宇宙地球環境科学のためのクラウドコンピューティング：
情報通信研究機構OneSpaceNetへの取組み (2)

グリードデータファーム (Gfarm) と JGN2+による大規模分散ストレージ・並列分散処理システムの構築

情報通信研究機構
村田 健史、亘 慎一、加藤 久雄、○森川 靖大、長妻 努、
石井 守、品川 祐之、久保 勇樹、
久保田 実、國武 学、秋岡 真樹、石橋 弘光、
田 光江、島津 浩哲、坪内 健、津川 卓也、
陣 英克、上本 純平、中溝 葵、永原 政人

2009/08/20- 21 データ科学ワークショップ@北海道 学術交流会館

はじめに

- 計算機シミュレーションの大規模化
  - 計算機性能の向上
  - 領域結合型地球環境シミュレータ
- 大規模データの解析・可視化処理のボトルネック
  - ディスク容量
  - ディスクI/O速度
  - 通信速度

本研究(＋α)の目標

- Gfarm と JGN2+ による大規模分散ストレージ・並列処理システムの構築
  - 数百TB～PB級大規模分散ストレージ
  - 分散ディスクでの並列処理
  - 10Gbps のL2ネットワーク

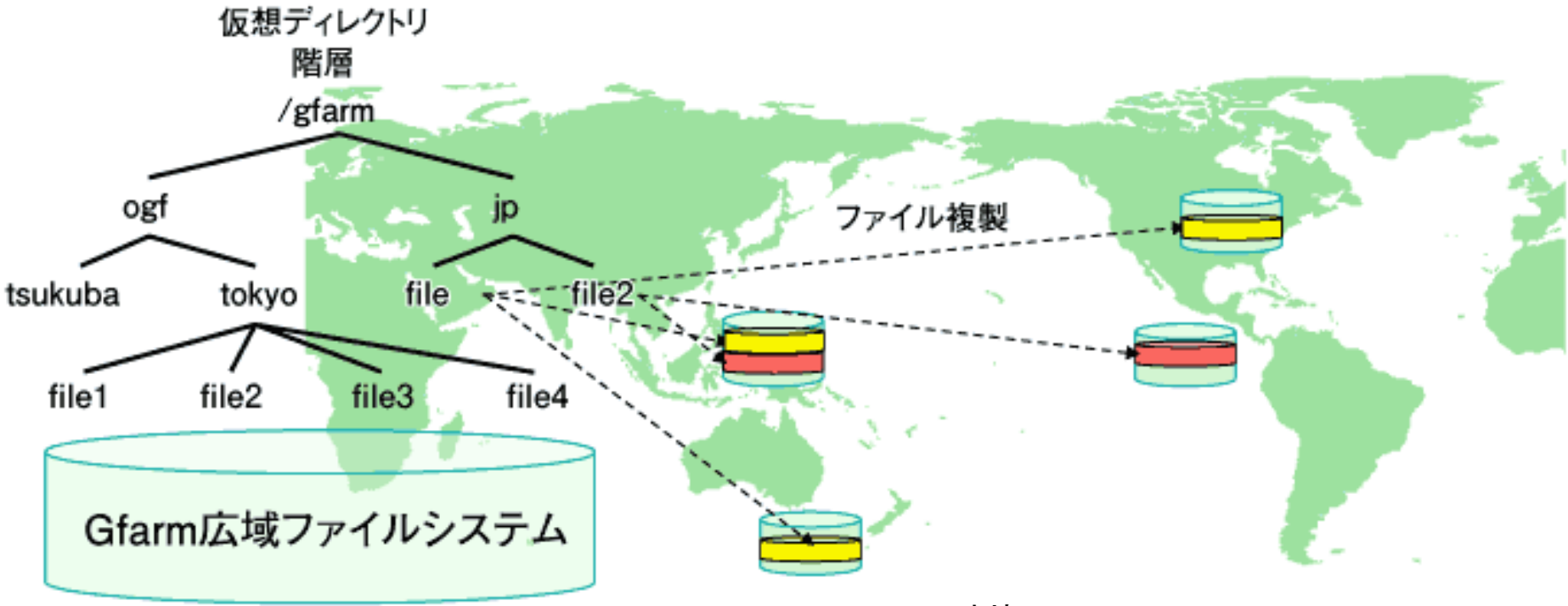
- スパコン＋解析可視化環境をセットで提供
  - OneSpaceNet

お品書き

- Gfarm ファイルシステム
- JGN2plus
- NICTにおけるストレージシステムの構想
- 運用までのシナリオ
