HPCI共用ストレージにおける データー貫性管理

Gfarmシンポジウム2015

原田 浩(理研　AICS)

HPCI共用ストレージ：概要

理研計算科学研究機構

- W1　storage
  - Gfarm
  - DDN　SFA10000(Total10PB)
  - メタデータサーバ
  - データサーバ16台
  - 10GbE　ネットワー
- 60 PB tape archive

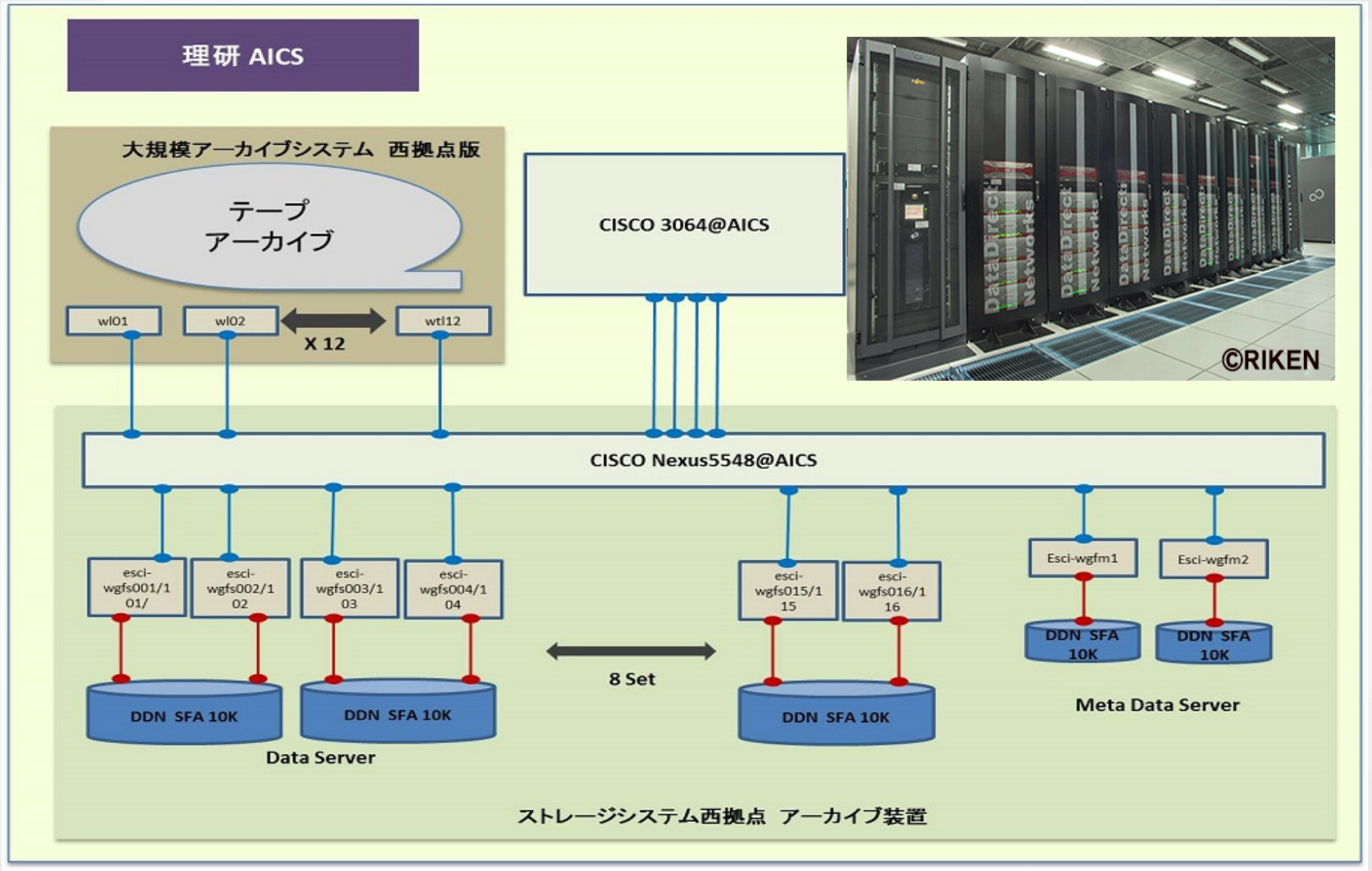
東京大学情報基盤センター

- E1 storage
  - Gfarm
  - DDN　SFA10000　9 セット（Total 8PB)
  - メタデータサーバ8台
  - データサーバ36台
  - ログイン4台
  - 10GbE　ネットワーク
- E2　storage
  - Gfarm
  - DDN SFA10000(Total 5.5PB)
  - メタデータサーバ2台
  - データサーバ8台
  - 10GbE　ネットワーク
- 20 PB tape archive（内部バックアップ利用）

