

# セッション1

MySQL Cluster が実現する  
超高速トランザクションと可用性

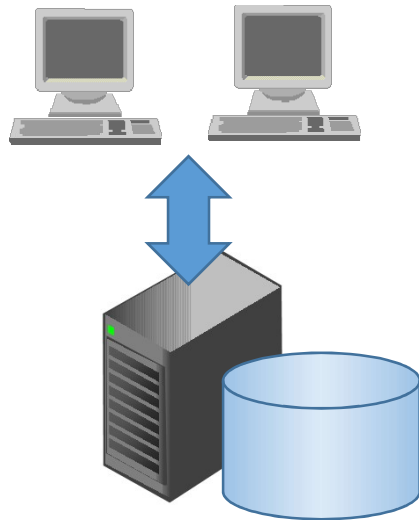
コネク特株式会社

これまでのRDBMSではなく  
分散データシステムなのか

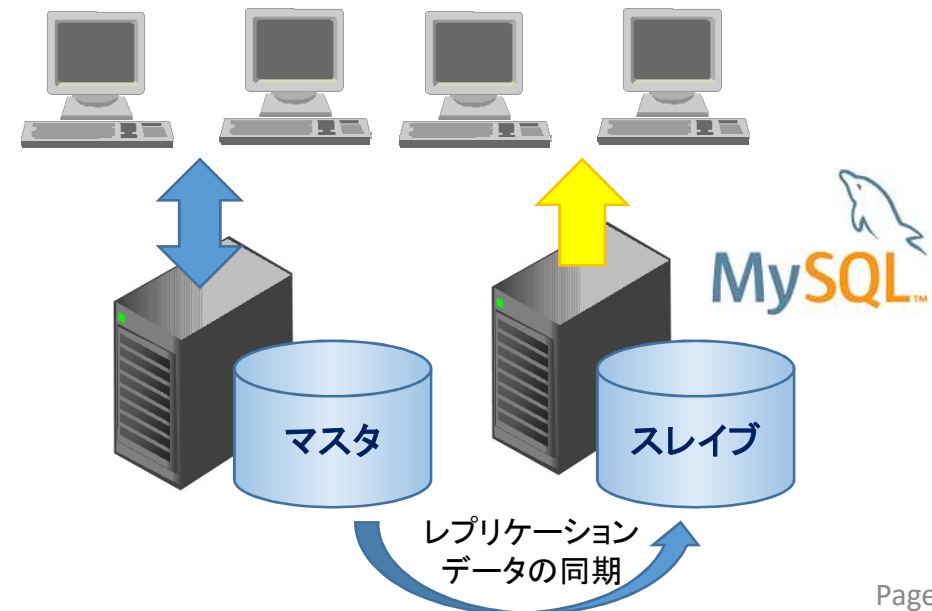
# なぜ分散データシステムなのか

- 処理量の増大
  - スレイブの増加による対応の限界
  - 論理的なマスタ分割による対応の限界
- 際限ない大容量化の限界
  - ディスク容量/メモリ容量/CPU性能/バックアップメディア
- ビッグデータに代表される大容量/高速処理の必要性
  - 大量データの収集と利用

## スタンドアロン



## レプリケーション



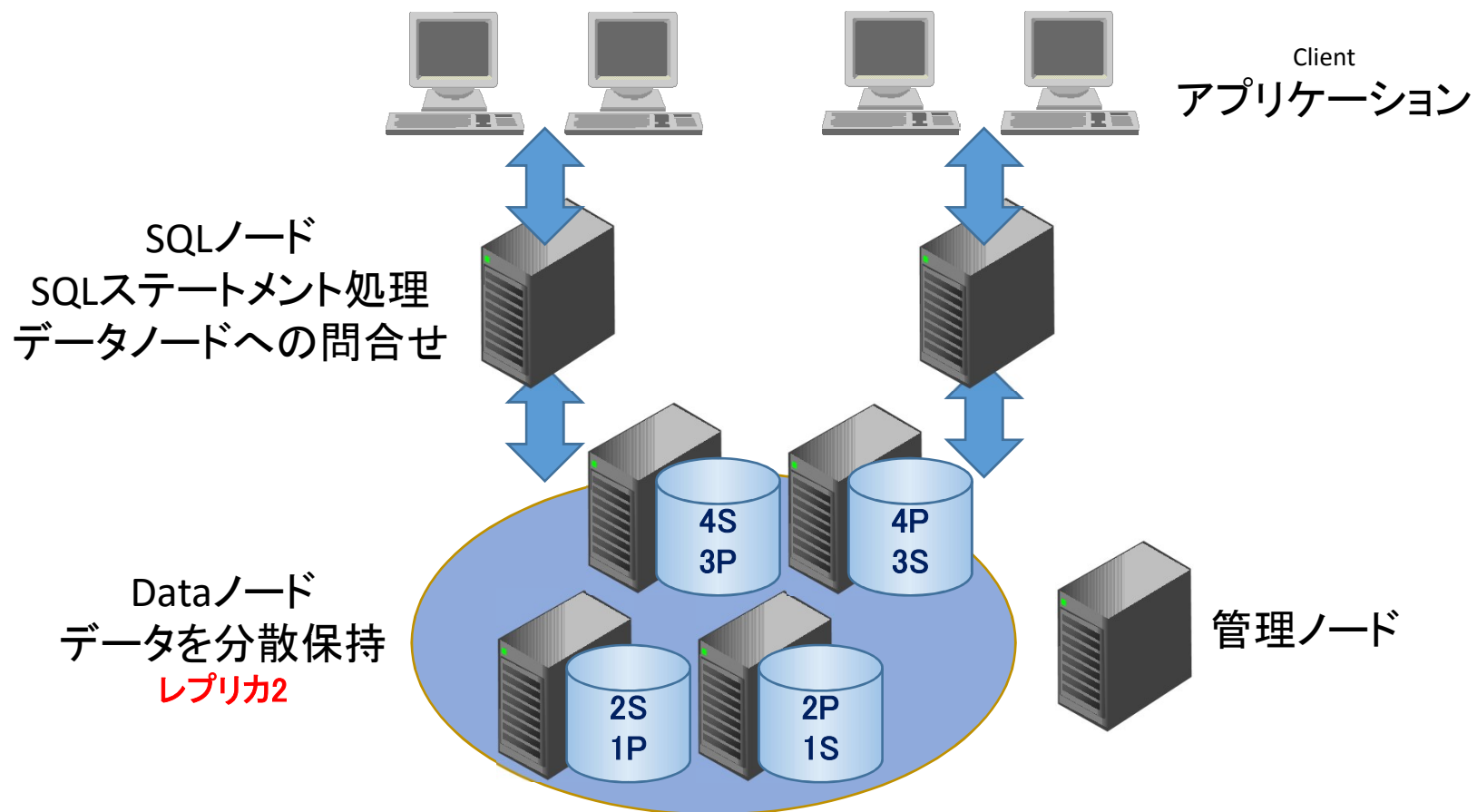
# これまでのRDBMSと何が違うのか

- 高速性
  - リクエストをノードに分散して処理
  - 空きノードの活用による高い応答性
    - ロードバランサによるラウンドロビンも可能
- 大容量
  - 各ノードのストレージを総合したデータ格納を実現
- 可用性
  - クラスタ全体で処理を担うため一部のノードダウンが影響しない
  - ノードにデータを分散するためデータの欠損を予防
  - MySQL ClusterとCassandraは、SPOF(単一障害点)を持ちません。