- 利用の手順

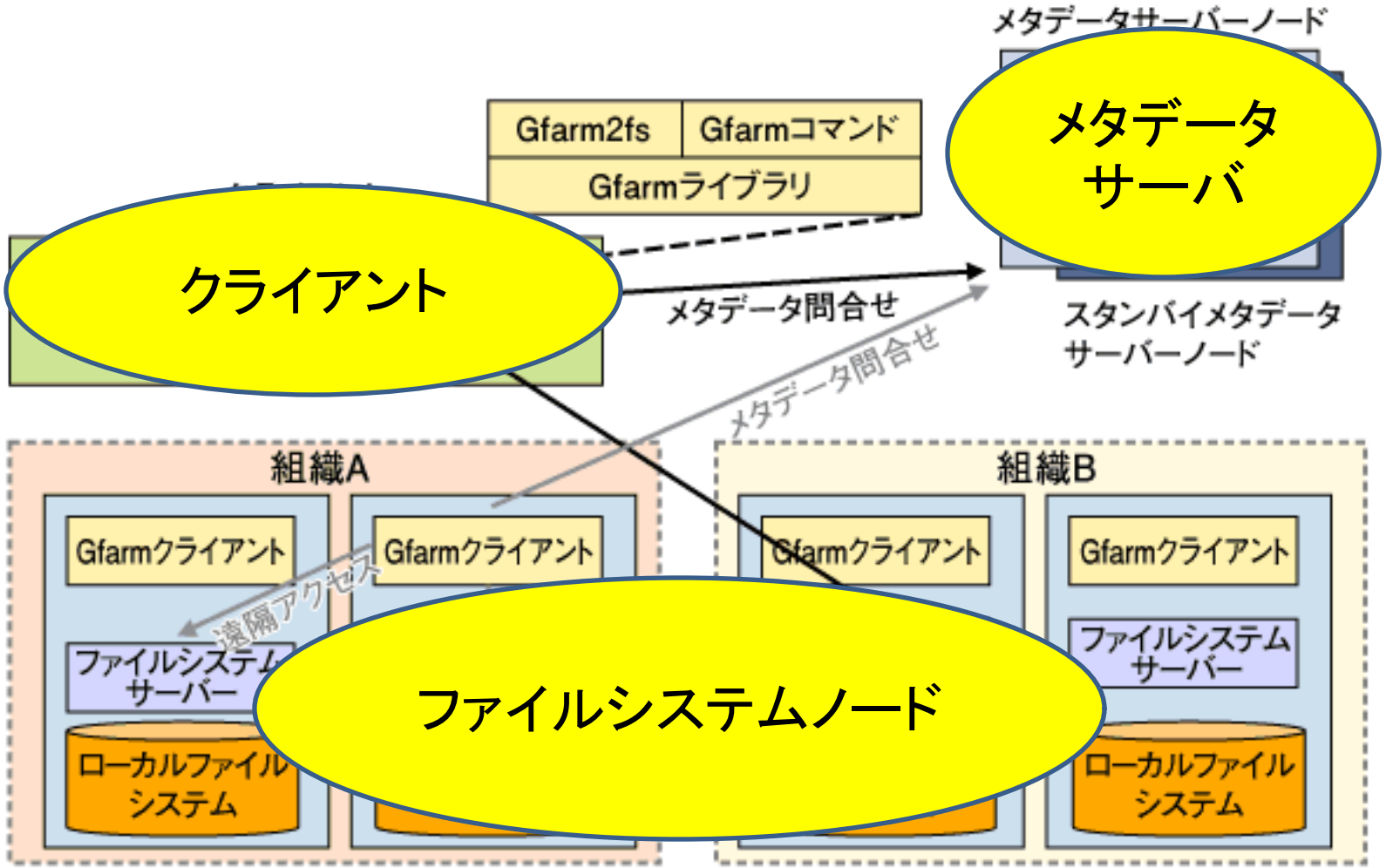
Gfarm 広域ファイルシステム

- 高速アクセス可能な広域の共有ファイルシステム
- 利用者はファイルの実際の格納場所を意識せず、仮想ディレクトリ階層へアクセス



http://thinkit.jp/article/772/1/ より

Gfarm ファイルシステム構成



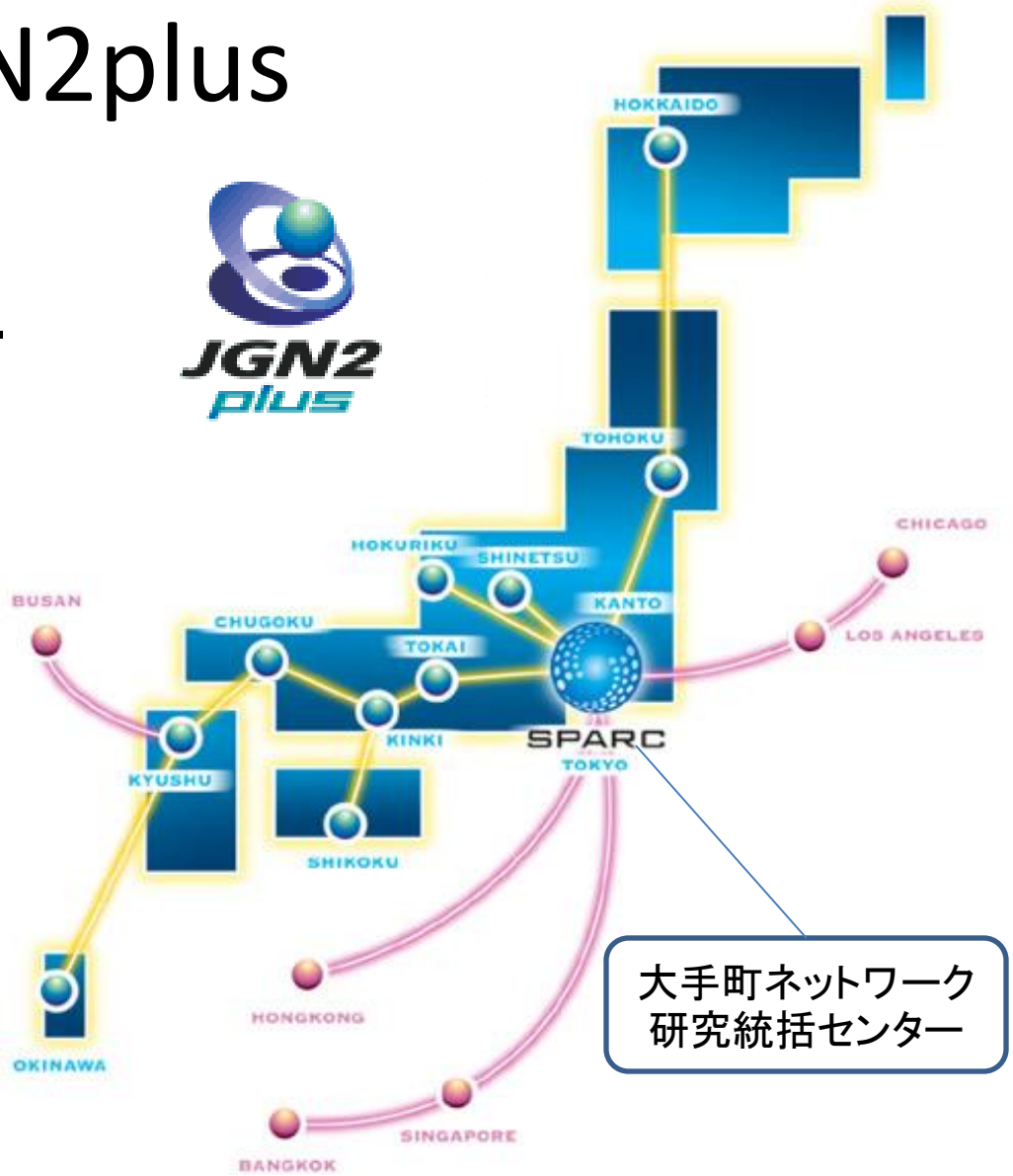
http://thinkit.jp/article/772/2/ より

Gfarm の特徴

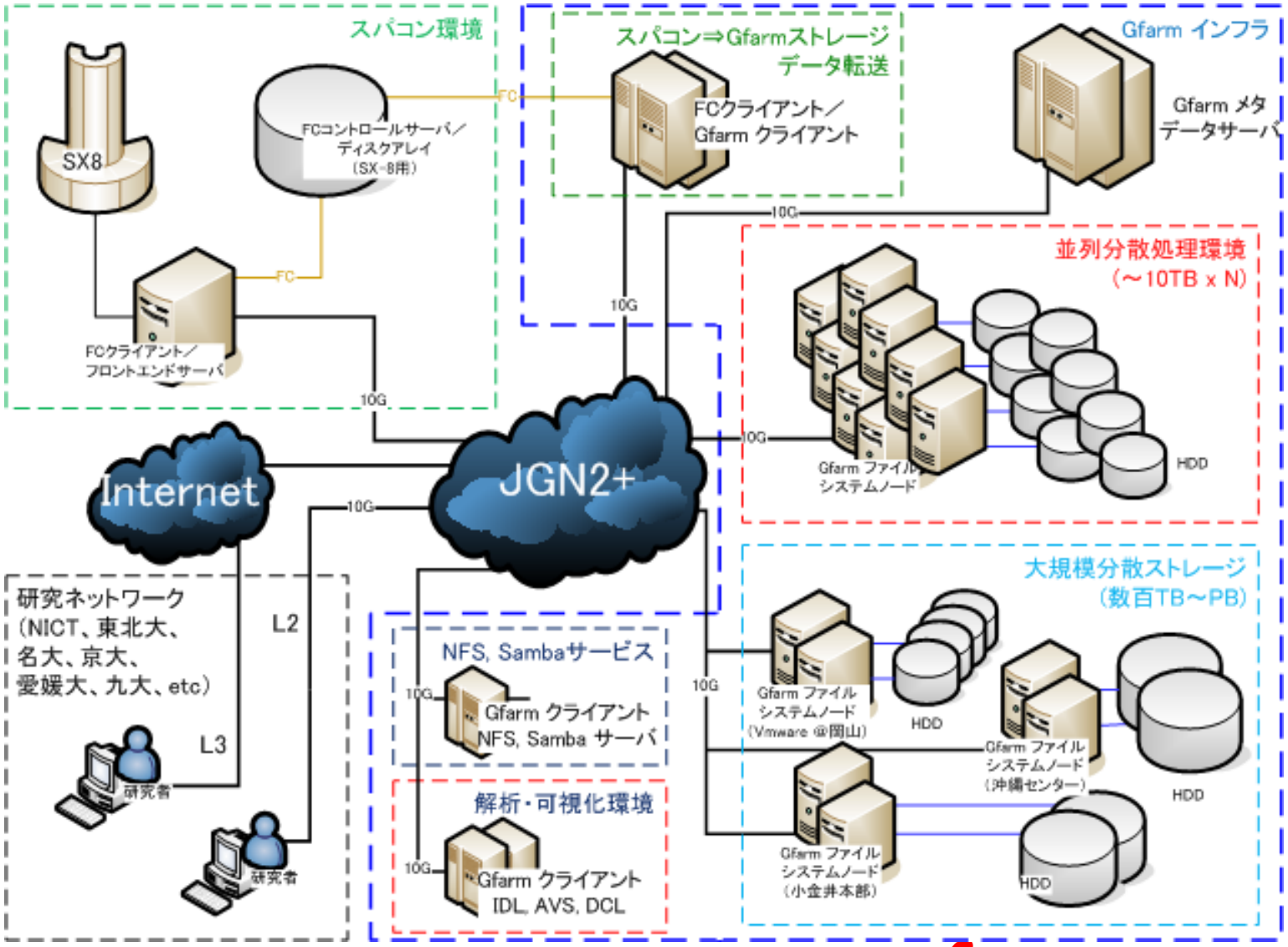
- 運用の容易さ
  - 複数ノードへのファイル複製により耐故障性を向上
  - システムを稼働したままノードの増設が可能
- ユーザビリティの高さ
  - GfarmFS の仲介により既存プログラムでファイル操作可能
  - Samba, NFS 経由でのアクセスも可能
- 並列分散処理のサポート
  - 各ノードへ実行ファイルをコピーして実行
- 分散ノードへの高速アクセス
  - 各ノードのネットワーク上の距離、CPU負荷の計測
  - ファイル複製を活用

JGN2plus

- NICT による研究開発テストベッドネットワーク
- NICT の各拠点と複数の大学間を結ぶ
- L2 Ethernet サービス (100Mbps～10Gbps) 提供

