HPCI共用ストレージにおける階層型ストレージの導入構想

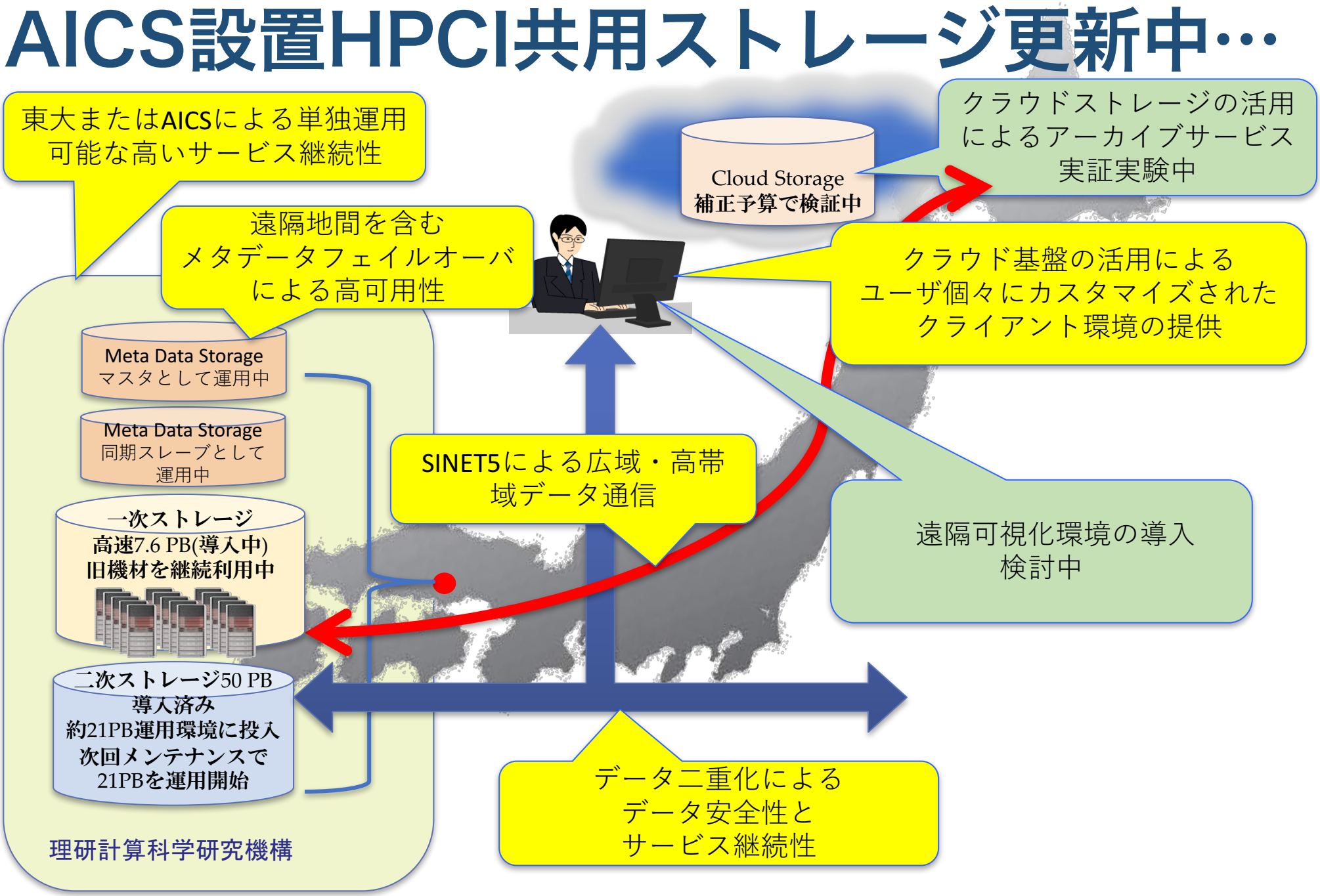
Gfarm日本征服計画

- H.Harada
- H.Kaneyama
- A.Kondo
- C.Shibano

2017/12/22

RIKEN AICS

AICS設置HPCI共用ストレージ更新中…



AICS共用ストレージ更新中…

- 容量：旧機材約10PBから新機材約50PBへ拡張
  - 10TBDisk、RAID6(8D2P) x 58セットのストレージを11セット → 10 x 8 x 58 x 11 = 51,040TB
- 東大、AICSのデータ二重化運用に向けてデータ移行中
  - 既存データの二重化に3ヶ月以上必要
- 階層型ストレージの導入
  - 1次書き込み先として高速ストレージを2月末導入予定
  - 書き込み、読み出しそれぞれ 30GB/sec以上
  - スプールサーバもIB(4xEDR)接続３台
  - 2月末までは1次ストレージとして旧機材を活用
  - ２次ストレージとして新機材を活用
- メタデータサーバの性能向上
  - 物理メモリは768GiBに強化
- ログイン環境の強化
  - UNCAI によるクラウド基盤の提供
  - 遠隔可視化サービスの提供(検討中)
  - GPGPU搭載計算機の提供

Gfarmの階層化支援機能…

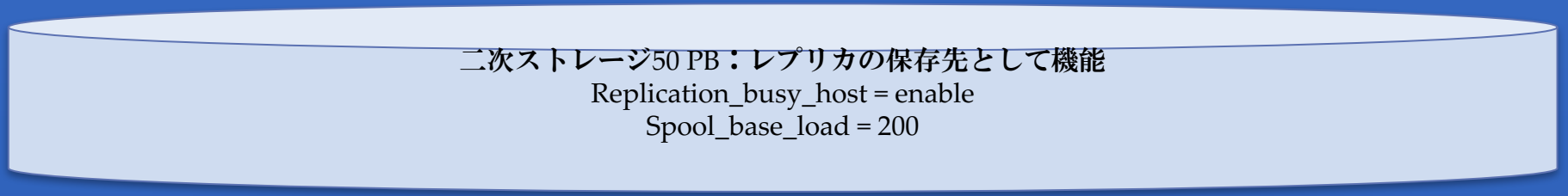
一次ストレージの制御

- 余剰レプリカ自動削除機能を用いて、キャッシュのような振る舞いを実現する
  - 余剰レプリカ削除機能：指定されたレプリカ数以上に存在するレプリカを自動削除する機能
- replica_check_remove_grace_used_space_ratio
  - 猶予容量使用率(%)を指定する
  - 0~100を指定可能
  - 各スプール領域に使用率が猶予容量使用率となるまで削除を猶予する
  - 例１：100TBのスプールサーバに対して80を指定すると、容量80TBを超えるまでは余剰レプリカは削除されない
  - 例２：100TBのスプールサーバに対して0を指定すると、全ての余剰レプリカが削除される
- replica_check_remove_grace_time
  - 余剰レプリカが削除されるまでの猶予期間を秒で指定する
  - アクセス時間から猶予期間(秒)経過した余剰レプリカが削除される
  - 例１：3600 x 24 x 365 を指定すると、最終アクセスから一年経過した余剰レプリカが削除される
  - 例２：0を指定すると、全ての余剰レプリカが削除される
- 削除対象となる余剰レプリカ
  - 使用率が猶予容量使用率以上のスプールに存在し、かつ、最終アクセスから猶予期間以上経過した余剰レプリカ
- スプールは常に、猶予期間内にアクセスされたレプリカで、かつ容量を猶予使用率に抑えることが可能

