Report TEG WLCG Data and Storage Management

Giacinto DONVITO
INFN-IGI

Outlook

- What are WLCG TEGs and why we need them?
  - How they are composed?
- What they deal about?
- How they have worked?
- Reports and Recommendations
- Conclusions and Future

What are WLCG TEGs and why we need them?

- Mandate:
  - To ***reassess the implementation*** of the grid infrastructures that we use ***in the light of the experience*** with LHC data, and technology evolution, but never forgetting the important successes and lessons, and ensuring that any evolution does not disrupt our successful operation.
- The work should:
  - Document a strategy for evolution of the technical implementation of the WLCG distributed computing infrastructure.
  - This strategy should provide a clear statement of needs for WLCG, which can also be used to provide input to any external middleware and infrastructure projects.
- Deliverables:
  - Assessment of the current situation with middleware, operations, and support structure.
  - ***Strategy document setting out a plan and needs for the next 2-5 years***.

How they are composed?

- Several groups – most relevant here are
  - Data Management (chairs: Brian Bockelman, Dirk Duellmann)
  - Storage Management (chairs: Wahid Bhimji, Daniele Bonacorsi)
- Several INFN peoples:
  - Daniele Bonacorsi, Giacinto Donvito, Luca Dell'Agnello, Riccardo Zappi, Vladimir Sapunenko
- DM-SM TEGs started their activity with a survey for all the members where each could provide feedback on the issues and success stories from his/her point of view
- Very detailed survey was asked to the 4 LHC experiments too

What they deal about?

- **Data Management**
  - DM.1 Review of the Data Management demonstrators from summer 2010.
  - DM.2 Dataset management and Data placement (policy-based or dynamic)
  - DM.3 Data federation strategies
  - DM.4 Transfers and WAN access protocols (HTTP, xrootd, gsiftp)
  - DM.5 Data transfer management (FTS)
  - DM.6 Understanding data accessibility and security requirements/ needs
  - DM.7 POOL
  - DM.8 ROOT, Proof
  - DM.9 Namespace management.
  - DM.10 Management of catalogues (LFC, future direcions)

What they deal about?

- **Storage Management**
  - SM.1 Experiment I/O usage patterns
  - SM.2 Requirements and evolution of storage systems
  - SM.3 Separation of archives and disk pools/ caches
  - SM.4 Storage system interfaces to Grid
  - SM.5 Filesystems/protocols (standards?)
  - SM.6 Security/access controls
  - SM.7 Site-run services.

What they deal about?

- Starting since the first meeting it was clear that the two TEGs overlaps in many areas
- Most of the meetings and the activities were carried on together
  - As the Face-to-face meeting held in Amsterdam (Jan 24th-25th)

Process

Information Gathering — **Nov 2011**

- Initial questionnaire
- Defined topics [TopicsDataStorageTEG]
- Soon Data / Storage TEG merged really..
- Questions to experiments:
  Experiment Presentations and Twikis
  [ALICE; ATLAS;CMS;LHCb]

Synthesis / Exploration / Orientation — **Jan 2012: Face-to-face**

- Storage Middleware presentations: 165687

- Face-to-face session for each topic plus broader discussions.

Feb 2012: GDB

"Emerging" recommendations

Refinement

Developed :

