## MIRACLE LINUXで構築する 高信頼データベース ソリューション



# 「高信頼DBクラスタ・ソリューション」 の検証結果のご紹介と RACとCLUSTERPROの導入ポイント

株式会社PFU

All Rights Reserved, Copyright PFU LIMITED 1996-2005

### **AGENDA**



- 5社で共同で検証した結果のご紹介
  - 一検証環境・検証概要
- RACとCLUSTERPROを導入する際のポイント
  - -SE RACとは
  - -SE RACの前提条件・注意事項
  - 一設計・構築時の留意事項
  - ー運用時の留意事項
- Oracle10g Release2 新機能
  - ー概要/Summary/SE RACでも使える新機能
- Q & A



# 5社で共同で検証した結果のご紹介





## 検証環境•検証概要



● 5社協業検証の実施 【期間】

(第一回) 2005年4月4日~4月15日

4ノード Oracle10g EE RACによる動作検証

(第二回) 2005年5月20日~6月3日

Oracle10g SE RACによる構築・動作検証

(第三回) 2005年9月22日~10月14日

現在実施中!

Oracle10g R2 SE RACによる構築・動作検証 64bit環境における構築・動作検証

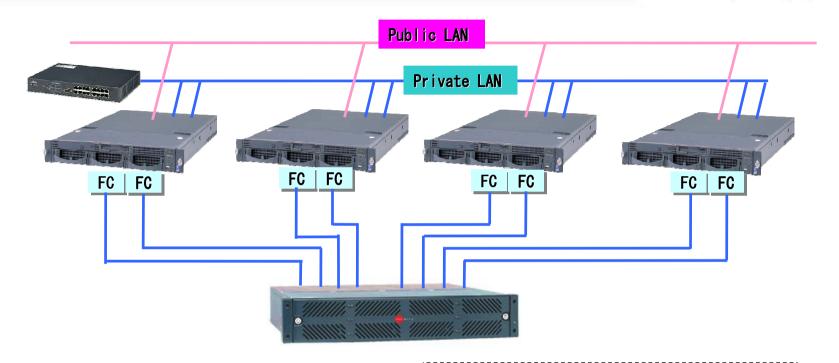
【場所】

富士通株式会社

Platform Solution Center (PSC) (浜松町貿易センタービル内)

## 検証環境 —4ノードEE RAC検証環境—





〈検証日時〉 2005年4月4日~4月15日

〈目的〉

Linuxソリューションの展開として、 各ベンダ製品間の動作検証 ・サーバ: PRIMERGY RX300

・ストレージ: SANnet II FC Type

• OS: MIRACLE LINUX V3.0

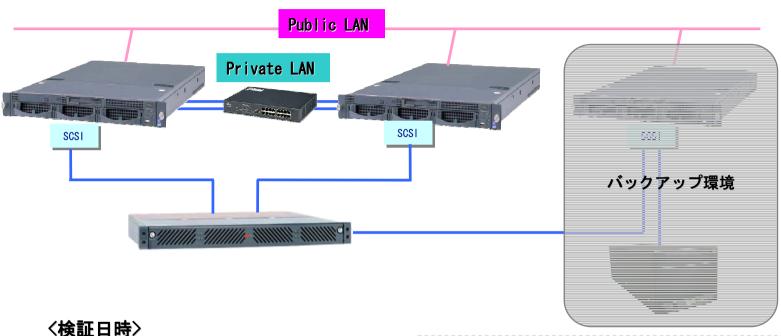
• DB: Oracle Database 10g EE (10.1.0.4)

Real Application Clusters

・クラスタ: CLUSTERPRO SE Ver3.1

## 検証環境 — 2ノードSE RAC検証環境—





2005年5月20日~6月3日

〈目的〉

Oracle10g SE RACを使用した安価な ソリューションの展開として、

構築手順確認と、動作検証

・サーバ: PRIMERGY RX300

・ストレージ: SANnet II Blade

• OS: MIRACLE LINUX V3.0

• DB: Oracle Database 10g SE (10.1.0.4)

