

NIFS LABCOMグループの 実験データシステム開発状況

自然科学研究機構 核融合科学研究所
高温プラズマ物理研究系 中西秀哉

核融合研究の状況

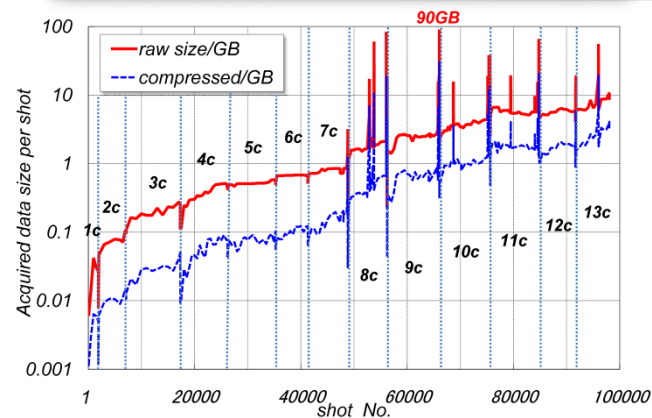
- 核融合科学研究所（岐阜県土岐市）では、ヘリカル磁場閉じ込め方式の核融合実験「**大型ヘリカル装置（LHD）**」を運用（1998年3月～）



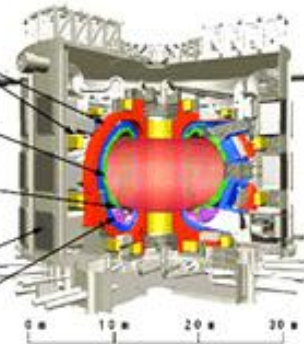
- LHD、九州大QUEST、筑波大GAMMA10で実験データのSINET仮想閉域網を介したマルチサイト実時間共有を開始（2008年6月～）
＜核融合バーチャルラボトリ＞



- 世界7極による**国際熱核融合実験炉（ITER）**が、2019年実験開始に向け仏カダラッシュに建設中
 - 2010年7月、ITER新機構長に本島修NIFS前所長が就任
 - 青森県六ヶ所村に、ITER遠隔実験センターが建設予定



超伝導コイル
 (■, ●)
 真空容器
 ブランケット
 クライオスタット
 ダイバータ



ITERの設計値	
核融合出力	: 500 MW
エネルギー増倍率	: Q 10 (~400 s) Q ~5 (定常運転)
大半径/小半径	: 6.2 m / 2.0 m
プラズマ電流	: 15 MA
最大磁場	
(コイル中心)	: 11.8 T
(プラズマ中心)	: 5.3 T
プラズマ体積	: 840 m ³
最大加熱パワー	: 73 MW