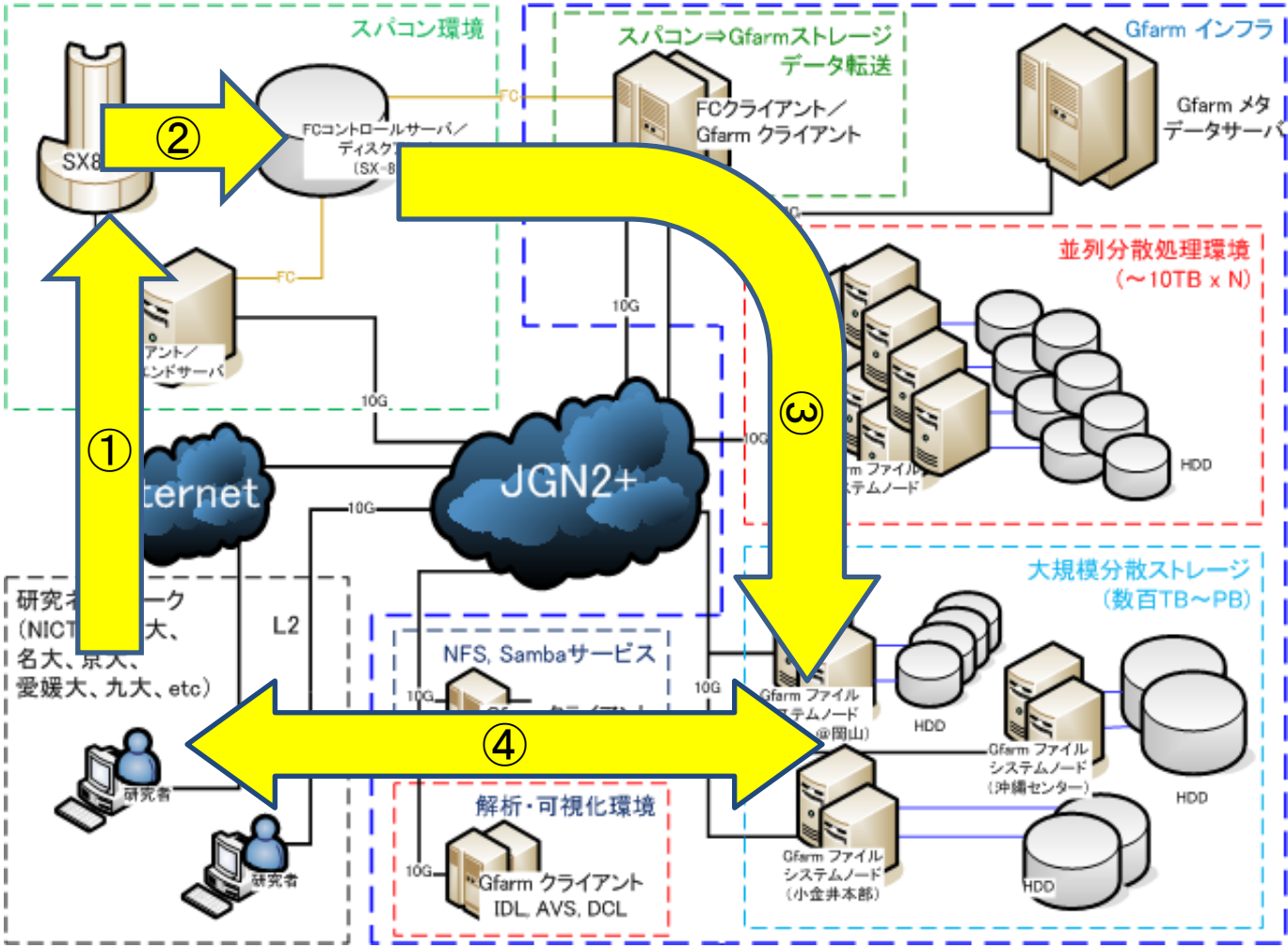


システム構想図


スパコン環境
SX8
FCコントロールサーバ／
ディスクアレイ
（SX-8用）
FC
FC
FCクライアント／
フロントエンドサーバ
10G
スパコン⇒Gfarmストレージ
データ転送
FCクライアント／
Gfarm クライアント
Gfarm インフラ
Gfarm メタ
データサーバ
10G
10G
並列分散処理環境
（～10TB x N）
10G
Gfarm ファイル
システムノード
HDD
Internet
JGN2+
10G
研究ネットワーク
（NICT、東北大、
名大、京大、
愛媛大、九大、etc）
L2
L3
研究者
研究者
NFS, Sambaサービス
10G
Gfarm クライアント
NFS, Samba サーバ
解析・可視化環境
10G
Gfarm クライアント
IDL, AVS, DCL
10G
大規模分散ストレージ
（数百TB～PB）
Gfarm ファイル
システムノード
（Vmware @岡山）
HDD
Gfarm ファイル
システムノード
（沖縄センター）
HDD
Gfarm ファイル
システムノード
（小金井本部）
HDD