# MySQL Cluster (NDBCLUSTER)



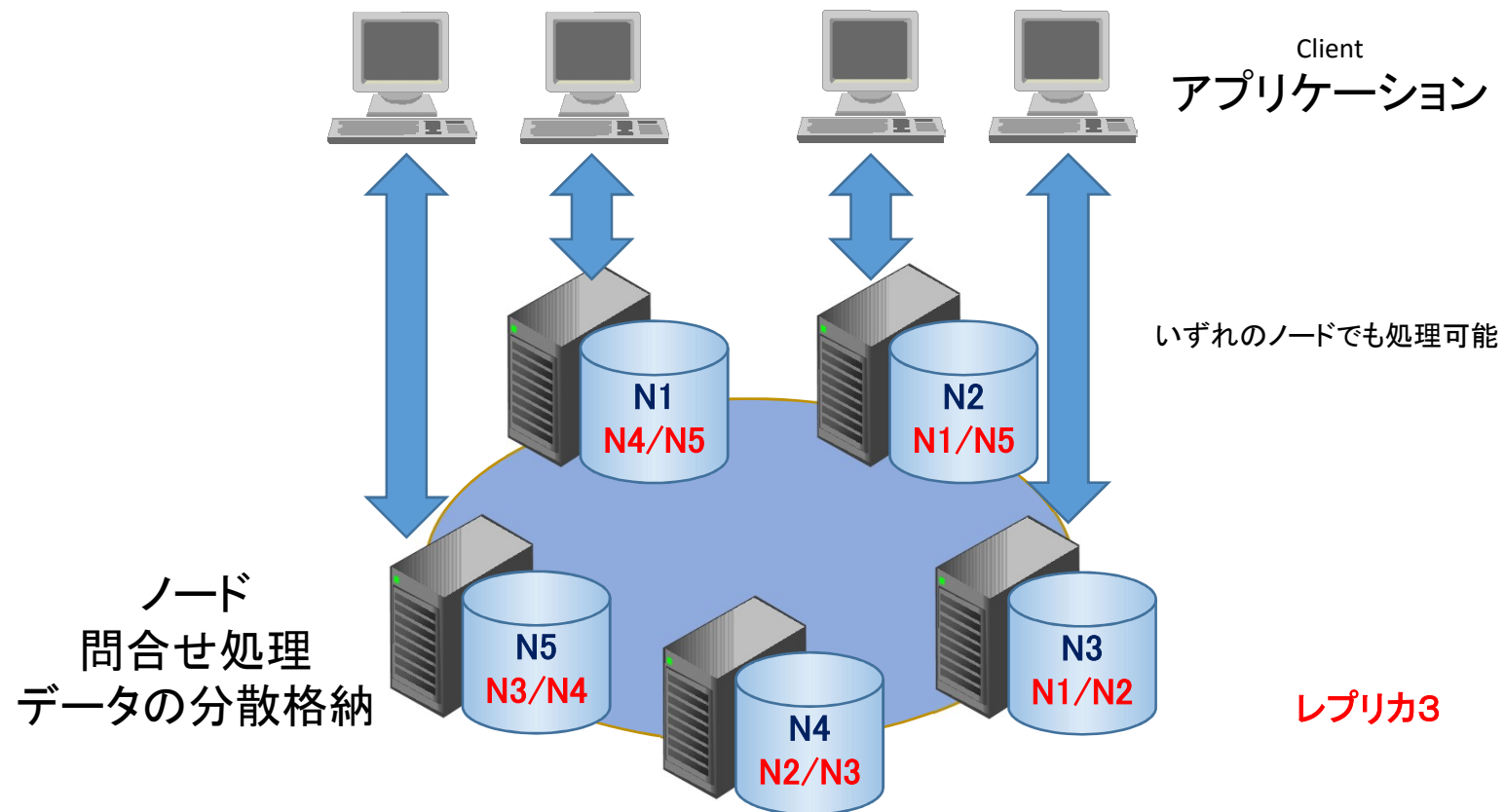
- RDBMSでありながらSPOFを持たず高速性とHAを両立しています。



# Cassandra



- 高速/大容量処理に定評があり、SPOFがありません。



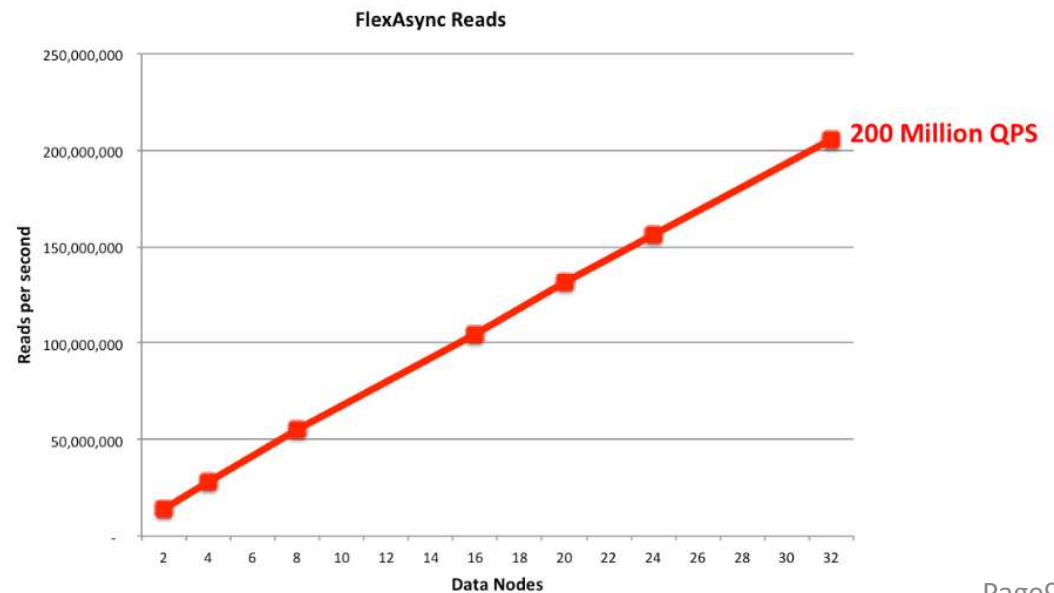
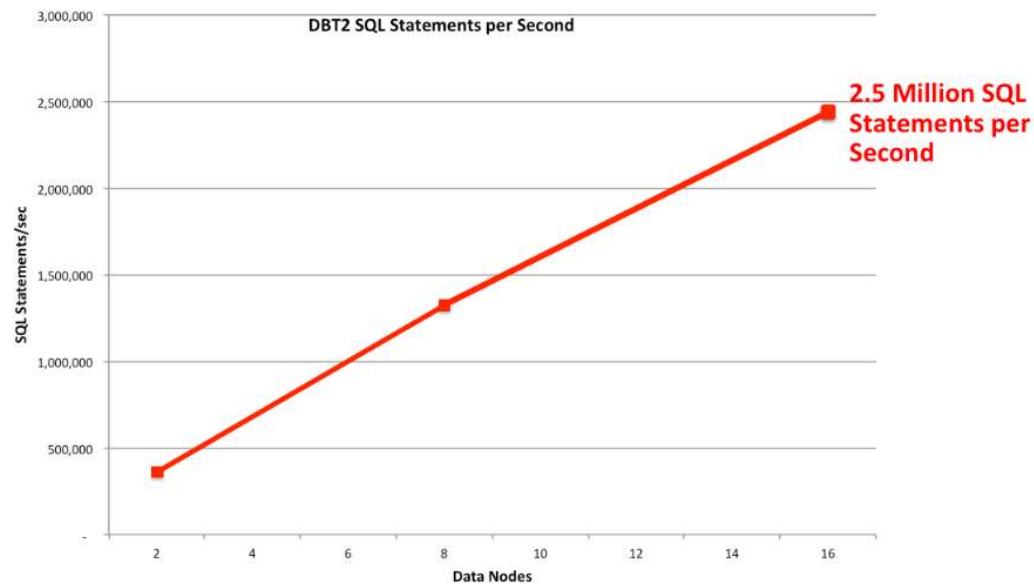
# MySQL Cluster 特徴と留意点

# どのような業務に向いているか(事例)

- CGEとはCarrier Grade Edition、NDBとはEricsson Network Database
  - 元々は、IP電話のアドレス変換を行うエンジンとして開発されました。
  - In Memory Databaseが主体となっています。
  - SQLベースのトランザクションに対応したリクエストを大量に高速で処理できます。
- Oracle社のMySQL Cluster顧客情報ページ
  - <https://www-jp.mysql.com/customers/cluster/>

