リアルストレージワークロード特徴抽出のためのデータ収集蓄積技術のご紹介

大江　和一*

1　概要

大規模データの長期収集，蓄積，分析の１事例として，我々が実施したストレージワークロード収集蓄積技術を紹介する．当社ではストレージシステムの研究開発を進めており，その中でも我々はリアルストレージワークロードにフィットするキャッシュ・階層ストレージ制御方式の研究を進めている．この研究に用いる目的で，3,000 user 前後のアクセスが常時発生する samba ワークロード（最大 4.4TB）のデータ収集を約 1.5 年行った．

本稿では，データ収集を行ったストレージシステムとワークロードの概要，データ収集システムの概要，運用実績，収集したワークロードログより分かること，1PB キャプチャを行う場合の試算，ワークロード収集システムの現場での運用，に関して説明する．

実際に運用を行っているストレージシステムを対象に長期間ワークロード収集した事例は余りなく，我々が遭遇した課題とその解決方法が今後似たようなデータ収集に取り組む皆様の少しでも参考になればと考えている．

2　発表内容

発表内容は 3 章の後のスライド資料に示す．

3　質疑応答

3.1　当日の質疑応答

twitter(#spro2013)に書き込まれた質疑応答内容と我々の記録に残っている質疑応答内容を紹介する．なお，質問・回答 5,6 は我々の記録にのみ残っている質疑応答内容である．

質問 1　　ワークロードの偏りは RAID の組み方やファイルシステムで変わるのでは？

回答 1　　ファイルシステムはそうかも知れない。RAID は分からない。

質問 2　　NAS 環境だとチャンクの使い回し云々があって consistent な計測はブロックレベルでは難しそうだが何とかなるのか？

*株式会社富士通研究所 ICT システム研究所

回答 2　　今回の計測では全部 Windows だがそこそこ一致していた

質問 3　　今回の計測データを製品の改良に活かすことを考えるとどうなのか

回答 3　　まだ取らなければいけない情報があるがなかなか難しい

質問 4　　この話を進めていくと上で何が起きているか分からないが DB 側で何とか最適化にしてあげる方向にいくか

回答 4　　まだ製品に入ってないが，自分の興味としてはそういうことをしたい

回答 4　　週ごとの差異なども見えるのでキャパシティプランニングなどに使えればよいと思っている

質問 5　　1 サンプルの分析で大丈夫でしょうか（samba のバグなどで偏りが出ている、とか）

回答 5　　MSR Cambridge を分析しても同様な結果が出ており，今回紹介したようなワークロードが少なからず存在する，と判断している．

質問 6　　ファイルシステムレベルやもっと上位のログは必要ない？

回答 6　　上位アプリ，ファイルシステムレベル，ブロックレベルで串刺しで取るのが理想ではある．実験室レベルでは出来るが運用環境ではなかなか難しいので，今回の収集でもブロックレベルのみです．逆にブロックレベルの分析でうまく一般化出来れば何処へも持っていくことができます．

3.2　twitter への書き込みを踏まえた補足説明

講演中に twitter に書き込んで頂いた内容を踏まえた補足説明を行う．

3.2.1　LBA レベルの分析で何が分かるのか

アプリケーションごとに大まかな特徴を明確にすることは出来，特徴に基づいたストレージ制御方法を検討することが出来ると考えている．今回ご紹介したファイルサーバを例に取ると，負荷の偏りが発生する LBA の幅と全 IO に占める割合の傾向が把握できれば，その負荷を対象にする階層制御方式等を提案可能である．

もちろん，同一アプリケーションにおいても OS やファイルシステムが異なると LBA レベルでの偏りなども異なってしまう可能性があり，正確なアプリケーションの分析と言えないが，ストレージ制御の観点からは十分有効な分析であると考えている．

