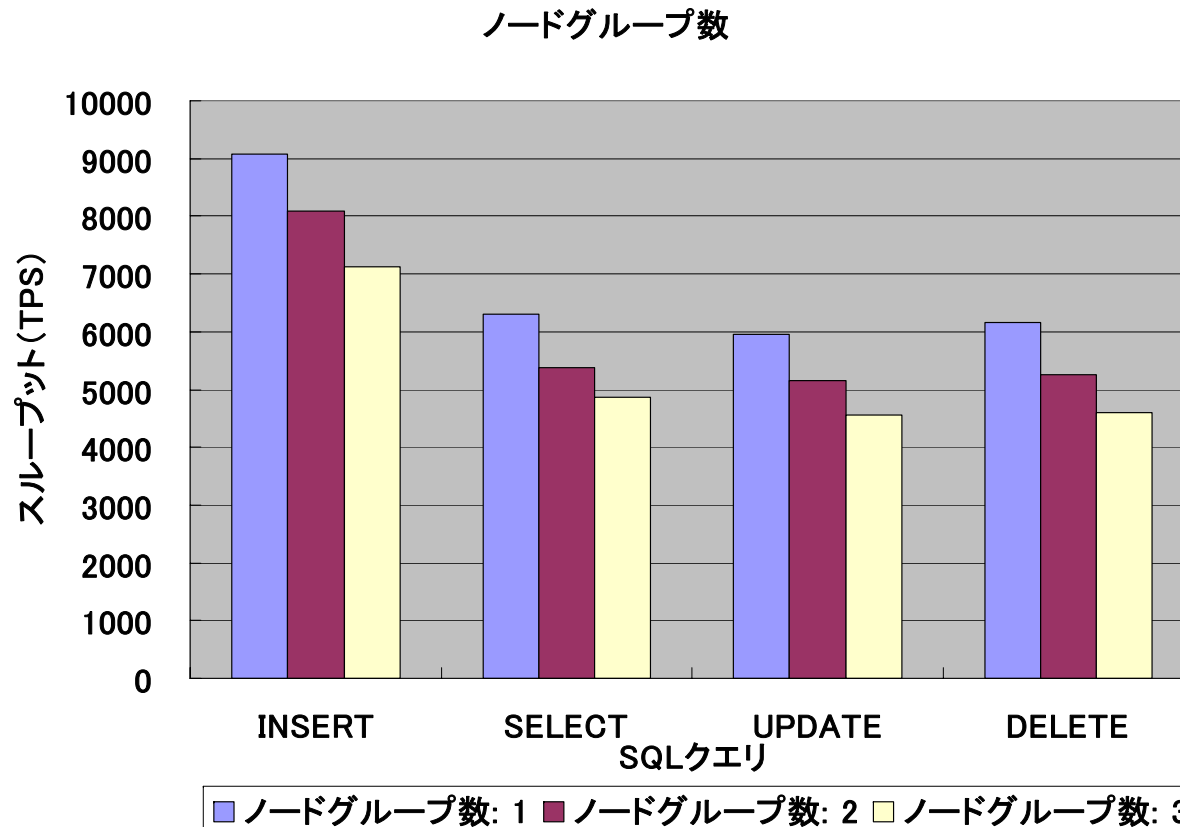


- SQL Node: 4
- ノードグループ数: 3~1
で計測
- Data Node数の変化
 - 合計3台、レプリカ数1、
ノードグループ数3
 - 合計2台、レプリカ数1、
ノードグループ数2
 - 合計1台、レプリカ数1、
ノードグループ数1



