Data Organization, Management, and Access

Steering Board #5

FEARLESS SCIENCE

$\Delta t = 6$ months

What's new with the DOMA group since last time?

The DOMA Team

The DOMA team is a distributed team working across UIUC, Morgridge, U Chicago, UCSD, UCSC, UNL, UW.

UNL

Plan to start in a few months

UIUC

U Chicago

Morgridge

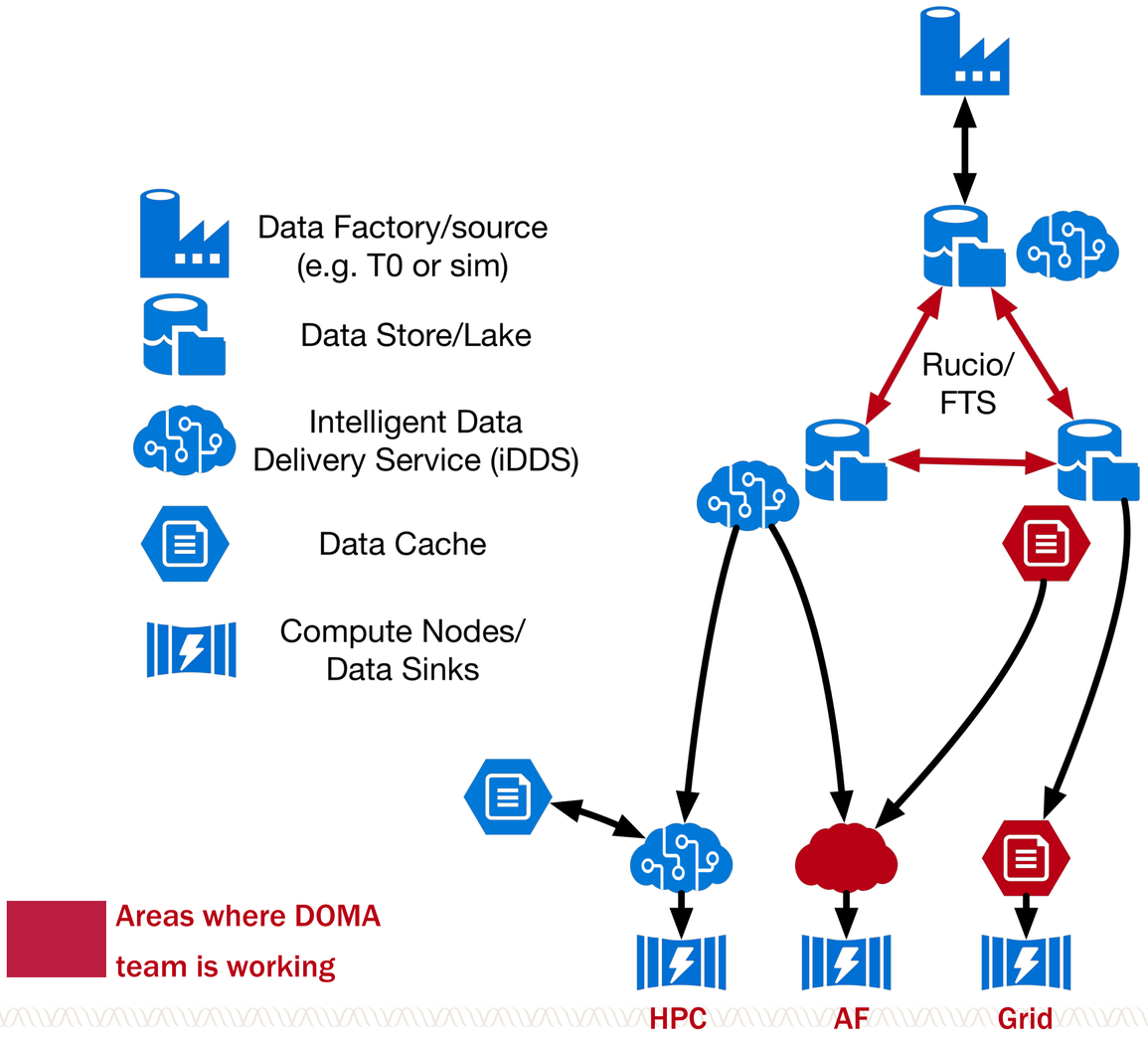
UCSD

Wisconsin

Joined late 2019

UCSC

Unifying DOMA Vision



The DOMA team is working across the following areas:

- **Modeling**: understanding how we utilize the data management systems today and potential impact of changes.
- **Organization**: What data is written to disk and how it is serialized.
- **Management**: Bulk data movement between storage facilities.
- **Access**: how "data sinks" – compute clusters or analysis systems – accesses the data.