核融合バーチャルラボトリ

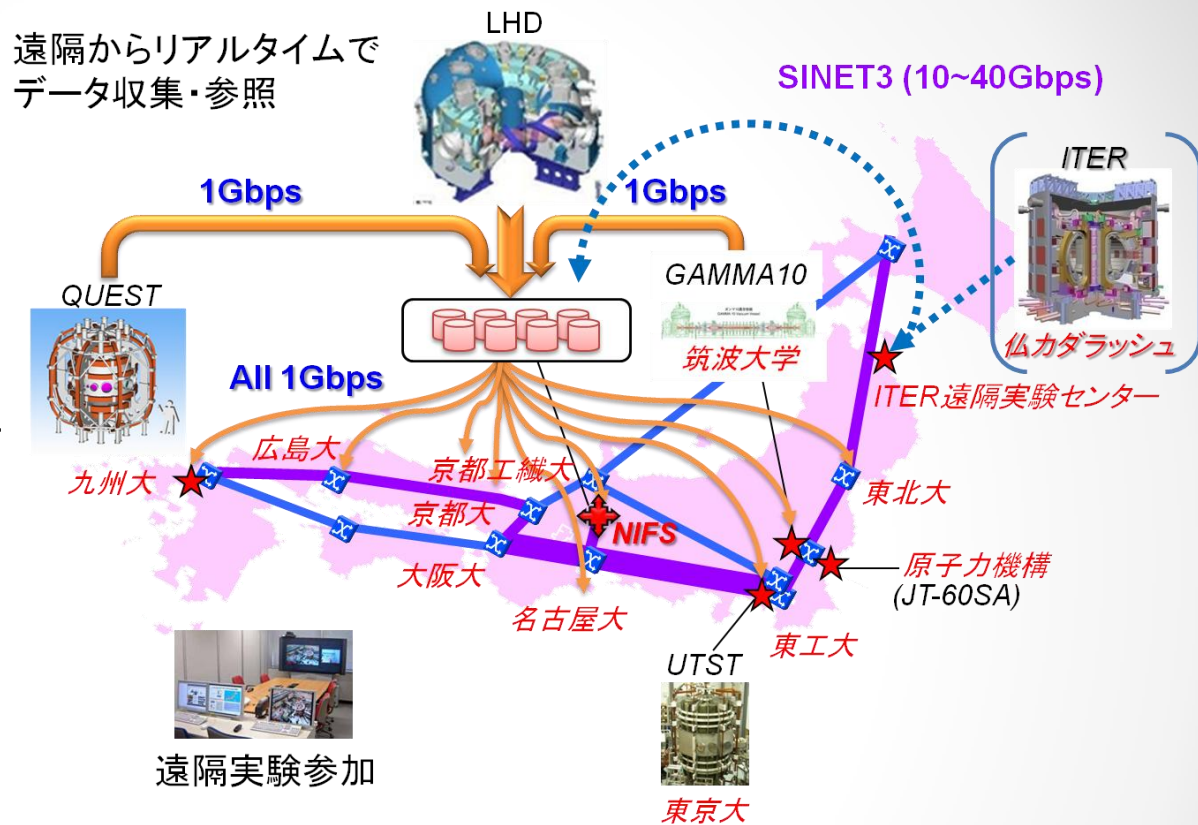
- SINET4上に核融合分野専用の仮想閉域網 (SNET)
 - 帯域1Gbps (大学側) ~ 10Gbps (NIFS側)
 - 国内双方向共同研究の通信バックボーン



- LHD実験データシステムを多サイト拡張 ⇒ 高可用性が不可欠に

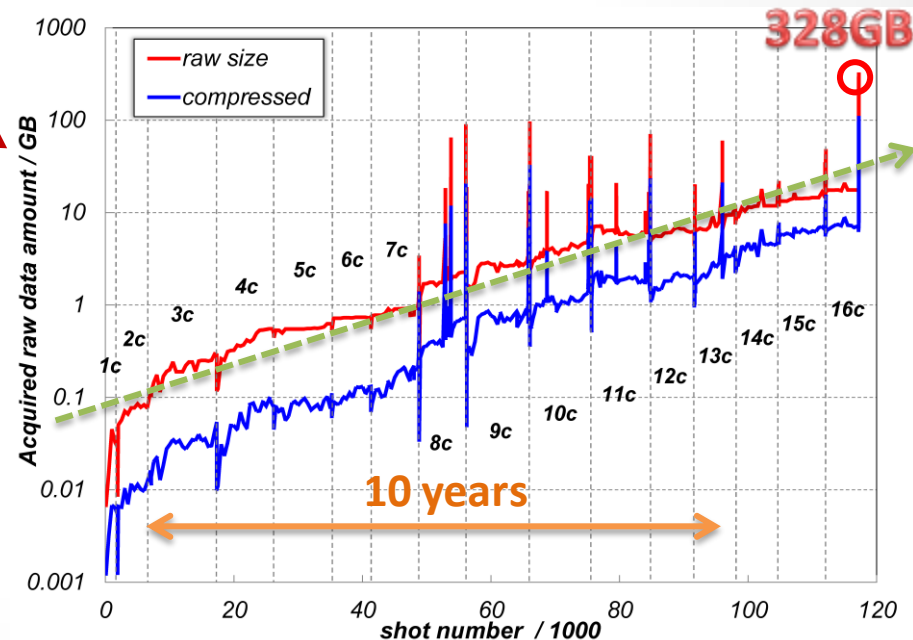
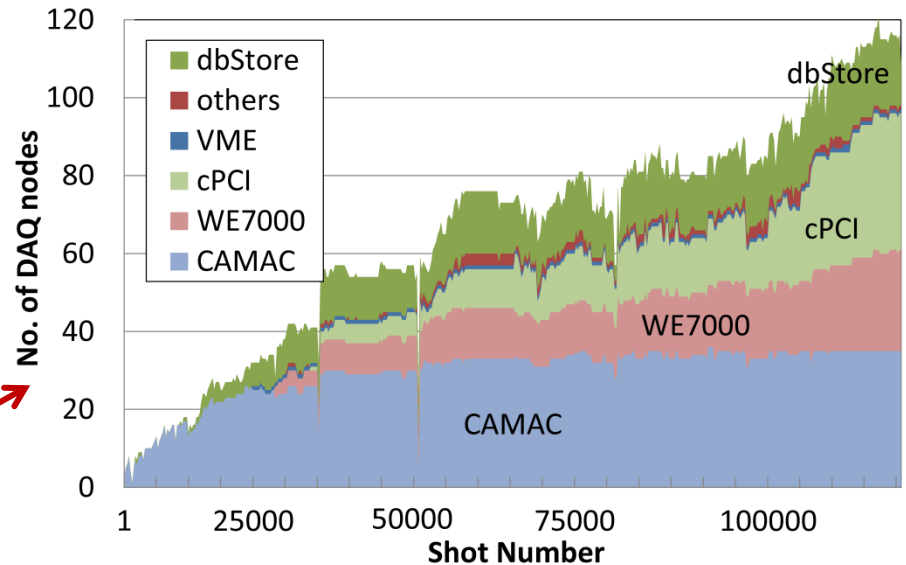