Data Node数(ノードグループ数): 結果

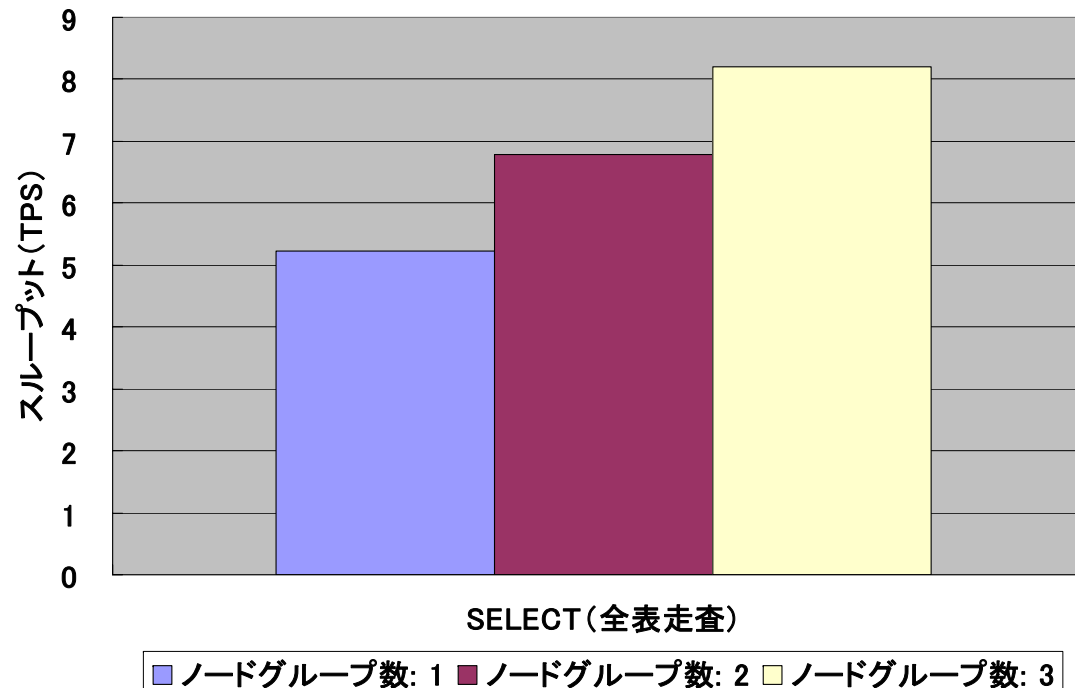
- ノードグループの数を増やす(=テーブルを分割する)と、
 - 簡単なSQLクエリの場合はスループットは悪くなった

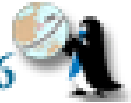


Data Node数(ノードグループ数): 結果

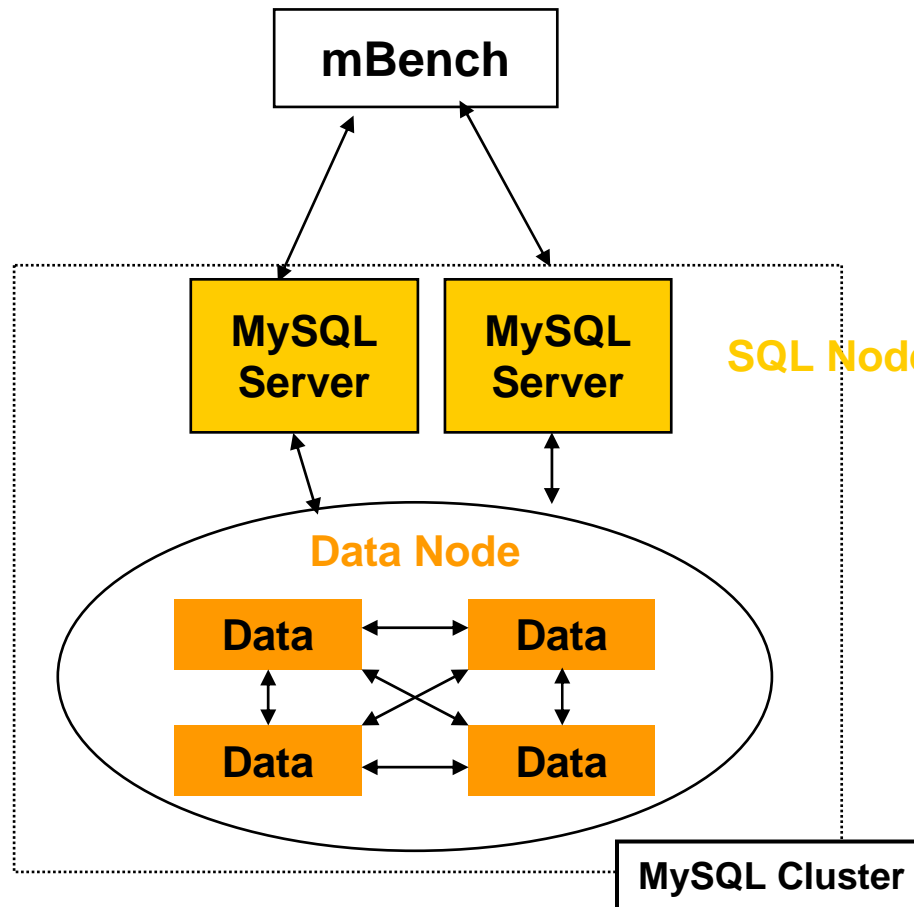
- ノードグループの数を増やす(=テーブルを分割する)と、、、
 - 簡単なSQLクエリの場合はスループットは悪くなった
 - 全表走査の場合はスループットは良くなった

