基于Rucio的高能物理网格数据管理

张玄同，张晓梅

高能物理研究所

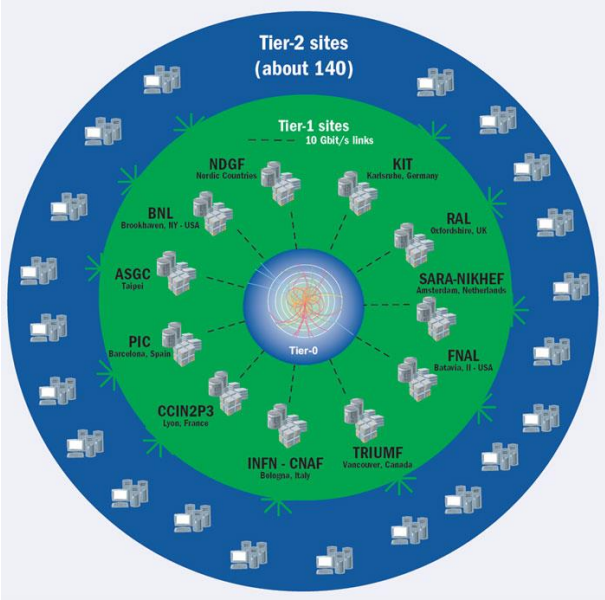
背景：国际网格数据管理现状

LHC国际网格(WLCG)，

- 联接全球计算资源，存储、分发和分析由LHC实验产生的数据。
- 欧洲核子研究组织（CERN）主导，40多个国家的170多家计算中心参与建设。
- 为各种不同的实验提供分布式计算和存储提供基础设施和标准解决方案。

国内WLCG应用现状，

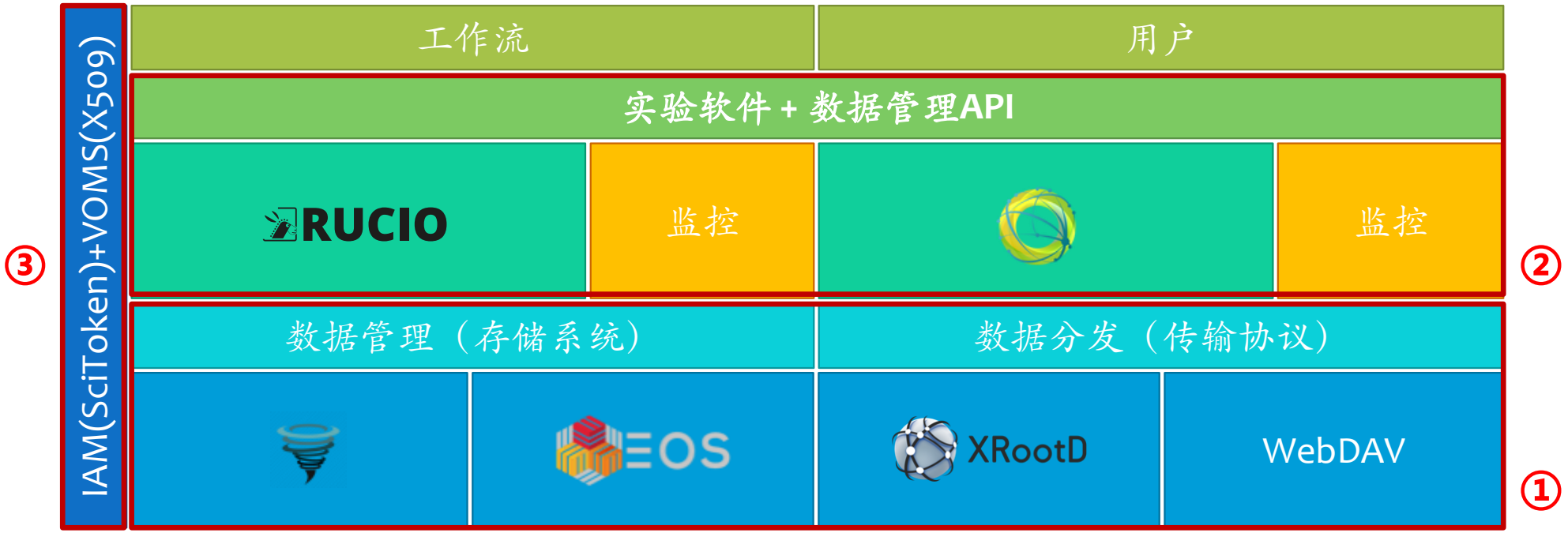
- BESIII，
  - 北京谱仪，采用DIRAC系统管理来自欧洲和亚洲的站点资源，低活跃。
- JUNO，
  - 位于江门的中微子观测站，5个主要计算和数据站点，高活跃。
  - 网格应用负责产生和分发每年2.4 PB原始数据，0.6 PB模拟数据。
- HERD，
  - 建设中的中国空间站上高能宇宙射线观测站，主要为中国-意大利站点组成，高活跃。
  - 负责分发每年~1 PB数据。
- CEPC，
  - 计划中的大型环形正负电子对撞机，正在设计阶段，低活跃。



背景：网格数据基础设施组成

网格数据基础设施组件：

① 数据存储层和数据协议层：站点数据的保存和站点数据统一URI。
② 数据管理系统和传输系统层：多站点数据命名空间和站点间数据分发管理。
③ 认证授权服务管理层：多站点用户授权管理。



背景：网格技术新变化

高能物理网格数据特点，也是需要解决的问题，

- 存储系统的**异构性**，
- **认证**方式和用户身份**差异性**，
- 数据**密集**、传输任务**频繁且持续**，
- 数据**副本分发**和保存策略。

网格技术基础环境的变化推动了网格的进化和拓展，

- **物理学家需求在变化，**
  - 数据访问和物理分析软件的直接集成——基于Root软件的XRootD协议的发展，
  - Web端在线数据访问和处理服务——SWAN和CERNBox的发展。
- **计算机硬件水平变化，**
  - 数据站点存储数据量扩张——Rucio等~100 PB、~EB级别数据管理软件的开发和应用，
  - 计算站点大规模计算资源调度——HTCondor-CE的发展。
- **工业界软件技术变化，**
  - 用户授权模型的进步——SciTokens授权方式取代GSI证书模式，
  - 云计算云存储的广泛使用——通用数据访问协议S3+WebDAV的发展。