- 遠隔の3実験 (LHD, QUEST, GAMMA10) のリアルタイムデータ収集ノード、データストア、データ参照クライアントがSNET上に広域分散
 - (遠隔) データ収集、データサービスの集中管理 (NIFS側)
 - 将来、ITER実験データの再配信にも



LHD Data Trend

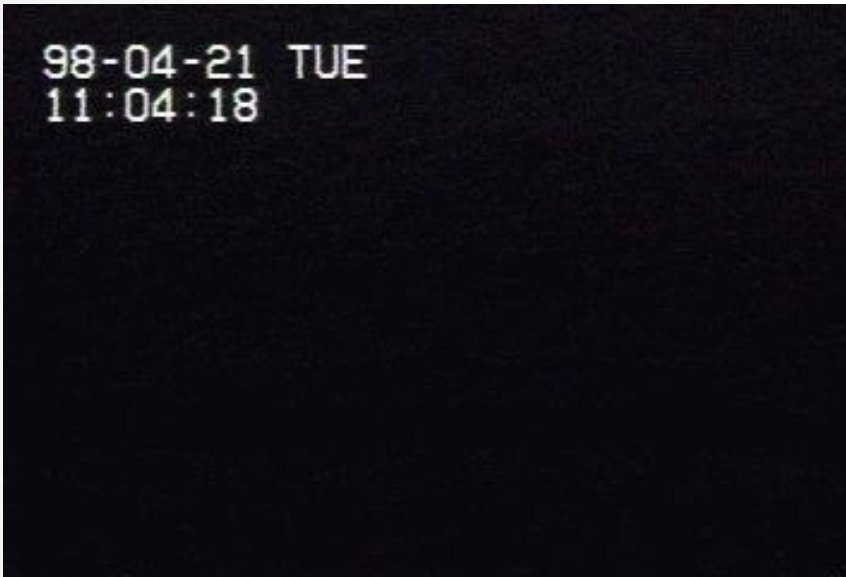
- LHD DAQ & archiving system has a massively distributed structure
 - ✓ having 110 DAQ nodes (2012)
 - ✓ operates in every 3 min. → 180 /day
- No. of DAQs continues growing almost **linearly**.
- Data amount continues growing **exponentially**.
 - ✓ totally acquires ~ 18 GB/short-pulse
- LHD shares the central storage with QUEST and GAMMA10.
- Easy scale-out** and **fault recovery on the fly** are mandatory for data storage.