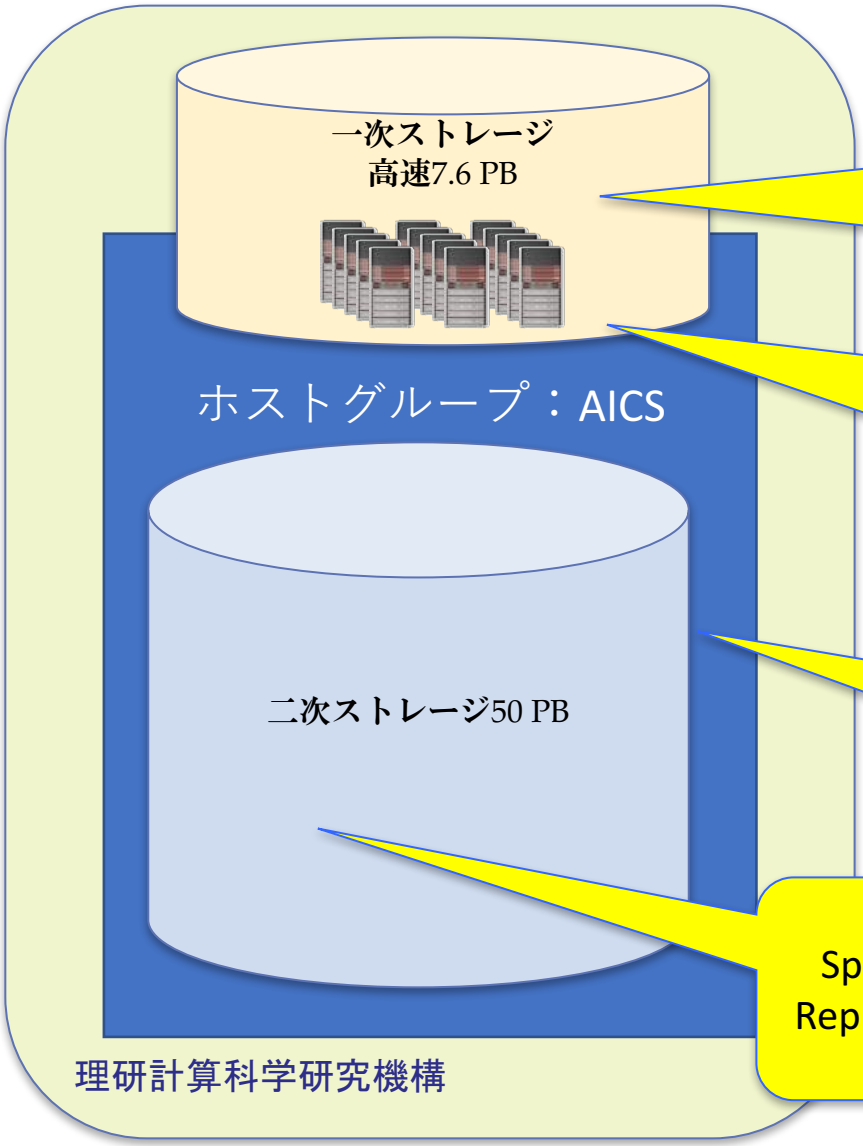
Gfarmの階層化支援機能…

二次ストレージへの一次書き込み抑制

- Gfarmの一次書き込み先選択アルゴリズムではbusy_hostは一次書き込み、レプリカ作成の対象外
  - Busy_host：schedule_busy_load_thresh< (load + spool_base_load) / cpu_core数
  - Schedule_busy_load_threshのデフォルト値は0.5
  - ただし、 replication_busy_host変数をenableに設定すると、レプリカ作成の対象となる
- spool_base_load
  - Gfsd(スプールサーバ)のロードに加算される値
  - 例：spool_base_load=100を設定されたスプールサーバはGfarmによって常にload 100以上と認識される
  - Gfarm書き込み先選択アルゴリズムによって、一次書き込み先に選択されなくなる
  - 二次ストレージとして運用するスプールサーバは高いspool_base_load値を設定する
- Replication_busy_host=enable、 spool_base_loadを十分大きい値に設定することで、レプリカ作成は可能だが、一次書き込みを抑制することが可能



AICS共用ストレージにおける階層化…



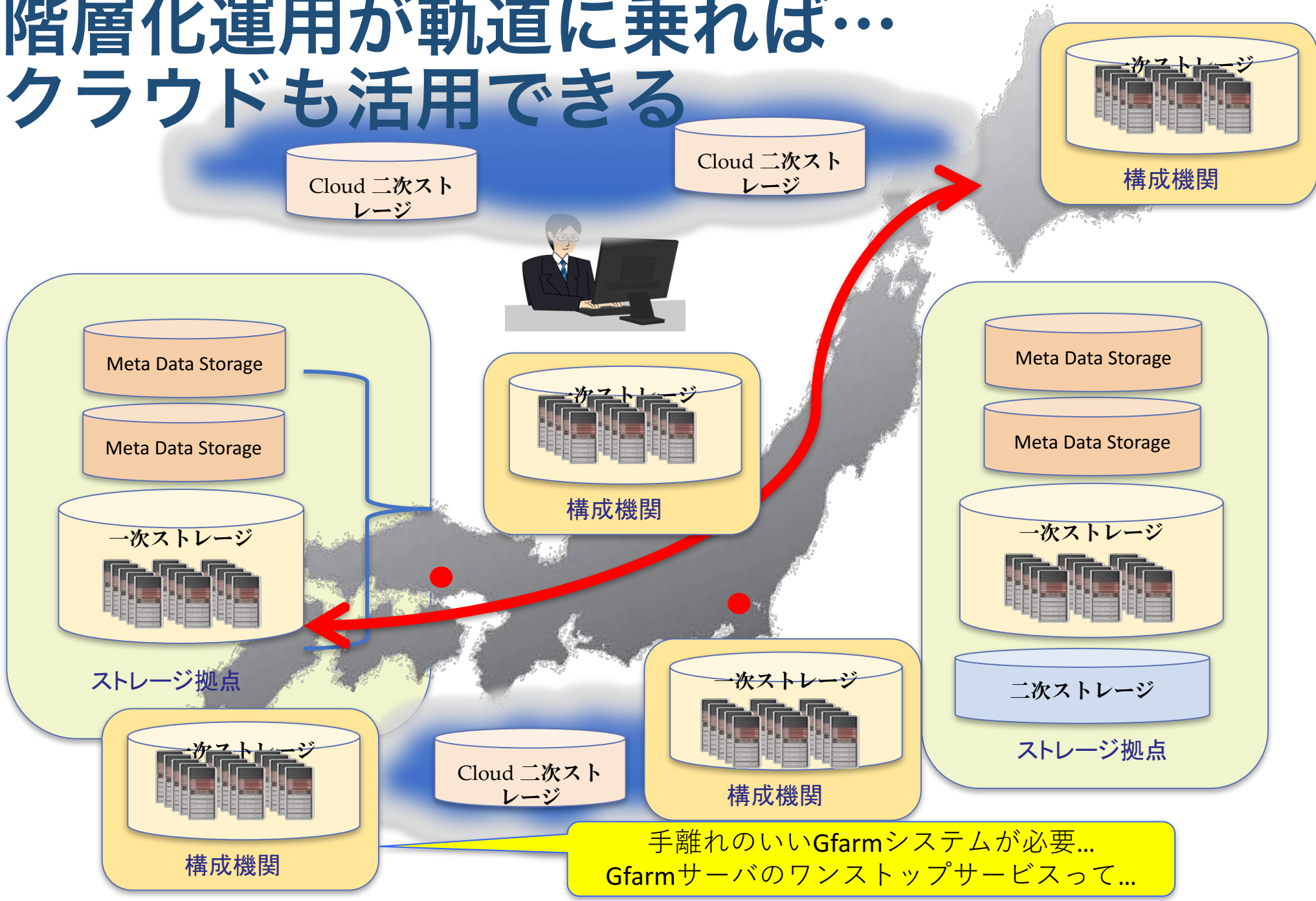
一次書き込み先として十分なアクセス帯域
書き込み、読み出し 30GB/S 以上
スプールサーバ3台による冗長構成
4xQDR IBによる高いネットワーク帯域確保