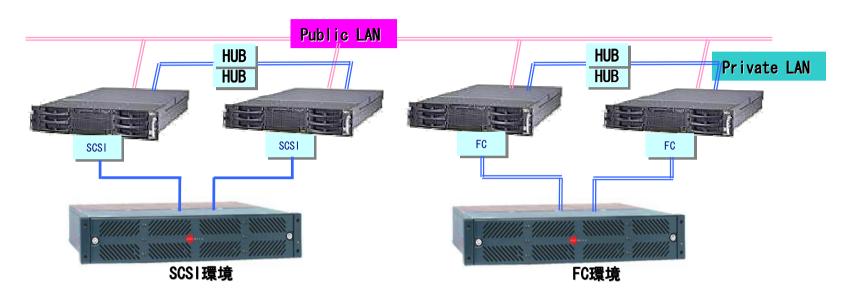
Real Application Clusters

・クラスタ: CLUSTERPRO SE Ver3.1

(バックアップ環境を除いた環境)

## 検証環境 —10gR2 & 64bit 検証環境—





〈検証日時〉

2005年9月22日~10月14日

〈目的〉

Oracle10g release 2を使用した 各ベンダ間の動作検証と新機能の確認 64bit環境での動作検証 ・サーバ: PRIMERGY RX300 S2

・ストレージ: SANnet II SCSI Type

& SANnet II FC Type

• OS : MIRACLE LINUX V4.0

• DB: Oracle Database 10g SE (10.2.0.1)

Real Application Clusters

・クラスタ: CLUSTERPRO SE Ver3.1



- ①4ノードEE RAC 検証概要
- 通常時運用テスト
  - -3ノード運用からの1ノード追加検証
  - 4 ノード運用時のDB高負荷時の動作検証
- 異常時運用テスト
  - ーストレージ・OS (サーバ)・データベース・クラスタ 各種異常時の動作検証とエラー検知機能の確認
  - ーノード交換(ノード削除→追加)
- ②2ノードSE RAC 検証概要
- 構築手順確認
  - ーストレージ・OS・データベース・クラスタの構築・設定
  - -ASMを使用したデータベース構築・設定
  - ーバックアップ環境の構築・設定
- 異常時運用テスト
  - 一前回検証時に問題のあった箇所の確認
  - -ASM領域のバックアップ・リカバリ手順確認
  - ーCRS・OCR領域・VotingDisk領域障害時の復旧手順確認



#### ● 異常運用テスト(共通)

項番	大項目	中項目
4	ストレージ	FCケーブル異常確認
'	ストレージ	
		コントローラ異常(擬似確認)
		SANpath(パス切替) 確認
		SANscape エラー検出
		SANscape Alertエラー検出
		ログの収集
2	サーバ(OS)	サーバ本体
		インターコネクト(NIC)の異常確認
		インターコネクトケーブルの異常確認
		パブリックLAN(NIC)の異常確認
		パブリックLAN(ケーブル)の異常確認
		FCケーブル異常確認
		システム高負荷
		2台目shutdown
3	Oracle	インスタンス消失
		CRS消失
		インスタンスストール
		リスナー消失
		リスナーストール
4	CLUSTERPRO	サスペンド
		プロセス消失
		業務移動(WebManagerの業務移動確認)
		メール送信
		ログ収集
5	DB負荷ツール	DBへの負荷をかける

さまざまな障害を想定し 動作を確認

DBに負荷をかけた状態 での稼動状況を確認



- ③10gR2 & 64bit 検証概要
- 構築手順確認
  - -10gR2 を使用した5社モデルの構築・動作検証
  - -64bit環境における構築・動作検証
- 異常時運用テスト
  - ーストレージ・OS (サーバ)・データベース・クラスタ 各種異常時の動作検証とエラー検知機能の確認
- 新機能確認
  - -Oracle10g R2 & CLUSTERPRO(3.1-5) 新機能確認

5社の最新モデルにてデータベース・サーバ環境を問題なく構築 出来ることを確認

あらゆる障害に対して検知・対処が適切に行われることを確認



#### ●主な問題点

- ·Oracleインスタンス・ストール、リスナー・ストール時の対処
  - ーストール状態の検知は、OracleよりCLUSTERPROのDB監視オプションがいち早く検知している
  - ⇒CLUSTERPROによるストール監視を行い、 ストール時はサーバ再起動を行う運用で回避可能
- ASMを使用したバックアップ・リカバリ手順
  - ーASMを使用すると、通常のRMANリカバリだけでは復旧できない
  - ⇒ディスク・グループの再作成後、リストア・リカバリで復旧可能