核融合プラズマ実験

10秒未満の短パルス実験

定常（長パルス）保持実験



98-04-21 TUE
11:04:18



80sec
Discharge

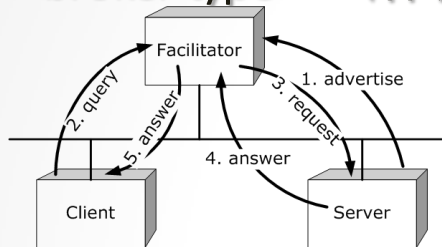
定常データ収集ではノード毎に100MB/s ~ の収集能力が必要

LHD実験データシステム

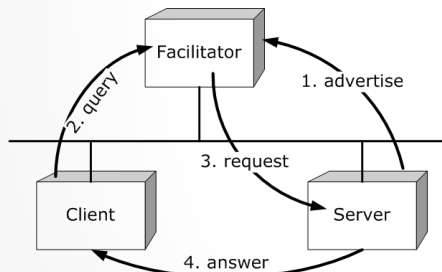
P2Pデータサービスにはデータの斡旋が必要

→ **Facilitator (Mediator) model**

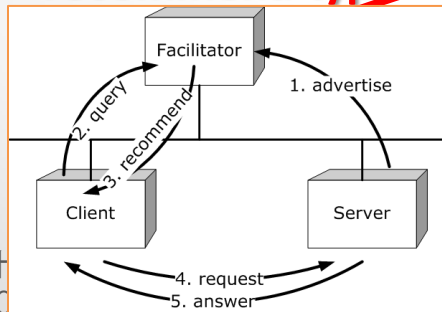
1. broker type (仲介)



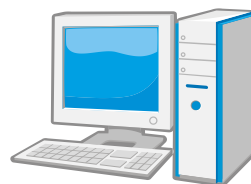
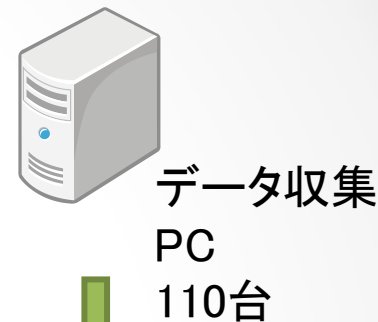
2. recruit type (斡旋)



3. recommend type (推薦)



計測器 ~ 110種類



データ参照
PC
100台



ディスクサーバ
< 10台



Facilitator

データライブラリ
4台



データ収集系のクラウド化

- 計測器1台に収集PC1台を用いて同時並行で収集・処理する大規模分散形態
- 100超の収集ノードは、ネットワーク上を流れる実験シーケンスに同期して処理進行
- 制御コマンド/ステータス授受はIPマルチキャスト