適切な
replica_check_remove_grace_used_space_ratio ＆
replica_check_remove_grace_time
設定による、一次書き込み先としての運用

ホストグループによるデータ二重化運用
AICS、東大それぞれにホストグループを設定して
レプリカは常に東大、AICSに二重化保存

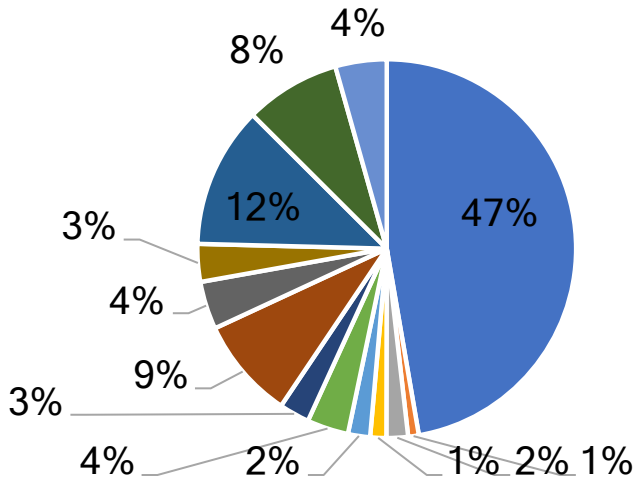
Spool_base_load設定による書き込み先選択の回避
Replication_busy_host設定によるレプリカ作成先選択

階層化運用が軌道に乗れば…
クラウドも活用できる



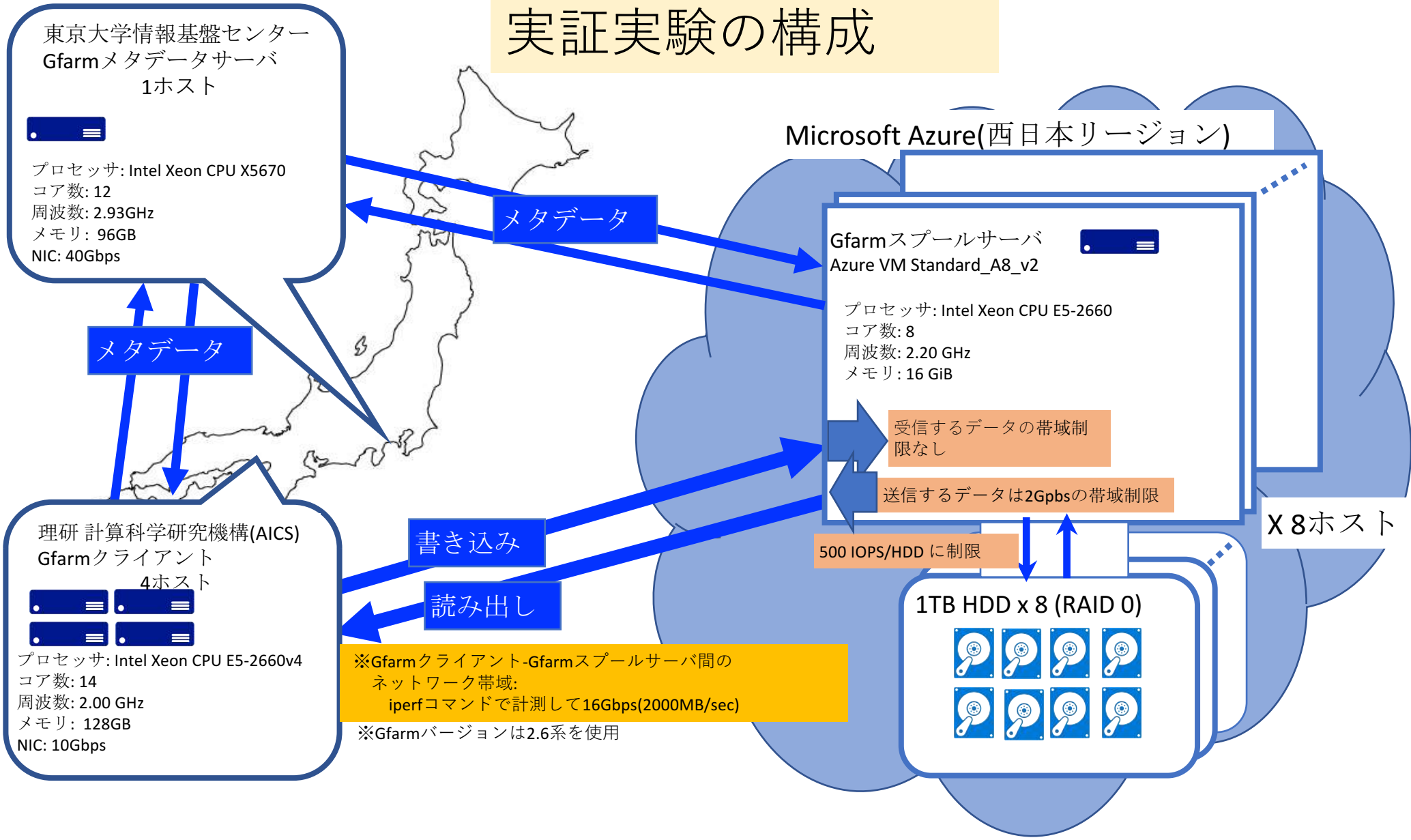
HPCI共用ストレージの利用状況

- 年々割当希望量が増大し2015年度からは割り当て希望量が提供資源量を大きく上回る。
- 予算や設備の制約によりフレキシブルに資源提供量を調節することも困難。
- 保存されているデータ
  - 研究成果の証左としてアーカイブされているデータ。
  - 2016年7月時点で1年以上アクセスされていないデータが47%もあります。