The strength of the team is we can have an end-to-end system view!

Community Building

More than just technical projects, DOMA work includes significant external collaborations:

- The international WLCG DOMA activity is the umbrella for WLCG-wide

activities in the DOMA area

  - Of the three active working groups (ACCESS, TPC, QoS), two are co-lead with effort in IRIS-HEP.
  - Through this, we are closely embedded with the U.S. LHC Ops programs.
- Skyhook DM project at UCSC has ties far outside the "traditional HEP' realm. For example, this allows us to interact with groups like the HDF Group, Ceph community.

Data movement with HTTP-TPC and token authorization has built a significant global team to move the WLCG forward - we move up to ~0.5PB / week between about 3 dozen sites.

Many of these ties (particularly UCSC) came out of the conceptualization process leading up to the institute's creation.

19 presentations to the community across the DOMA area.

WLCG Third Party Copy Working Group

The WLCG DOMA TPC working group, co-lead by Bockelman, is making steady progress on an alternate data movement platform for WLCG.

We've been recently working close with the AAI Working Group and put together a WLCG JWT Hackathon.

Objective:

- Demonstrate full-stack HTTP **X509-free data transfer management** with tokens issued by IAM and compliant with the WLCG JWT profile focusing on scope-based authZ
  - Full stack, from Rucio to FTS to storage.

WLCG JWT Hackathon

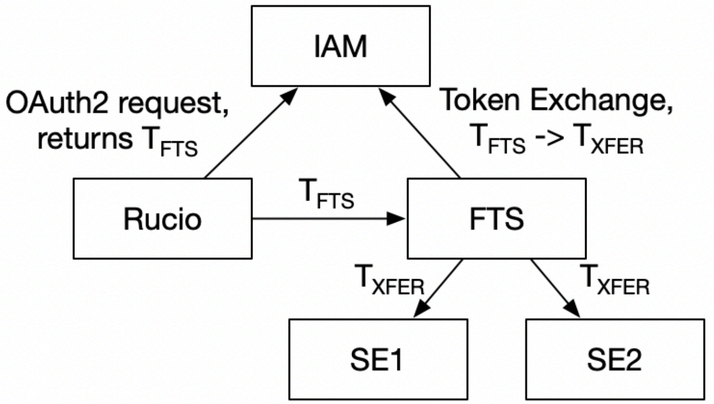
See GDB Update.

WLCG JWT Hackathon

So, how'd we do?

- Working fully token-based transfers managed by FTS against XRootD, StoRM, dCache
- Bug fixes/enhancements in token-based authentication and authorization support in IAM, FTS, XRooTD
- Initial support for WLCG profile and token-based auth in DPM and EOS
- Knowledge exchange! And more … see Google doc.

Every major WLCG storage system has at least a prototype using HTTP-TPC.



Scaling the Third-Party Copy Work

DOMA-TPC smoke test, started 2020-02-14T12:00+0100, took 60:26.