FUJITSU
shaping tomorrow with you

リアルストレージワークロード特徴抽出のためのデータ収集蓄積技術のご紹介

2013年8月25日
（株）富士通研究所　大江　和一

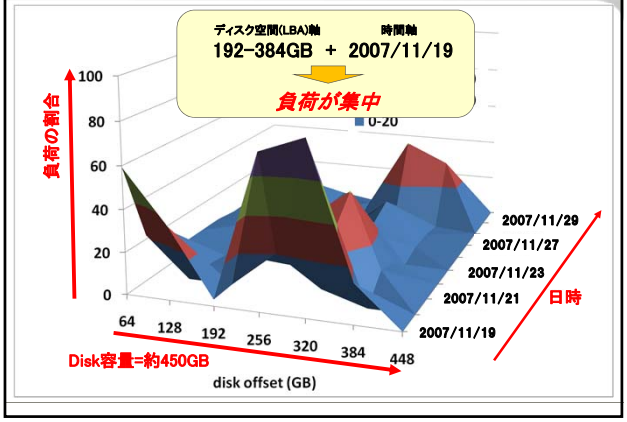
：講演後に書き込んだ注釈

2013年夏のプログラミング・シンポジウム
ビューティフルデータ　講演資料

Agenda

1. ワークロード収集の概要
2. 収集したワークロードログより分かること
3. 1PBキャプチャを行う場合の試算
4. ワークロード収集システムの現場での運用について
5. まとめ

例えば、こういうことを知りたい



リアルストレージワークロード収集の意義

目的：　時間的・空間的局所性を分析⇒一般化

ストレージ制御に関する新たなシステム提案が出来る、かも

- **従来のstatic tieringでは性能向上しないワークロードへの方式提案**
- **新たなcacheアルゴリズムの提案**
- **省電力の観点からのSSD/HDD tiering方式提案、など**

- ストレージへのデータ量：　爆発的に増大
  - ⇒　ストレージアクセス履歴量も膨大に …
- どうやって解決するのかが共通の課題
  - ストレージアクセス履歴保存に必要なディスク容量
  - ストレージアクセス履歴の増加に対応可能なシステム性能

⇒　我々の経験が少しでも参考になれば…

関連研究

- 主なストレージワークロード分析論文
  - 仮想環境でのストレージワークロード分析
    - Storage Workload Characterization and Consolidation in Virtualized Environments; Ajay Gulati, Chethan Kumar, Irfan Ahmad (VPACT 2009)
    - An Analysis of Disk Performance in VMWare ESX Server Virtual Machines; Ifan Ahmad, Jennifer M. Anderson, Anne M. Holler, Rajit Kambo, Vikram Makhija (6th Annual Workshop on Workload Characterizetion, 2003)
  - Windows Serverのワークロード分析
    - Caracterization of Storage Workload Traces from Production Windows Servers; Swaroop Kavalanekar, Bruce Worthington, Qi Zhang, Vishal Sharda (ISWC2008)
  - UNIXディスクアクセスパターン分析
    - UNIX disk access patterns; Chris Ruemmler and John Wilkers (USENIX Winter 1993 Technical Conference)

LBAごとの偏りまで踏み込んだ長期間のワークロード分析論文を見つけられなかった

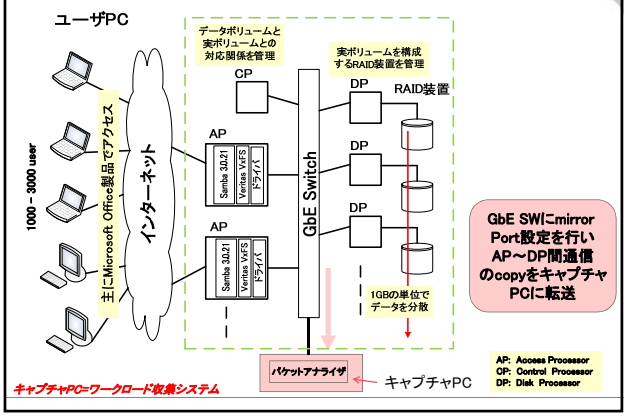
- 一般公開されているストレージワークロード
  - MSR Cambridge トレースデータ(http://iotta.snia.org/traces/388)
    - Write off-loading: Practical Power Management for Enterprise Storage (USENIX FAST'08)で紹介
    - ⇒　プロジェクト開始時点では未公開