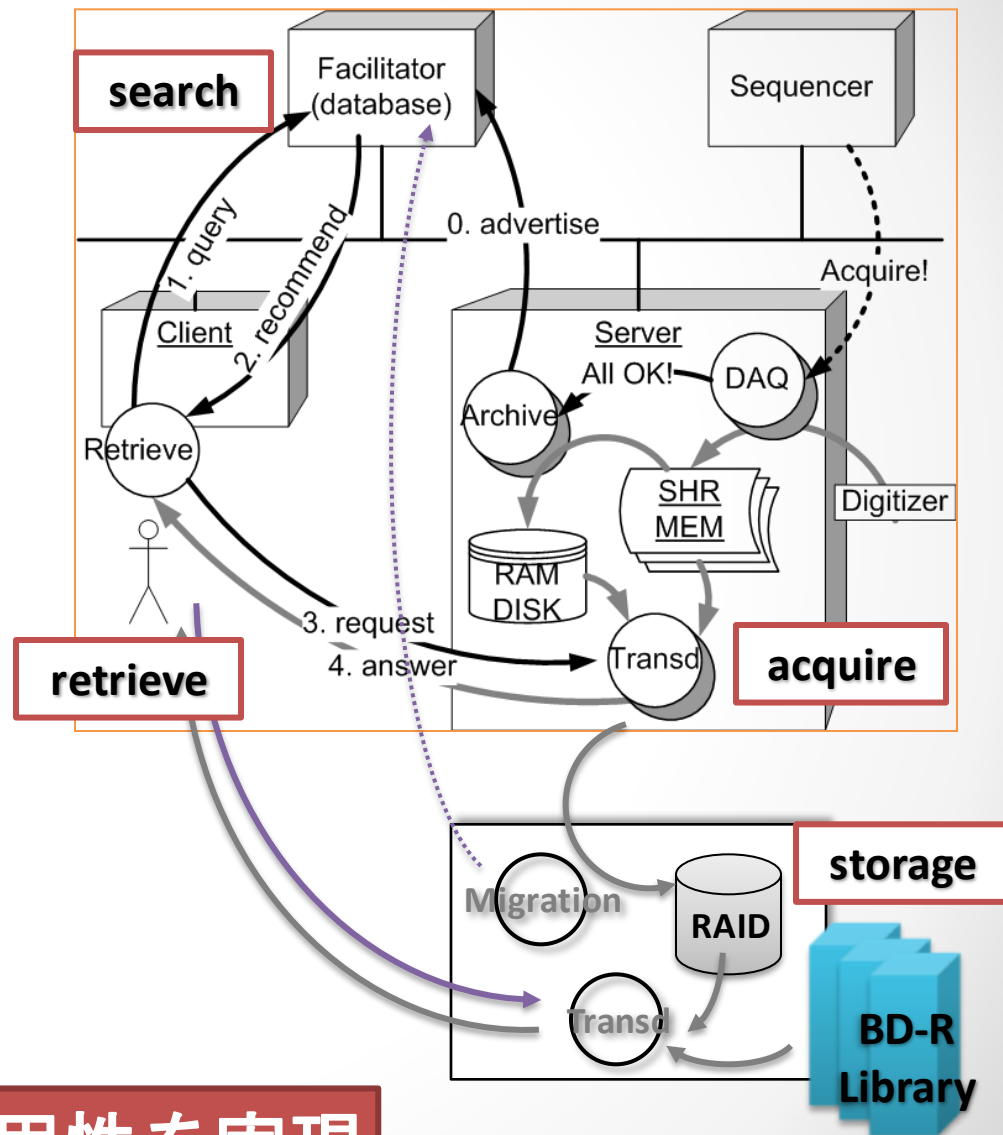


- PXE+DHCPによるディスクレス・ネットワークブート構成
- 実験中でも任意ノードの挿抜・交換が可能



- 各ノード間で殆ど依存関係のないクラウド形態

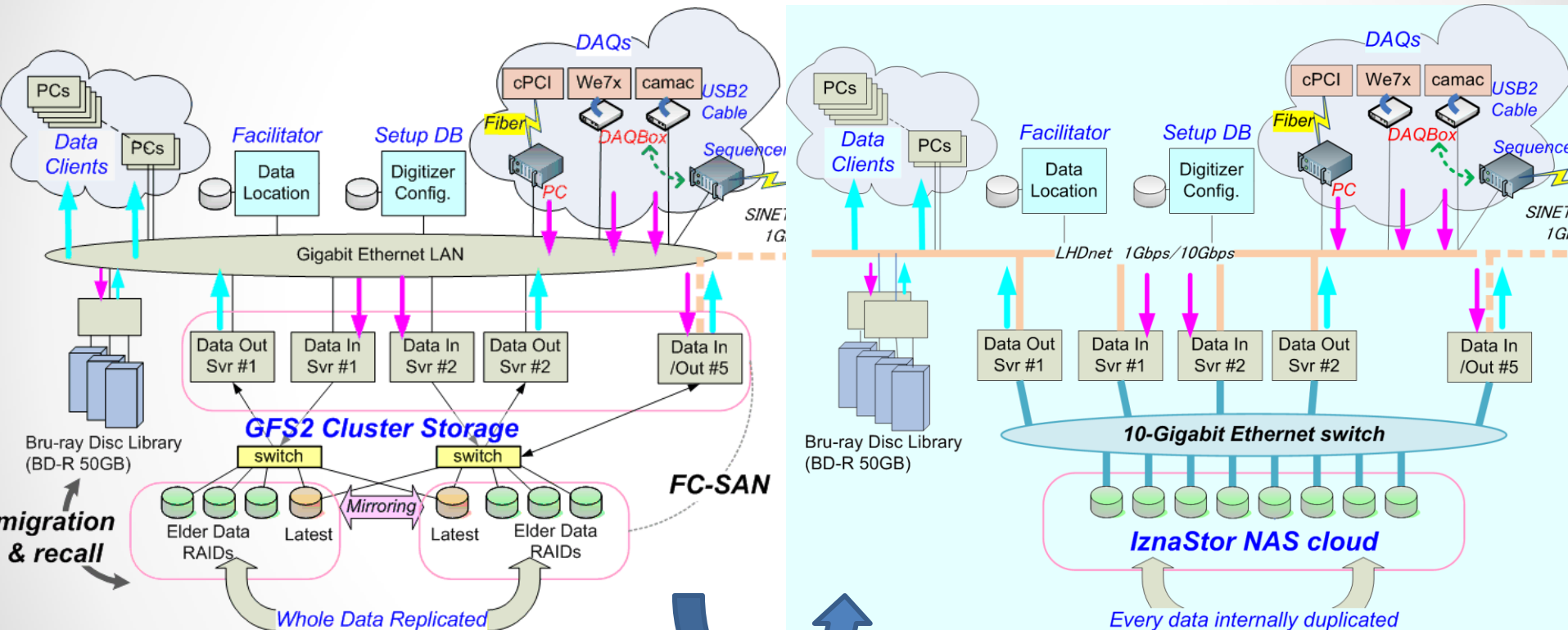
高可用性を実現



データストアのクラウド化

4Gbps FC-SAN/RH GFS2による
ミラー形クラスタストレージ

分散Key-Value Storeによる
10Gbps NASクラウドストレージ



- ノード間が密結合なクラスタ構成
- 拡張性、保守性の面で難

- クラウド技術による高可用性
- スケールアウトを容易に

Recent upgrade → “GlusterFS” distributed FS

- Background