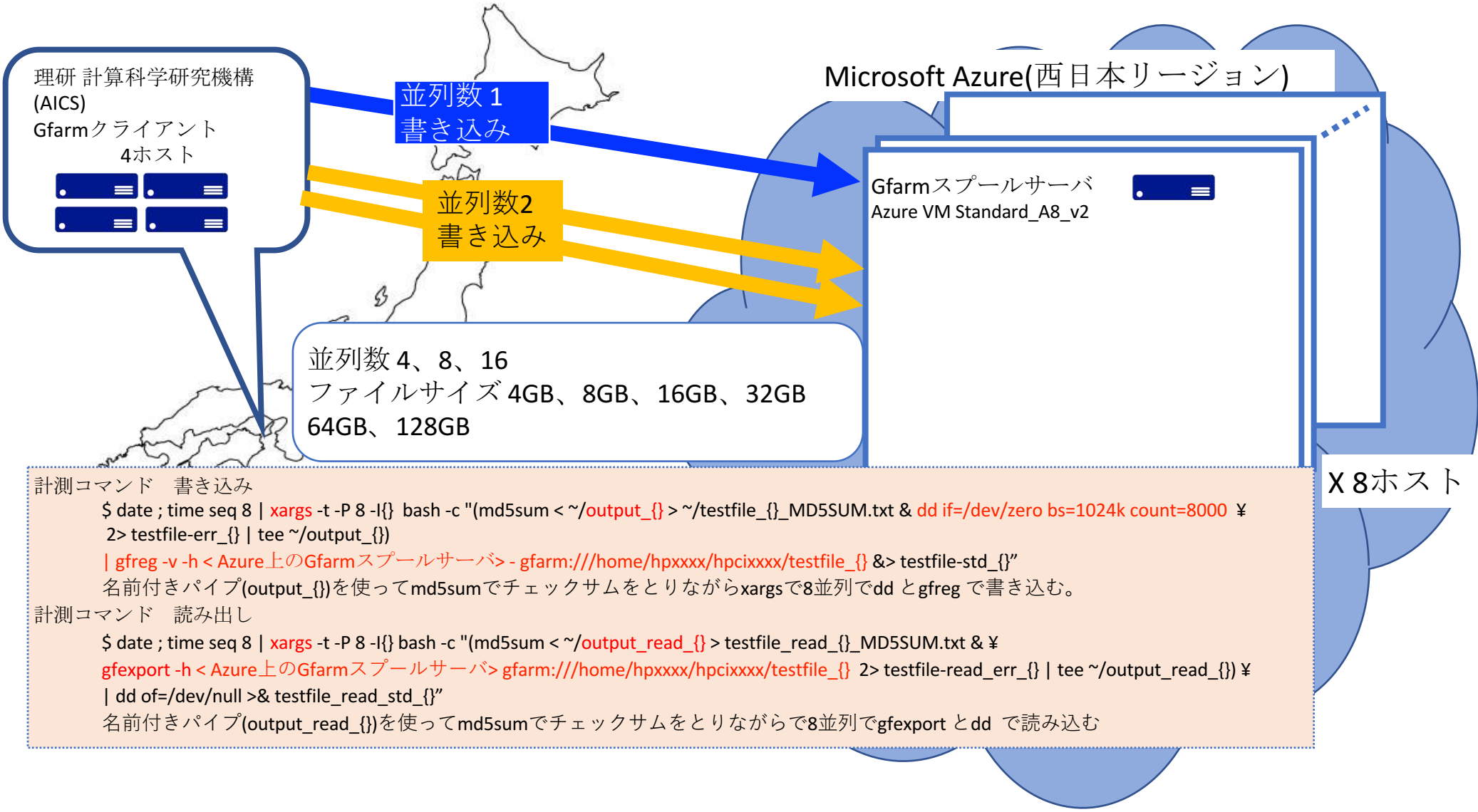


共用ストレージ東拠点資料
「機材利用状況2016/09」より引用

実証実験の構成


東京大学情報基盤センター
Gfarmメタデータサーバ
1ホスト
プロセッサ: Intel Xeon CPU X5670
コア数: 12
周波数: 2.93GHz
メモリ: 96GB
NIC: 40Gbps
メタデータ
メタデータ
Microsoft Azure(西日本リージョン)
Gfarmスプールサーバ
Azure VM Standard_A8_v2
プロセッサ: Intel Xeon CPU E5-2660
コア数: 8
周波数: 2.20 GHz
メモリ: 16 GiB
受信するデータの帯域制
限なし
送信するデータは2Gpbsの帯域制限
500 IOPS/HDD に制限
1TB HDD x 8 (RAID 0)
X 8ホスト
理研 計算科学研究機構(AICS)
Gfarmクライアント
4ホスト
プロセッサ: Intel Xeon CPU E5-2660v4
コア数: 14
周波数: 2.00 GHz
メモリ: 128GB
NIC: 10Gbps
書き込み
読み出し
※Gfarmクライアント-Gfarmスプールサーバ間の
ネットワーク帯域:
iperfコマンドで計測して16Gbps(2000MB/sec)
※Gfarmバージョンは2.6系を使用

Gfarmスプールサーバ1ホストの転送帯域測定



計測コマンド　書き込み

$ date ; time seq 8 | xargs -t -P 8 -I{} bash -c "(md5sum < ~/output_{} > ~/testfile_{}_MD5SUM.txt & dd if=/dev/zero bs=1024k count=8000 ¥
2> testfile-err_{} | tee ~/output_{})
| gfreg -v -h < Azure上のGfarmスプールサーバ> - gfarm:///home/hpxxxx/hpcixxxx/testfile_{} &> testfile-std_{}”