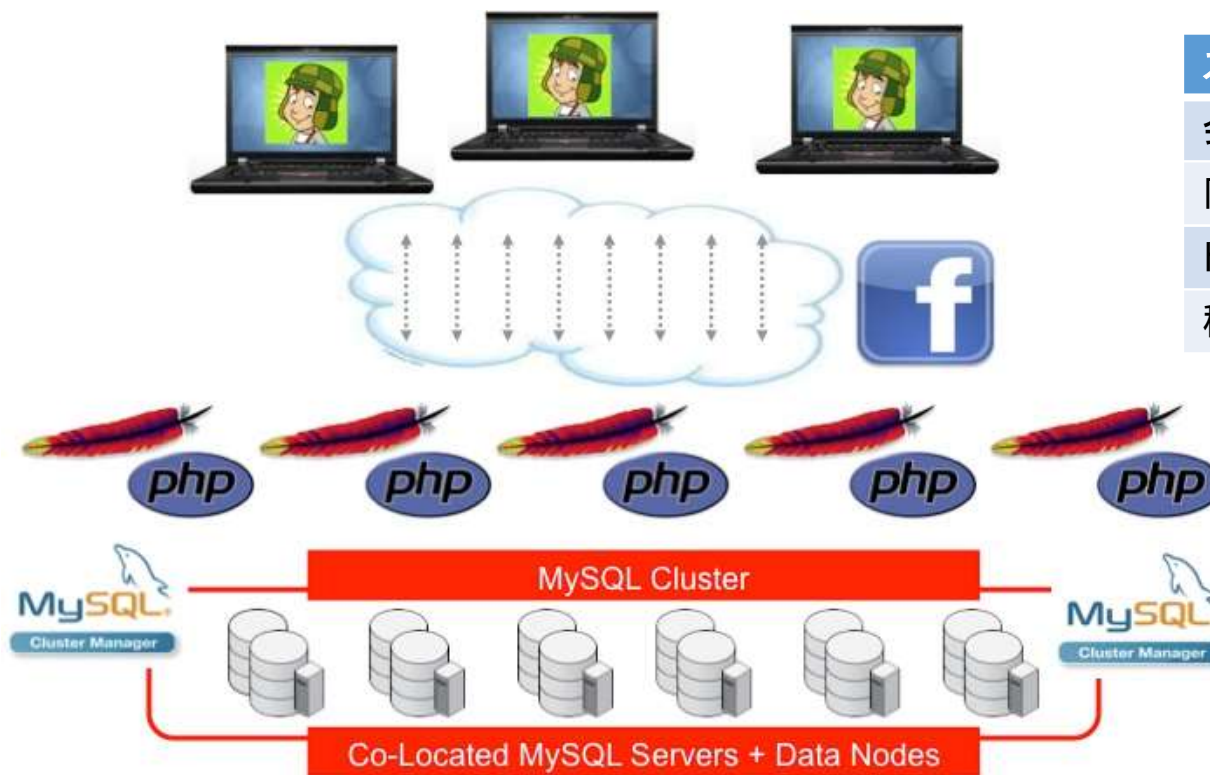


# ベンチマーク: SQLで2.5万、NoSQLで200万QPS



# FacebookGame「PLAYFUL PLYA」

- 大量リクエストを高速処理する必要のあるインターネットゲームにMySQL Clusterが活用されています。

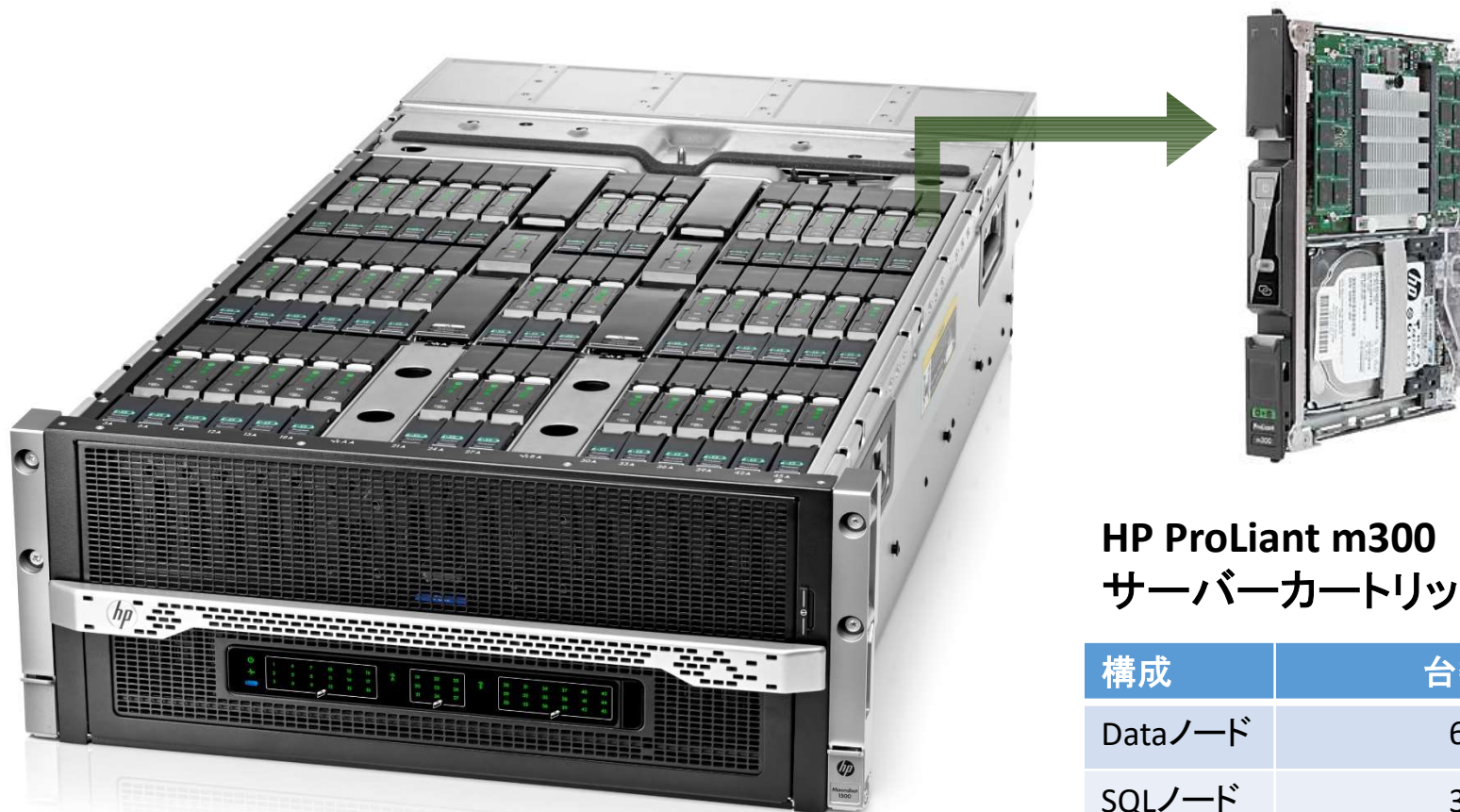


スケール	スペック
会員数/増加数	300万人/3万人
同時ユーザ数	1万人
トランザクション	1万(1秒間)
稼働率	99.999%

構成	台数
Dataノード	12
SQLノード	12
管理ノード	2
スペック	CPU : x86 24core メモリ : 64GB OS : Linux

出典 : Oracle社のMySQL Cluster顧客情報ページ  
[https://blogs.oracle.com/MySQL/entry/mysql\\_cluster\\_powers\\_el\\_chavo](https://blogs.oracle.com/MySQL/entry/mysql_cluster_powers_el_chavo)