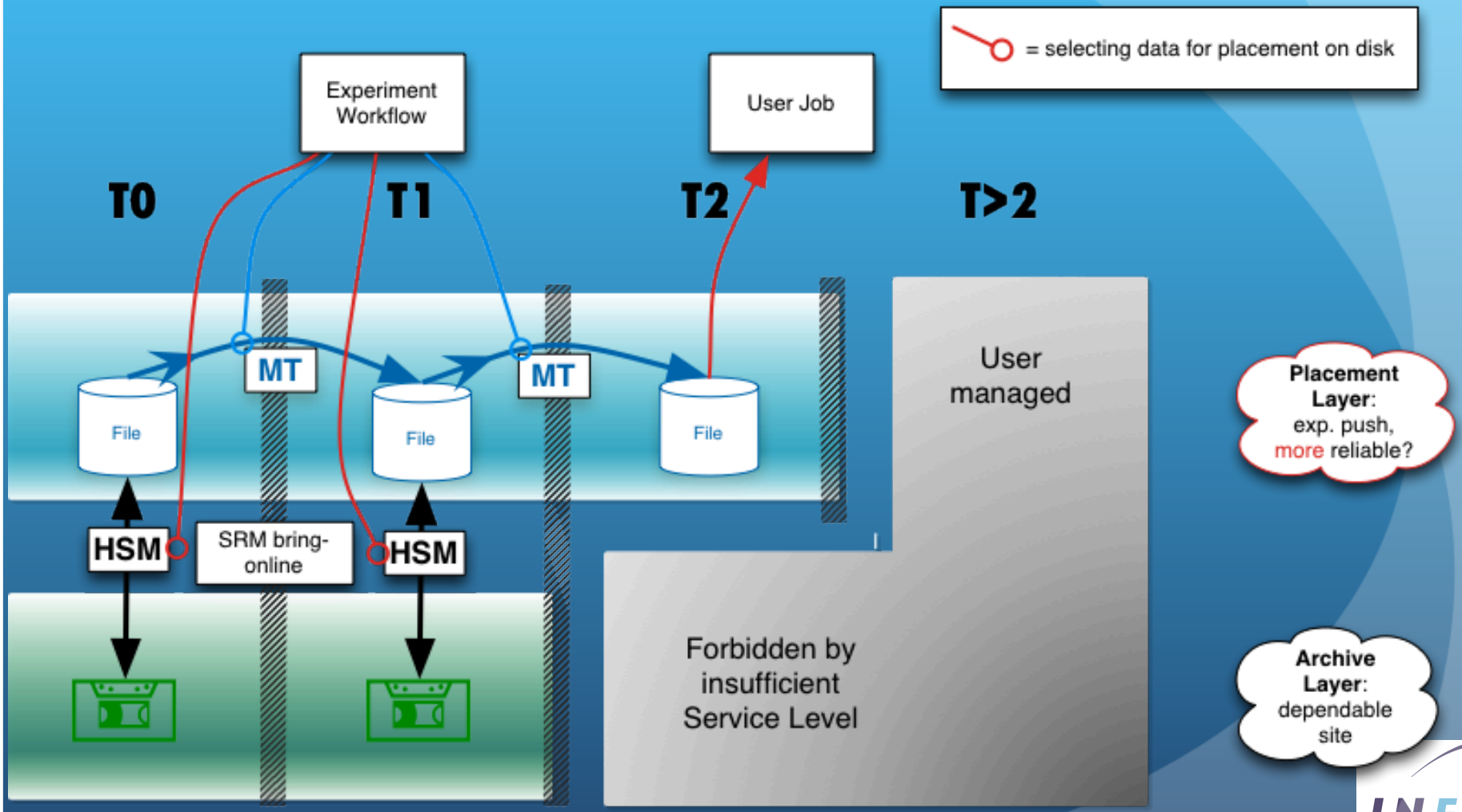
- Layer Diagram: Overarching picture
- Recommendations under each topic

See:
Final and draft report

Apr 2011

Recommendations

Data Placement



Archive Layer

- The archive layer has two different responsibilities:
  - ***Cost-effective scaling*** in storage volume, not client access bandwidth
  - ***Increasing reliability*** by provisioning a separate copy of files on a different storage media to decrease the risk of data loss (due to ***software or operational mistakes***)
- In the future it may happen that the tape as a medium may get progressively replaced by disk-based solutions (reducing latency, but increasing the power budget),
  - But it is not guaranteed, as tape densities are increasing
  - In the long term, the majority of the archive layer will likely move with the consumer market to random-access bulk storage
  - ***We expect that the WLCG storage volume will shrink relative to typical market volumes***