問題点をいち早く洗い出し、適切な対処法・回避法を検証し、 安定・安心のシステムをご提供可能に

> ※本検証に関する報告書は、PSC様から掲載予定です http://primeserver.fujitsu.com/partner/rov

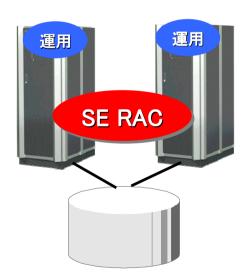


# RACとCLUSTERPROを導入する際のポイント

#### SE RAC とは



- Oracle Database 10g Standard Editionでは、 特別なライセンス無しで Real Application Clustersを利用可能 (Oracle 9i以前では利用できません)
- 中・小規模のシステムでも 低価格で高度な可用性を実現可能
- いくつかの制限事項あり(後述)





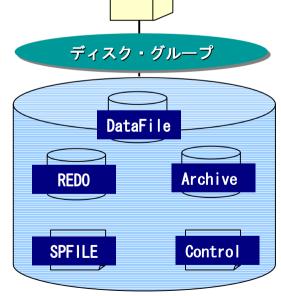


- クラスタ構成をとる全サーバの搭載プロセッサ数の合計 が4プロセッサまで(2CPUサーバの場合2台)
- ◆ クラスタウェアとして
   CRS (Cluster Ready Services)の使用が必須
   ※CRSとは、プラットフォームに特化しない、Oracle社が提供するクラスタウェアです
- ストレージマネージャとして ASM (Automatic Storage Management) の使用が必須 ※ASMについては、次項にて説明
- Enterprise Edition特有の機能は使用できない (DataGuard、オンライン表再定義、ファイングレイン監査、 Partitioning、Flashback Database/Table ...etc)

#### 設計・構築時の留意事項 —ASMのディスク・グループ配置—



- ASM(Automatic Storage Management)とは
  - -複数の物理ディスクを集めて 仮想的に1つのディスクとして利用する機能
  - ーミラーリング・ストライピング・リバランシング
    - の機能を備える
    - ※ハードウェアRAIDが可能であれば、 ミラーリングはハード側で行うのが推奨 (今回の検証もRAID1を使用)
  - ーディスクの追加・削除が EM(Enterprise Manager) やSQLで 容易に実行・管理できる

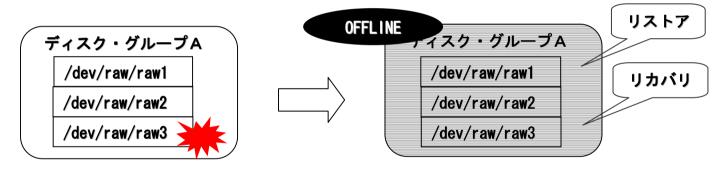




#### 設計・構築時の留意事項 —ASMのディスク・グループ配置—



- ディスク・グループはどう設計すべきか?
  - 1つのディスク・グループでデータベース全体を管理することも可能→ただし1つのデバイスが完全に破壊されるとディスク・グループ全体のリストア→リカバリが必要となってしまう



→したがって、ある程度管理面を考慮した構成にする必要がある



#### 設計・構築時の留意事項 —SE RACとCLUSTERPROの連携—



- SE RACの制限事項
  - ークラスタウェアとしてCRSの使用が必須
    - →「他社製クラスタウェアからOracleを管理しないこと」 というライセンス上の制限
    - →CRSからはデータベースの管理のみ可能
    - →<u>データベース以外のサーバ監視&管理はどうするか?</u>

- SE RACとCLUSTERPROの連携
  - -CLUSTERPROから、Oracle (CRS) の監視やCRSでは監視できないが、 サーバ内の障害(サーバダウン、OSパニック、ディスク故障etc) の監視を行い、速やかなフェイル・オーバを実現可能