ノードグループ数(60,000行の全表走査)





b. インターコネクト: 論理構成



□ 3種類のインターコネクトで計測

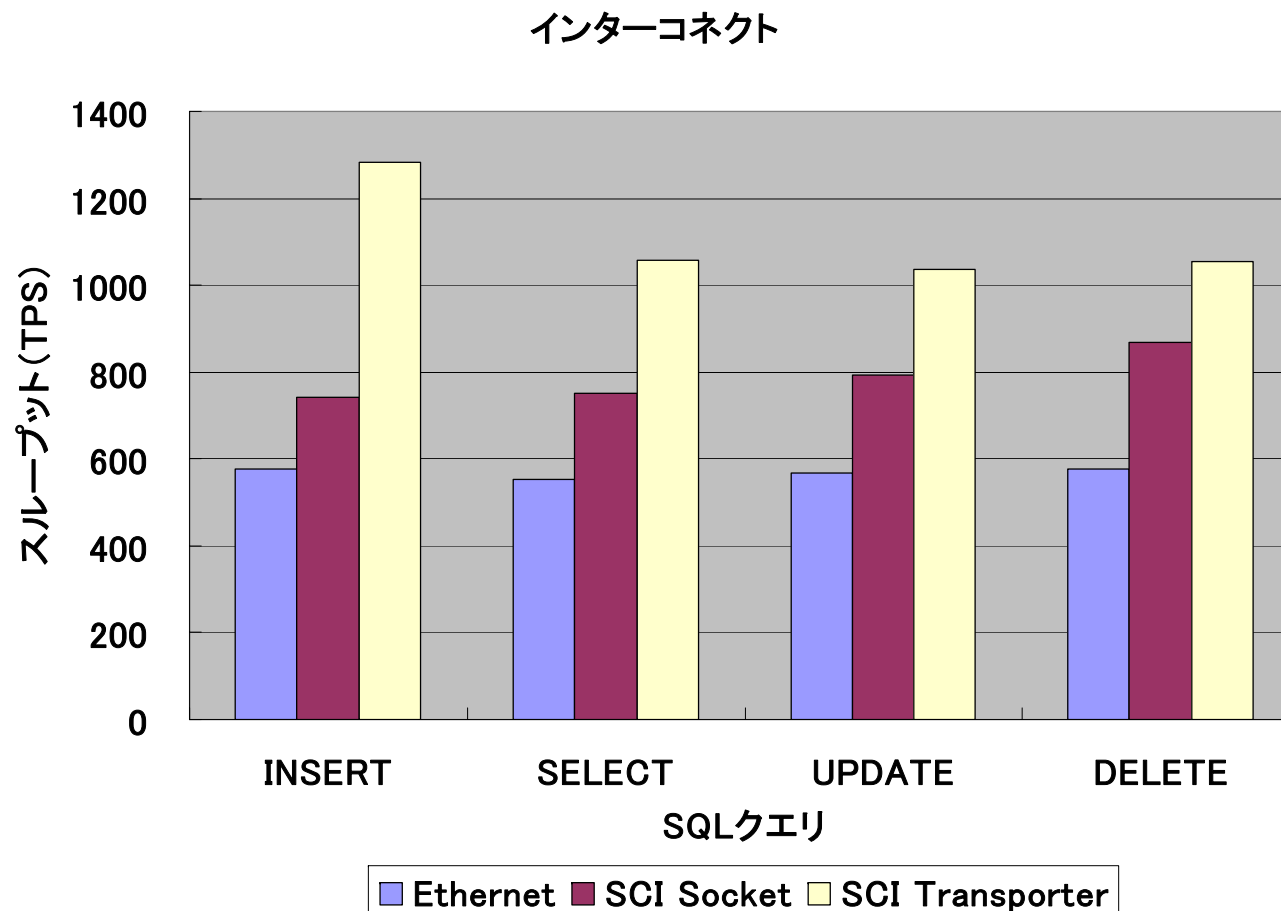
- Gb Ethernet
- SCI Socket
- SCI Transporter