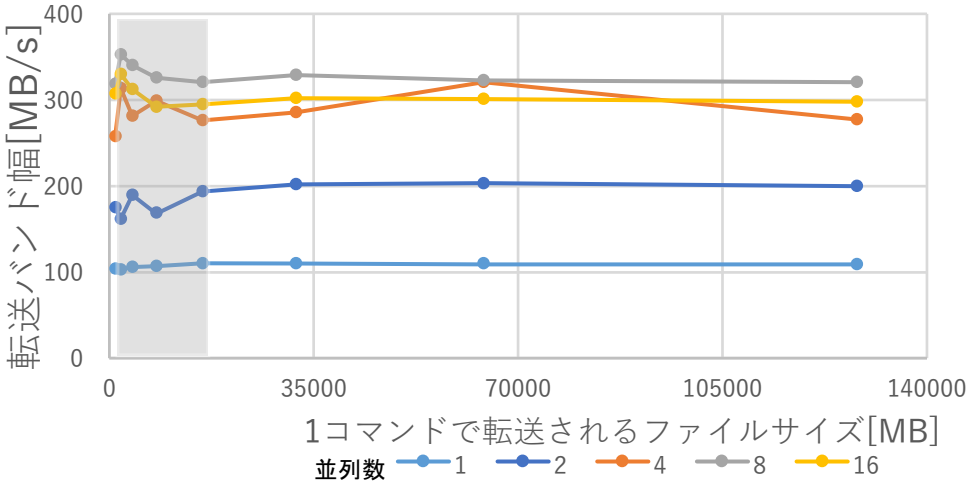
名前付きパイプ(output_{})を使ってmd5sumでチェックサムをとりながらxargsで8並列でdd とgfreg で書き込む。

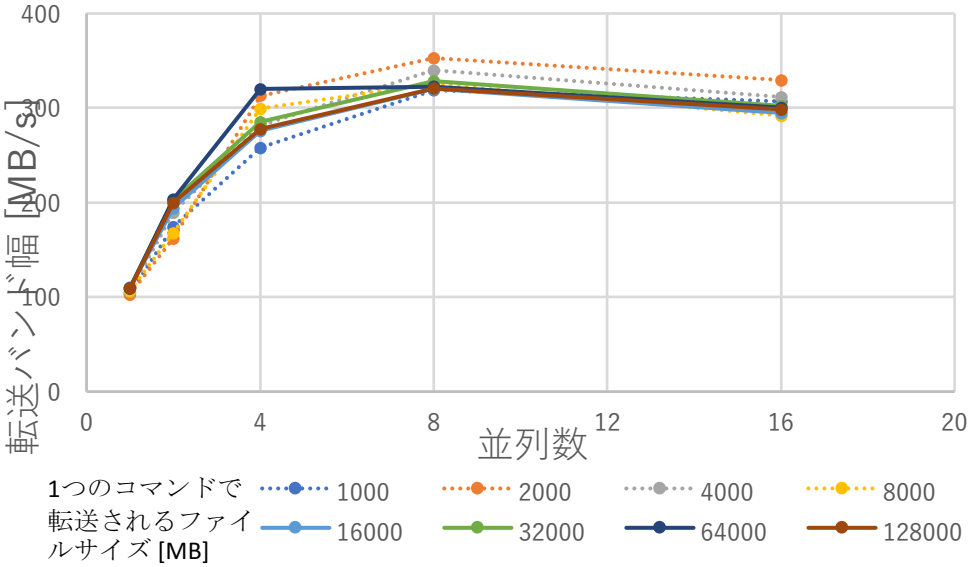
計測コマンド　読み出し

$ date ; time seq 8 | xargs -t -P 8 -I{} bash -c "(md5sum < ~/output_read_{} > testfile_read_{}_MD5SUM.txt & ¥
gfexport -h < Azure上のGfarmスプールサーバ> gfarm:///home/hpxxxx/hpcixxxx/testfile_{} 2> testfile-read_err_{} | tee ~/output_read_{}) ¥
| dd of=/dev/null >& testfile_read_std_{}”

名前付きパイプ(output_read_{})を使ってmd5sumでチェックサムをとりながらで8並列でgfexport とdd で読み込む

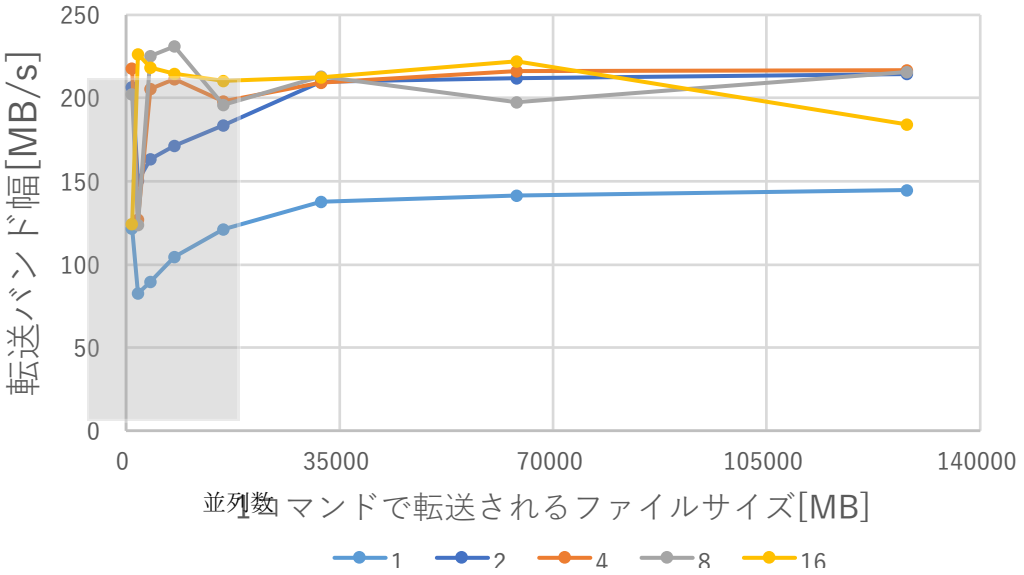
Gfarmクライアント(計算科学研究機構(AICS))からAzure 上のGfarmスプールサーバ(1ホスト)への書き込み