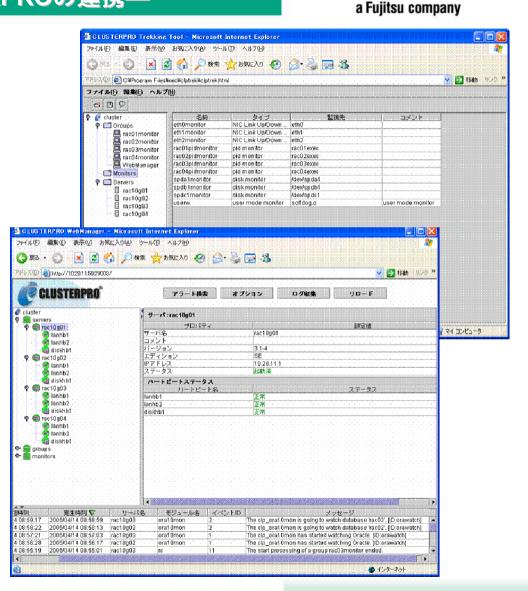




#### 設計・構築時の留意事項 —SE RACとCLUSTERPROの連携—



- CLUSTERPROの設定(1)
  - 「Trekking Tool」により、
    - クラスタ(サーバ)設定
    - ・監視対象・監視間隔の設定
    - 異常時の最終動作等を簡単に設定
  - 「Web Manager」により、
    - ・クラスタの状態監視
    - サーバ/グループの起動/停止
    - ・ログ収集 等が容易に可能

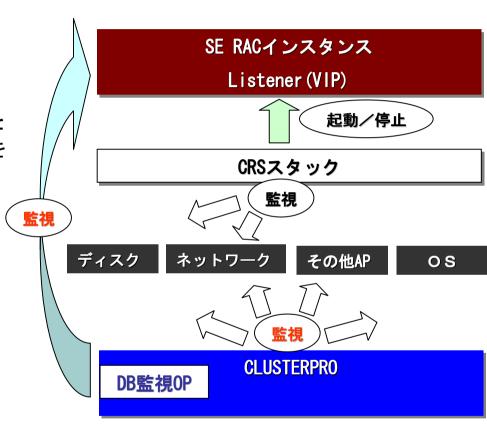


#### 設計・構築時の留意事項 —SE RACとCLUSTERPROの連携—



#### ● CLUSTERPROの設定(2)

- 「DB監視オプション」を使用すると インスタンスの監視及び起動・停止を CLUSTERPROから実行可能
- -Oracle10gに対応した監視スクリプトが提供されているため、導入も容易に可能(カスタマイズも可)
- →ただしSE RACの場合は、起動
- 停止は設定せず、インスタンスの 監視のみを行う設定をする (SE RACの制限事項により)





- ASM領域のバックアップ
  - -Oracleが提供しているのはRMAN(Recovery Manager)を使用する方法のみ (一部のストレージではストレージ機能の使用が可能)
  - -ASM領域にバックアップファイルを置くことも可能
  - ※バックアップスクリプトの例(RMAN部分のみ抜粋)

```
su - oracle -c "rman target sys/serac@serac" > ./rman.log 2>&1 << END:RMANBACKUP
run{
configure channel device type disk format '/oraback/%U';
configure controlfile autobackup on:
configure controlfile autobackup format for device type disk to '/oraback/%F';
backup full database plus archivelog;
sql "create pfile=''/oraback/initserac.ora.bak'' from spfile";
}
exit
END:RMANBACKUP</pre>
```



- ASM領域のリカバリ
  - -基本的には通常のRMANを使用したリカバリで対処可能
  - ーただし、ASM特有の対処が必要な場合もある
- (例) ディスクグループ内の 1 デバイスの全壊 SYSTEMやユーザ表領域が含まれるディスク・グループ内のデバイス に対して不正な書き込みを行い、障害を発生させた



以下のエラーが発生して、インスタンスダウン!!

Errors in file /opt/app/oracle/admin/serac/bdump/serac1\_lmon\_8637.trc:

ORA-00202: control file: '+ASMDG1/serac/control01.ctl'

⇒次ページ以降のリカバリ対処が必要



- ●ディスク・グループ内の1デバイス全壊の対処
  - ① リカバリ対象ディスク・グループを削除→再作成

再作成の例

SQL> create diskgroup ASMDG1 external redundancy disk

2 '/dev/raw/raw4','/dev/raw/raw5';

ディスク・グループが作成されました。

- ② データベースをnomountで起動
- ③ ディスク・グループ2にある正常な制御ファイルをディスク・グループ1に複製

RMAN> restore controlfile from '+ASMDG2/serac/control02.ctl';

restoreが開始されました(開始時間: 05-09-16)

