広島大学におけるキャンパスグリッド事例

広島大学情報メディア教育研究センター
庄司　文由

SS研　Grid Computing 2004 WG 第4回会合

1

広島大学および情報メディア教育研究センターについて

- 広島大学
  - 学生数：約15000、教職員数：約3000
  - 中四国の基幹校（と自分たちは思っている）
  - 1.5流らしい
  - SuperSINETあり
- 情報メディア教育研究センター(2001～)
  - 総合情報処理センター（1981～2001）＋ 情報教育研究センター(1996～2001)
  - スタッフ：常勤20名
  - 情報プロパーは少ない(HPCはゼロ)
  - 予算は厳しい
  - システム（メール・Web・教育端末・HPC）は一括で5年レンタル（直近のリプレースは2005年）

2

なぜキャンパスグリッド？

- 進んでやろうとしたわけではない
  - 独法化後の予算配分事情
  - 何か新しいこと・ユメがあること？
  - お金がかからない

⬇

Grid技術を使った教育用端末の計算資源化

- お金はかからない
- 新しいしユメもある（かもしれない）
- 石井センター長（当時）のプッシュ

3

広島大学の端末環境

開室時間：平日9:00～22:00、土曜9:00～17:00まで
長期休暇はほとんどが閉室

PC約700台の潜在的計算資源

4

教育用端末の遊休時間を利用することについて

メリット

- **巨大な計算資源の創出**
  - HPC資源は32GFLOPS（～2005）、730GFLOPS（2005～）
  - 理論性能はクロック周波数x2x台数（3TFLOPS以上）
- **導入コストが低い**
  - 既存端末＋α
- **計算資源化しやすい**
  - スペック/運用ルールが均一
  - 地理的にある程度集約されている
- **新規アプリケーションへの柔軟な対応**
  - スカラアプリ、分散並列アプリユーザーの新規開拓

5

端末の利用方法について

配慮した点

- ✓ 既存サービス（教育用端末環境）に影響が出てはいけない
- ✓ “使えるモノ”にしたい
  - ✓ 一発屋は嫌だ
  - ✓ 高信頼性
  - ✓ 柔軟性

方針

- ✓ 別々のOS
- ✓ 丸ごとクラスタはしない
- ✓ 夜間・休日・長期休暇に限る

6

別々のOS

- 責任問題の回避
- デュアルブート(ディスクレス)システムは導入事例が多い
- 株式会社ミントウェーブのVIDシステムを採用
  - 教育用端末時
    - Linux/Windowsのデュアルブート
    - 電源断がデフォルト
  - クラスタ/グリッド時
    - 専用Linux
    - ローカルディスクをワーク領域に
  - 端末40～50台に対し、VID-IO/NFSサーバ1台

メリット

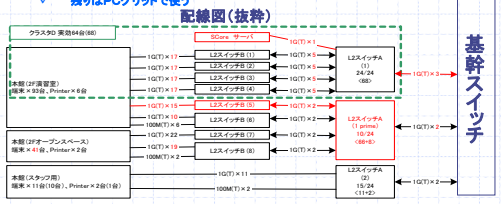
- 原理的に1台分のボリュームイメージがあればよいので、ディスク容量が大幅に節約できる
- 端末個々のディスク障害が運用に影響しない
- 一度の作業で全端末分の設定変更が可能
- 複数のイメージを切り替えることで、クラスタのOSを瞬時に切り替えることができる

7

丸ごとクラスタはしない

- Myrinet/InfiniBandは無理だし、ギガで組むと大規模構成は非現実的
- どうせ全体で使うのは最初のベンチマークだけ
- 数十台規模でも、ある程度ニーズには応えられる
- 端末のギガNICは教育用・クラスタ/グリッド用で共用する
- ギガビットスイッチ(L2)24ポートを使う
- クラスタ構成のところはカスケードは2段までで、アップリンクが多くなるように
- 残りはPCグリッドで使う

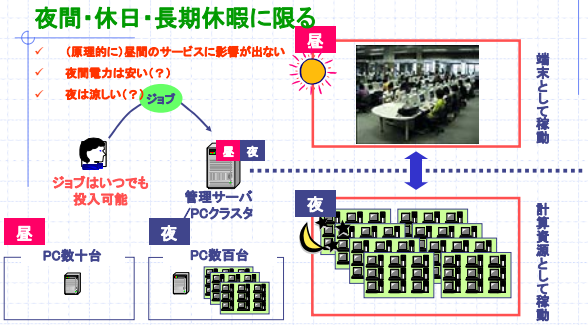
配線図(抜粋)



8

夜間・休日・長期休暇に限る

- (原理的に)昼間のサービスに影響が出ない
- 夜間電力は安い(?)
- 夜は涼しい(?)