- Tokyo Tech　storage
  - Gfarm
  - DDN　SFA10000(Total 4
  - データサーバ4台
  - 10GbE　ネットワーク

東京工業大学学術国際情報センター

HPCI共用ストレージ：AICS設置機材概要



発端

- 2015年3月末にユーザの通知から大規模データ破損障害が発覚
- 障害(被害)規模の特定が完了したのは、5月初旬
- 2015年のGW明けまで、約2週間サービス停止して東大・東工大・AICSの全ファイルのチェックサム照合を実行し障害規模を特定した
- 5課題、5ユーザ922ファイルを消失(復元不可能)
  - オリジナルファイルが障害発生サーバに書き込まれた
- データ破損152,392ファイル(共用ストレージのオリジナルから復元可能)
  - レプリカが障害発生サーバに書き込まれた
  - オリジナルは正常
- 障害発生時期は、2014年8月20日から10月29日
- 障害発生から約7か月後に発覚・その後障害規模把握に一月要した
- (障害発覚当時は)明確な原因は判明せず
- 再計算のための資源割当が求められる
  - 再計算資源量が判明したのが5月末
  - 「京」740万ノード時間を再計算に提供、、、、(小規模のスパコンセンター1年分の資源量に匹敵)

発端（続き）

典型的なサイレントなデータ破損障害

- ユーザがファイル書き込み時にエラー検出不可能なデータ破損障害
  - OSレベルでストレージコントローラへの書込み成功し、かつユーザプロセスがclose()処理に成功
  - ユーザプロセスにはエラーが返らない
  - Close()処理後にOSレベルがバッファフラッシュ→ストレージコントローラ→ディスク書込み時までに何らかの障害が発生
- 運用・管理者も障害把握は困難
- 運用・管理者側はユーザのデータ正当性を判断できない

- 障害発覚までに長期間
  - 被害（障害）規模の拡大
  - 障害状況保存ができない→原因究明がとても困難
  - 被害救済が困難
  - ユーザの信用失墜
    - 発表済み論文のデータが破壊されてしまったら…

対策（続き）

- データ破損が発生した場合でも早期発見・早期対策
  - ユーザの信用維持
  - 一定期間内に、データ消失ファイルを発見できれば、ユーザが保持するオリジナルファイルから復旧可能
    - ユーザが一定期間オリジナルファイルの保存必要
  - 再計算資源の節約
  - 状況保全→原因追求
- データ消失ファイルとデータ破損ファイルの２本立てチェック
  - データ破損ファイル＝メタデータにファイルが登録されているが、データファイルがデータ破損している。 正しいデータファイル(レプリカ)が一つ以上存在すれば、ファイル復旧可能
  - データ消失ファイル＝メタデータにファイルが登録されているが、全データファイルがデータ破損しており正常なデータファイルが無い。ファイル復旧不可能。
- Gfarm2.6の新機能が有効
  - レプリカ作成時のチェックサム照合

対策（続き）

- データ消失ファイル検出
  - Gfarmのreplicacheck
  - データ消失ファイル＝メタデータにファイルが登録されているが、有効なデータファイルが一つも無い
  - HPCI共用ストレージでは1日2回、消失ファイルリスト(i-node番号)とオーナーのUIDが運用関係者に自動送信
  - 消失ファイルのリストは手順に従って管理
  - 消失ファイルの見落とし事案もあったので、消失ファイルはメタデータから消去する方針で管理

システム　利用者支援

共用ストレージ

消失ファイル管理

updated by Seiichiro Naka (2015/11/20)

消失ファイルの初期化

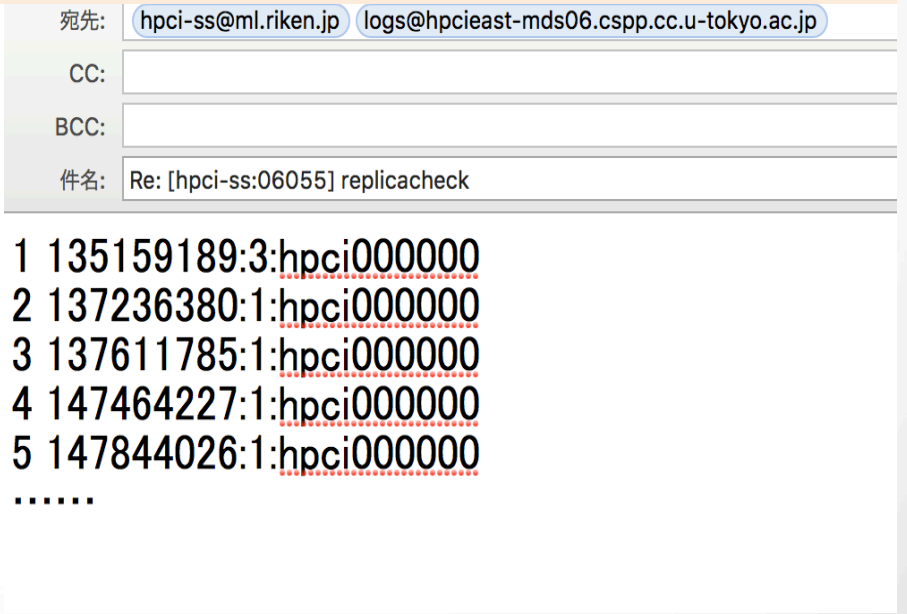
1.replicacheckでファイルの実体（replica）が一つも無いファイルの管理方法を定める。