自分たちで必要なデータを採取することに

ワークロード収集の概要

- 収集システム開発＆ワークロード収集を行った期間
  - 2007.10～2010.12
    - 継続してログ収集が出来たのは、最後の1.5年間
- 収集を行ったストレージシステム
  - スケールアウト型分散ストレージシステム
    - 研究所開発システムがベース　⇒　仕様入手＆検証が容易
    - 実運用システム　⇒　まとまったユーザ数のワークロードが収集可能
  - ワークロード収集を行った主なストレージシステム
    - Samba+backup（35TB）：　数ヶ月単位のワークロードを収集
    - Samba（4.4TB）：　連続して約1.5年分のワークロードを収集
- 収集システムの課題と解決方法
  - 障害原因とならないこと
    - ⇒　GbE SW mirroring機能を用いた主系とは別PC上のパケットキャプチャ
  - ごく普通の1U IA serverで運用＆頻繁にデータ回収出来ない
    - ⇒　生トレースの保存は諦め、1GB・1分単位の統計情報を圧縮・保存

ワークロード収集を行った分散ストレージシステム

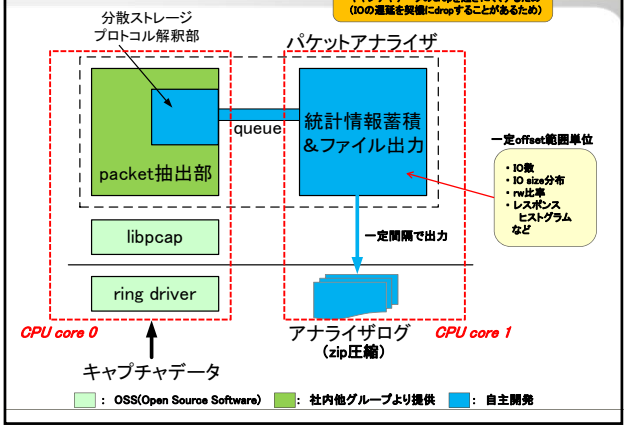


キャプチャPCの課題と解決方法

- ストレージシステム側との関係：
  - いかなるケースでもストレージシステム側の障害原因とならないこと
    - ⇒ ***GbE SW mirroring機能を利用することで別PCで分析***
- キャプチャPCの要件：
  - 設置面積： 1U程度に収まること
  - マシンスペック： 比較的安価なサーバを用いてもログ収集できること
  - データ回収： 数ヶ月程度はスタンドアロンで動き続けること
    - ⇒ ***生トレースの保存は諦め、1GB・1分単位で統計情報を抽出し、圧縮・保存***
- パケットアナライザ*開発方針：
  - 性能： 収集ワークロードに耐えれる程度の性能（むやみに高性能を狙わない）
  - 実装： 自主開発コードは必要最小限に＆長期間安定動作
    - ⇒ ***収集ワークロード負荷は分からないので、考えられる対策を行った上であとは実際に運用してみて対策を考えることに…***

*: キャプチャPC上で動かすソフトウエア。このソフトウエアが収集したワークロードの統計処理・ファイルへの統計情報出力を行う

パケットアナライザの実装



パケットアナライザの動作条件

複数のPCプラットフォームにおける動作確認結果

- 2.33 GHz x 2、L2 cache size = 4MB、DDR2(533 MHz) ⇒ drop率 = 0.02%以下
- 2.40 GHz x 2、L2 cache size = *512KB*、*DDR(533 MHz)*
- 2.93 GHz x *1*、L2 cache size = *256KB*、DDR2(533 MHz) ⇒ **drop率 = 40 – 50%**
  - L2 関連の命令/データアクセス回数が大きく上昇 ⇒ CPU/メモリ帯域ネック