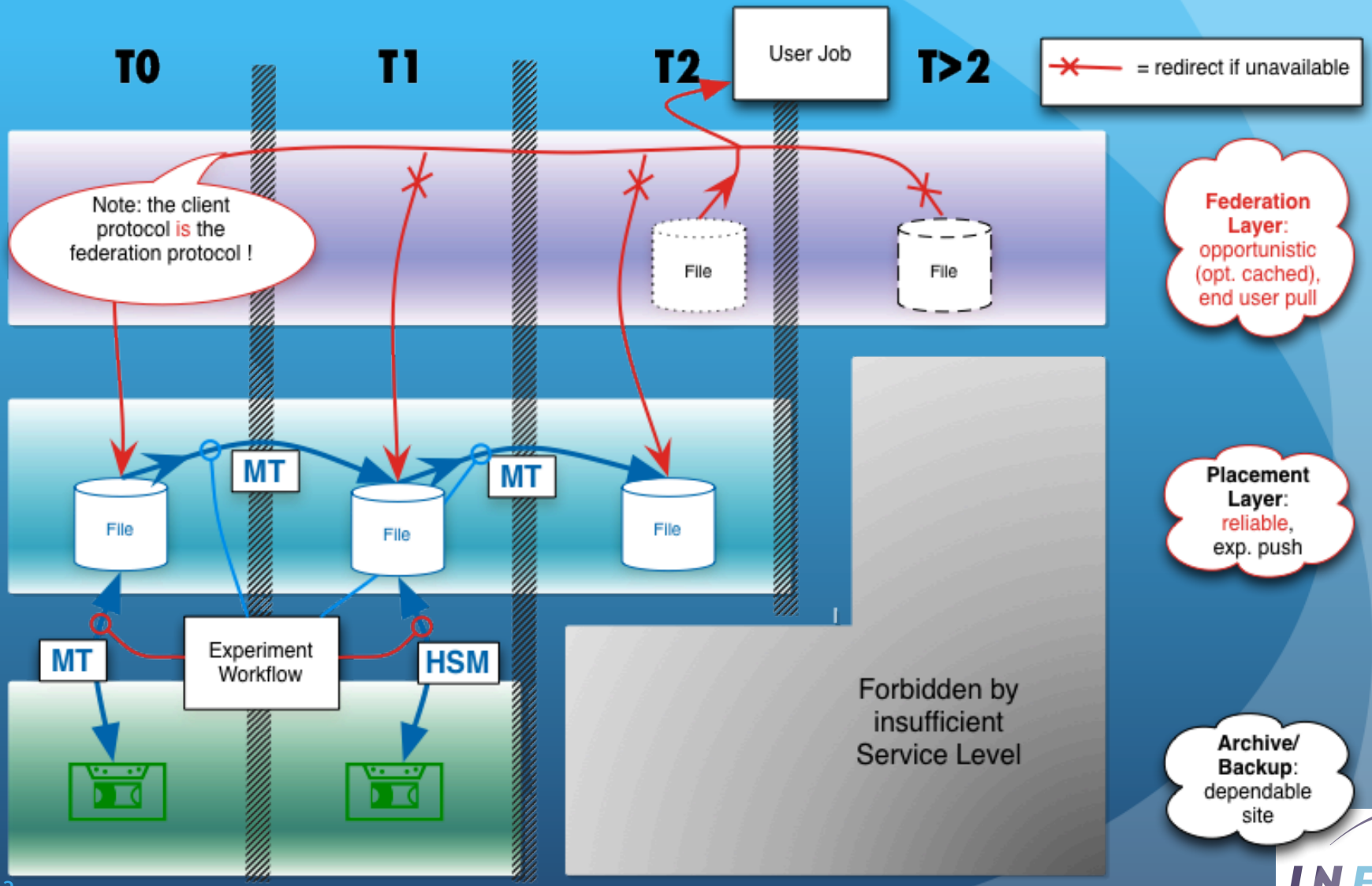
Placement Layer

- ***Using placed data*** has the important benefit of not introducing ***additional transfer latency*** or wide-area network (WAN) activity for jobs that process this data.
- ***Predicting*** which ***datasets*** will be ***popular*** is a non trivial task that has resulted initially in an inefficient use of the available storage space and network capacity.
- Therefore, experiments have introduced an active ***monitoring system***, which collects and aggregates information on ***file popularity***
- Traditionally, the disk pools on the placement layer have been integrated via a hierarchical storage management system (HSM) with local archive resources to transparently manage the files available on disk
- The WLCG usage pattern involves ***data scans***, which are well-known to ***perform poorly on LRU caches***

Placement Layer

- At this point in time, all experiments can work with both scenarios - HSM coupled storage or split disk and archive components
- ***some experiments have a preference to extend the split architecture*** to become a strategic direction over the next years