2.上記replicacheckのエラーファイルは１日２回、蛯原さんから hpci-ssへ自動メイル送信される

3.repliucacheckの見逃しを避けるため、エラー出力されているファイルは下記の手順で一旦削除する

消失ファイル初期化手順

1. すでに課題実施者から削除承認されているファイルの削除を9月29日課題検討会議で合意したい
   →9/29に合意済み、96ファイル消去の準備進める
2. 2.1 で合意が得られたファイルは 10月15日までに消去し、課題実施者へメイル連絡
3. 削除主体は運用者(AICS)



宛先: hpci-ss@ml.riken.jp　logs@hpcieast-mds06.cspp.cc.u-tokyo.ac.jp

CC:

BCC:

件名: Re: [hpci-ss:06055] replicacheck

1 135159189:3:hpci000000
2 137236380:1:hpci000000
3 137611785:1:hpci000000
4 147464227:1:hpci000000
5 147844026:1:hpci000000
……

対策（続き）

データ破損ファイル検出

- GfarmのGfspooldigestコマンド
- メタデータ上のチェックサムとデータファイルのチェックサムを算出して比較
- メタデータ上のチェックサムは、書き込み時データから算出
- チェックサム比較の結果、差異があるファイルはデータ破損

- 書き込みが行われた全ファイルに対して、チェックサム比較を定期実行することで、データ破損障害を検知可能
- 実行状況を管理表で共有
- 実行結果は運用関係者にメイルで毎日送信
- AICSは、毎週、全ファイルサーバについて前1週間分の全書き込みファイルに対してデータ破損チェックを実行
- 東大も一週間分、書込みファイルをチェック
- ユーザには2週間のオリジナルデータ保存を要請

共用ストレージ

定期ファイル検査（gfspooldigest）管理表

updated by Jun Ebihara (昨日の 10:29:29)

スプール	5月	6月	7月	8月	9月	10月	11月	12月	1月	2月	3月
esci-epgfd101	21	22,29	6,13,20,27	3,9,17,24,31	9,14,21,28	5,12,19					
esci-epgfd102	21	22,29	6,13,20,27	3,9,17,24,31	9,14,21,28	5,12,19					
esci-epgfd103	21	22,29	6,13,20,27	3,9,17,24,31	9,14,21,28	5,12,19					
esci-epgfd104	21	22,29	6,13,20,27	3,9,17,24,31	9,14,21,28	5,12,19					
esci-epgfd105	21	22,29	6,13,20,27	3,9,17,24,31	9,14,21,28	5,12,19					
esci-epgfd106	21	22,29	6,13,20,27	3,9,17,24,31	9,14,21,28	5,12,19					
esci-epgfd107	21	22,29	6,13,20,27	3,9,17,24,31	9,14,21,28	5,12,19					