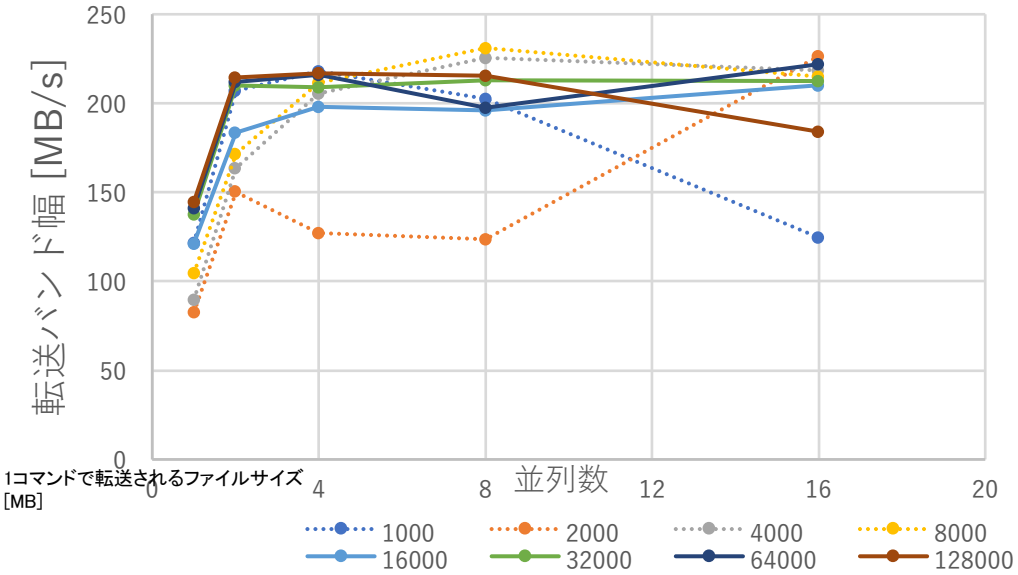




- 横軸が「1コマンドで転送されるファイルサイズ」と「並列数」のグラフを作成
- 1つのコマンドで実行するファイルサイズは16GB以上のサステインな帯域を採用する。(グラフの灰色部分、点線は採用しない)
- 1並列で最大 109MB/sec
- 8並列で最大 328MB/sec
- 16並列では8並列より転送バンド幅は低下している。

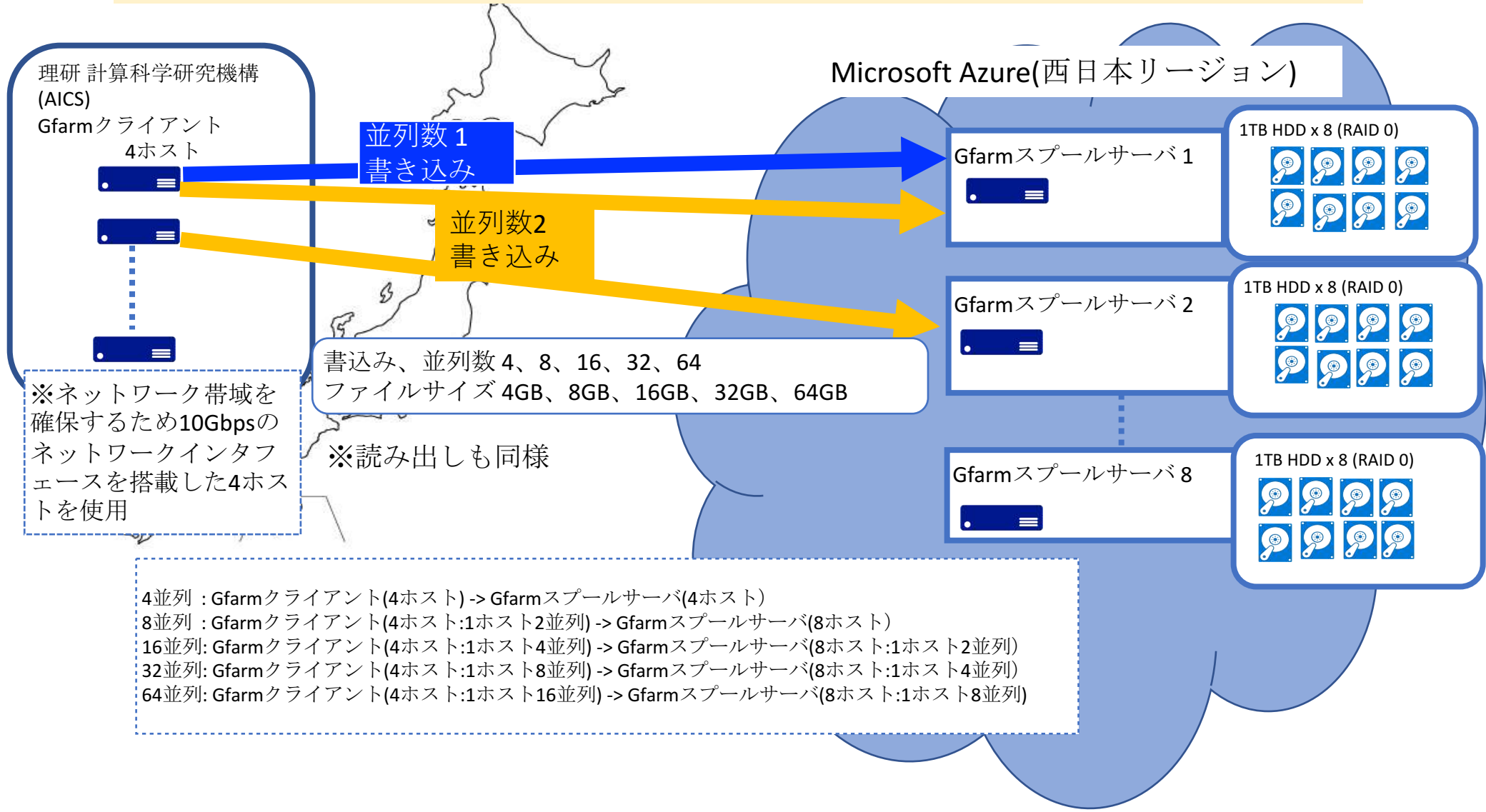
Gfarmクライアント(計算科学研究機構(AICS))でAzure上のGfarmスプールサーバ(1ホスト)からの読み出し