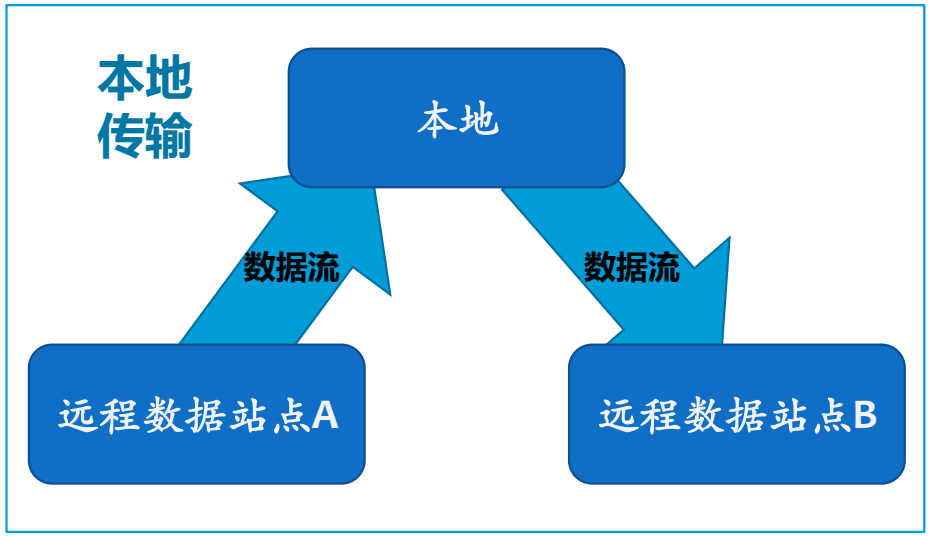
随着新技术和网格基础设施的充分结合，形成了高能物理独特的国际资源互联共享框架。

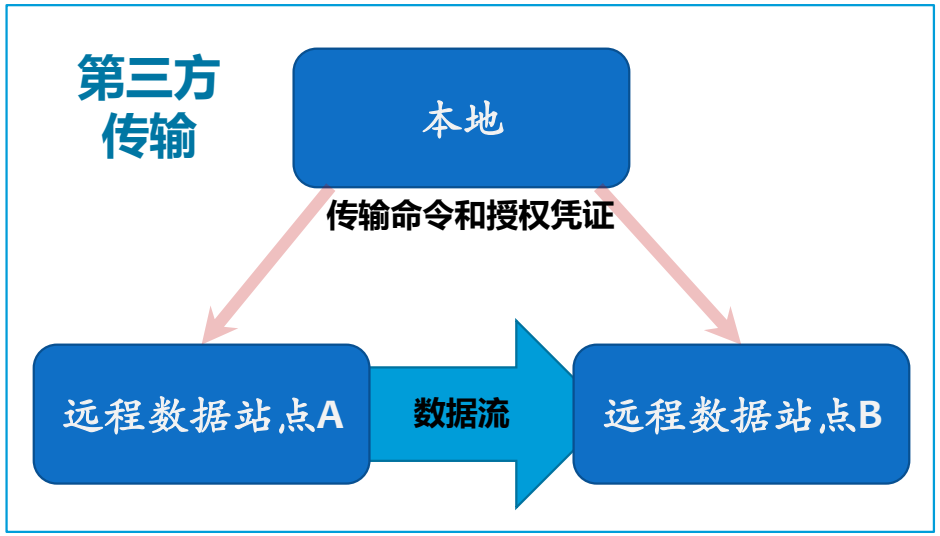
新技术：第三方传输协议

解决网格授权模型下数据在远程站点间直接传输的需求，提高传输效率和数据安全。

- 用户发送传输命令并携带授权凭证到传输的两端站点，站点间自行建立传输。
- 第三方传输（Third-party-copy，TPC）需要的基础设施，
  - 双方数据站点都支持相同的数据协议，
  - 双方站点承认相同的授权凭证，
  - 数据协议支持第三方传输。
- 支持TPC的协议：XRootD、WebDAV。