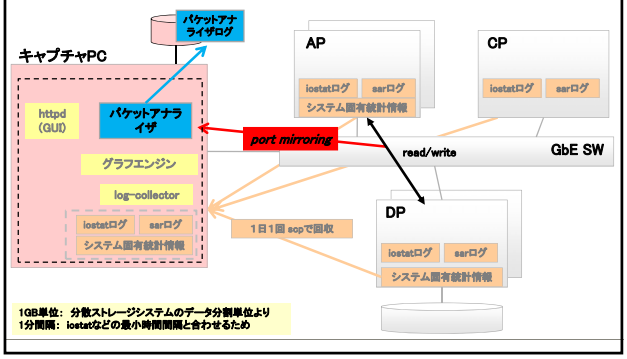
メモリアクセスが大量に発生するアプリであるため、L2 cacheやFSB帯域がある程度大きくないとキャプチャ出来ない

	要求spec.
CPU クロック	2.33 GHz以上 (Xeon 5148相当以上)
CPU数(コア数)	2 以上
L2 cache size	4 MB以上
FSB	1333 MHz以上
Mem	DDR2(533MHz) 32GB以上
disk	300 GB程度
	評価結果
drop率	7000 iops ⇒ 0.02%以下
データ量(圧縮後)	200 KB/回 ⇒ 約8GB/月

事前評価：
2TBキャプチャ時=20KB前後
（1GB単位＆1分間隔を設定）
⇒ 10倍のデータ量で見積り

構築したキャプチャPC

キャプチャデータを1GB単位で統計処理 ⇒ 1分間隔でファイルへ



主な運用実績*

長期（最大17ヶ月）のワークロードログ回収に成功

	system 1 (2007.10 start)	system2 (2009.8 start)
アプリケーション	Samba+backup(35TB)（Sambaは約2TB）	Samba(4.4TB)
統計データの最大サイズ(KB/回)	150	50
統計データの最大サイズ(GB/月)	**6.2**	**2.1**
平均drop率	1%未満	1%未満
平均user数(Samba)	**1000**	**3000**
最大iops	1500	2000
キャプチャPCの運用月数	30	17
回収出来たキャプチャデータ(ヶ月)	(1-2ヶ月) x 3	**17**

*: 上記以外にもう1システムで運用実績(30ヶ月)があるが、データが残っていないため未掲載
**: 最初の1年間はSambaのみ(2TB)で運用。その後、35TBに容量を拡張。

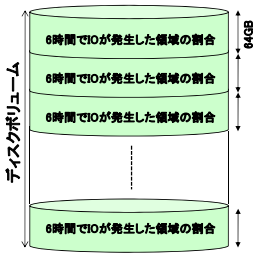
- ワークロード: 1000user/35TB、3000user/4.4TB
- ログデータ量： 最大6.2 GB/月
- 回収出来た期間： 連続して17ヶ月が最大

ワークロードログより分かること(1)

「ワークロードログより分かること(2)、(3)」の解釈方法

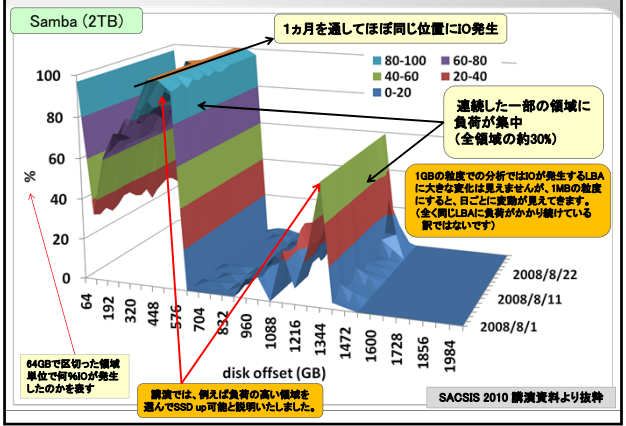
■ 分析の観点:
- IOが発生する領域の局所性
- IOが発生する領域の時間変動