大規模分散ストレージ利用

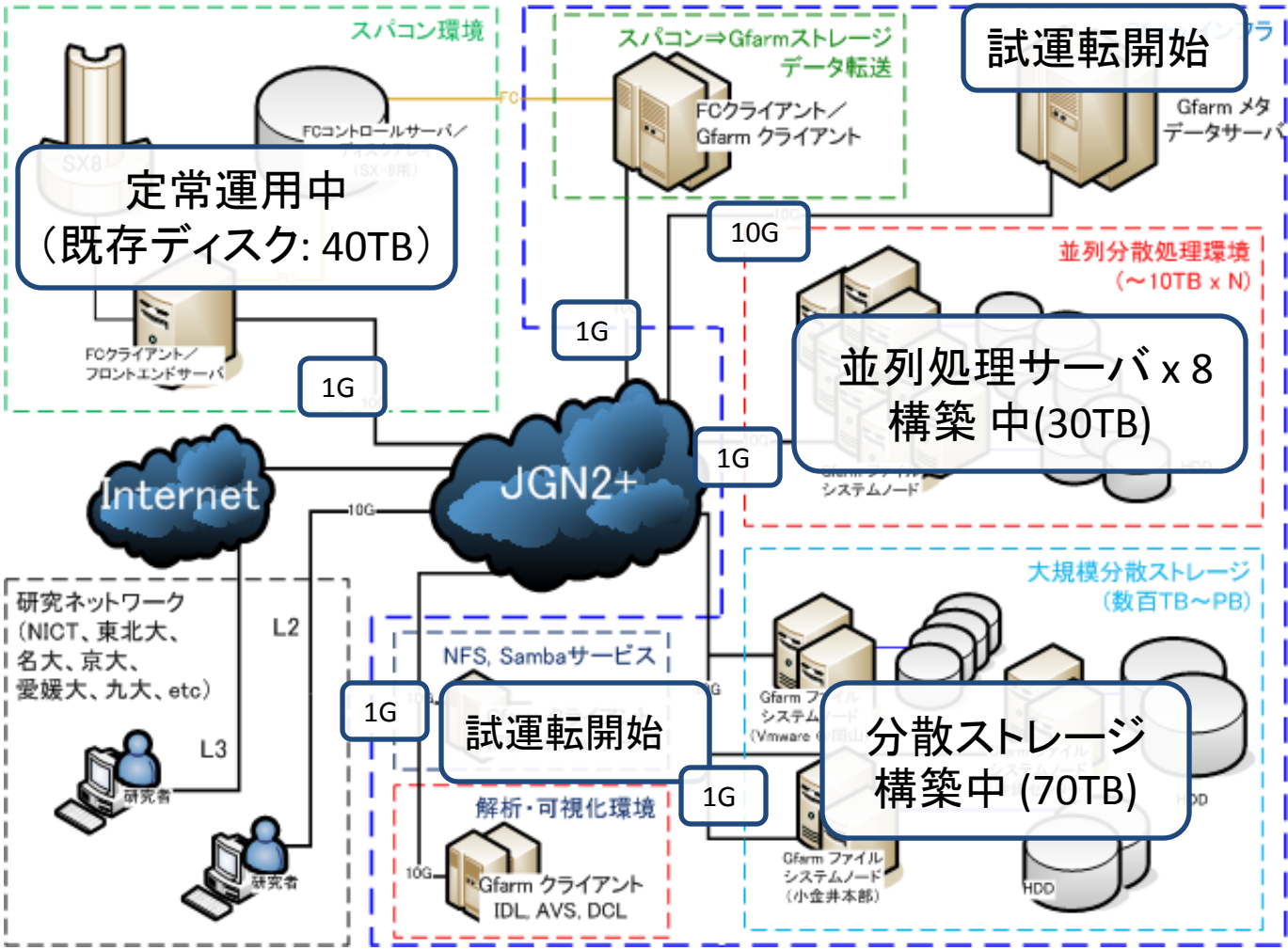
① ジョブ投入
② 演算・データ出力
③ 大規模分散ストレージへのデータ転送
④ 解析・可視化