XRootD DAV

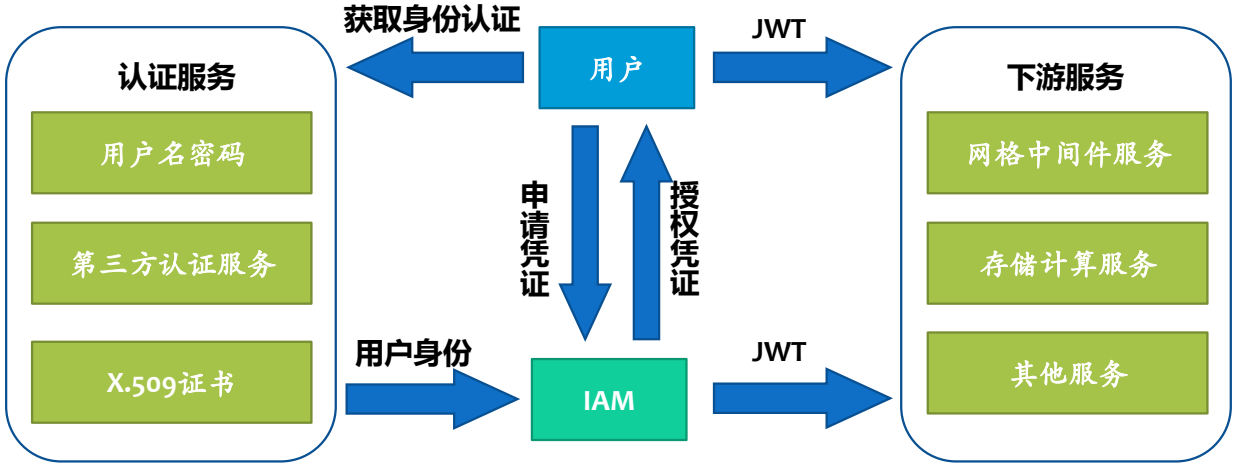




新技术：身份认证授权系统Indigo-IAM

身份授权管理系统（IAM），WLCG下一代网格资源授权服务，

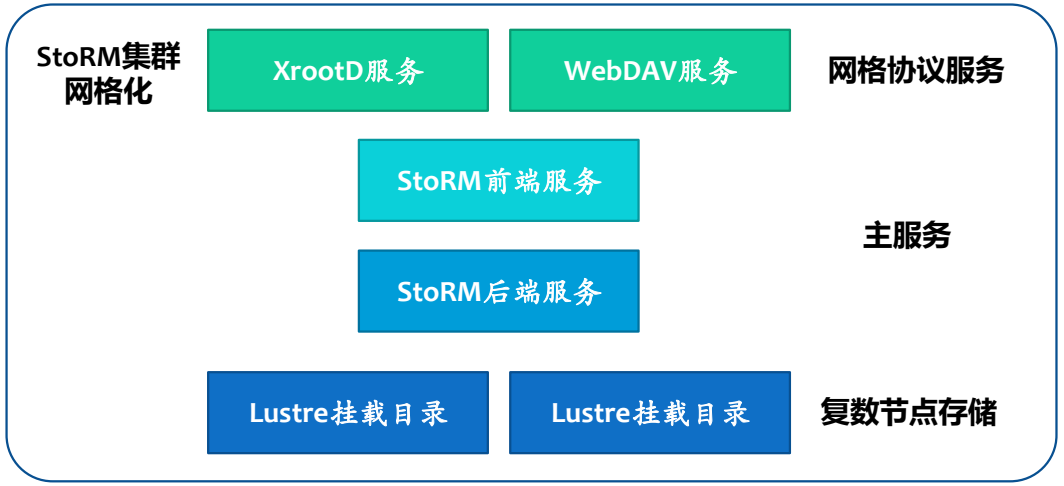
- 2000年早期，X.509是唯一成熟的安全方案，但现在基于JWT的OIDC授权模型逐渐流行。
- 通用授权服务，**与工业界认证授权模型接轨，降低数据访问和接入的门槛。**
- 基于开源的MetriID开发，扩展X509-VOMS凭证授权服务，为多类型下游服务提供类似SSO的授权服务。
- 用户身份验证，
  - 用户注册IAM账号后，可关联多认证服务，包括，
    - 用户名密码，SAML，OpenID，X.509等。
- 身份管理，
  - 继承自VOMS的用户角色(role)，通常根据用户在某个VO内所处的组进行区分，如，
    - Role=Production, Role=Computing等。
- 授权凭证，
  - 支持OIDC Token和VOMS，Token使用JSON Web Tokens(JWT)格式，包括，
    - OpenID提供的标准用户信息：username，email等，
    - WLCG Scope，包括Role和网格资源授权信息：storage.read:/, storage.write:/, compute.create, compute.cancel等。
  - JWT内WLCG相关Scope格式由WLCG与资源开发组协调后约定。
- 下游服务获取和刷新凭证，
  - 与普通OIDC服务一致，需要作为SP服务注册在IAM内以便获取自动授权。

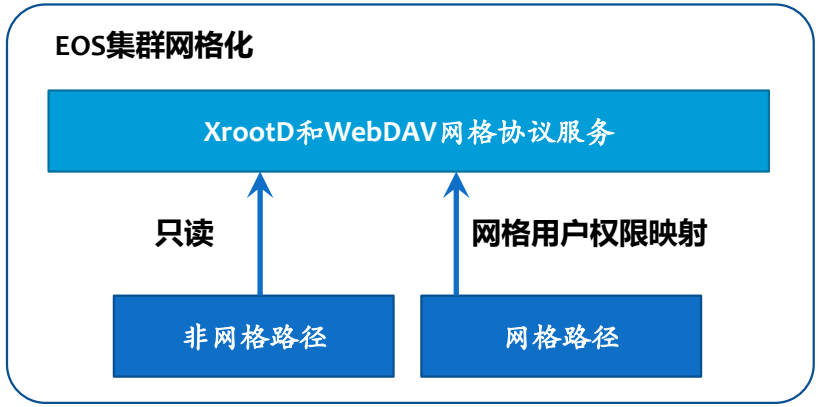


新技术：存储系统网格化技术

底层存储系统为网格数据提供2种存储模式：

1. **先符合POSIX标准，再支持网格协议：**
   - Linux磁盘文件系统+协议，
     - 小站点常见模式，适合T3级别站点单存储节点方案。
   - Lustre+StoRM协议层，
     - T2-T3级别中小型节点常见模式，可扩展多存储节点。
     - 优点：和本地用户结合紧密，也可以存储网格数据。
     - 缺点：存储和协议分割，存储节点继续扩大的情况下，网格协议性能下降。
2. **直接支持（原生支持）网格协议：**
   - DPM，
     - Disk Pool Manager，WLCG为LHC网格站点开发的文件系统，将于2024年停止支持。
   - dCache
     - DESY为LHC实验T1站点开发的开源分布式文件系统，
     - 除了原生底层，还支持连接到其他三级存储系统，
     - 嵌入XrootD、WebDAV、GridFtp等网格协议。
   - EOS
     - CERN为LHC Run2开发的开源分布式文件系统，支持EB量级文件管理。
     - 深度整合XrootD协议，与高能物理数据结合更好。





特性：Rucio系统介绍

Rucio系统设计目的，

- 核心服务为数据密集型科学合作组的**异构存储站点的统一数据分发和数据管理**。
- 最初为ATLAS实验设计，现已推广应用在高能物理、天文、生物等多个学科的科学数据管理上。

Rucio系统特性，相比于传统高能物理数据管理系统。

- 服务的高伸缩性，
  - 所有功能均为模块化容器化设计，可以通过打开或关闭相关容器增加或缩减相关服务的功能。
  - 从主服务到客户端都支持docker/k8s的容器化部署。
- 组件高可扩展性，
  - 允许用户自行定制数据命名空间、数据分发策略、数据管理策略等。
  - 开源、基于Python开发并提供丰富的API接口，满足实验软件开发需求。
- 后台异步数据处理设计，支持海量数据检索和传输，**~billion级别文件搜索**、**EB级别数据管理**、**PB级别数据传输管理**。
- **规则订阅式数据管理**，用户通过数据订阅和创建接近**自然语言**的规则，表达数据管理指令。