- ✓ The prior “cloud storage” was good to scale-out the capacity easily.
- ✓ When a node failure happened, however, it took **many days** to recover the lost replicas and re-sync meta-data among all nodes.



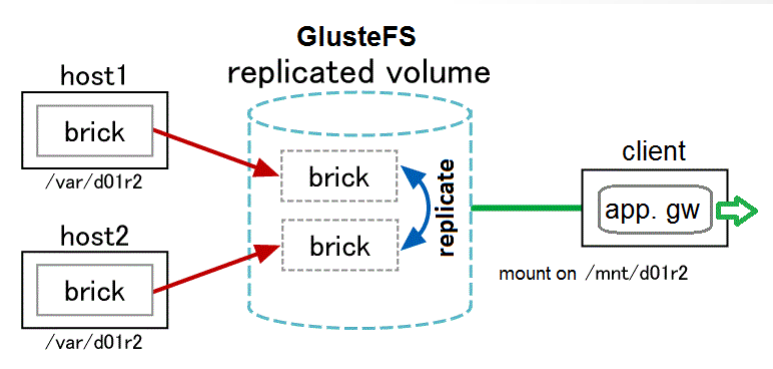
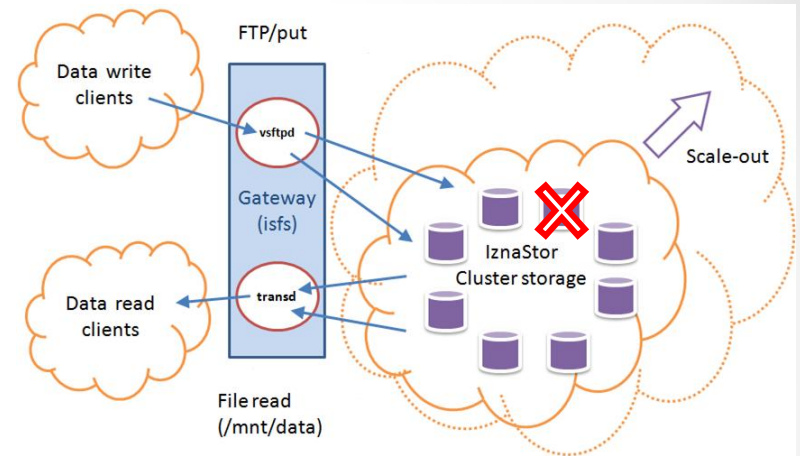
- ✓ **Slow recovery** is a serious problem, which is caused by a large capacity of each RAID volume having tens of TBs nowadays.