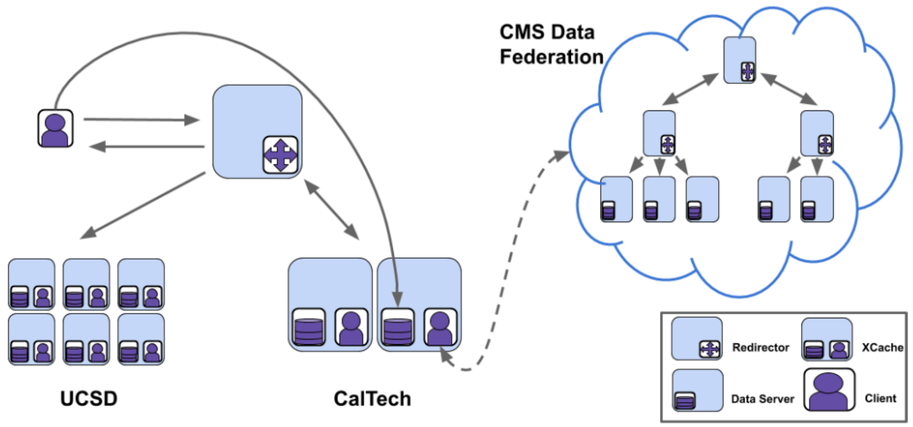
SOUND ENDPOINTS

SCORE	ENDPOINT	SOFTWARE	WORK-AROUNDS
20	AGLT2	dCache	[in 00:56]
20	BEIJING	DPM	[in 03:08]
20	BRUSSELS	dCache	[in 00:22]
20	CALTECH2	xrootd-R/HDFS	[L] [in 02:14]
20	CALTECH	xrootd-D/HDFS	[in 01:48]
20	CERN-DYN-S3	DynaFed/S3	[in 00:44]
20	CERN-TRUNK	DPM	[in 01:16]
20	DESY-DOMA	dCache	[in 00:25]
20	DESY-PROM	dCache	[in 00:24]
20	FNAL	dCache	[in 00:51]
20	IN2P3	dCache	[in 00:22]
20	IN2P3-TEST	dCache	[in 00:28]
20	INFN-T1	StoRM	[in 00:49]
20	KIT	dCache [2]	[in 00:30]
20	LRZ-LMU	dCache	[in 00:45]
20	NDGF	dCache [2]	[in 00:38]
20	NDGF-PREPROD	dCache [2]	[in 00:46]
20	NEBRASKA2	xrootd-R/HDFS	[in 02:45]
20	NEBRASKA	xrootd-D/HDFS	[in 05:51]
20	PIC-PROD	dCache	[in 01:52]
20	PRAGUELCG2	DPM	[in 00:30]
20	PURDUE	xrootd-D/HDFS	[in 00:54]
20	SARA	dCache [2]	[in 02:31]
20	SARA-test	dCache	[in 00:19]
20	TOKYO-LCG2	DPM	[in 01:25]
20	TRIUMF-DYNAFED	DynaFed/S3	[in 01:50]
20	TRIUMF-PPS	dCache	[in 00:54]
20	UKI-BRUNEL	DPM	[in 00:53]
20	UKI-IC	dCache	[in 00:23]
20	UKI-LANCS	DPM	[in 01:11]
20	UKI-MAN	DPM	[in 00:44]
20	UKI-MAN-PROD	DPM	[in 00:40]
20	UKI-QMUL-DEV	StoRM	[in 00:49]
20	UKI-QMUL-PROD	StoRM	[in 00:43]
20	UNI-BONN	xrootd-R/CephFS	[in 00:33]

We have a goal of migrating 30% of the traffic at one U.S. LHC site over to non-GridFTP yet this spring.

- This is now coming into focus (although unlikely to meet the 1 March goal).
- U.S.LHC Sites in the WLCG nightly HTTP-TPC tests:
  - AGLT2
  - Caltech
  - Florida
  - Nebraska
  - Purdue
  - SLAC
- Others are known to function (MWT2, UCSD) but aren't in the WLCG tests.
- Conspicuously missing are the U.S. LHC Tier-1 sites (BNL & FNAL).
- **This is an area where we could use some help from the operations programs!**
  - Still need to push to get these transfers in production in Rucio and PhEDEx.

XCache – Cache-based data access



The "SoCal" cache setup – a distributed caching instance setup that delivers analysis data to San Diego and Caltech. Petabytes moved – and usage continues to grow

A significant portion of the data delivery needs can be improved by a site-local cache:

- **Analysis** has a traditional "cache-friendly" workload with many repeated data reads.
- **Production** benefits from the latency-hiding effects of having a nearby data source.

The "SoCal" (UCSD & Caltech) cache continues to expand; now covers both CMS MINIAOD & NanoAOD.

- In the last few months, effort has gone toward operational concerns: e.g., developing scripts to **validate data in the caches**.
  - Generic ROOT file corruption detection.