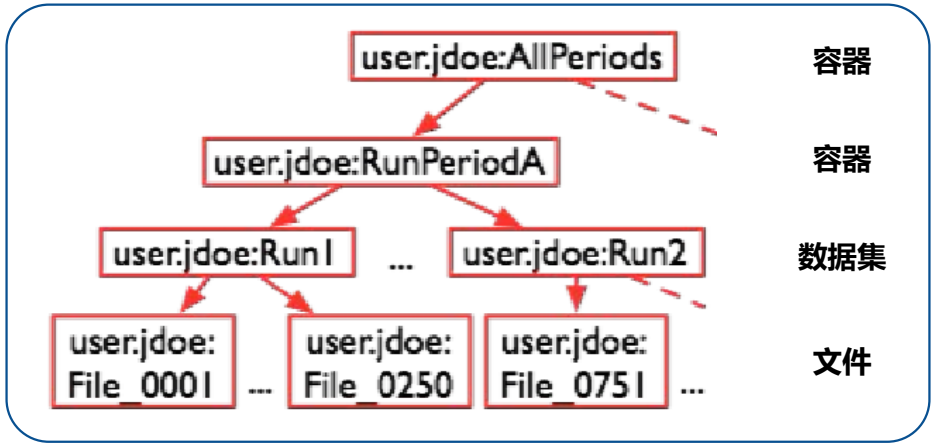
特性：数据命名空间

数据身份（Data Identifier，DID）

Detsim.calib.2023:detsim_run23.root

Scope : Name

- **DID用来命名，**
  - 文件，数据集（包含文件），容器（包含数据集和其他容器）。
- **DID规则，**
  - 每个DID都是唯一的，独占的，
  - 每个文件DID对应一个网格文件和其在所有站点上的所有副本，
  - 数据集或容器DID和网格文件没有对应关系，只用作数据管理，
  - DID的命名规则支持用户自定义，不需要遵守目录式文件结构规则。
- **DID元数据规则，**
  - 文件DID默认元数据包含文件副本状态、数据可用可见状态等。
  - 数据集和容器DID默认元数据包含集合本身可否增减文件状态、集合内所有文件副本状态等。
  - 支持用户自定义DID元数据，
  - 支持元数据检索。



特性：数据分发管理

Rucio的数据分发基于数据订阅和订阅后数据的副本规则，

- **为数据创建订阅，**
  - 根据DID元数据订阅，
    - 如标记为DataType=RAW的数据作为订阅规则。
  - 根据DID名称Filter订阅，
    - 如DID名称里带RAW的，就会满足filter=RAW，并且进行订阅。
  - 当满足条件的DID被创建，上述订阅会自动触发相关副本规则。
- **为订阅数据创建副本规则，**
  - 副本规则用于触发数据分发过程，使用接近自然语言的方式描述参数。
    - 订阅的数据、需要存储副本的站点描述、副本份数、规则有效时间。
  - 数据加入订阅后由副本规则管理，
    - 如果数据没有被任何规则管理，数据将会被删除，
    - 如果数据被多个规则管理，部分规则的失效不会造成数据的丢失。
    - 订阅后的数据发生部分副本丢失时会触发传输进行修复，全部副本丢失的，会标记不可用。

副本规则举例

- 2 copies of user.alice:myanalysis at `country=US` with 48 hours of lifetime
- 1 copy of user.bob:myoutput at `CERN` until January
- 1 copy of user.carol:testdata at `country=DE&type=tape` with no lifetime

特性：异步数据分发

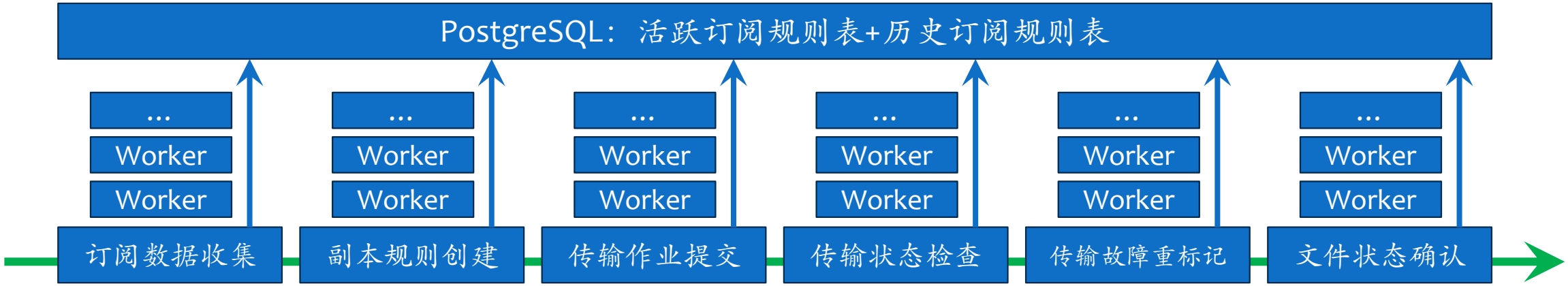
对数据数据分发周期的步骤进行分解，每个步骤状态使用数据库跟踪，实现数据分发的异步执行。