Placement with Federation



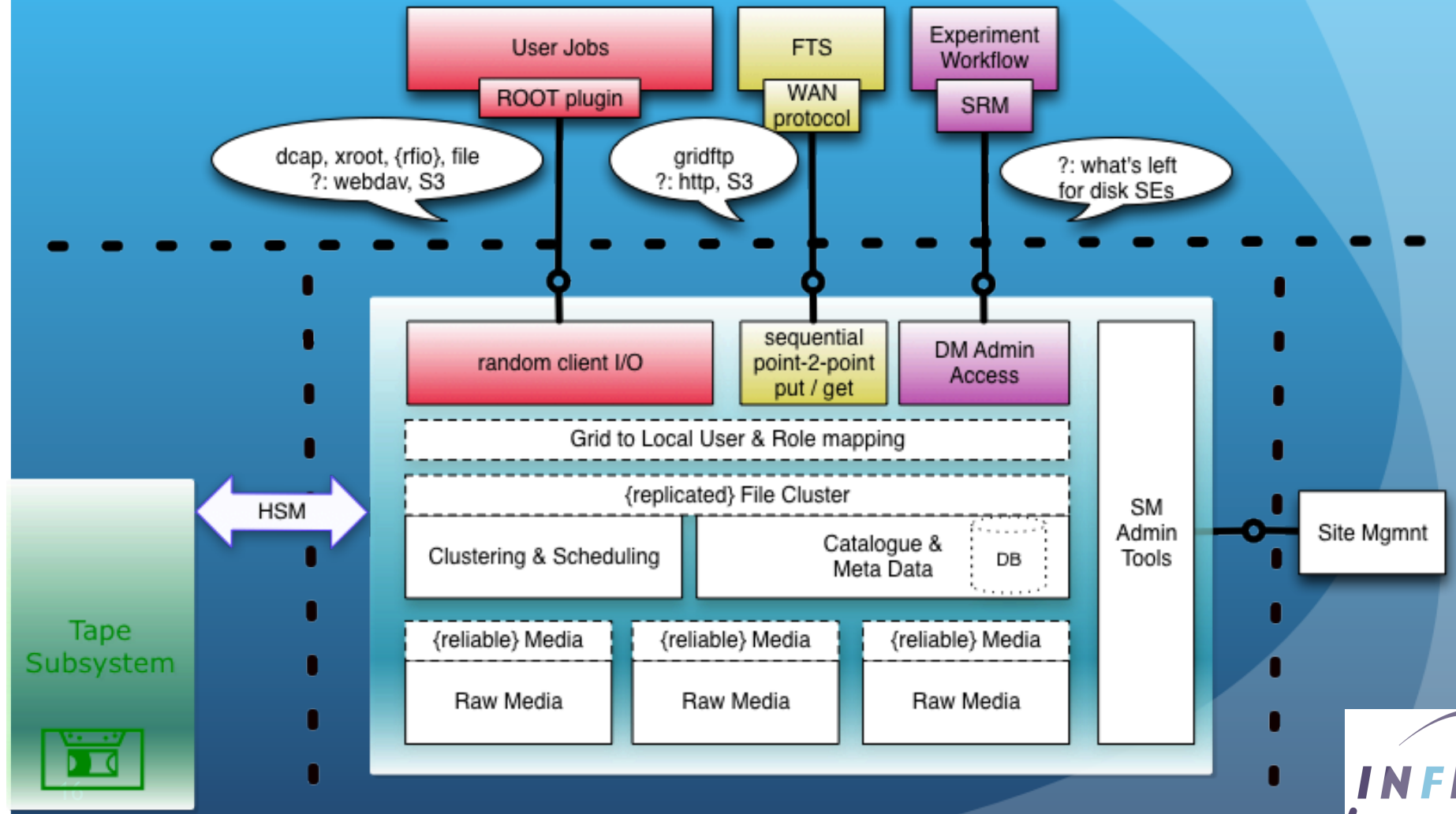
Federation Layer

- The pre-placement data movement has been more recently complemented by the federation approach
- the ***redirection capabilities of the client*** access protocol are used to increase data availability beyond the level a single site can provide
- Requirements:
  - A ***common client access protocol***
  - A deterministic mapping from site-local file names for replicas to the experiment's namespace
  - Federated data is read-only and replicas of a file are identical
- To hide ***local unavailability*** due to service problems for a small subset of jobs, which would otherwise fail and be resubmitted