チャネル: ORA\_DISK\_1が割り当てられました

チャネルORA\_DISK\_1: SID=146 インスタンス=serac1 devtype=DISK

チャネルORA DISK 1: 制御ファイル・コピーをコピーしました

出力ファイル名=+ASMDG1/serac/control01.ctl

出力ファイル名=+ASMDG2/serac/control02.ctl

出力ファイル名=+ASMDG1/serac/control03.ctl

restoreが完了しました(完了時間: 05-09-16)



④ データベースの マウント→リストア→リカバリ→オープン

RMAN> recover database;

リカバリの例

recoverが開始されました(開始時間: 05-09-16)

チャネルORA\_DISK\_1の使用

メディア・リカバリを開始しています

メディア・リカバリが完了しました。経過時間: 00:00:01

recoverが完了しました(完了時間: 05-09-16)

- ⑤ ディスク・グループ1にあるRedoログメンバーを削除→再作成
- ⑥ クラスタ(RAC)として起動

⇒リカバリ完了!!

※その他ディスク・グループの配置等の条件よって手順が変わってくる ⇒リカバリには十分なテストが必要

# 運用時の留意事項



—DISKの追加・削除—

■ ASMを使用するとDISK(デバイス)の追加・削除等のメンテナンスが容易に実行可能





2 デバイスで構成されている ディスク・グループ

> 1 デバイスを追加 (SQLでも実行可能)

簡単に追加ができ、さらにデータのリバランシングもされている







- CRSの再インストールが必要となるケース (他の方法で可能なケースもあります)
  - -CRSのインストール失敗
  - -0CR領域の修復
  - -Voting Disk領域の修復
  - ーサーバのパブリックIP、プライベートIP、ホスト名の変更
    - ※OCR領域…CRS稼動に必要な情報を格納する共有ディスク上のレジストリ
      ※Voting Disk領域…ハートビートやサブクラスタ決定に使用される共有
      ディスク上の領域
- 再インストール手順(1)
  - ① \$0RACLE\_HOME/crs/installにてrootdelete.sh、rootdeinstall.sh を実行
    - →Oracle、Listener、CRSデーモンの停止を確認



- 再インストール手順(2)
  - ② rootユーザにて、OS起動時のCRS起動を制御するため下記を実行

```
# rm -f /etc/init. d/init. crs
# rm -f /etc/init. d/init. crsd
# rm -f /etc/init. d/init. evmd
# rm -f /etc/rc2. d/K96init. crs
# rm -f /etc/rc2. d/S96init. crs
# rm -f /etc/rc3. d/K96init. crs
# rm -f /etc/rc3. d/K96init. crs
# rm -f /etc/rc5. d/S96init. crs
# rm -f /etc/inittab. crs
# rm -f /etc/inittab. crs
```

※ファイルが既に存在しない場合もあり



- 再インストール手順(3)
  - ③ /etc/oracle、または/var/opt/oracle配下の「ocr. loc」を削除
  - ④ \$CRS\_HOME配下を削除→OUIからCRSの削除

# rm -Rf \$ORA\_CRS\_HOME

- ※①~④を全てのノードで実行
- ⑤ OCR、VotingDisk領域をクリーンアップ

# dd if=/dev/zero of=/dev/raw/raw1 bs=4096

# dd if=/dev/zero of=/dev/raw/raw2 bs=4096

- ⑥ CRSのインストール
- ⑦ VIPCA、NETCAでVIPとListenerを作成

名前	型	タート	状態	ホスト
ora01. Isni		ONLINE	ONLINE	serac01
orac01. rsiii		ONLINE	ONLINE	serac01
orac01.ons		ONLINE	ONLINE	serac01
orac01.vi;	application	ONLINE	ONLINE	serac01
ora02. Isni	application	ONLINE	ONLINE	serac02
orac02. gsc	application	ONLINE	ONLINE	serac02
ora c02. ons	application	ONLINE	ONLINE	serac02
orac02.vip	application	ONLINE	ONLINE	serac02

⑦実行後のcts\_statコマンドの状態→



- 再インストール手順(4)
  - ⑧ srvctlコマンドでデータベース&インスタンスを登録 (例) データベースseracの登録
    - \$ srvctl add database -d serac -o /opt/app/oracle/product/10.1.0/db
  - ⑨ srvctlコマンドでASMインスタンスを登録
    - (例) ASMインスタンス+ASM1の登録
    - \$ srvctl add asm -n serac01 -i +ASM1 -o /opt/app/oracle/product/10.1.0/db