使用数据库跟踪数据规则状态，

- 区分活跃规则表和历史规则表，降低数据库查询压力，
- 默认采用PostgreSQL，保证数据库性能。

步骤状态更新使用守护程序(Daemon Worker)管理，

- 采用心跳(Heart beat)管理每个Worker，
- 守护程序可根据需求自由扩展，
- Worker使用K8s容器化管理。

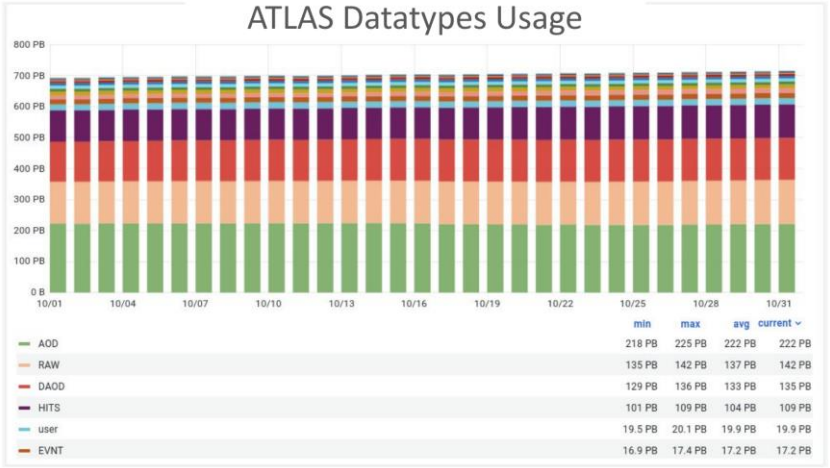


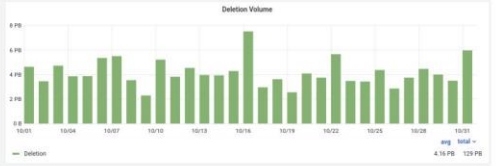
分发数据生命周期举例

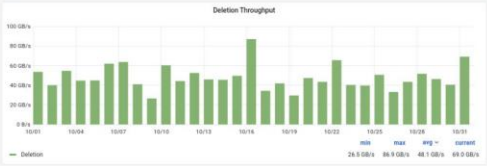
特性：数据管理性能表现

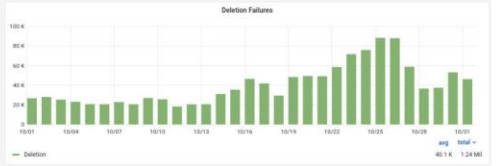
目前最大规模Rucio服务为ATLAS实验:

- 1B+文件, 700+PB数据,
- ~50GB/s数据传输管理，同时~400个服务会话,
- 每日成功传输任务~300Million，每日数据传输量~4PB。









应用：HERD实验文件命名空间

Rucio DID根据实验文件管理需求定制，满足类UNIX的文件路径命名。

- 符合直观文件名定义，文件目录结构和存储站点文件目录结构保持一致。

```
+-----------------------------------------------------------------------------+-------------------+
| SCOPE:NAME                                                                  | [DID TYPE]        |
|-----------------------------------------------------------------------------+-------------------|
| temp:/herd/user/z/zhangxt                                                   | DIDType.CONTAINER |
| temp:/herd/user/z/zhangxt/                                                  | DIDType.DATASET   |
| temp:/herd/user/z/zhangxt/opt/herd/proton-center-E2.7-1_20TeV-34621161.0.root | DIDType.FILE      |
| temp:/herd/user/z/zhangxt/output1-test.g4mac.root                           | DIDType.FILE      |
+-----------------------------------------------------------------------------+-------------------+
```

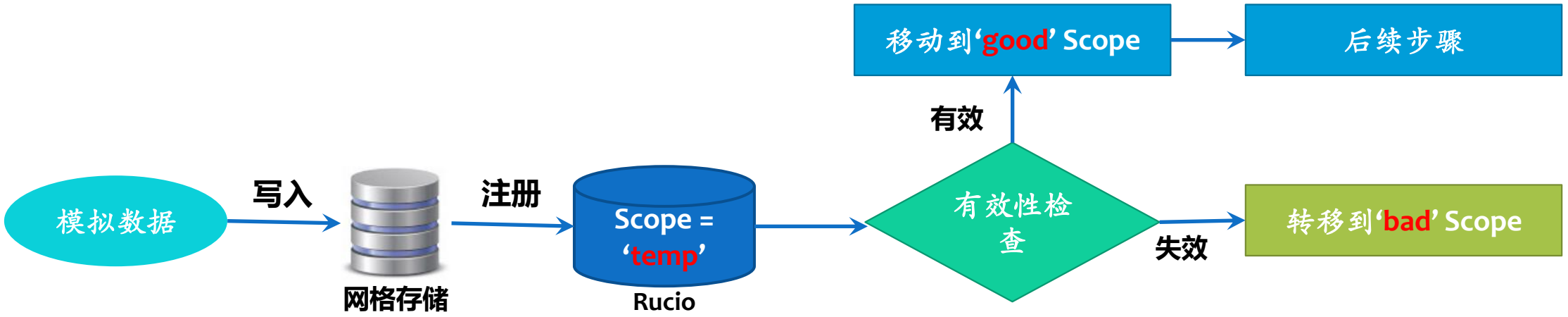
自动创建以目录路径命名的数据集和容器。

Rucio DID	HERD实验Rucio DID策略
Name	类UNIX的文件路径
Scope	定义为工作流数据状态，如temp、valid、corrupt等
Dataset	某个目录下的所有文件，以'/'结尾
Container	某个目录下所有子目录，同Dataset，但不以'/'结尾