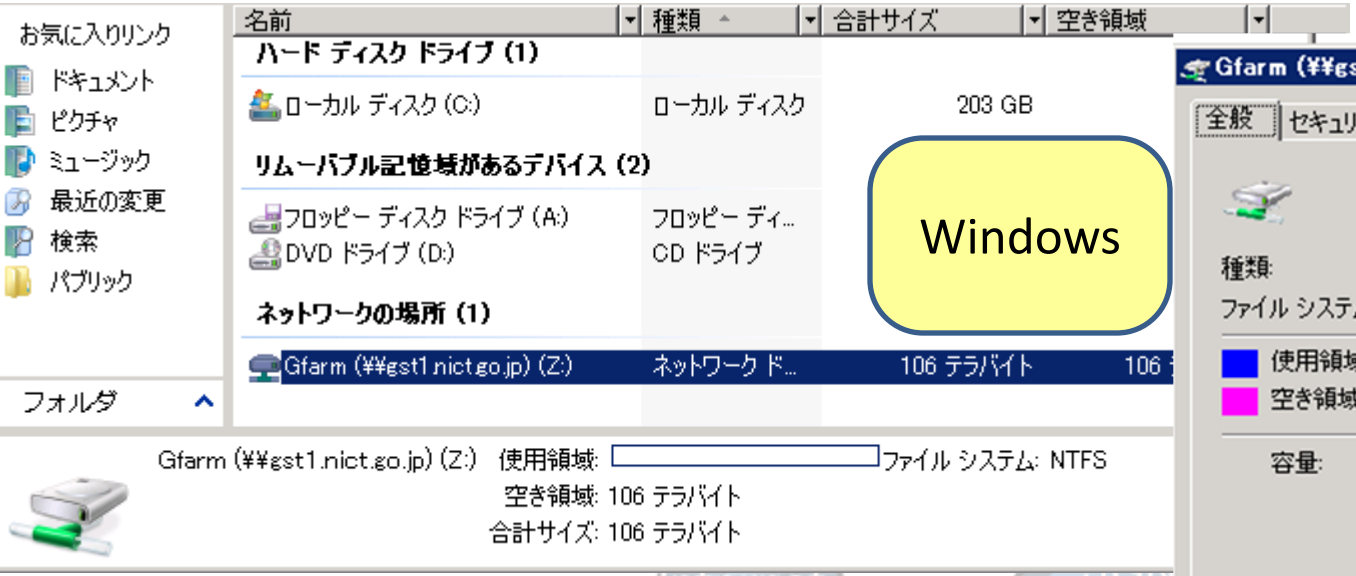
ミドルウェアを意識しない大規模ストレージ利用

2009/08/20 時点での構築状況

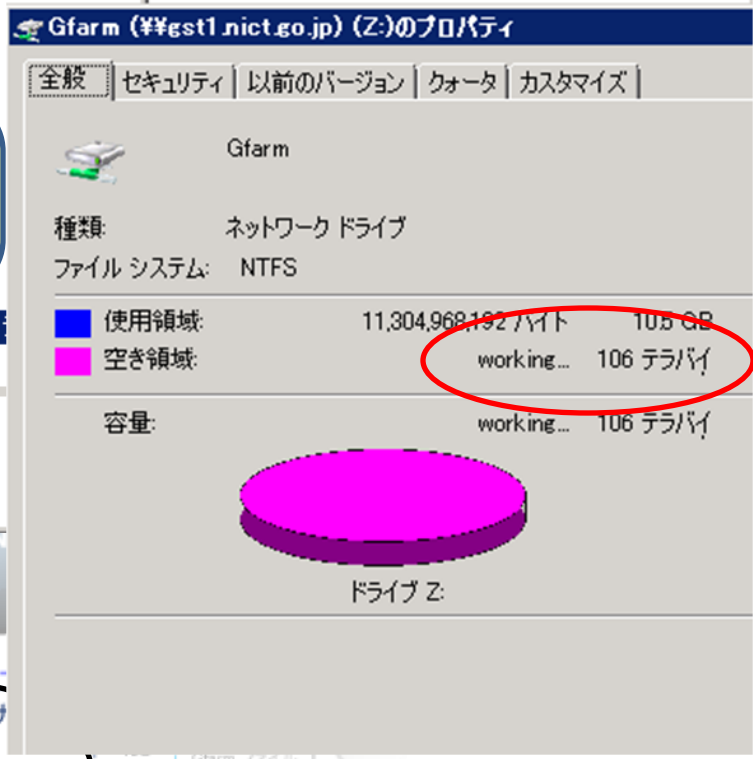
- ToDo
  - 解析・可視化環境へのGfarm導入
  - ディスク増設
  - I/O速度検証
  - 耐障害性検証
  - スパコン⇒ストレージ転送システム構築
  - 運用体制確立
  - 10G 化
  - 並列分散処理システム構築