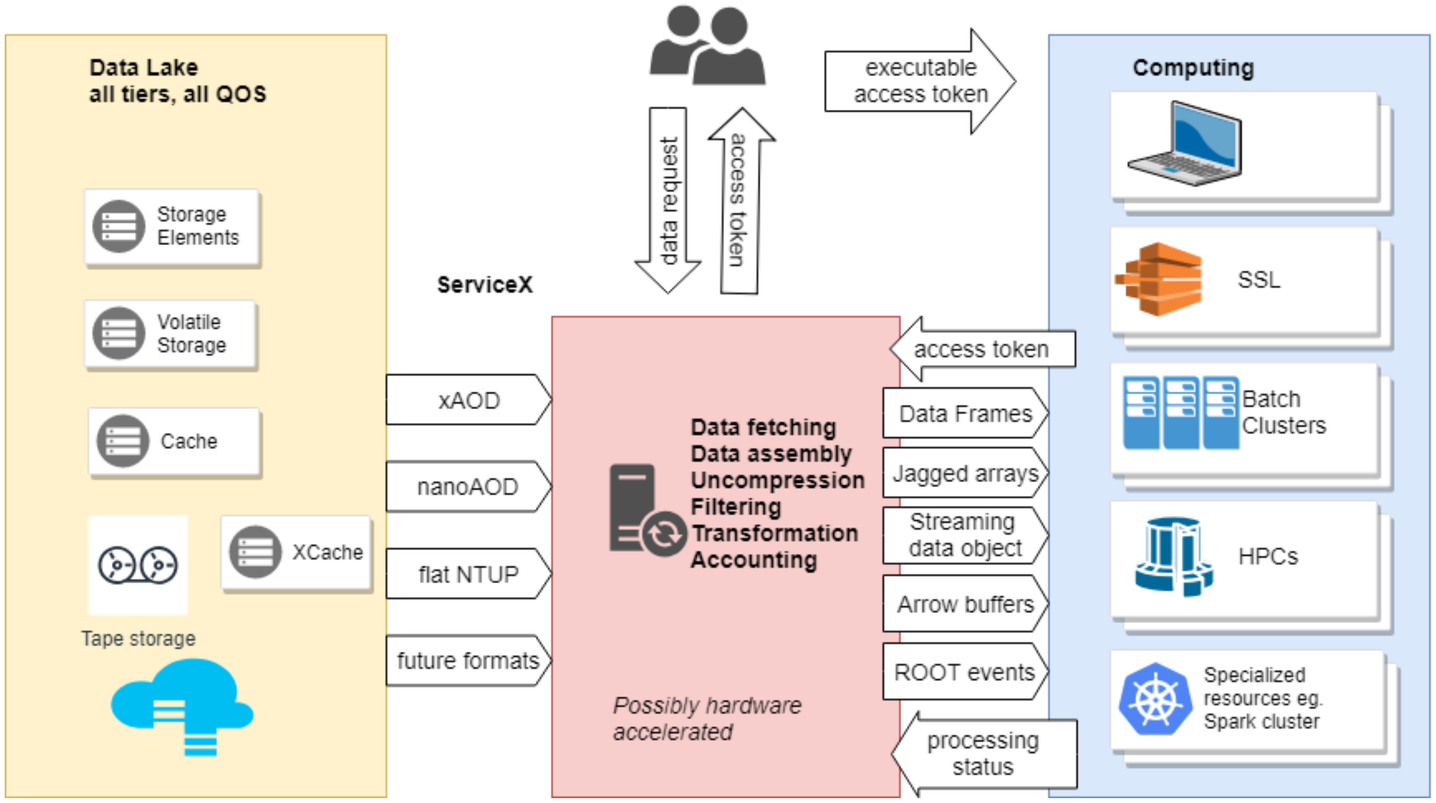
ServiceX and IDDS

Users specify needed events/columns and desired output format:

- Request by metadata tags (real/sim data, year, energy, run number, …)
- Any required preselection.