应用： HERD实验模拟数据工作流

为HERD实验开发了面向实验软件的Rucio管理接口，

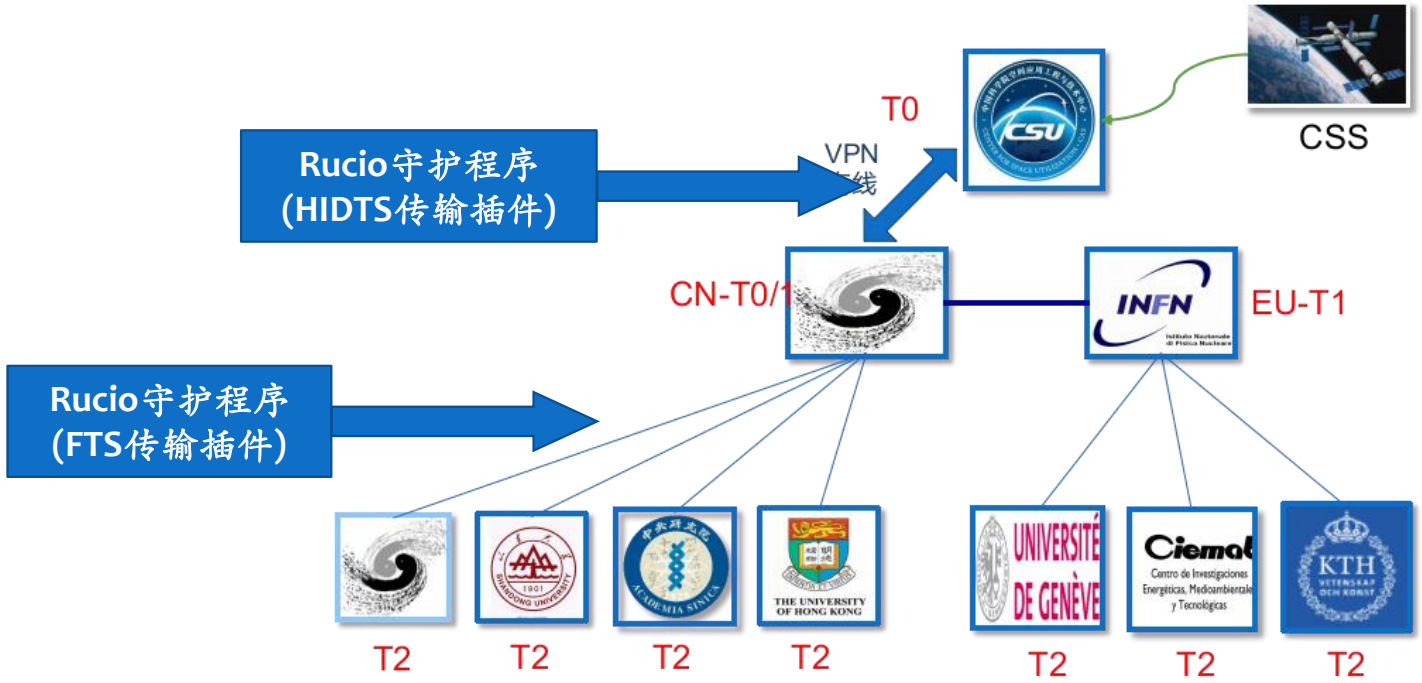
- **Rucio Scope定义为工作流数据状态，**
  - 如temp、valid、corrupt等。
- **HERD实验模拟数据工作流，**
  1. 原始模拟数据上传到存储站点并注册进Rucio并标记为temp scope，
  2. 通过Rucio API获取数据，执行数据有效性检查程序，
  3. 有效则转移数据到valid scope并进行后续步骤，
  4. 失效则转移到corrupt scope，等待进一步检查和删除。



应用： HERD实验原始数据传输

HERD实验计划开发非网格站点传输组件，

- IHEP HIDTS是IHEP存储站点间非网格存储站点传输服务。
- 与网格传输系统FTS类似，可支持通过RESTFUL API的方式提交传输作业。
- HERD实验从CSU到高能所这一段的数据预传输是非网格传输，可以使用此插件进行传输。
- 正在开发中。



应用：基于DIRAC-Rucio整合的分布式计算系统

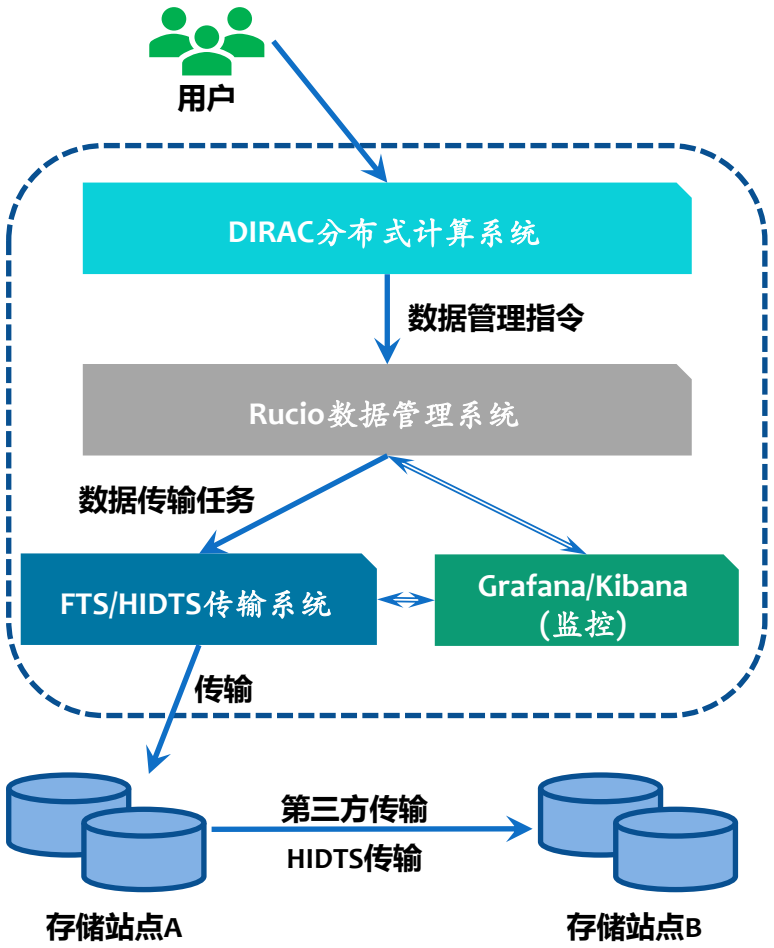
DIRAC系统是一套基于网格的分布式计算系统，

- 提供从作业管理到数据管理的整体解决方案，
- 作业管理使用Pilot作业设计，
  通过在站点启动Pilot作业的方式接管作业提交。
- 也提供功能相对简单的数据管理系统DIRAC-DFC。

在JUNO等采用DIRAC系统的实验中，

- 使用Rucio替换DIRAC-DFC，
- 对Rucio的文件命名策略按照DIRAC-DFC的方式定制，
- 订阅式数据分发和DIRAC的批数据分发系统做功能映射。

测试服务已建立，正在对相应功能进行测试和调整。



应用： 2022年Rucio测试系统运行状况

JUNO数据传输任务：

- IHEP StoRM -> JINR EOS, ~70 TB, ~10 Million文件,
- 因为传输文件都比较小（数MB甚至数KB），最大速度大概~20 MB/s,
- 文件注册速度~90,000 files/s,

Source	Destination	VO	Submitted	Active	Staging	S.Active	Archiving	Finished	Failed	Cancel	Rate (last 1h)	Thr.
srm://storm.ihep.ac.cn	root://eos.jinr.ru	juno	1896662	64	-	-	-	1877	53	24201	97.25 %	8.48 MiB/s
davs://storm.ihep.ac.cn	davs://eos.jinr.ru	juno	1253931	-	-	-	-	4021	-	24099	100.00 %	9.40 MiB/s