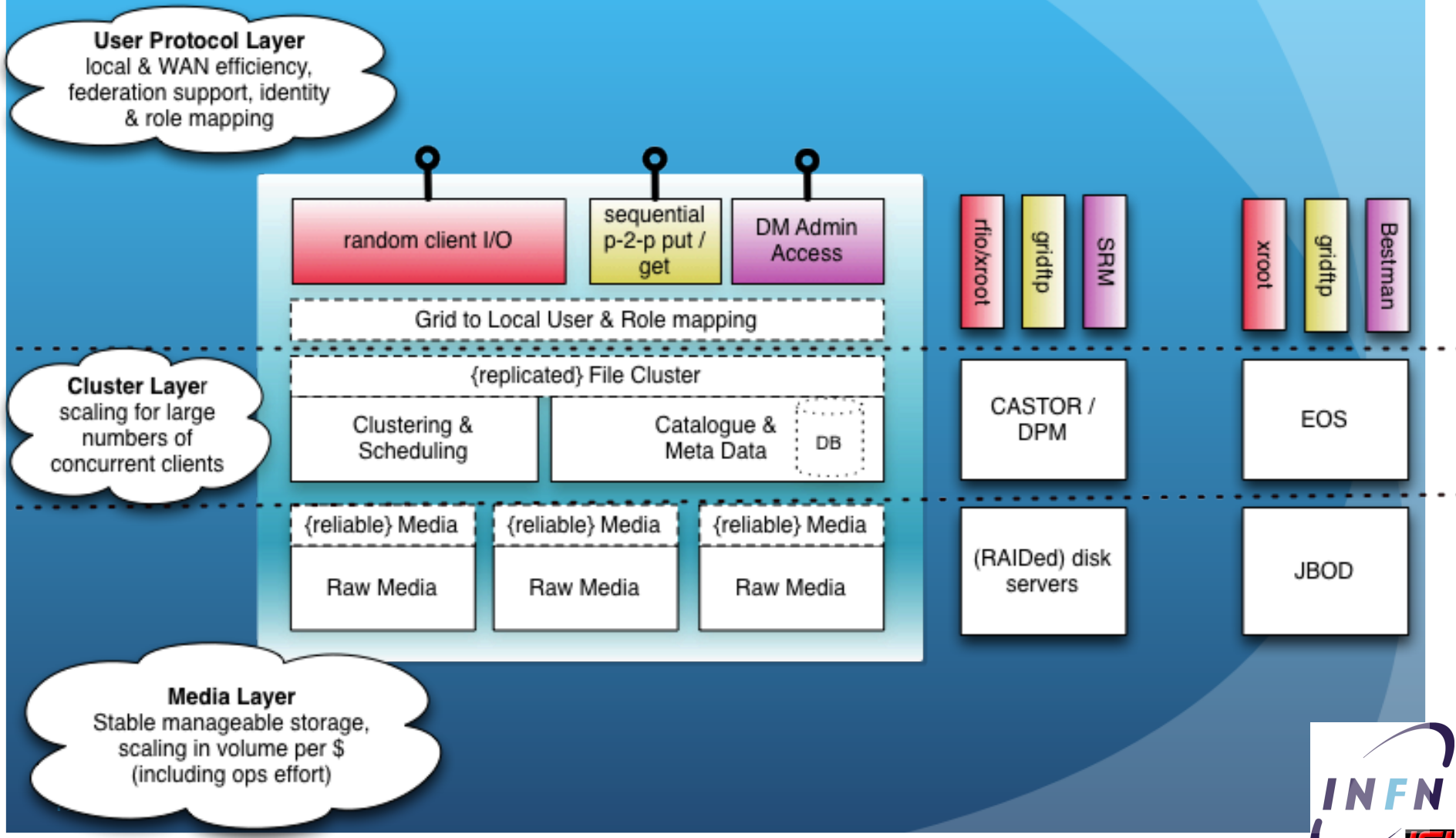
Recommendation Federation

- Focused work on an ***http plugin for xrootd***
- Establish a ***monitoring*** of the ***aggregate*** network ***bandwidth*** used via federation mechanisms
- Launch and keep alive storage working groups to follow up a list of technical topics
  - Detailing the process of publishing new data into the read-only placement layer
  - Investigating a more strict separation of read-only and read-write data
  - Feasibility of moving a significant fraction of the current (read-only) data to world readable access
  - Investigating federation as repair mechanism of placed data