# ナレッジベース「障害情報システム」



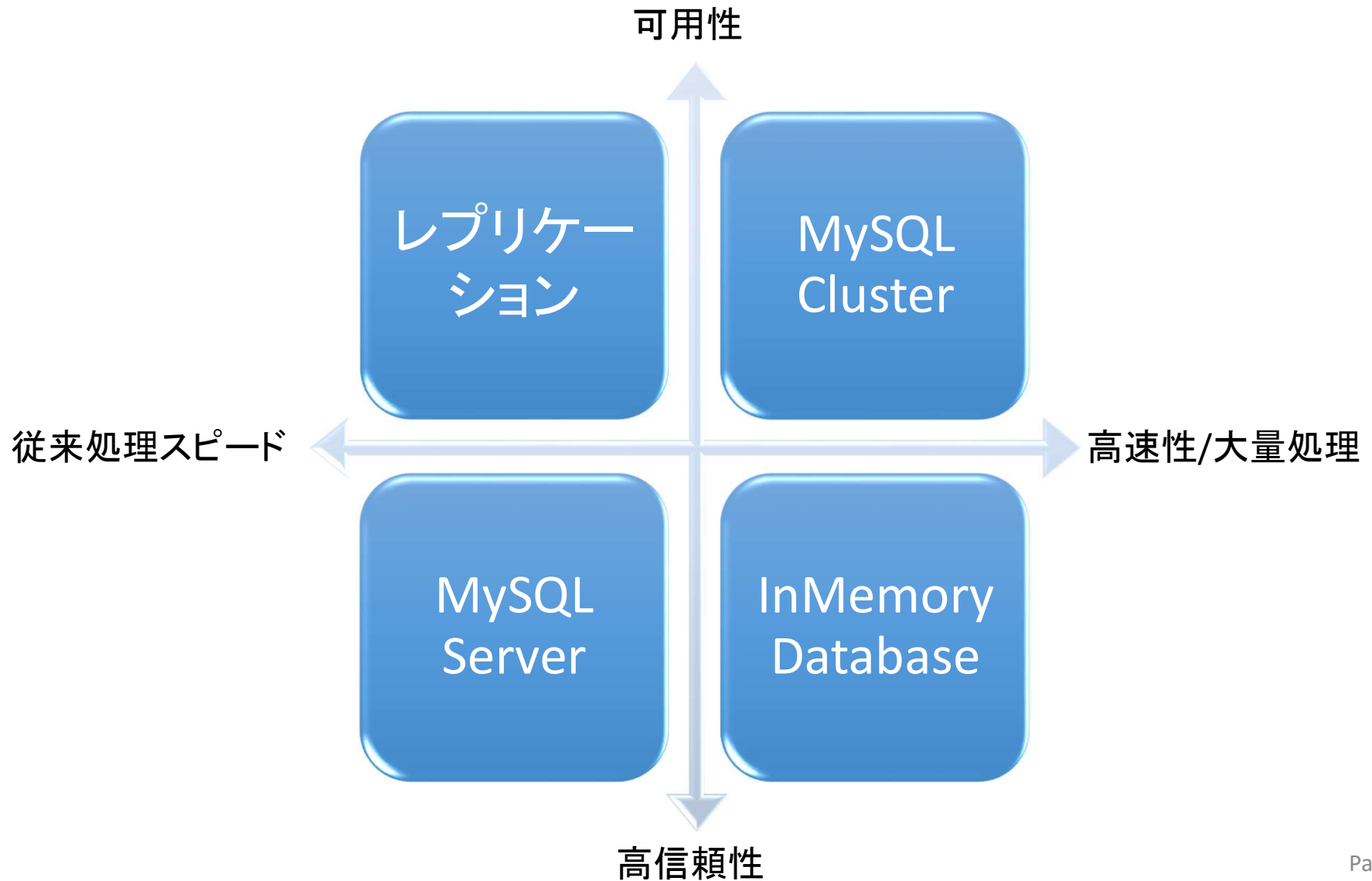
HP Moonshot 1500 シャーシ

HP ProLiant m300  
サーバーカートリッジ

構成	台数
Dataノード	6
SQLノード	3
管理ノード	2
スペック	HP Moonshot System

出典: 日本HP お役様事例ページ  
<http://h50146.www5.hp.com/products/servers/news/casestudy/jeis/>

# 同様製品との比較(ポジション)



# OSS(Community)とコマーシャル(Commercial)の違い

- MySQLコミュニティ
  - MySQLは、中心となる企業がすべての権利を掌握して、GPLライセンスと商用ライセンスを提供するデュアル・ライセンスとなっています。
  - MySQLのすべての権利は、MySQL ABからサン・マイクロシステムズを経てOracleが商標権ならびに著作権を保持しています。
- MySQLのライセンス
  - GPL (General Public Licenseに基づく提供)
  - 商用ライセンス(BugFixおよびサポートなど提供するサブスクリプションライセンス)

• MySQL Standard Edition (1-4socket Server)	240,000円
• MySQL Enterprise Edition (1-4socket Server)	600,000円
• MySQL Cluster Carrier Grade Edition (1-4socket Server )	1,200,000円
• MySQL Standard Edition (5+ socket Server)	480,000円
• MySQL Enterprise Edition (5+ socket Server )	1,200,000円
• MySQL Cluster Carrier Grade Edition (5+ socket Server	2,400,000円

# MySQL Cluster 歴史

- Ericssons社によって、1990年代にテレコム／IP環境向けの高可用性クラスターデータベースが開発されました。
- Ericsson社は、2000年にベンチャー企業としてAlzato社を設立し、Alzato社が高可用性ネットワークデータベースシステム「NDB Cluster」として製品化いたしました。
- MySQL社は、2003年9月にAlzato社を取得し、「NDB Cluster」を「MySQL」のストレージエンジンの一つとして取り入れ、新製品「MySQL Cluster」として発表しました。現在は、MySQL5.7対応の7.5が開発中となっています。

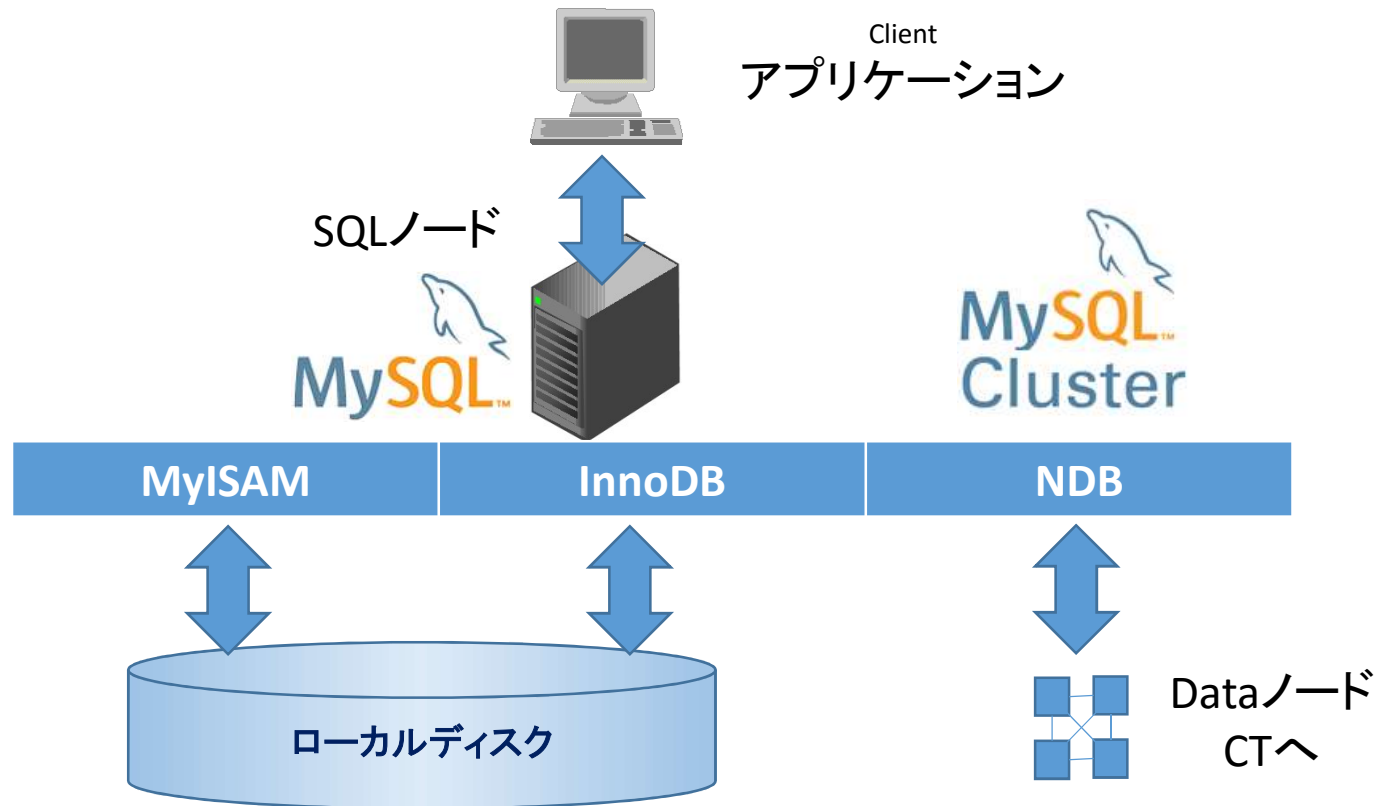
バージョン	最終バージョン	リリース日	プレミアサポート終了日	エクステンドサポート終了日	特徴
Cluster 6.1 6.1.0(5.1.14) 2006/12/20	6.1.23	2007/11/20	2013-03-31		
Cluster 6.2 6.2.5GA(5.1.22) 2007/09/06	6.2.18	2009/06/01			
Cluster 6.3 6.3.8GA (5.1.23) 2008/01/29	6.3.51	2013/02/01			
Cluster 7.0 7.0.5GA(5.1.32) 2009/04/20	7.0.37	2013/02/01	2014-04-30		マルチスレッド化
Cluster 7.1 7.1.3GA(5.1.44) 2010/04/12	7.1.37	2015/09/08	2015-04-30		
Cluster 7.2 7.2.4GA(5.5.19) 2012/02/15	7.2.23(5.5.47)	2016/01/19	2017-2-28	2020-2-28	5.5対応/Join高速化
	7.2.24(5.5.47)	予定			
Cluster 7.3 7.3.2GA(5.6.11) 2013/06/18	7.3.12(5.6.28)	2016/01/19	2018-6-30	2021-6-30	5.6対応/Join高速化
	7.3.13(5.6.28)	予定			
Cluster 7.4 7.4.4GA (5.6.23) 2015/02/26	7.4.10(5.6.28)	2016/01/29	2020-2-29	2023-2-28	外部キーサポート
	7.4.11(5.6.28)	予定			
Cluster 7.5 Developer Milestone	7.5.0(5.7.10)	2016/02/05			
	7.5.1(5.7.11)				
	7.5.2(5.7.11)				