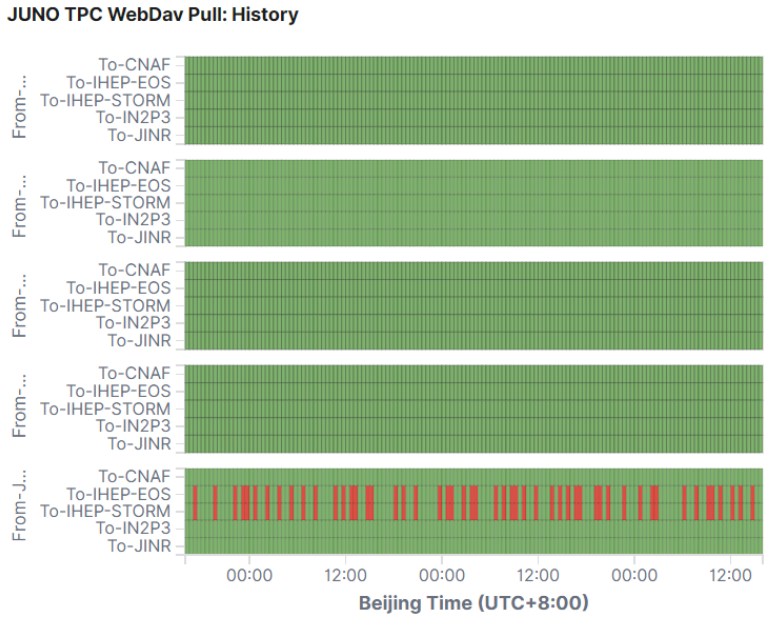
HERD工作流测试任务：

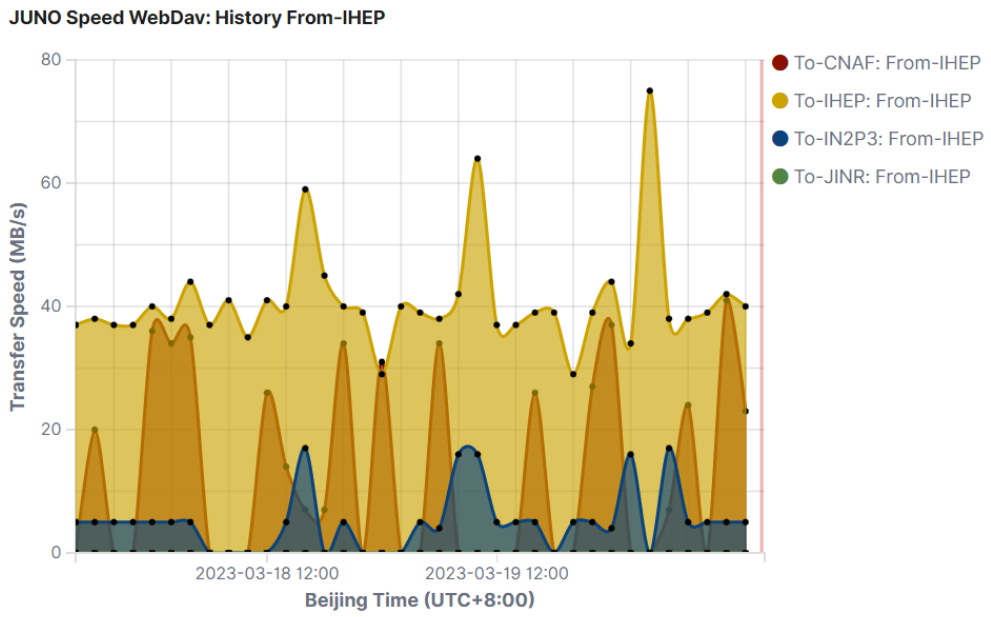
- 已经完成HERD实验工作流RucioAPI软件框架的开发，
- 基本的面向实验软件的数据管理和分发功能已经实现，优化和其他功能开发还在进行。
- 同物理用户和工作流数据库同事紧密合作。

应用：第三方传输性能主动探测监控

基于GFAL传输工具开发的主动Rucio站点第三方传输功能主动探测型监控服务。

- 收集上传、下载、读取、删除等基础功能测试，每30分钟一次。
- 第三方传输，WebDAV/XrootD两个协议，Pull/Push/Streamed等模式传输测试，每30分钟一次。
- 传输速度测试，每个站点之间每20分钟一次。
- 数据使用Elasticsearch收集，使用Kibana监控台展示。





总结

网格在高能物理和IT技术的新需求形势下也发展了很多新组件和新技术。

Rucio系统是未来WLCG上多个实验采用的数据管理系统，目前也已经广泛使用在高能物理领域。

系统具有高扩展性和模块化的特点，支持EB级别大规模数据管理，支持规则订阅式数据分发。

IHEP在应用Rucio系统方面进行了很多尝试，HERD，JUNO实验等都已先后进入测试使用的阶段，目前使用状况良好。

IHEP根据现有系统和服务的特点，开发了多个Rucio组件，面向IHEP主导的实验提供了更多本土化服务。

谢谢！

目录

网格数据协议

网格数据底层系统不统一，网格协议实现异构数据访问和传输。

WLCG网格协议，

- **所有网格文件都有一个独一无二的URI，授权用户可以通过URI访问相应网格文件，无需关注数据所在文件系统类型。**
- **高能物理常用协议：GridFTP、XrootD、WebDAV，**
- **GridFTP，**
  - 网格基础数据访问协议，支持第三方传输，
  - 因为其开源版本停止维护，WLCG已于2022年全面停止使用GridFTP。
- **XrootD，**
  - 基于高能物理数据分析软件ROOT开发的网格数据协议，
  - 深度嵌入ROOT，高能物理用户可以直接在其数据分析程序内通过XrootD协议远程“创读写改”网格文件。
- **WebDAV，**
  - HTTP协议的扩展，云数据访问常用协议之一，
  - 轻量、可靠，支持使用HTTP通用工具访问网格数据，支持第三方传输。

网格数据传输

解决高能物理单文件数据较大断点续传，和不同协议的传输工具不同，学习成本高的问题。

数据传输工具，

- **协议传输工具，同类型协议间传输所用工具，**
  - Xrd(XrootD)，Davix(WebDAV)等。
- **GFAL(Grid File Access Library)，传输中间层，**
  - 为不同网格/云协议提供通用接口的API工具，降低用户访问不同协议时的复杂度。