- テストラン/小規模ジョブおよび大規模ジョブの投入は24時間可能
- 大規模ジョブの実行は夜間のみ

9

問題点

理想的には、

	昼	夜
コンパイル	○	○
小規模ジョブ(投入/実行)	○	○
大規模ジョブ(投入)	○	○
大規模ジョブ(実行)	×	○

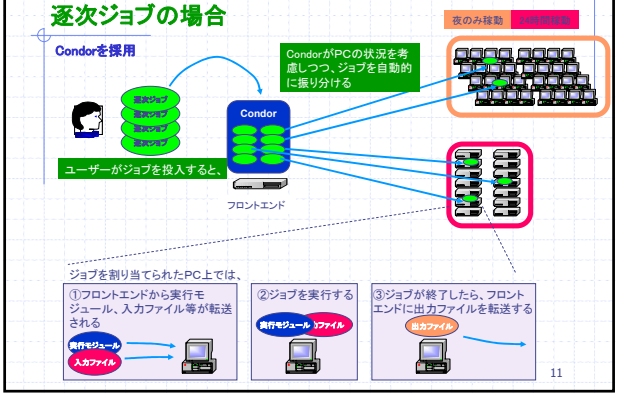
技術的問題点として、

- 長時間ジョブをどう扱うか(逐次/並列)
  - 一晩で終わらないジョブは永遠に終わらない?
- 端末個々の障害対策(逐次/並列)
  - サービスの質を考えると本質的
- 並列ジョブと逐次ジョブの資源分配
  - 潜在的なニーズは 逐次 >> 並列 のはず
  - 静的分割はしたくない

10

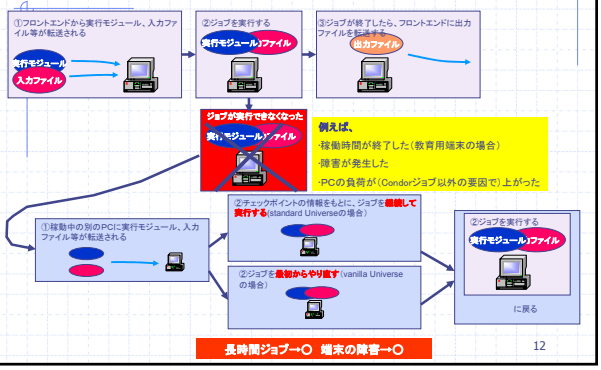
逐次ジョブの場合

Condorを採用

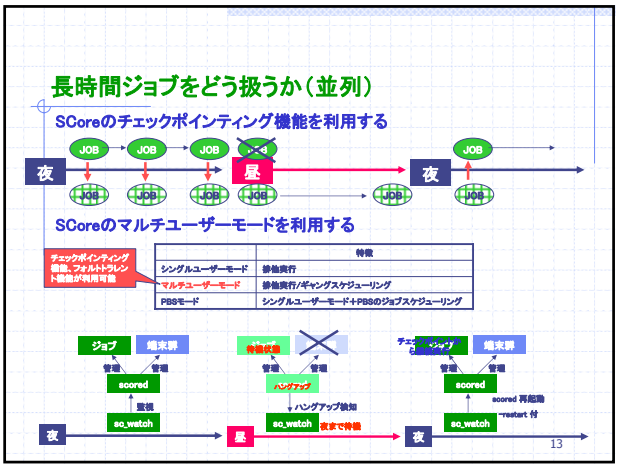


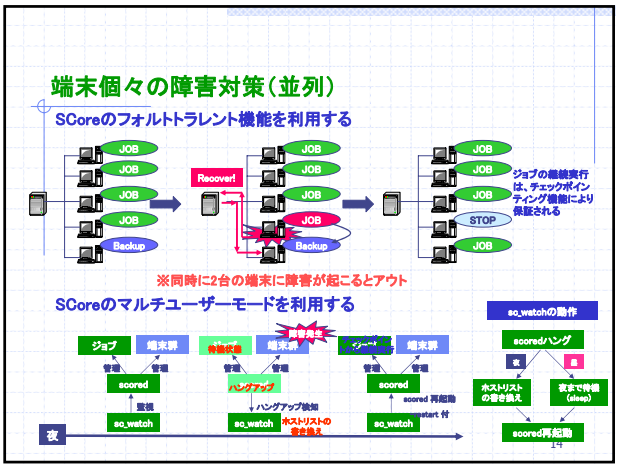
11

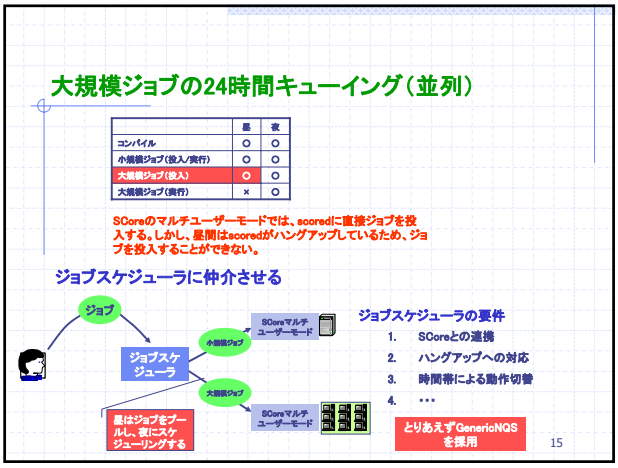
逐次ジョブの場合(2)