# 何が優れているのか

- 高速なトランザクション処理が可能
  - NoSQLベースのデータシステムで不向きな処理
  - Dataノードの追加によりスケールアウト
  - NoSQLインターフェイスの提供(memcached互換API)
  - MySQLとの共存
    - データ処理を行うDataノードとリクエストを受けるSQLノードが分離しているので、処理内容に応じて台数を調整することが可能

# MySQLノードの役割

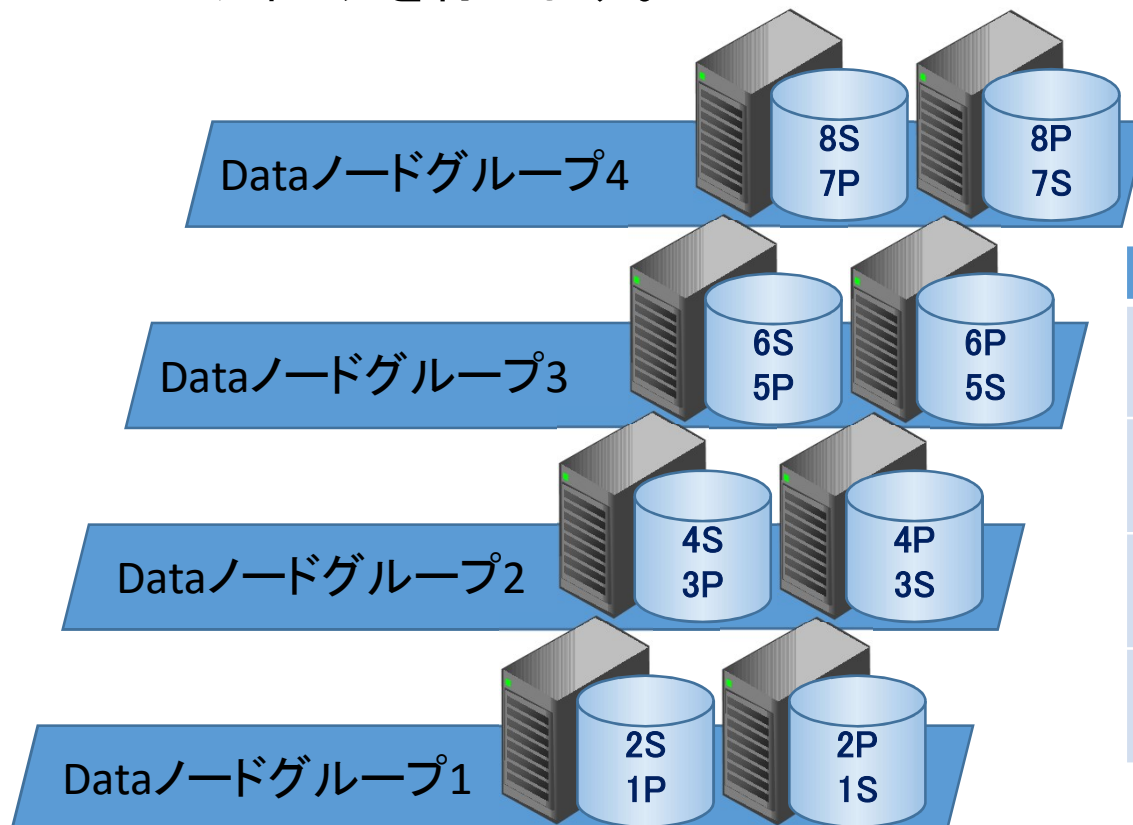
- SQLノードは、通常のストレージエンジンに加えてNDBを搭載しています。
- SQLステートメントを解釈して、使用するストレージエンジンを選択します。





# データ格納モデル

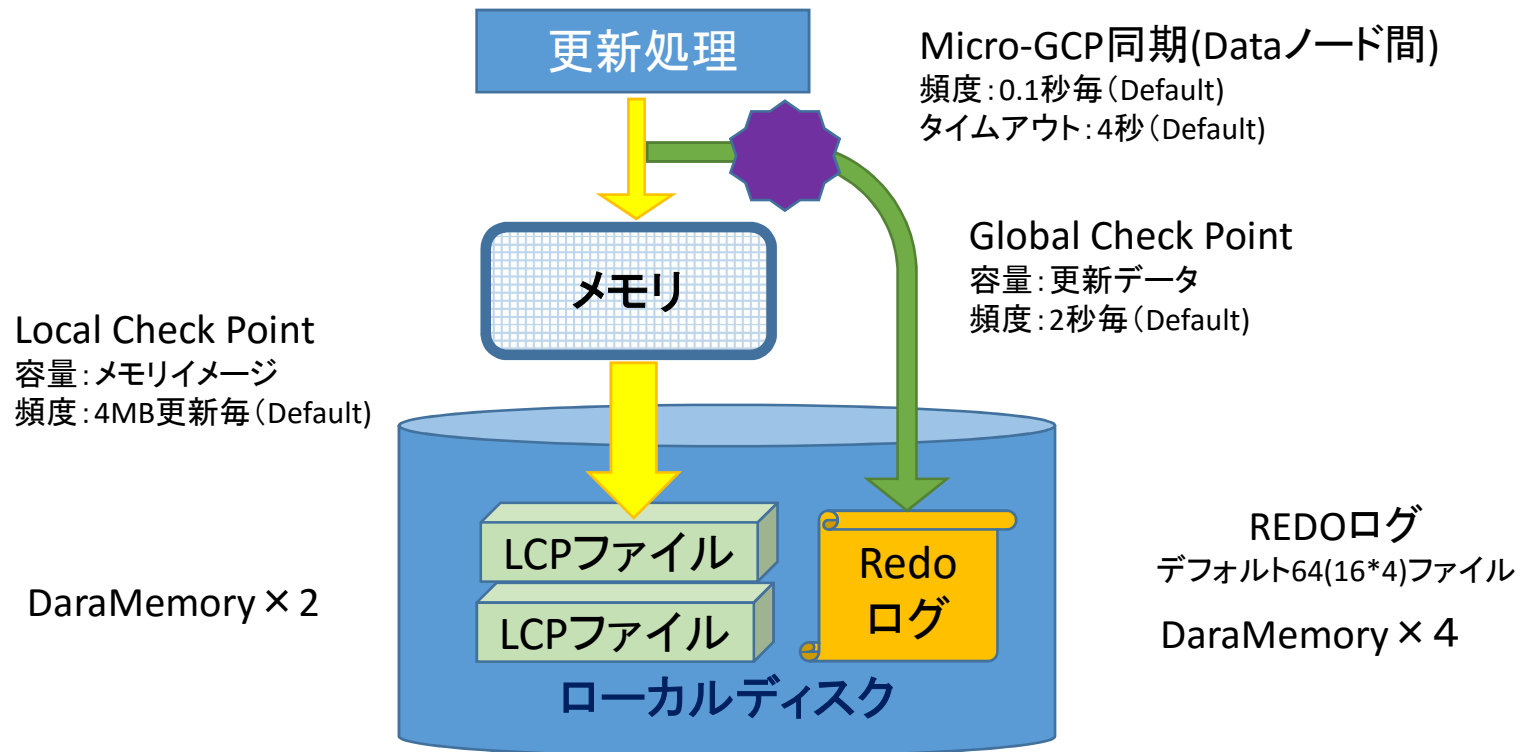
- データ全体を自動的にシャーディングしてノードに分散します。
- ノードグループ毎に片方のノードが起動していれば正常稼働します。
- 後からノードグループを追加した場合は、テーブルの再作成でシャーディングを行います。