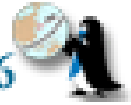
□ SCIとは

- IEEE1596-1992で定義された通信規格
- ハイパフォーマンスクラスタリングや科学技術計算用のクラスタリングシステムで多く採用
- 専用のハードウェア及びドライバーが必要

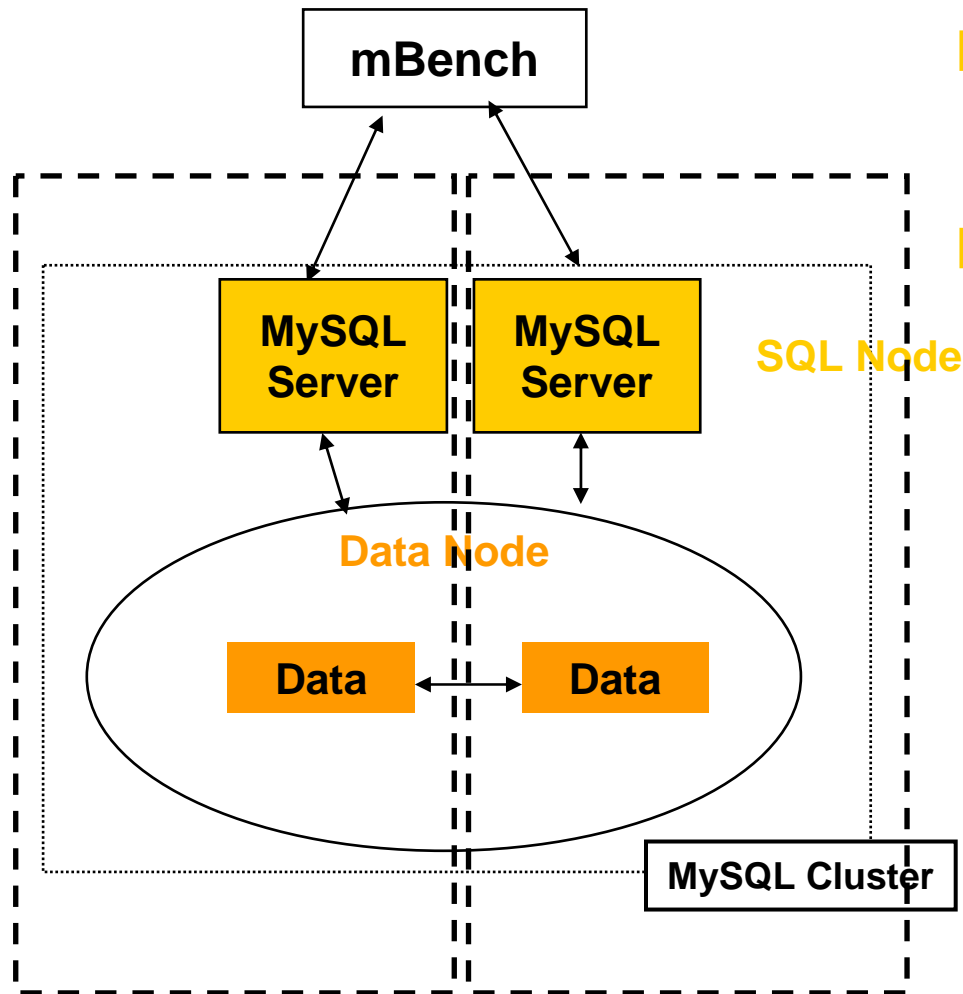
インターコネクト：結果

- SCI Transporterでは80～120%スループットが向上





c. 32ビット vs. 64ビット: 論理構成

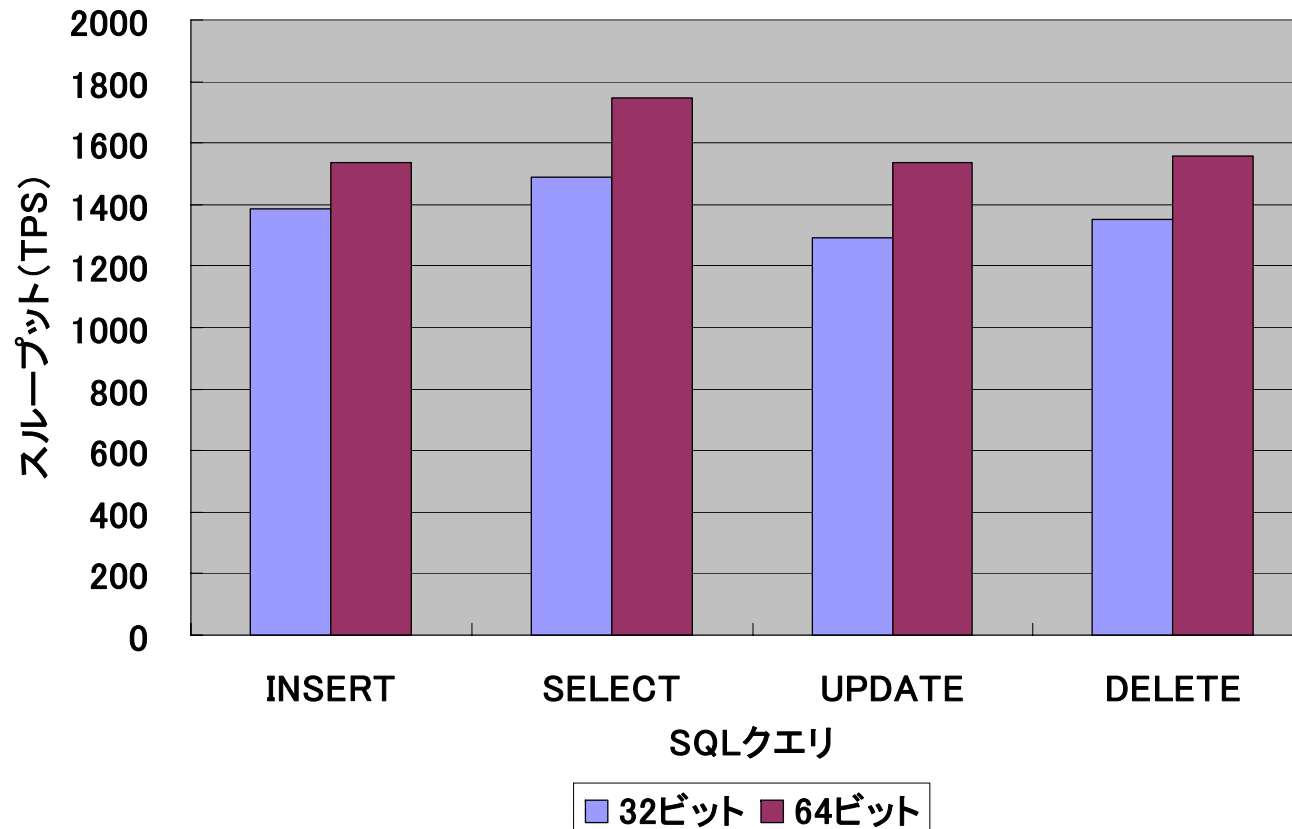


- 1サーバにSQL Node, Data Nodeを構成
- 2サーバでクラスタ構成

32ビット vs. 64ビット: 結果

- 64ビット環境ではスループットが10~20%向上

32ビット vs. 64ビット



d. HA性能

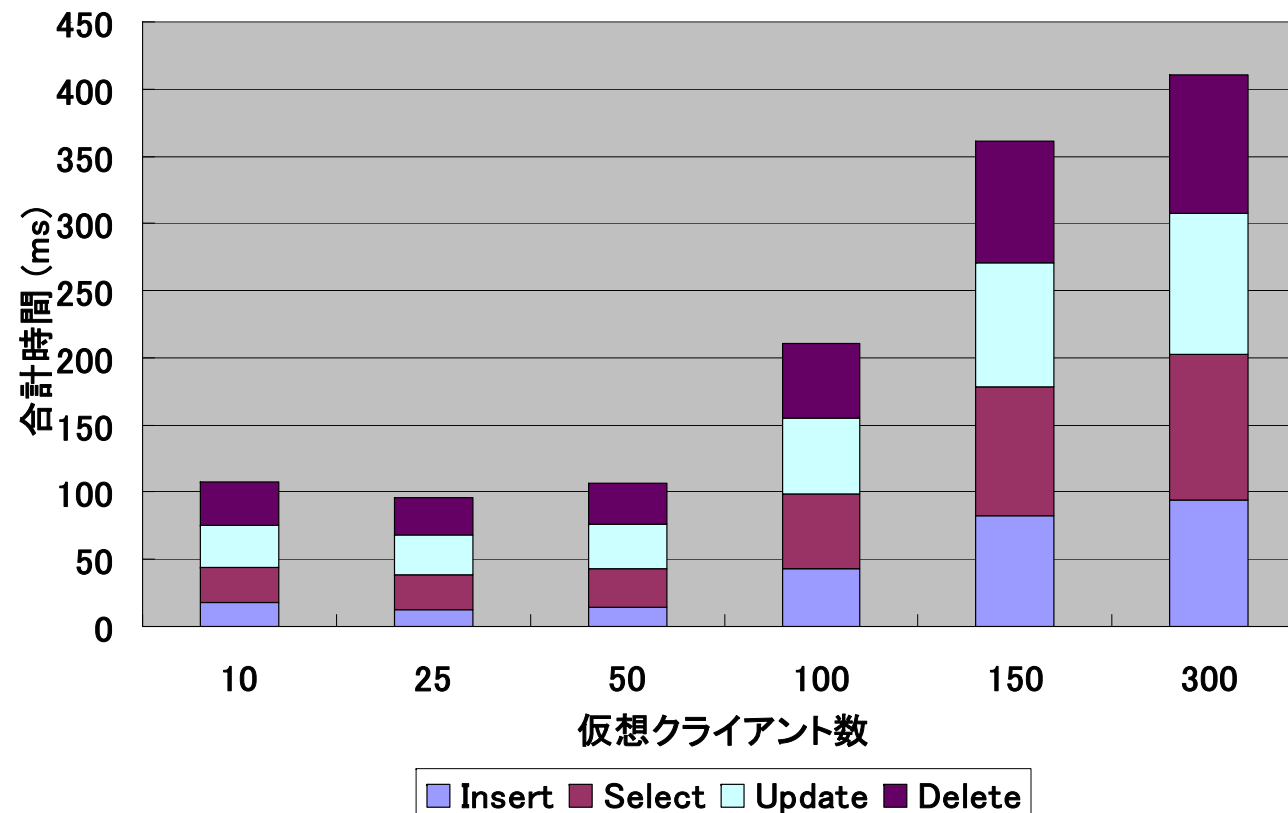
- 耐障害性を検証した結果、高いサービス継続性/セッション継続性を維持していることが分かった。

	サービス継続性	セッション継続性	トランザクション継続性
SQL Nodeプロセス障害	○	○	○
Data Nodeプロセス障害	○	○	×