- ✓ To speed up the recovery, each storage element had better **have a smaller size** to get replicated again.
→ **No RAID** but a **mirrored pair of 2 HDDs**

- Requirements & Survey

- ✓ Candidate software should provide **faster fault recovery** and still a **high scalability**.
- Open source software: **“GlusterFS”**.



Storage Separation for Faster I/O

- GlusterFS can provide 'distributed', 'replicated', and 'striped' volumes, and their combinations like 'distributed replicated' one.

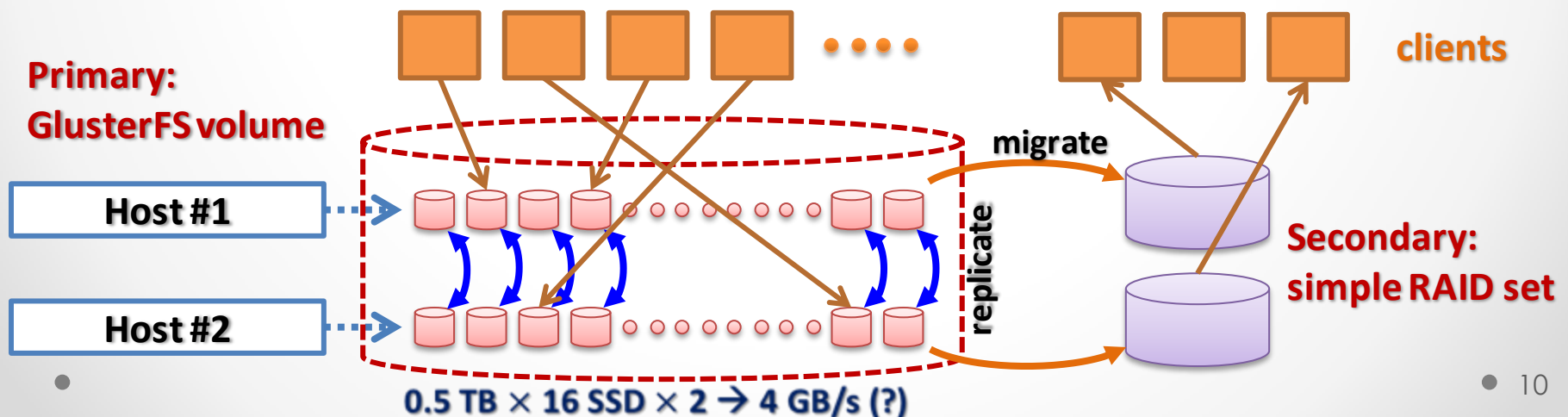
→ For the data safety, replication-based one is preferable.

- GlusterFS is none of the parallel (striping) filesystem so that making a replica may need longer time than writing a single data file.

✓ We need more I/O speed for fast DAQs and steady-state experiments.