■ 分析方法:
- ディスクボリュームを64GB単位に分割
- 64GBの領域を1GB単位のサブ領域に分割
- 64GBの領域中1IOでもアクセスが発生したサブ領域の数を集計
- 発生するIOを6時間ごとに蓄積(*)
(負荷が最も高い12:00-18:00を選択)



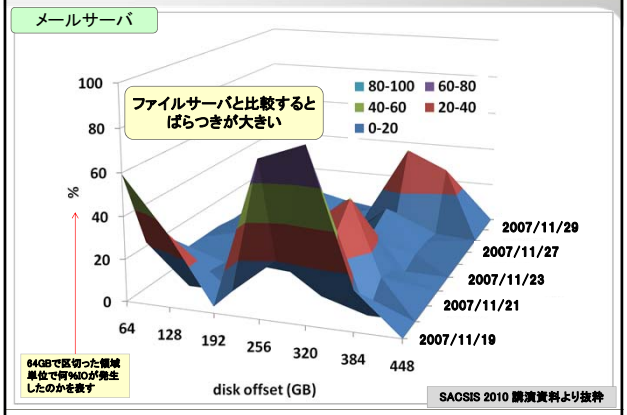
SACSIS 2010 講演資料より抜粋

ワークロードログより分かること(2)



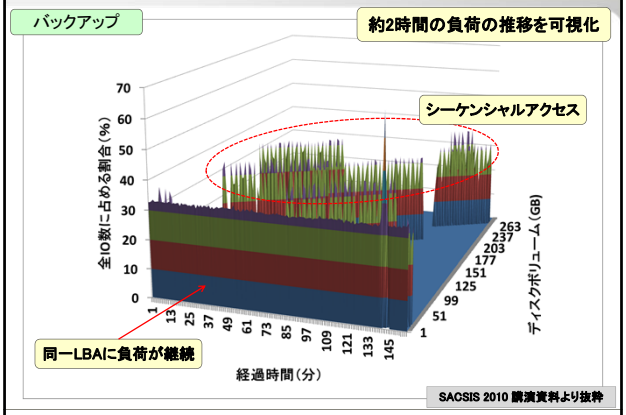
SACSIS 2010 講演資料より抜粋

ワークロードログより分かること(3)



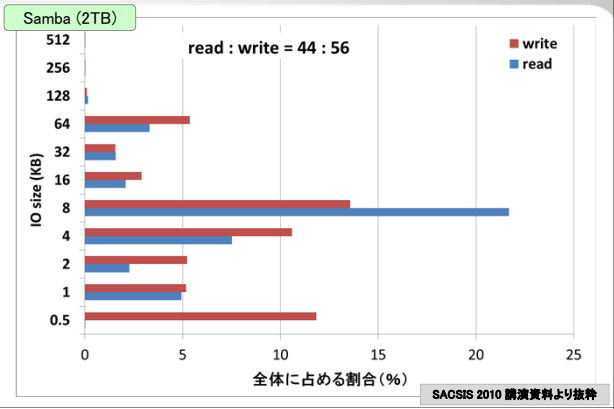
SACSIS 2010 講演資料より抜粋

ワークロードログより分かること(4)



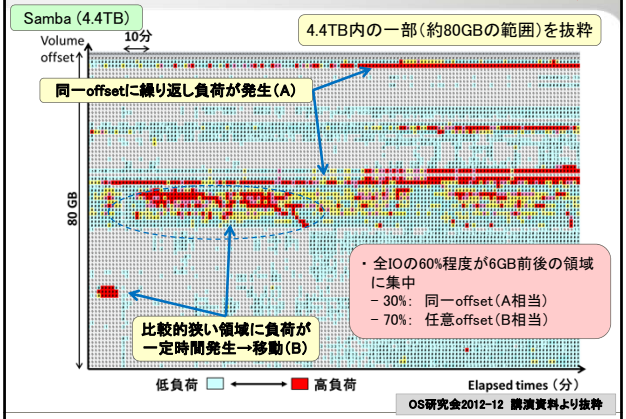
SACSIS 2010 講演資料より抜粋

ワークロードログより分かること(5)