数据传输系统，

- **FTS(File Transfer System)，**
  - WLCG数据批量传输任务管理系统，管理有限网络资源下的大批量数据传输。
  - 基于GFAL开发，提供高效率数据传输调度管理，优化数据传输时网络使用率。

网格权限认证授权服务

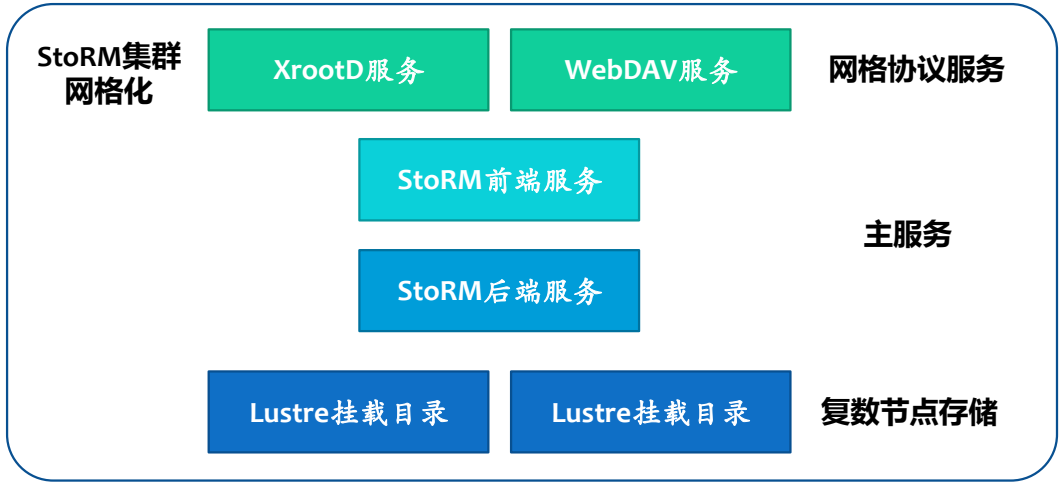
网格提供了多用户环境，需要远程用户与本地用户映射和授权的规则。

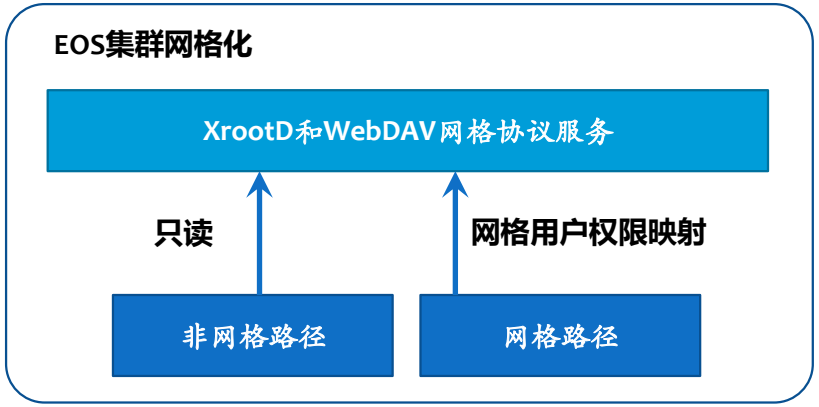
1. **基于VO和X.509证书的认证授权管理，**
   - 网格安全基础结构（GSI），
     - 用户申请X.509证书作为唯一身份证明，
     - 持有X.509加入虚拟组织（VO）并获取资源授权，
     - 使用证书通过VOMS产生限时加密凭证，凭借凭证访问资源。
   - WLCG将于2025年完全放弃GSI，
     - 授权模型仍然足够安全，
     - 但支持这一模型的资源不够多，用户学习成本高。
2. **基于OIDC(OAuth2)的认证授权管理，**
   - OIDC认证授权模型，
     - 用户注册或者通过第三方授权得到账户作为VO身份证明，
     - 访问资源时主动向授权服务器请求身份验证和授权，
     - 用户使用凭证（密码或第三方授权）获取资源授权凭证，
     - 凭证可自动发送到资源或自动刷新。
   - 逐步取代GSI模型，2025年后成为WLCG唯一的授权系统，
     - 支持未来所有类型网格资源授权，
     - 支持云计算等其他类型资源，
     - 更加流行和方便使用。

存储系统网格化

网格提供了两种存储系统网格接入方式，网格中间件接入和原生网格存储系统，

- **网格中间件，**
  1. 存储提供POSIX标准的数据访问方式，
     - 如FUSE等技术实现的文件目录等。
  2. 使用网格协议实现的单节点或中小型复数节点存储站点。
     - XrootD协议集群，
     - StoRM集群。
- **原生网格存储系统，EOS等EB级别数据管理系统，**
  3. 设置网格用户与本地用户授权映射实现站点网格化。
     - 常用VOMS或WLCG Tokens对用户身份进行本地化映射。





底层存储系统

底层存储系统为网格数据提供2种存储模式：

1. **先符合POSIX标准，再支持网格协议：**
   - Linux磁盘文件系统+协议，
     - 小站点常见模式，适合T3级别站点单存储节点方案。
   - Lustre+StoRM协议层，
     - T2-T3级别中小型节点常见模式，可扩展多存储节点。
     - 优点：和本地用户结合紧密，也可以存储网格数据。
     - 缺点：存储和协议分割，存储节点继续扩大的情况下，网格协议性能下降。
2. **直接支持（原生支持）网格协议：**
   - DPM，
     - Disk Pool Manager，WLCG为LHC网格站点开发的文件系统，将于2024年停止支持。
   - dCache
     - DESY为LHC实验T1站点开发的开源分布式文件系统，
     - 除了原生底层，还支持连接到其他三级存储系统，
     - 嵌入XrootD、WebDAV、GridFtp等网格协议。
   - EOS
     - CERN为LHC Run2开发的开源分布式文件系统，支持EB量级文件管理。
     - 深度整合XrootD协议，与高能物理数据结合更好。

组件框架

守护程序组件使用heartbeat用来定期处理：

- 用户订阅的数据的标记和收集，
- 数据状态和数据所处规则对比，对比不同时触发传输，
- 传输任务和传输结果收集。