#### <注意>

ASMインスタンスの起動完了後に、データベースインスタンスが起動するように 依存関係を必ず設定します

- (例) インスタンスserac1とASMインスタンス+ASM1の依存関係設定
  - \$ srvctl modify instance -d serac -i serac1 -s +ASM1
- ⑩ srvctlコマンドでサービスの登録
- (例) インスタンスserac1、serac2をサービスserac\_srvに登録
  - \$ srvctl add service -d serac -s serac\_srv -r serac1, serac2



- 再インストール手順(5)
  - ⑧ srvctlコマンドで各リソースを起動して完了

名前 型	タート	、状態	ホスト
ora. serac. db application	ONLINE	ONLINE	serac02
orac1.inst application	ONLINE	ONLINE	serac01
orac2.inst application	ONLINE	ONLINE	serac02
orasrv.cs application	ONLINE	ONLINE	serac01
oraac1.srv application	ONLINE	ONLINE	serac01
oraac2.srv application	ONLINE	ONLINE	serac02
oraSM1.asm application	ONLINE	ONLINE	serac01
oraO1.lsnr application	ONLINE	ONLINE	serac01
oracO1.gsd application	ONLINE	ONLINE	serac01
oracO1.ons application	ONLINE	ONLINE	serac01
oracO1.vip application	ONLINE	ONLINE	serac01
oraSM2.asm application	ONLINE	ONLINE	serac02
ora02.lsnr application	ONLINE	ONLINE	serac02
oracO2.gsd application	ONLINE	ONLINE	serac02
oracO2.ons application	ONLINE	ONLINE	serac02
oracO2.vip application	ONLINE	ONLINE	serac02

## 運用時の留意事項

—OCR/VotingDisk領域のリカバリ—



a Fujitsu company

● OCR領域のバックアップ

10g Release2 より

冗長化可能!!

-OCR領域はOracleが自動的にバックアップ採取している (例) OCRバックアップの確認

\$ ocrconfig	-showbackup	
serac01-ic	2005/06/02 06:51:21	/opt/app/oracle/product/10.1.0/crs/cdata/crs
serac01-ic	2005/06/02 02:51:21	/opt/app/oracle/product/10.1.0/crs/cdata/crs
serac01-ic	2005/06/01 22:51:21	/opt/app/oracle/product/10.1.0/crs/cdata/crs
serac02-ic	2005/06/01 03:53:06	/opt/app/oracle/product/10.1.0/crs/cdata/crs
serac02-ic	2005/05/30 23:38:38	/opt/app/oracle/product/10.1.0/crs/cdata/crs

- 「ocrconfig -export ...」コマンドで手動採取も可能
- 0CR領域のリカバリ
  - -0CR領域の破壊や不整合が発生した場合、下記の方法でリカバリが可能
    - 1.0CRを修復する
    - 2. バックアップからOCRを戻す ★本セミナーではこの方法を紹介
    - 3. CRSを再インストールする

#### 運用時の留意事項 —OCR/VotingDisk領域のリカバリ—



- 0CR領域のリカバリ〈手順〉
  - (例) 0CR領域デバイスの全壊
    →srvctl、crs\_statコマンド等はエラーが返る

\$ crs\_stat -t

PROC-22: OCRバックエンドに無効な形式があります

CRS-0184: CRSデーモンと通信できません。

- ① 障害や不整合が発生する前のバックアップが存在することを確認
  - \$ ocrconfig -show backup
- ② CRSを無効にして停止(もしくはシングルユーザモードで起動) 【全ノード】
  - # /etc/init. d/init. crs disable
  - # /etc/init. d/init. crs stop
- ③ CRSが起動していないことを確認後、ocroonfigコマンドでOCR領域をリストア
  - # ocrconfig -restore /opt/app/oracle/product/10.1.0/crs/cdata/crs/backup00.ocr
- ④ CRSを有効にして起動【全ノード】
  - # /etc/init.d/init.crs enable
  - # /etc/init.d/init.crs start

## 運用時の留意事項





a Fujitsu company

● Voting Disk領域のバックアップ -

10g Release2 より 冗長化可能!!