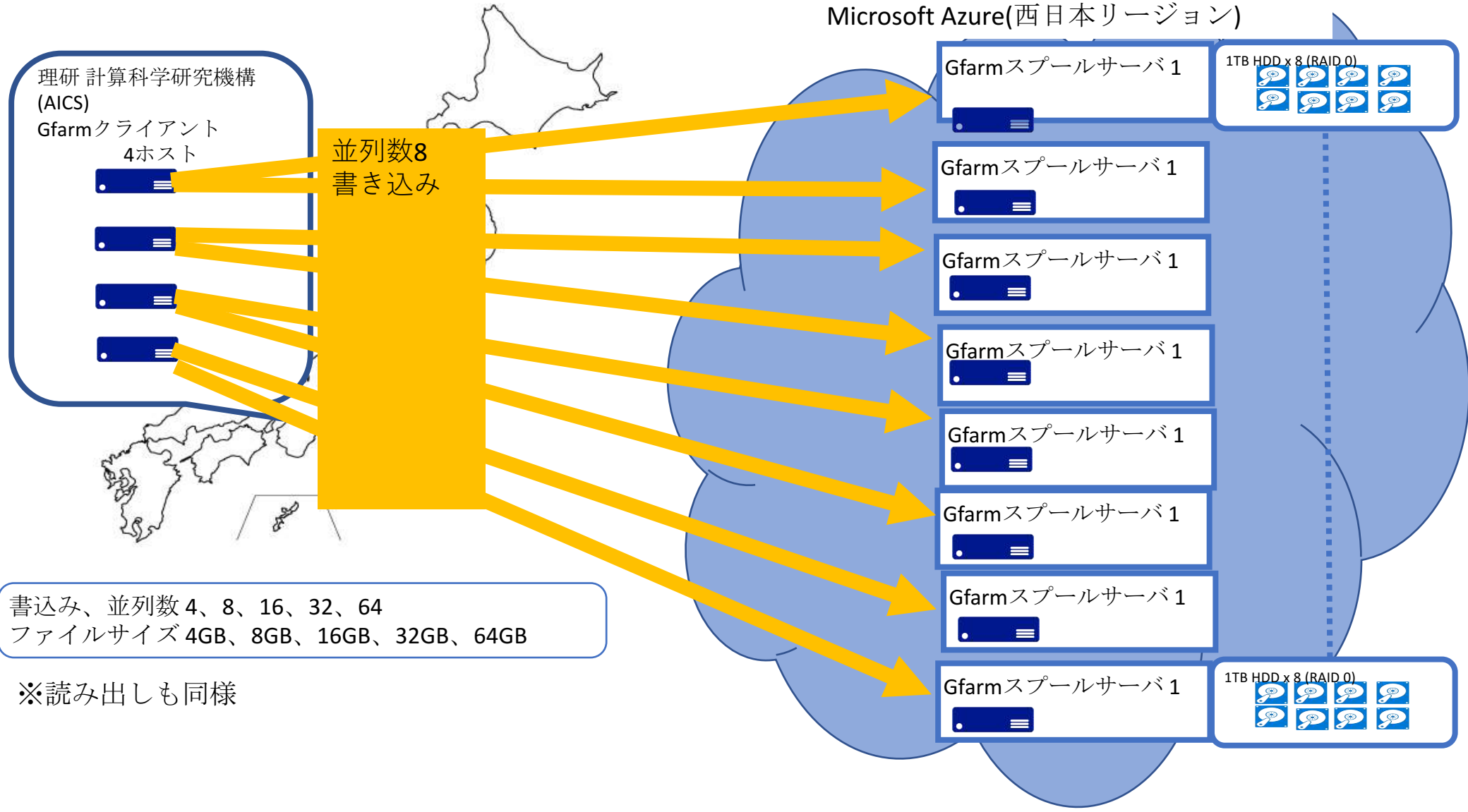
- 横軸が「1コマンドで転送されるファイルサイズ」と「並列数」のグラフを作成
- 1つのコマンドで実行するファイルサイズは16GB以上のサステインな帯域を採用する。(グラフの灰色部分、点線は採用しない)
- 1並列で最大144MB/sec
- 16並列で最大 221MB/sec
- Azure VMからの送信データは2Gbps(250MB/sec)に帯域制限にされている。最大221MB/secは帯域制限の影響であると考えられる。

Gfarmスプールサーバ8ホストの転送帯域測定
1、2並列

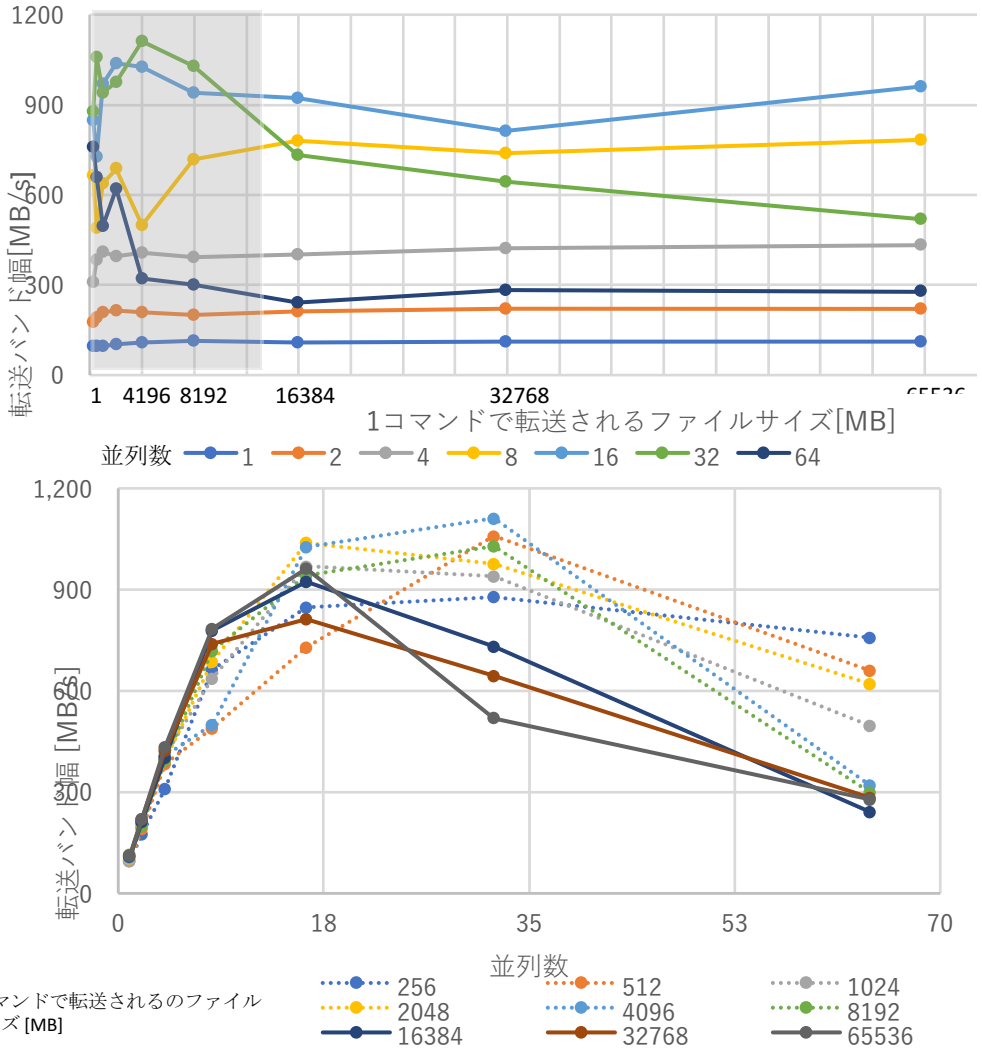


4並列 : Gfarmクライアント(4ホスト) -> Gfarmスプールサーバ(4ホスト)
8並列 : Gfarmクライアント(4ホスト:1ホスト2並列) -> Gfarmスプールサーバ(8ホスト)
16並列: Gfarmクライアント(4ホスト:1ホスト4並列) -> Gfarmスプールサーバ(8ホスト:1ホスト2並列)
32並列: Gfarmクライアント(4ホスト:1ホスト8並列) -> Gfarmスプールサーバ(8ホスト:1ホスト4並列)
64並列: Gfarmクライアント(4ホスト:1ホスト16並列) -> Gfarmスプールサーバ(8ホスト:1ホスト8並列)