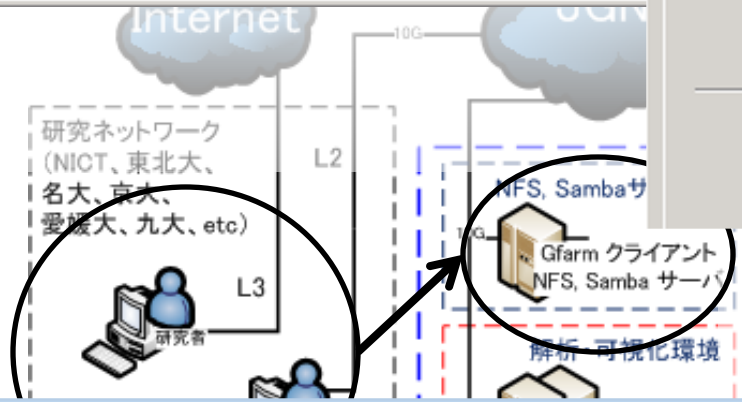


2009/08/20 時点での構築状況



Windows



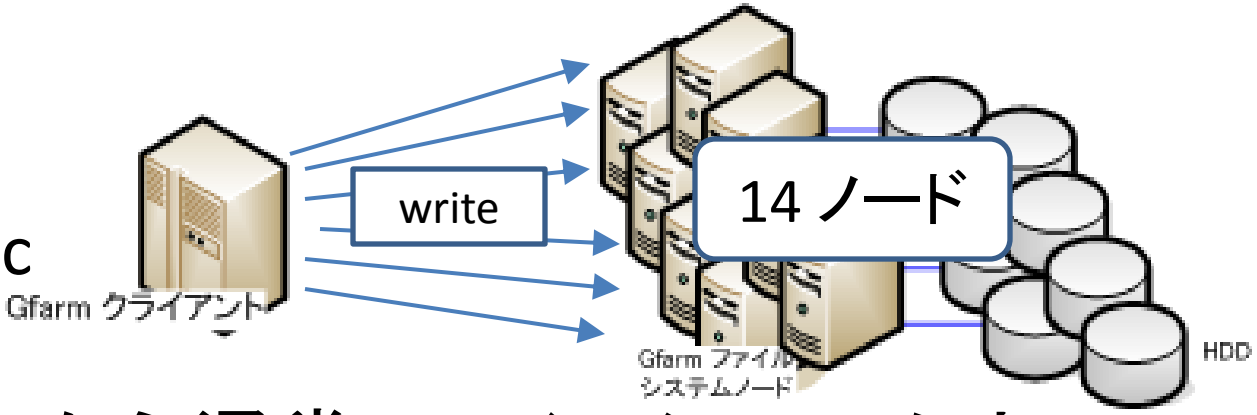


Linux

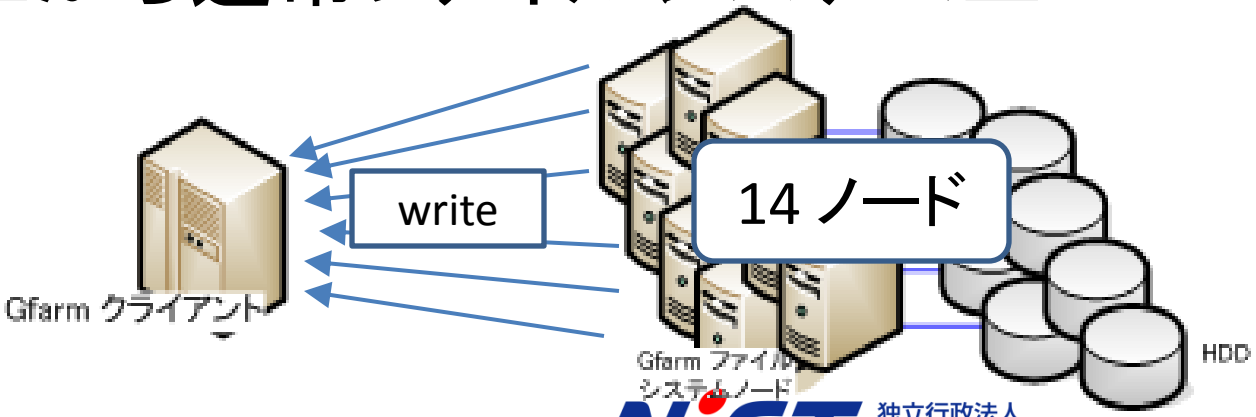
```
seg-user@gst1:~> df
Filesystem           1K-ブロック      使用    使用可 使用% マウント位置
/dev/sda2             32890776   4375740  26844276  15% /
udev                   2026012        92   2025920   1% /dev
gfarm2fs          114196153016  12679360 114183473656   1% /home/seg-user/seg-space
seg-user@gst1:~>
```

I/O・通信速度計測中

- 単一ホスト（メモリ上）から分散ディスクへ書込
  - 100MB x 40
    書込: 20 sec
    ≒ 200 MB/sec



- 分散ディスク上から通常ファイルシステム上へのデータ書込
  - 100MB x 40
    読込: 50 sec
    ≒ 80 MB/sec



大規模ストレージ
運用開始までのスケジュール

- 08月：～100TB ディスクの構築
  - 以後、定期的にディスクの増強
- 09月：運用体制の検討・スパコンとの連携
- 10月：仮運用開始
  - OneSpaceNet ユーザであれば利用可能
- 11月：I/O速度性能、耐障害性の検討
- 12月：運用体制の見直し
- 01月：数百TBストレージ本運用開始

利用に際して（再掲）

- 利用目的
  - 宇宙環境、地球環境、電磁波計測、ICT技術の研究
- 利用手順
  - 利用申請書提出
  - 機器接続申請、JGN2+利用申請 （JGN2+直結の場合）
- 年1回で成果発表会
- OneSpaceNet-Admin@ml.nict.go.jp までお気軽にご相談ください。
  - 近々申請用Webサイト整備の予定

まとめ

- Gfarm と JGN2+によるL2ネットワーク上での数百TBストレージと並列処理環境を構築中
- 大規模シミュレーションデータの解析・可視化環境の基盤へ
- 今年中の運用開始を目指す
- テストベッド環境として使ってくれる方募集
  - OneSpaceNet-Admin@ml.nict.go.jp まで