Management Nodeプロセス障害	○	○	○
SQL Nodeノード障害	○	○	○
Data Nodeノード障害	○	○	×
SQL Node - Data Node間通信障害	○	○	×
Data Node - Data Node間通信障害	○	○	×

補足: コネクション数

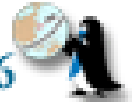
- 1 SQL Nodeへの接続数は50以下にするとスループットが良い

仮想クライアント数とクエリ完了の合計時間



4. まとめ

- MySQL Cluster最適構成の指標



MySQL Cluster最適構成の指標

- ノード構成は
 - SQL Nodeを増やして
 - Data Nodeはなるべく少なく
 - レプリカ数は2~で可用性向上
 - ノードグループ数を増やす事で全表走査などの性能向上
- インターコネクトはSCIを利用して
- 64ビット環境で
- コネクション数を50以下に

おまけ: MySQL Users Conference 2006

- MySQL 5.1
 - ディスクベースクラスタ
 - パーティショニング
 - Global Replication
- 事例
 - 8x8.com: アメリカのIP電話
 - PagineGialle.it: イタリアのイエローページサイト
- BoF
 - 某著名E-Commerceサイトが32台のMySQL Clusterを検証中

SCSのMySQL関連サービス

- MySQL設計構築サービス
 - MySQL Cluster
 - レプリケーション
- オフィシャルトレーニング
- OEM向サービス
 - OEMライセンス
 - テクニカルサポート

- URL: <http://www.scs.co.jp/mysql/>
- E-Mail: mysql@scs.co.jp

さいごに

- 本評価は、「IPA(独立行政法人情報処理推進機構) オープンソースソフトウェア活用基盤整備事業」の一環として実施しました。
- 日本OSS推進フォーラム - 開発基盤WG - DB層のレポートとしてWebで公開中 (<http://www.ipa.go.jp/software/open/forum/>)
 - 環境
 - 設定ファイル一式
 - 結果データ
 - ベンチマークmBenchソースコード一式 (sourceforge.net)
- iPedia (<http://ossipedia.ipa.go.jp/>) でも一部公開中
- 本評価では、日立様及びMIRACLE LINUX様の協力を得ました。ありがとうございます。