Gfarmスプールサーバ8ホストの転送帯域測定 8並列

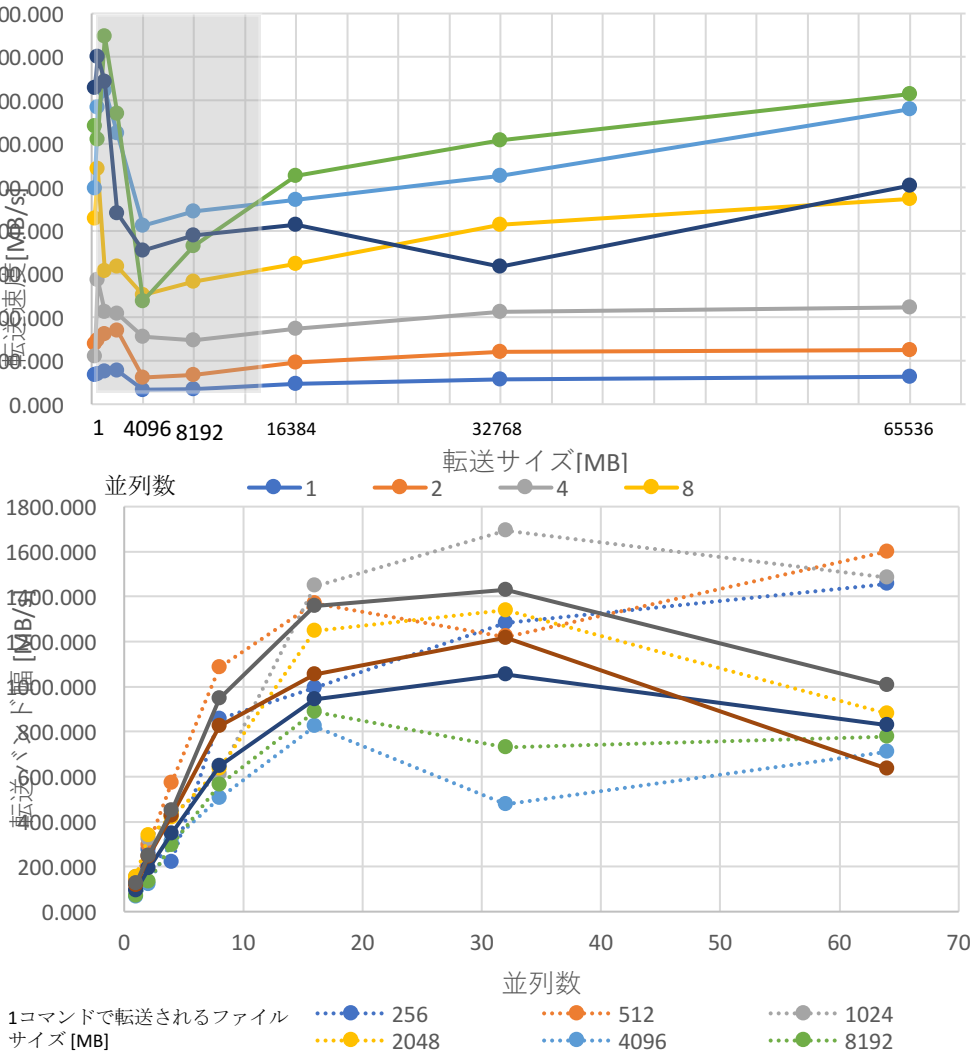


Gfarmクライアント(計算科学研究機構(AICS))でAzure 上のGfarmスプールサーバ(8ホスト)への書き込み



- 横軸が「1コマンドで転送されるファイルサイズ」と「並列数」のグラフを作成
- 1つのコマンドで実行するファイルサイズは16GB以上のサステインな帯域を採用する。(グラフの灰色部分、点線は採用しない)
- 16並列 最大 962MB/sec
- 32並列では最大で732MB/secであり、16並列より転送バンド幅は低下している。
- 64並列では最大で283MB/secであり、32並列よりさらに転送バンド幅は低下している。

Gfarmクライアント(計算科学研究機構(AICS))でAzure 上のGfarmスプールサーバ(8ホスト)からの読み出し



- 横軸が「1コマンドで転送されるファイルサイズ」と「並列数」のグラフを作成
- 1つのコマンドで実行するファイルサイズは16GB以上のサステインな帯域を採用する。(グラフの灰色部分、点線は採用しない)
- 32並列で最大 1430MB/sec
- 64並列では最大1006MB/secであり、32並列より転送バンド幅は低下している。

Gfarmスプールサーバ8ホストの転送帯域測定考察

- 書き込み最大: 16並列の962MB/sec
  - 16並列の場合にはGfarmスプールサーバ1ホストあたり2並列x 8ホストであるが、 Gfarmスプールサーバ1ホストの計測では2並列、 64GBで200MB/secでていた。
  - Gfarmスプールサーバ1ホスト200MB/sec x 8ホスト= 1600MB/sec出てもよいと考えられるが962MB/secにとどまっている。
  - クライアントサーバー間のネットワーク帯域は、実測値で16Gbps(2000MB/sec)なのでこれがボトルネックになっているとは考えられない。
  - どこでボトルネックになているかは現在のところ不明で今後の課題となっている。
- 読み出し最大: 32並列の1430MB/sec
  - 32並列の場合にはGfarmスプールサーバ1ホストあたり4並列x8ホストであるが、 Gfarmスプールサーバ1ホストの計測では4並列、64GBで432MB/secでていた。
  - Gfarmスプールサーバ1ホスト432MB/sec x 8ホスト= 3456MB/sec出てもよいと考えられるが、1つのAzure VMから出ていくデータの2Gbps(250MB/sec) x 8ホスト = 2000MB/secの帯域制限までも利用できていない。
  - クライアントサーバー間のネットワーク帯域は、実測値で16Gbps(2000MB/sec)までも利用できていない。
  - どこでボトルネックになているかは現在のところ不明で今後の課題となっている。
- **実は、、、同じクラスのインスタンスでも異なるプロセッサが採用されていた。**
- **プロセッサによって性能差が発生していた（遅いgfsdが３台あった！）**
- **カーネル更新によって性能を均等化できそう。**
- **再計測を予定**
- **山あり谷ありでしたが、クラウドは技術的には実用可能。かつアーカイブ向けのクラウドストレージも実用段階に突入。**
- **あとは予算の年度区切り問題が解決してくれれば。。。。**
- **階層化とクラウドを活用できれば、、、、**
  - **各計算機センターからこれまで以上に高速なデータ保存が可能**
  - **提供容量をフレキシブルに調整可能**
  - **より安全なデータ多重化サービスの提供**