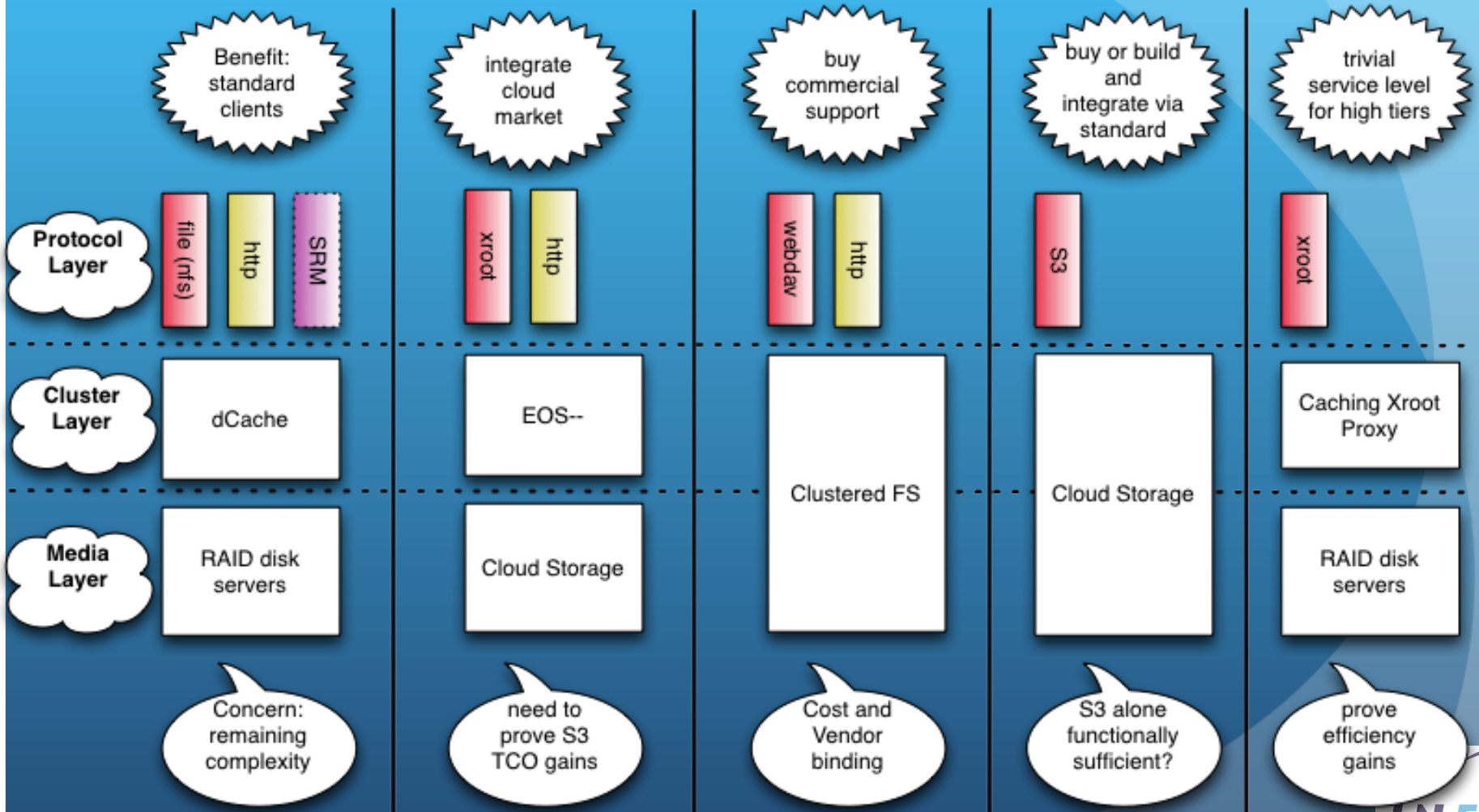
Storage Element Components



Examples of current SE's



Examples of (possible) future SE's



Catalogues and Namespaces

- We observe that ***each LHC experiment*** has implemented ***its own cataloguing software*** for the dataset namespace
- For the file namespace, ALICE and CMS again have unique, internally-developed cataloguing software. However, LHCb and ATLAS currently share a common piece of software ***LFC***, supported by EMI
  - both plan to ***remove their use of this software*** in the medium term.
- The WLCG should plan for the LFC to become experiment-specific software, then eventually unused as an experiment catalogue in the medium-term. Particularly, we should advise for EMI (and subsequent projects) of this fact.
  - In the meantime, maintenance will likely be needed.
  - We believe the ***LFC may be repurposed*** by subsequent EMI projects; i.e., ***as a central redirector*** for a federation system.