グループ	パーティション	偶数ノード	奇数ノード
4	8	S	P
	7	P	S
3	6	S	P
	5	P	S
2	4	S	P
	3	P	S
1	2	S	P
	1	P	S

# データを保持する仕組み(GCPとLCP)

- MySQL Clusterは、Dataノード間の同期とOnMemoryDatabase保持のために次のような処理を行います。
  - Dataノード間は、一定周期で更新内容をEpoch(Micro-GCP)として記録します。同期信号(Micro-GCP同期)がタイムアウトとなった場合、当該Dataノードは、ダウンとみなされます。
  - Epochを一定周期でディスクに書き出す処理がGCP(Global Check Point)です。
  - メモリの内容をファイルとして保存するのが、LCP(Local Check Point)です。



# 留意しなければならない点

- 複雑なSQLステートメントを避ける
  - 複数テーブル処理が苦手、特にJOINはInnoDBよりも低速
- ロック/トランザクションのタイムアウト
  - 高速性と可用性を維持するためにロックを取得できない(デフォルト1.2秒間)とエラーとなる
  - トランザクションはフェールオーバーされる
- テーブルの構造
  - PRIMARY keyが必須
- ネットワークの重要性
  - 切断による障害
  - 暗号化などのセキュリティが行われていない
- 仮想環境の見えないボトルネック
- つまり、MySQL Clusterを意識した設計が必要です。

# MySQL Cluster サンプル構成の構築

# サンプル構成

- 1サーバ上(VMでも可)にMySQL Clusterを構築
  - これができれば、個別に作成することは難しくありません。
- ベース環境
  - OS: CentOS6
    - CPU: 2
    - メモリ: 8GB
    - ディスク: 10GB
  - 使用ID root
- MySQL Clusterノード構成
  - 管理ノード: 1
  - Dataノード: 4
  - SQLノード: 4

ノード	使用ディレクトリ	使用ポート
管理ノード	/var/lib/mysql-cluster/mgm	1186 *default
Dataノード1	/var/lib/mysql-cluster/data1	エフェメラル ポート
Dataノード2	/var/lib/mysql-cluster/data2	
Dataノード3	/var/lib/mysql-cluster/data3	
Dataノード4	/var/lib/mysql-cluster/data4	
SQLノード1	/var/lib/mysql-cluster/sql1	10001
SQLノード2	/var/lib/mysql-cluster/sql2	10002
SQLノード3	/var/lib/mysql-cluster/sql3	10003
SQLノード4	/var/lib/mysql-cluster/sql4	10004

\* SQLノードのデフォルト使用ポートは3306です。

\*エフェメラルポートは、規定範囲内のポートを動的に使用する仕組みです。範囲として32768～61000(デフォルト)が使用されます。

# インストール手順概略

- 事前準備
  - SELinuxの解除(停止)
  - Iptablesの無効化(停止)
- 1 MySQL-Clusterのインストール
  - rpmダウンロード
  - 展開
  - rpmインストール①
  - mysqlユーティリティ削除
  - rpmインストール②
  - rpmインストール③(SQLノードのみ)
  - ディレクトリ作成
  - config.ini配置
  - SQLノードのデータディレクトリ初期化
- 2 最初の起動
  - 管理ノードを初期起動
  - Dataノードを初期起動
  - Dataノードと管理ノードの起動確認
  - SQLノードの起動
  - SQLノードの起動確認
  - SQLノードへログイン確認
  - NDBエンジン実装を確認
  - SQLノードを停止
  - 管理ノードとDataノードを停止
- 3 通常起動
  - 管理ノードの通常起動
  - Dataノードの通常起動
  - SQLノードの通常起動
  - SQLノードを停止
  - 管理ノードとDataノードを停止

# config.ini

## [MGM DEFAULT]

Portnumber = 1186

## [MGM]

NodeId = 1

DataDir = /var/lib/mysql-cluster/mgm

HostName = 127.0.0.1

## [TCP DEFAULT]

## [MYSQLD DEFAULT]

## [NDBD DEFAULT]

NoOfReplicas = 2

NoOfFragmentLogFiles = 4

DiskPageBufferMemory = 4M

## [NDBD]

NodeId = 3

Hostname = 127.0.0.1

DataDir = /var/lib/mysql-cluster/data1

## [NDBD]

NodeId = 4

Hostname = 127.0.0.1

DataDir = /var/lib/mysql-cluster/data2

## [NDBD]

NodeId = 5

Hostname = 127.0.0.1

DataDir = /var/lib/mysql-cluster/data3

## [NDBD]

NodeId = 6

Hostname = 127.0.0.1

DataDir = /var/lib/mysql-cluster/data4

## [MYSQLD]

NodeId = 50

[MYSQLD] \*同様に10行以上

# MySQL Cluster 監視ツール

MySQL Enterprise Monitorによる監視



# コマンドベースの監視

管理ノードおよびDataノードに関しては、管理コマンドで監視します。

- 各ノードの状況

```
[root@dhcp-165 ~]# ndb_mgm -e show
Connected to Management Server at: localhost:1186
Cluster Configuration
-----
[ndbd(NDB)] 4 node(s)
id=3      @127.0.0.1 (mysql-5.6.28 ndb-7.4.10, Nodegroup: 0, *)
id=4      @127.0.0.1 (mysql-5.6.28 ndb-7.4.10, Nodegroup: 0)
id=5      @127.0.0.1 (mysql-5.6.28 ndb-7.4.10, Nodegroup: 1)
id=6      @127.0.0.1 (mysql-5.6.28 ndb-7.4.10, Nodegroup: 1)

[ndb_mgmd(MGM)] 1 node(s)
id=1      @127.0.0.1 (mysql-5.6.28 ndb-7.4.10)

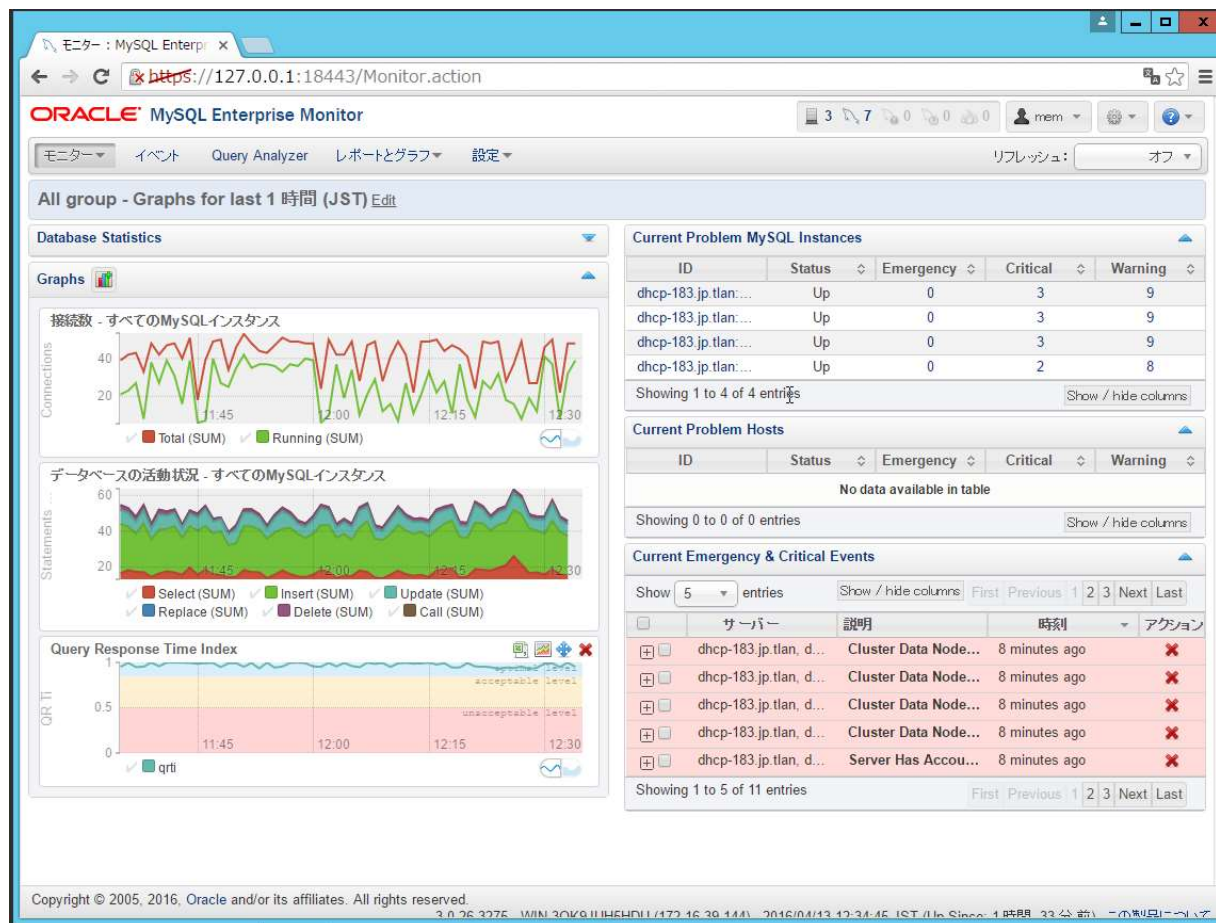
[mysqld(API)] 21 node(s)
id=50     @127.0.0.1 (mysql-5.6.28 ndb-7.4.10)
id=51     @127.0.0.1 (mysql-5.6.28 ndb-7.4.10)
id=52     @127.0.0.1 (mysql-5.6.28 ndb-7.4.10)
id=53     @127.0.0.1 (mysql-5.6.28 ndb-7.4.10)
id=54     @127.0.0.1 (mysql-5.6.28 ndb-7.4.10)
id=55     @127.0.0.1 (mysql-5.6.28 ndb-7.4.10)
id=56     @127.0.0.1 (mysql-5.6.28 ndb-7.4.10)
```