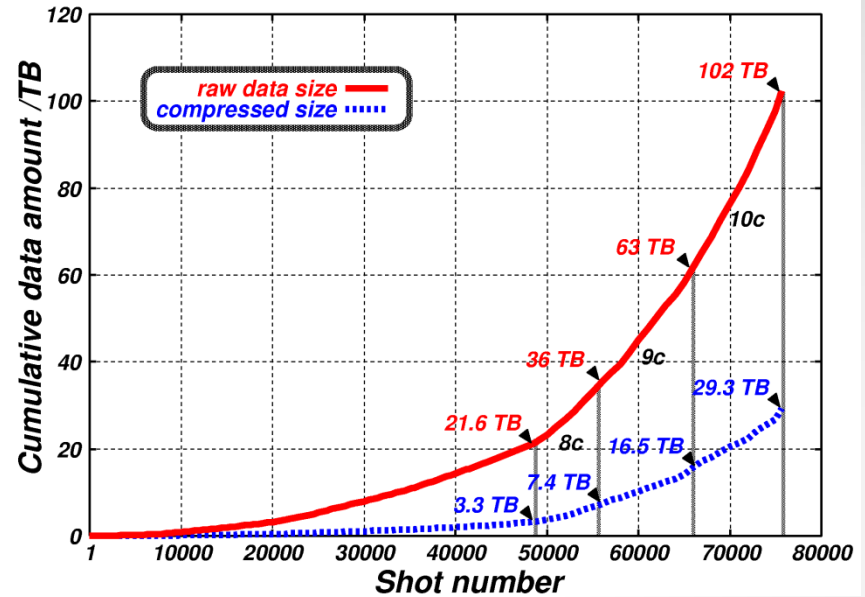
- On top of the archived storage, an additional tier adopting GlusterFS 'distributed replicated' volume has been installed with using many SSDs.



Data Lifecycle Management

- In case of continuously growing data like LHD, **80 %** of data are younger than **3 years old**.
- **Elder data** being accessed less frequently occupies only a small part (**20 %**) of the whole storage.

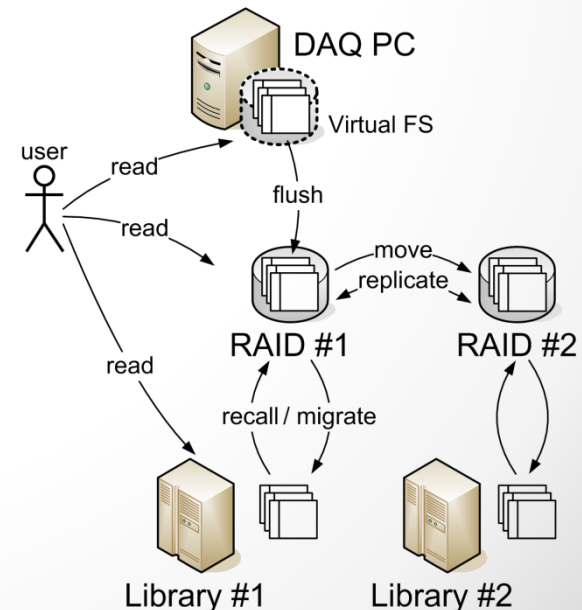
→ No major effect of reducing the storage size



- We have already implemented and tested a **recall mechanism** in LHD. It is based on the access history in indexing database.



- However, we never put it on practical use until now.



まとめと展望

- LHDデータ収集システムでは、大規模並行分散形態をとって、約110種のプラズマ計測に対応しており、同分野内でのデータ収集量世界記録も樹立している。
 - ✓ 分散配備するPCはゼロスピンドルによるメンテナンスフリー化とともに、遠隔電源制御ユニット・ネットワークKVM等を用いて、完全遠隔操作を可能にしている。
 - ✓ 九州大、筑波大などの遠隔サイトでも同システムを遠隔運用、データをリアルタイム共有するとともに、大学側のシステム運用負担の低減にも貢献している。
- データ収集クラウドとストレージの非クラウド化を両立
 - ✓ クラウド技術による多ノード運転の省力化と、非クラウドのシンプル&高速なSSDベースの高速フロントエンド・ストレージの組合せ
- 今後の開発課題
 - ✓ ノードあたり～1GB/sの高速リアルタイムデータ収集： 高速度カメラ, etc.
 - ✓ データ解析のためのワークフロー記述・処理プラットフォーム
 - ✓ ITER遠隔実験に向けた遠距離高速データ伝送と遠隔実験システム
 - ✓ 高放射線環境に耐える計測フロントエンドの検討
 - ✓ インテリジェント・センサー： SoC, FPGA, etc.