SACSIS 2010 講演資料より抜粋

ワークロードログより分かること(6)

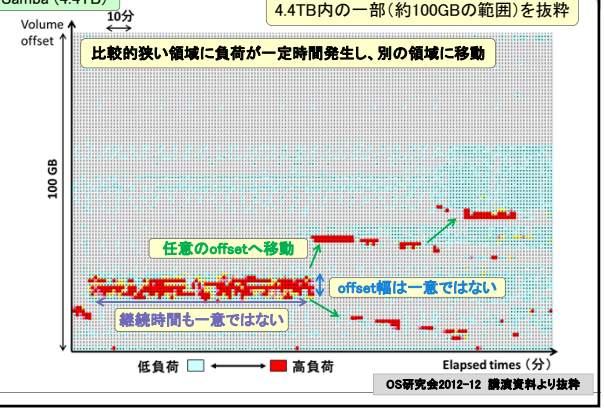


OS研究会2012-12 講演資料より抜粋

ワークロードログより分かること(7)

Samba (4.4TB)

4.4TB内の一部(約100GBの範囲)を抜粋



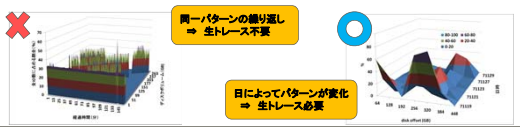
OS研究会2012-12 講演資料より抜粋

ワークロードログより分かること(まとめ)

- 分かること(最大「1GB・1分間」の粒度で)
  - IOの時間的・空間的局所性の把握
  - IO size分布やrw比の内訳

新たなストレージシステム検討の材料に …

- 今後の課題
  - (部分的な)生トレースデータの採取
    - 必要な理由: キャッシュアルゴリズムの検討や提案方式の評価に必要
    - 課題: 保存データ量を大幅に増やさず必要な生トレースデータを集める方法
    - 解決方針: 生トレースが必要なワークロードを選んで採取する仕組みの確立



1PBキャプチャを行う場合の試算(1)

Disk 容量・帯域

key	value	comment
1TB当たりの統計データ量(KB) 1回分	12	圧縮後の容量。圧縮で約1/10に。
1PB当たりの統計データ量(MB) 1回分	12	
スループット(MB/sec)	4.2	「生データwrite->read + 圧縮データwrite」を1分以内で
1PB当たりの統計データ量(GB) 1日分	16.875	1分間隔でデータ保存
1PB当たりの統計データ量(GB) 1ヶ月	506.25	
HDD 1本当たりの性能(MB/sec)	0.8	io size=4KB, ランダムアクセスを仮定(200 iops)
RAID0+1構成でのHDDの必要数	10.5	最悪ケースを想定してもHDDを10本準備出来れば必要性能を賄えることに

- **Disk容量:**
  - **約500GB/月 ⇒ 許容範囲**
- **Disk帯域:**
  - **HDD x 10 ⇒ 許容範囲**
  - **SSDとHDDを組み合わせた階層ストレージでもよいかも**
    例、その日の統計データはSSDへ ⇒ 日が変わったらHDDへ

現行ストレージ製品を組み合わせることでなんとかなりそう

1PBキャプチャを行う場合の試算(2)

キャプチャ性能

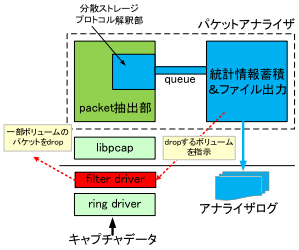
key	value	comment
1TB当たりの最大iops	454.5	運用実測値
1PB当たりの最大iops (A)	465454.5	
キャプチャシステムで保障可能最大iops (B)	7000	
A/B (倍)	66.5	
平均io size=8KBとした場合の帯域(MB/sec) (C)	54.7	7000 iopsでもGbE帯域の半分程度
1PB時の必要帯域(GB/sec) (D)	3.6	40GbEで収容可能
DDR2、メモリバス=533MHz (GB/sec) (E)	8.5	運用システムのメモリ性能
DDR3、メモリバス=1333MHz (GB/sec) (F)	21.3	最新のメモリ性能
F/E (倍)	2.5	メモリ性能は約2.5倍に向上