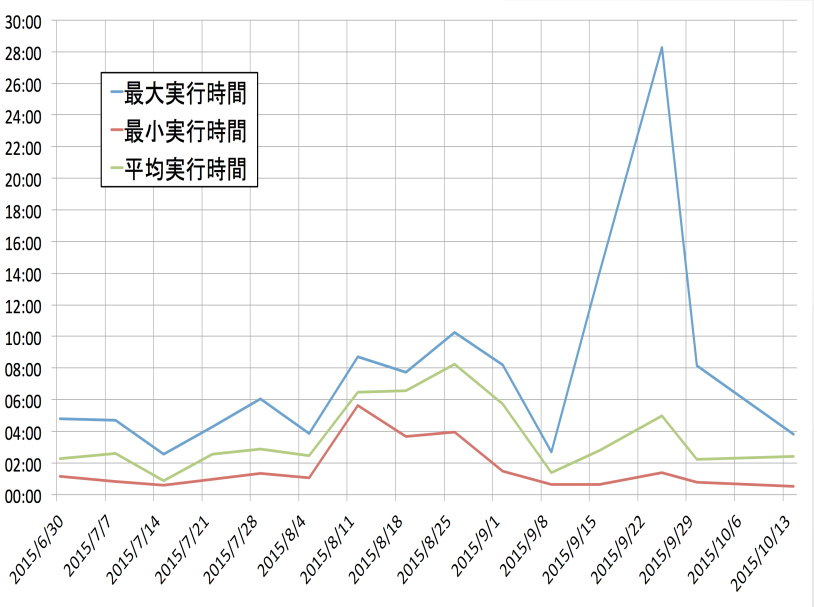
対策（続き）

- gfspooldigestコマンドは、スプール領域上の保存ファイルの破損を確認するコマンド

  $ gfspooldigest -M 1 -m 8 -G -h <host名> -r <スプール領域>

  - チェック対象(Readするファイル)は、-M, -mオプションで指定可能。
    上記の例は、昨日～8日前までに更新・作成されたファイルのみをチェック。
  - 破損ファイルは自動的にlost+foundに移動(Gオプション)。

- データ破損ファイル検出のAICSにおける定期実行
  - cronで定期的に実行中
  - 30スプールを4日間に分割（8+8+7+7）
  - AICSでは、Zabbix Pluginを用意して実施状況を監視
    - gfspooldigestのプロセスが上がる または 無くなるとinformationのアラートが上がり管理者にメールが送信される仕組み
  - 破損ファイルの検出有無についてはZabbixとチェック用スクリプトで監視
    - ログに出力されたアラートを拾って、管理者にメールが送信される仕組み
  - 各スプールサーバにおけるgfspooldigestの実行時間が右グラフ



効果

- 2015年10月27日にストレージコントローラのHW交換実施
- 2015年10月28日に10月27日更新ファイルに対するデータ破損検出
- 10月28日に該当ストレージコントローラのログチェック後に緊急停止
- 10月27日16時23分から10月28日17時28分までに該当サーバに書きこまれたファイルがデータ破損
- 3課題・3ユーザ9ファイルが消失
- 再計算資源は不要だった
- 早期発見・早期対策の結果、データ破損の原因も解明できた
- 東工大と同じ現象であることも確認できた
- IBのsrpdデーモンの二重起動によるMAX_SECTOR_SIZEの誤設定が原因
- Srpdデーモンの二重起動障害に対する対策・再発防止策も実行できた

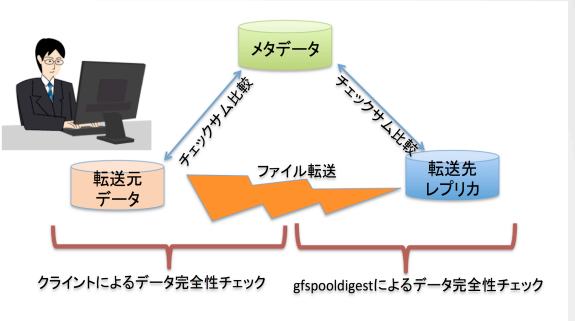
まとめと課題

- 2015年3月に、大量のデータ破損障害が発覚
- 典型的なサイレントなデータ破損障害であった
- 2015年4月末～5月初旬に全ファイルのデーター貫性検査（チェックサム照合）
- 2015年5月からデータ破損ファイル検知（データー貫性チェック）を定期実行
- 2015年10月末に、データ破損を検知
- 被害ファイル(消失ファイル)は、9ファイルで済んだ
- 早期検知・早期対策の結果
- データ破損の原因も解明できた
- トータル容量22.5PBの大規模分散ファイルシステムにおいて、データー貫性チェックをオンラインで定期実行して充分効果的であることが実証できた
- Gfarmの自動チェックサム算出機能はサイレントなデータ破損障害に対して有効
- 障害発生時の被害状況把握にも有効

まとめと課題

課題

- データ破損ファイル、消失ファイル管理の自動化・省力化
- データー貫性チェックの負荷軽減と実運用影響の最小化
  - 次期Gfarmで自動化される？
- 運用側障害対応手順が明確になっていなかった
  - 被害最小化
  - 被害規模特定方法
  - 原因追求



- サイレントなデータ障害は検知されていないだけで、実際には発生しているのでは？
  - 10PB級のファイルシステムにおいて定期的にデーター貫性チェックを実施している運用機関はマレだろう
- 複数機関による共同運用システムにおける危機管理
- HPCIでは技術的なリード＝サブWG、運用者（構成機関）＝各スパコンセンター、ユーザ対応窓口＝ヘルプデスク、全体的なコーディネート＝サービス連携委員会・作業部会、、、、迅速・的確な意思決定
- 障害規模が大きいほど困難
- 管理者と作業者の分離