Archive/Disk Separation

- All of the LHC experiments seem to be working fine with (or towards) ***splitting disk caches from tape archives***. ALICE, ATLAS and LHCb are split, while CMS has a work plan in progress.
- The experiments require a separation between archives and disk caches.
- At a first glance, ***FTS seems to provide the needed functionalities*** to transfer data from the disk cache to the disk buffer in front of the tape system. We hope to verify this in the next version, but thought should be given to if FTS is the most appropriate tool for scheduling, or if a different concept or architecture should be developed.
- ***None*** of the experiments ***want to "drop"*** the useful functionality of ***HSM*** in managing a disk buffer in front of the archive
- Management of archive storage in the model described potentially moves from within a single storage system and involves the transfer layer. Experiment workflows may need to be adapted accordingly and so tools such as FTS should support the required features as in the MT recommendations

Managed Transfer

- Flexible ***integration with the experiment workflow*** management systems, supporting different scheduling strategies
- Ability to manage a large number of endpoints, including the management of ***fair-share and priorities***.
- Management of the ***staging process*** to move data from the Archive to the Placement Layer.
- ***Fault tolerant behavior***, resuming interrupted transfers, retries and the use of replicas (in case source files are not available, but replicas on other sites exist).
- The ability to handle back-pressure.
- Support ***for sites not providing an SRM*** interface.
- Support for additional transfer protocols, such as ***HTTP and xrootd***.
- Detailed monitoring of transfers to allow optimization of storage endpoints and networks.

Storage Management Interfaces

- Not all storage is manageable through SRM – particularly storage implementations ***outside the HEP niche*** do not integrate SRM.
- Not all Storage Element implementations provide the ***complete*** implementation of SRM specification.
- Not all of the SRM v2.2 specification has proved useful
- Standards also exist that provide comparable levels of functionality: CDMI is an emerging standard interface for managing storage and, in addition, WebDAV provides many of the functionality of SRM
- Maintain ***SRM at archive sites***
- Experiments, middleware experts and sites should agree on ***alternatives to be considered*** for testing and deployment

Site Storage Performance

- ***Benchmarking and I/O requirement gathering***
- Protocol support and evolution: Both remote I/O (direct reading from local storage) and streaming of the file (copy to WN) should be supported in the short/medium term. The trend to ***move towards remote I/O*** should be encouraged by both experiments and storage solution providers
- I/O error management and resilience
- ***Future technology review***
- High-throughput computing research: Possibilities for much higher throughput computing should be investigated. ***This research should not be restricted to ROOT data structures*** and should fully utilize cutting edge industry technologies, such as Hadoop data processing or successors, building on existing exploration activity

Site Storage Operations

- Site involvement in protocol and requirement evolution
- ***Expectations on data availability and access and the handling of data losses***
  - Where additional reliability is required, ***sites shall use 'smart' techniques and redundancy on block or file level*** to setup and operate this storage
- Improved activity monitoring:
  - ***Monitoring of files accesses, access frequency***
  - Catalogue level monitoring
- ***Storage accounting***. We recommend that WLCG agrees to use the EMI StAR accounting record.

POOL persistency

- POOL has become experiment-specific software, and will become unnecessary in the medium-term. ***No future development is foreseen***.

Security

- ***Remove backdoors from CASTOR***
- Check actual permissions implemented by storage systems
- ***Resolve issues raised with data ownership***

Conclusions and Future

- Reports have been delivered
  - https://espace.cern.ch/WLCG-document-repository/Boards/MB under "Technical Evolution Strategy" folder
- Goal now is to have initial summary for WLCG workshop at CHEP
- At CHEP:
  - Summary of important recommendations – and initial proposal of priorities
  - Summary of areas where open questions have not been (fully) addressed
  - Summary of areas where more work needs to happen, or discussions need to finish
  - Initial proposal of working groups that WLCG should set up (for GDB and pre-GDB slots)
  - Proposal for those general topics that could be dealt with at HEPiX (for example)