現行キャプチャPC: PC内処理(パケットアナライザ等)がネック?
キャプチャ性能を67倍に: ネットワークは1GbE⇒40GbEで解決
課題: 3.6GB/secの帯域をPC内で処理できるのか?
- メモリ性能向上率=2.5倍 ⇒ PCを最新に更新しただけでは対応できないかも

⇒ ①PC内処理の最適化、 ②複数PCに負荷分散

1PBキャプチャを行う場合の試算(3)

パケットアナライザ側の工夫



- 一部ボリュームのパケットを意図的にdrop
  - 毎日同じ負荷の繰り返し
    ⇒自律的にdropを判断
  - システム管理者からの指示
- 余った帯域で生トレースの部分採取
  - 帯域を増やす機能であるが、必要なので

GbE Switchのfiltering機能と併用する方法も …

ワークロード収集システムの現場での運用(1)

運用前準備

- 運用方法
  - 分散ストレージシステム: 運用管理配下 ⇒ ***キャプチャPCも運用管理対象***
    - キャプチャPC運用開始後: システム運用者以外は操作不可になる場合がある
- システム運用者からの要望とその対処
  - 運用中の分散ストレージシステムの性能・信頼性低下が起きないこと:
    - 机上説明
      - キャプチャPCが分散ストレージ側に影響を与えないことを説明
    - 実証実験
      - GbE mirror port設定あり/なしでネットワーク性能比較 ⇒ ***性能低下せず***
        (運用機材と同じ機材(Switch設定などもそろえて)で検証)
      - キャプチャPC側でデータ滞留を意図的に起こし、その影響調査
        - キャプチャ中にキャプチャPC panic ⇒ ***分散ストレージシステムの動作に影響せず***

実運用システムへ適用OK

ワークロード収集システムの現場での運用（2）

運用

- インストール
  - 運用者のインストール作業に立ち会ってのインストール作業サポートも必要
    - OSのカスタマイズ、GbE Switchの設定、パケットアナライザの初期化、など
- 統計データ（アナライザログ）の回収
  - 運用者の作業に同行して回収
    - 管理用ネットワーク帯域が細い、あるいは、潤沢に使えない場合がある
    - 遠地に設置してある場合には、頻繁な回収は困難
    - 運用者側の業務スケジュールなどにも依存するので事前の計画が重要に
- 日々の運用
  - 運用手順を簡略化できないと、設定誤りのため正しく動作しないケースも
    - 例、ボリュームを追加 ⇒ 運用者がパケットアナライザの設定変更が必要に
      ⇒ 出来る限り自動化しておく必要性を痛感
  - バージョンアップ
    - 運用者のバージョンアップ作業に立ち会う必要性がある場合も
  - 障害状況
    - 3年間ほぼ安定して動作させることができた

まとめ

- ストレージワークロード収集の概要
  - 4.4TBと35TBの分散ストレージシステムが対象
  - 最大17ヶ月間のワークロードログ収集に成功(4.4TB)
    - 「1GB・1分間」の粒度 ⇒ 最大6.2 GB/月のデータ量
- 収集したワークロードログより分かること
  - 時間的・空間的局所性の特徴抽出が可能
    ⇒ 新たなストレージシステム検討可能
- 1PBキャプチャを行う場合の試算
  - 4.4TB分散ストレージシステムのログ量が単純にスケールする前提
  - PCのメモリ・バス帯域ネックになる可能性
    ⇒ ①キャプチャソフトウエアの最適化、②複数PCに分散してキャプチャ
- ワークロード収集システムの現場運用
  - 本番（分散ストレージシステム）系の障害原因とならないこと
  - 「エラーケースまで含めた検証済み手順の確立＋現場で容易な運用」が重要

FUJITSU

shaping tomorrow with you