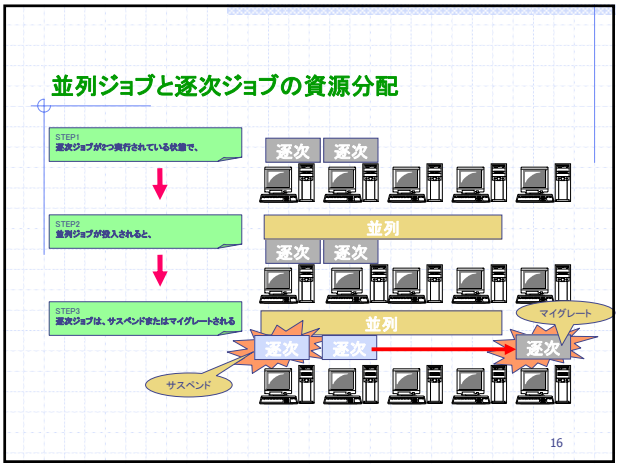


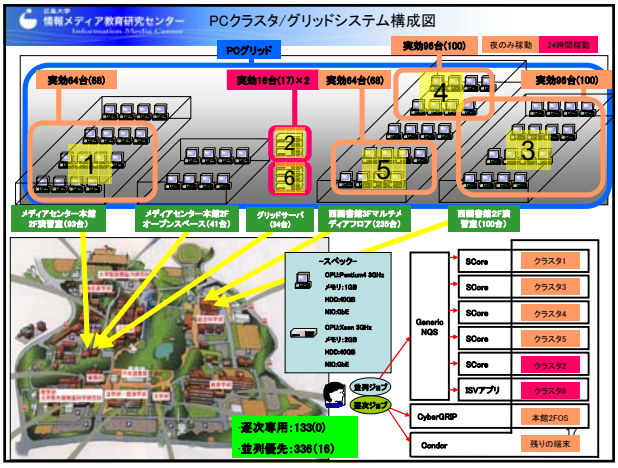
12

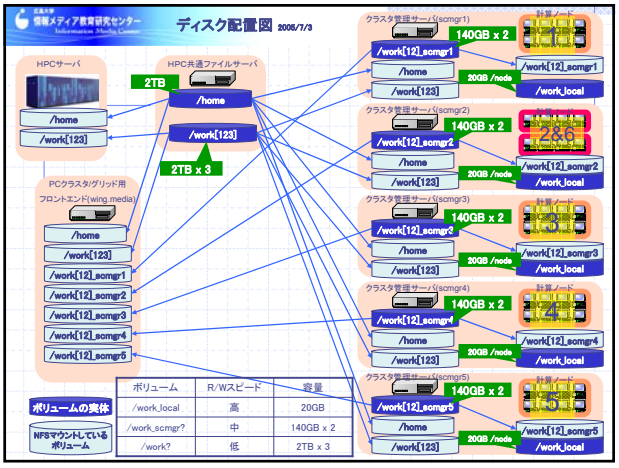

長時間ジョブをどう扱うか（並列）
SCoreのチェックポインティング機能を利用する
夜
昼
夜
JOB
SCoreのマルチユーザーモードを利用する
チェックポインティング機能、フォルトトレラント機能が利用可能
特徴
シングルユーザーモード
排他実行
マルチユーザーモード
排他実行/ギャングスケジューリング
PBSモード
シングルユーザーモード+PBSのジョブスケジューリング
ジョブ
端末群
scored
sc_watch
ハングアップ
ハングアップ検知
scored 再起動
-restart 付
13


端末個々の障害対策（並列）
SCoreのフォルトトレラント機能を利用する
JOB
Backup
Recover!
STOP
ジョブの継続実行は、チェックポインティング機能により保証される
※同時に2台の端末に障害が起こるとアウト
SCoreのマルチユーザーモードを利用する
sc_watchの動作
scoredハング
ホストリストの書き換え
夜まで待機(sleep)
scored再起動
夜
14