-Voting Disk領域はノードの追加・削除に応じて手動でのバックアップ 採取が必要

(例) Voting Disk領域のバックアップ

# dd if=/dev/raw/raw2 of=./voting0602.vot bs=4096

- Voting Disk領域のリカバリ
  - -Voting Disk領域の破壊や不整合が発生した場合、下記の方法でリカバリが可能
    - 1. バックアップからVoting Diskを戻す ★本セミナーではこの方法を紹介
    - 2. CRSを再インストールする

〈手順〉

- ① CRSの無効化・停止
- (OCR領域のリカバリと同様)
- ② バックアップのリストア

# dd if=./voting0602.vot of=/dev/raw/raw2 bs=4096

- ③ CRSの有効化・起動
- (OCR領域のリカバリと同様)

## 運用時の留意事項





- DB障害発生時の動作(CLUSTERPROとの連携)
  - -DBに障害が発生した場合、もしくはその他の障害でDB起動・停止の必要がある場合は、SE RACではOS停止・起動で対処可能。
    EE RACであれば、CLUSTERPROからのCRS起動(10.1.0.4~)・停止も可能。
- CRS監視とCLUSTERPRO監視

	CRS監視	CLUSTERPR0監視
共有DISKストール	IMR機能で15分で検知	2分(デフォルト)で検知
インスタンスやリスナー	IMR機能で15分で検知	2分(デフォルト)で検知
のストール	(下表参照)	
Oracle以外のAPの監視	不可	可能

注) IMR (Instance Membersip Recovery) 機能…クラスタの全てのメンバーがアクティブ であることを保障するための機能

※インスタンス・ストール検証時の動作

ノード再起動	既存セッション	新規セッション	CLUSTERPRO
行われない	応答なし	接続不可	2分(デフォルト)で検知

Oracle(CRS)のみで対処できない障害もCLUSTERPROですばやく検知





~ 概要 ~

● 2005年9月7日より「Oracle10g Release 2」を提供開始

-2005年 9月 7日 Linux x86

-2005年10月18日 Solaris Operationg System (SPARC)

HP-UX PA-RISC (64bit)

AIX5L Based Systems

-2005年11月 1日 Microsoft Windows (32-bit)

-2005年12月 6日 Linux x86-64

Linux Itanium

IBM Power Based Linux

[日本オラクル(株)ホームページ ニュースリリース〈2005/9/1〉より]

5社協業検証においても、いちはやく検証開始





- Grid Computing
  - -RACの拡張
  - ーローリング・アップグレード
  - -Oracle Data Guardの自動フェイルオーバー
- 管理性
  - -自動管理データベースの拡張
- セキュリティ
  - -Transparent Data Encryption
  - ーバックアップファイルの暗号化
  - 一外部パスワードの格納
- ビジネス・インテリジェンス
  - -0LAP
  - -Data Mining

Enterprise Editionでしか 使用できない機能が多い

~ SE RACでも使える新機能 ~





**● ASMCMDコマンドラインインターフェース** 

-ASMのファイル及びディレクトリを簡単に表示したり操作したりできる

[oracle@linchan11 oracle]\$ asmcmd
ASMCMD> Is
ASMDG1/
ASMDG2/
ASMCMD> cd ASMDG1
ASMCMD> pwd
+ASMDG1

● Cluster Verification Utility(クラスタ検証ユーティリティ)

-事前チェックを実行して、システムがRACのインストールのための基準

を満たしているかどうかを検証する

まだ誤検出がある場合があるので注意!

● 0CR, Voting Disk領域の冗長化 ーさらなる信頼性の向上



## 最後に



RACシステムを構築する前に、以下の点に留意して開発を進めてください。

- ノード間でのアクセス競合が発生しないようにアクセスを分割させる (複数ノード間の単一テーブルへのアクセスは極力発生させない)
  - ーアプリケーションによるノード間のデータ分割
  - ーデータによるノード間のデータ分割
  - ※キャッシュ・フュージョン機能によって性能問題は改善されましたが、頻繁な キャッシュ・フュージョンはシステムに影響を与える場合がありますのでご注意ください。
- フェイルオーバ形式を検討する
  - -TAF機能を使用するか
  - -VIP(仮想IP)などのクラスタ機能を使用するか
- **COMMIT間隔を不用に長くしない**
- • •

DBサーバ構築時には、ぜひ PFU にご相談ください!





・すべての名称並びに製品についての商標はそれぞれの所有者の商標または登録商標です。