- Dataノードの使用率

```
[root@dhcp-165 ~]# ndb_mgm -e 'all report memoryusage'
Connected to Management Server at: localhost:1186
Node 3: Data usage is 1%(26 32K pages of total 2560)
Node 3: Index usage is 0%(23 8K pages of total 2336)
Node 4: Data usage is 1%(26 32K pages of total 2560)
Node 4: Index usage is 0%(23 8K pages of total 2336)
Node 5: Data usage is 1%(30 32K pages of total 2560)
Node 5: Index usage is 1%(24 8K pages of total 2336)
Node 6: Data usage is 1%(30 32K pages of total 2560)
Node 6: Index usage is 1%(24 8K pages of total 2336)
```

# MySQL Cluster の監視①

- SQLノードは、MySQL Enterprise Monitorから監視が可能です。  
標準的なMySQLサーバと同様の監視が可能です。
- SQLノードを通じてDataノードの状況も監視できます。



# MySQL Cluster の監視②

- イベントの取得と対応

The screenshot displays the Oracle MySQL Enterprise Monitor web interface. The browser address bar shows the URL: `https://127.0.0.1:18443/Events.action?showAllAssets_control=True&assetSelection=%5B%7B%22id%3A%22%2C%22assetClass%3A%22%22%7D%5D`. The interface includes a navigation bar with tabs for 'モニター', 'イベント', 'Query Analyzer', 'レポートとグラフ', and '設定'. The 'イベント' (Events) tab is active.

On the left, a tree view shows the asset hierarchy: 'All' > 'MySQL Cluster(172.16.39.144)' > 'dhcp-183.jp.tlan' > 'dhcp-183.jp.tlan:100'. The main panel displays event filters: '時間範囲' (All Time), '状態' (Open), 'Current Status' (すべて), and 'ワーストステータス' (重大). Below these are buttons for 'フィルタの適用', 'デフォルトとして保存', and 'Reset to Default'.

The event list shows two entries:

現在	最重要	サーバー	説明	時刻	アクション
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	dhcp-183.jp.tlan, dhcp-183.jp...	Cluster Data Node Has Been...	5 minutes ago	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	dhcp-183.jp.tlan, dhcp-183.jp...	Cluster Data Nodes Not Run...	about a minute ago	<input checked="" type="checkbox"/>

Below the table, it says 'Showing 1 to 2 of 2 entries'. At the bottom, there is a button '選択したイベントをクローズ' (Close selected events).

Copyright © 2005, 2016, Oracle and/or its affiliates. All rights reserved.

# MySQL Cluster の監視③

dhcp-183.jp.tlan, dhcp-183.jp... Cluster Data Nodes Not Run... 2 minutes ago

**Topic:** Cluster Data Nodes Not Running

**Categories:** クラスター **Advisor:** Clusterデータノードが起動していません

**Current State:** オープン **Closed By:** Closed: **ワーストステータス:** Critical

**Auto-Closes by Default:** No **Worst Alarm Time:** 2016/04/13 11:54:09

**Notes:**

メモの入力はありません。

**Details:**

**Problem Description**  
Indicates how many data nodes are not running.

**Advice**  
Only 4 out of 4 data nodes are running. View which data nodes are unavailable by running "ndb\_mgm -e show"

**Recommended Action**  
ndb\_mgm -e show

**Links and Further Reading**  
None specified.

**Expression**  
(%Ndb\_number\_of\_data\_nodes% - %Ndb\_number\_of\_ready\_data\_nodes%) >= THRESHOLD

**Evaluated Expression**  
(4 - 4) >= 1

## 内容(デフォルトチェック間隔)

Cluster データノードのデータメモリの空き容量が少なくなっています(5分)

Cluster データノードが再起動しました(5分)

Cluster データノードのインデックスメモリの空き容量が少なくなっています(5分)

Cluster データノード Redo バッファスペースが少なくなっています(5分)

Cluster データノード Redo ログスペースが少なくなっています(5分)

Cluster データノード Undo バッファスペースが少なくなっています(5分)

Cluster データノード Undo ログスペースが少なくなっています(5分)

Cluster データノードが起動していません(5分)

Cluster DiskPageBuffer のヒット率が低いです(5分)

Cluster が停止しました(2分)

dhcp-183.jp.tlan, dhcp-183.jp... Cluster Data Node Has Been... 5 m

**Topic:** Cluster Data Node Has Been Restarted

**Categories:** クラスター **Advisor:** Clusterデータノードが再起動しました

**Current State:** オープン **Closed By:** Closed: **ワーストステータス:** Critical

**Auto-Closes by Default:** No **Worst Alarm Time:** 2016/04/13 11:15:11

**Notes:**

メモの入力はありません。

**Details:**

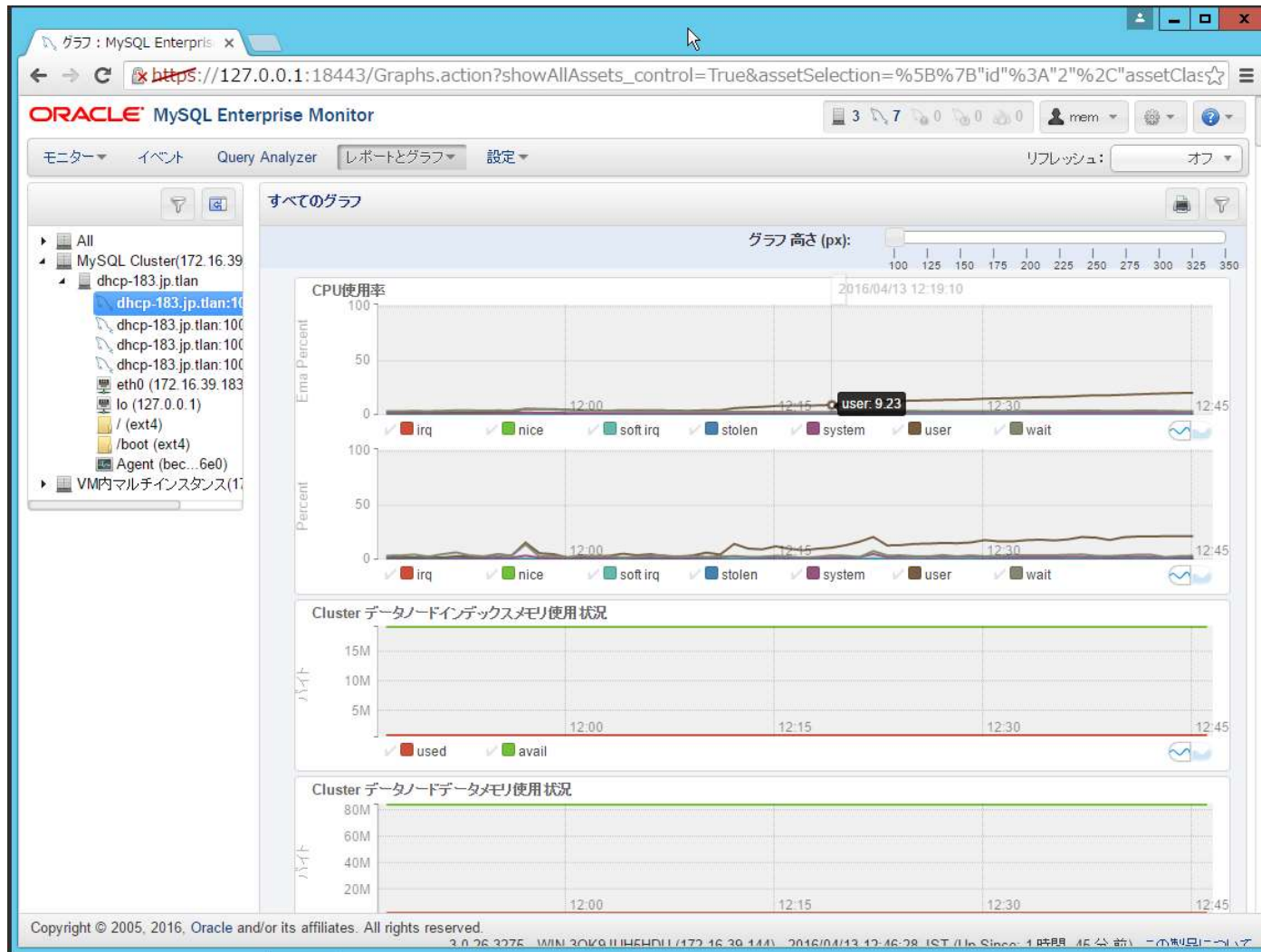
**Problem Description**  
To perform useful work, the cluster data nodes must be up-and-running continuously. It is normal for a production system to run continuously for weeks, months, or longer. If a data node has been restarted recently, it may be the result of planned maintenance, but it may also be due to an unplanned event that should be investigated.