大規模ジョブの24時間キューイング（並列）
昼
夜
コンパイル
小規模ジョブ（投入/実行）
大規模ジョブ（投入）
大規模ジョブ（実行）
SCoreのマルチユーザーモードでは、scoredに直接ジョブを投入する。しかし、昼間はscoredがハングアップしているため、ジョブを投入することができない。
ジョブスケジューラに仲介させる
ジョブ
ジョブスケジューラ
小規模ジョブ
大規模ジョブ
SCoreマルチユーザーモード
昼はジョブをプールし、夜にスケジューリングする
ジョブスケジューラの要件
1. SCoreとの連携
2. ハングアップへの対応
3. 時間帯による動作切替
4. ・・・
とりあえずGenericNQSを採用
15


並列ジョブと逐次ジョブの資源分配
STEP1
逐次ジョブがつつ実行されている状態で、
STEP2
並列ジョブが投入されると、
STEP3
逐次ジョブは、サスペンドまたはマイグレートされる
逐次
並列
マイグレート
サスペンド
16


PCクラスタ/グリッドシステム構成図
PCグリッド
実効96台(100)
夜のみ稼動
24時間稼動
実効64台(68)
実効16台(17)×2
実効64台(68)
実効96台(100)
メディアセンター本館2F実習室(69台)
メディアセンター本館2F オープンスペース(41台)
グリッドサーバ (34台)
西新宿館3Fマルチメディアプロア(235台)
西新宿館2F実習室(100台)
-スペック-
CPU:Pentium4 3GHz
メモリ:1GB
HDD:40GB
NIC:GbE
CPU:Xeon 3GHz
メモリ:2GB
HDD:40GB
NIC:GbE
Generic NQS
SCore
ISVアプリ
CyberGRIP
Condor
本館2FOS
残りの端末
逐次専用:133(0)
並列優先:336(16)
17


ディスク配置図 2005/7/3
HPCサーバ
HPC共通ファイルサーバ
2TB
2TB x 3
/home
/work[123]
PCクラスタ/グリッド用フロントエンド(wing.media)
クラスタ管理サーバ(scmgr1)
クラスタ管理サーバ(scmgr2)
クラスタ管理サーバ(scmgr3)
クラスタ管理サーバ(scmgr4)
クラスタ管理サーバ(scmgr5)
140GB x 2
20GB /node
/work_local
ボリュームの実体
NFSマウントしているボリューム
ボリューム
R/Wスピード
容量
/work_local
高
20GB
/work_scmgr?
中
140GB x 2
/work?
低
2TB x 3

運用ルール

端末室

設置場所	平日	西図書館閉館日	土曜	日曜・祝祭日	長期休暇	
西図書3F MF	8:30～21:00	NOT available	8:30～17:00	NOT available	168台	NOT available
					67台	8:30～ 17:00
西図書2F演習	8:30～21:00		8:30～17:00	NOT available		
本館2F演習	8:30～22:00			NOT available		
本館2F OS	8:30～22:00					
GRIDサーバ	NOT available					

設置場所	台数	ID	平日	西図書館閉館日	土曜	日曜・祝祭日	長期休暇
西図書3F MF	235	4	～8:00 , 21:30～	Full time	～8:00 , 17:30～	Full time	Full time
		5					
		---					～8:00 , 17:30～
西図書2F演習	100	3	～8:00 , 21:30～		～8:00 , 17:30～	Full time	
本館2F演習	93	1	～8:00 , 22:30～			Full time	

本館2F OS	41	---	～8:00 , 22:30～				
GRIDサーバ	34	2	Full time				
		6					

✓ VIDシステムの運用スケジューラで端末毎に起動時間/起動OSを管理

19

まとめ

- 耐障害性を重視した設計
  - SCore/Condorの冗長機能を活用
- 運用時間の不規則性を隠蔽
  - ジョブキューイングは24時間可能
  - 実行は運用時間に依る
  - ジョブは休止時間を挟んで継続実行可能
- 高柔軟性
  - 並列（MPI）ジョブ・逐次ジョブに対応

20

今後の課題

- どう使ってもらうか
  - 現状は数人・・・・
    - 講習会/説明会はこれから
    - コンサルティングが重要
    - 新しい利用方法に期待
    - 期間限定で占有利用？
    - 逐次ユーザーの爆発的増加に期待
- ディスクI/O問題
  - クラスタファイルシステム?
    - PVFS, lustre, GFS, etc.
- 空調・電気代
  - 大学側の締付けが厳しい
  - せめて空調は入れて欲しいのだが・・・・
  - ジョブ投入で電源ON
- ソフトウェア的改善
  - NAREGIスーパースケジューラ
  - SunGridEngine, UNICORE, etc.
- 教育用端末の未来
  - ノートPCの必携化
- 故障の頻度
  - クラスタ1について（5/23～6/28）
    - 起動できなかった端末：9台
    - 最多で2台（ジョブの継続ためには1台まで）
  - ジョブ実行中の障害はいまのところ無し
  - 夏が心配

21