**Advice**

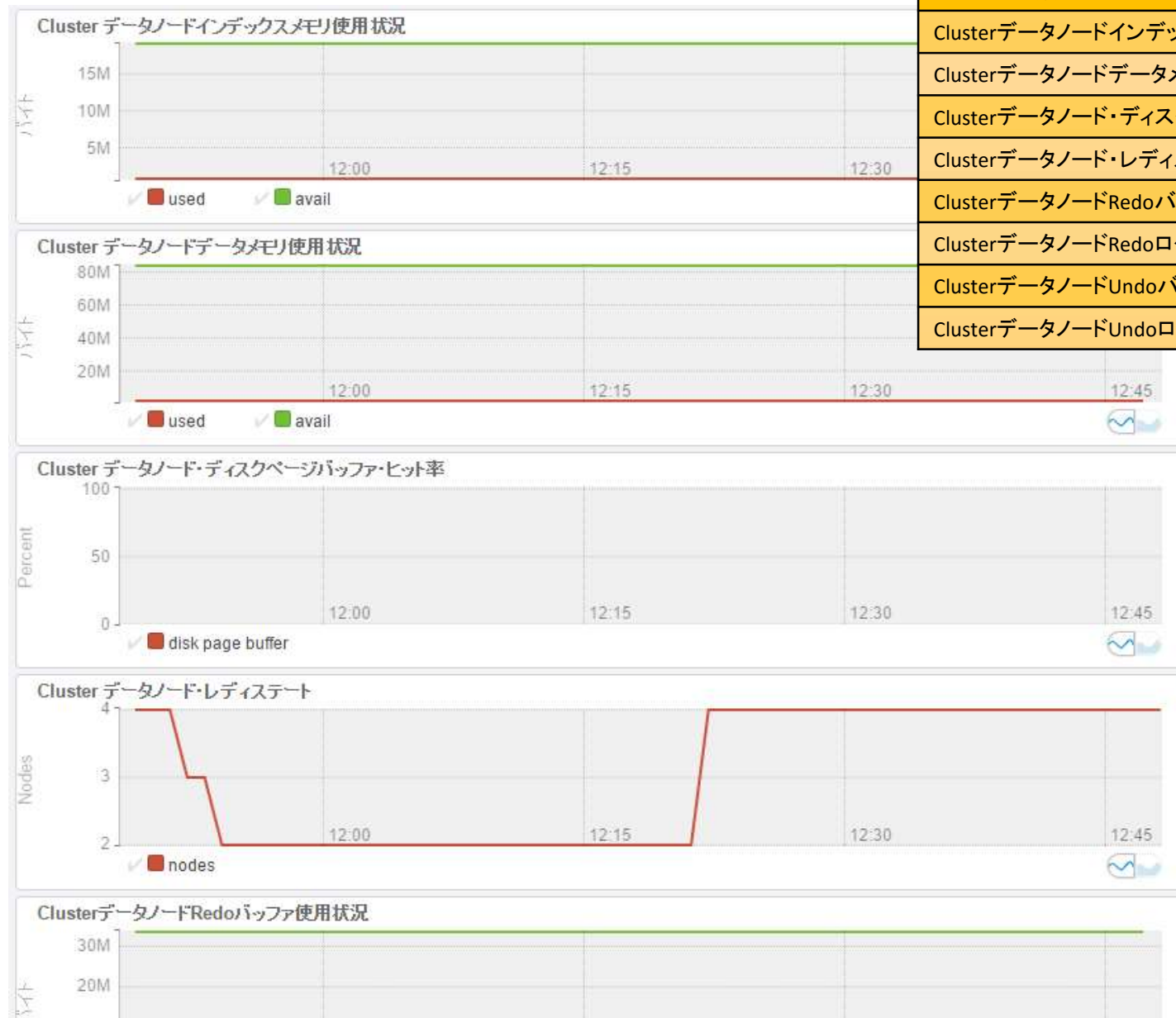
Read more ▾

# MySQL Cluster の監視④

- 取得したデータは、グラフとして表示できます。



# MySQL Cluster の監視⑤



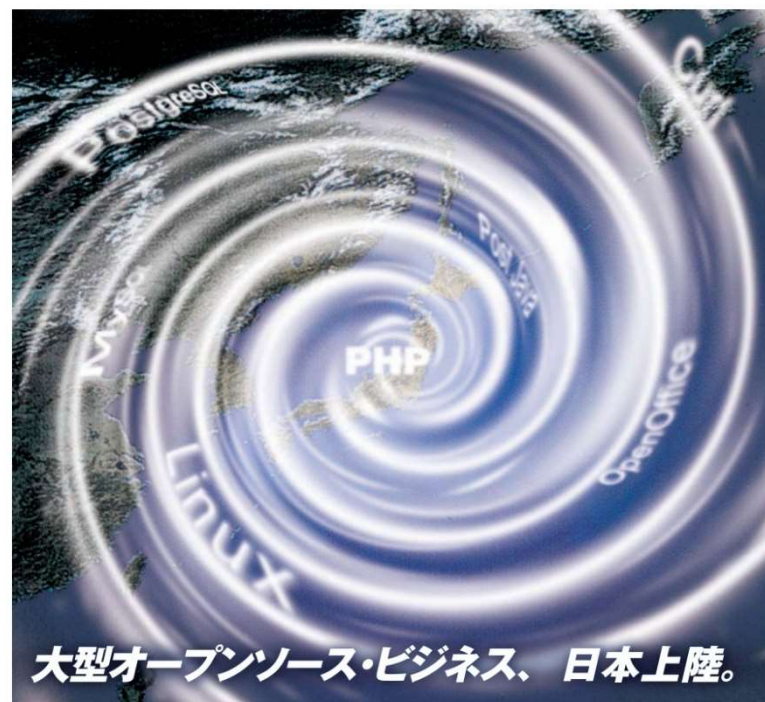
グラフの種類
Clusterデータノードインデックスメモリ使用状況
Clusterデータノードデータメモリ使用状況
Clusterデータノード・ディスクスペースバッファ・ヒット率
Clusterデータノード・レディステート
ClusterデータノードRedoバッファ使用状況
ClusterデータノードRedoログスペース使用状況
ClusterデータノードUndoバッファ使用状況
ClusterデータノードUndoログスペース使用状況



# konekto

企業向けにOSSを提供して15年

コネクト株式会社  
〒107-0052  
東京都港区赤坂4-8-14  
赤坂坂東ビル8F  
TEL:03-6434-7918  
FAX:03-6890-2220



大型オープンソース・ビジネス、日本上陸。

since 2001