ServiceX:

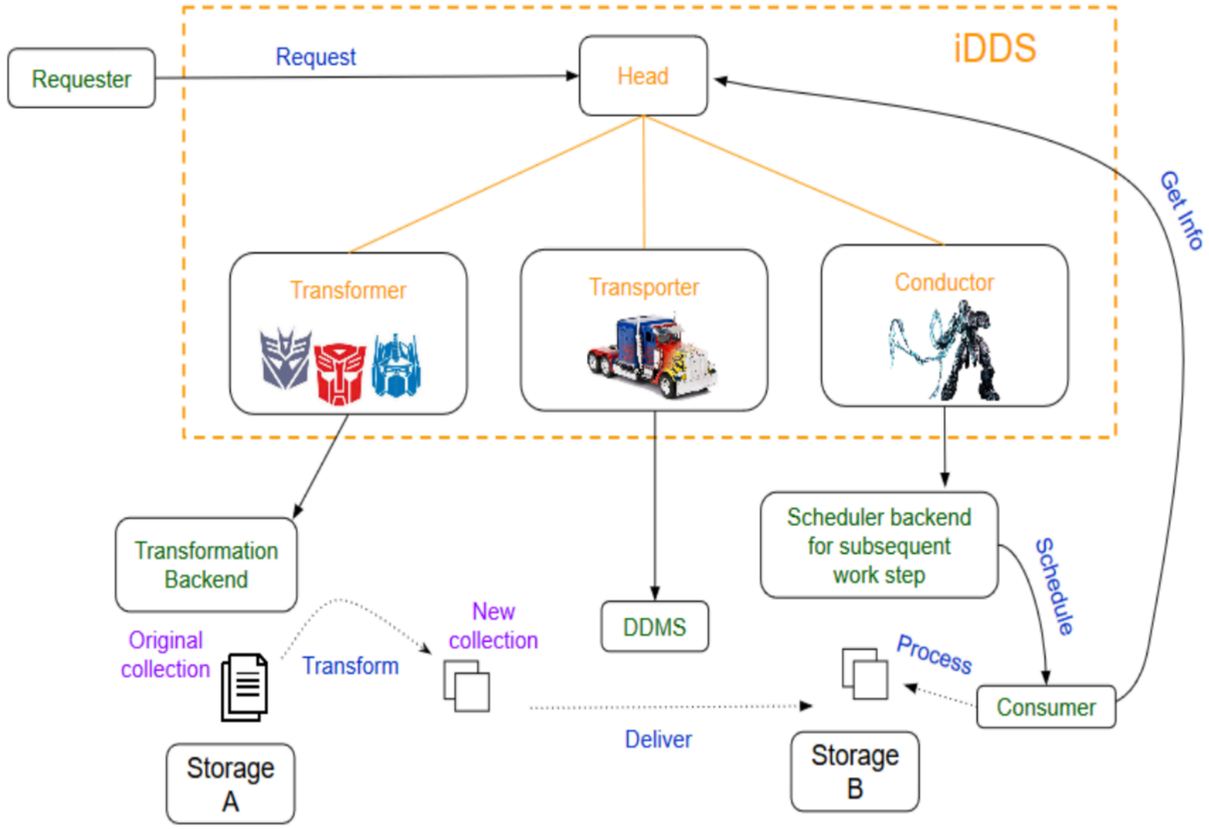
- Queries backend (Rucio) to find data
- Gives unique token to identify request
- Access data from storage through XCache
- Validates request
- Perform data transformations
- Keeps track of data delivered and processed; ensures bookkeeping.



ServiceX: Progress Since Last Time

- Demonstrated basic functionality at the IRIS-HEP Annual Retreat.
- Migrated service completely to the SSL.
- Worked to increase reliability and tail-processing: can request 10s of TBs to be processed and reliably (i.e., all the events) get the requested columns delivered.
- Integrated with the `func_adl` declarative analysis language from the Analysis Systems team.
- Work to enable a use case for Peter Onyisi's group at U. Texas.
  - These are our 'friendly guinea pigs' – someone willing to use the system in anger.
- CMS instance is all ready modulo access to CMS Rucio instance (behind CERN firewall).

ServiceX and IDDS



ServiceX is meant to quickly process on-disk data and deliver **columns** to analysis users.

More generally, iDDS (Wen Guan, UW) aims to reliably deliver events and files to be processed.

- Even when this requires longer-term management, such as pulling data from tape.
- Meant to integrate closely with an external processing system (PanDA).
- Evolution of the ATLAS Event Streaming Service (ESS).
- Working on first demonstrations for ATLAS Tape Carousel.

Skyhook DM

See CHEP presentation and internal presentation by Jeff LeFevre



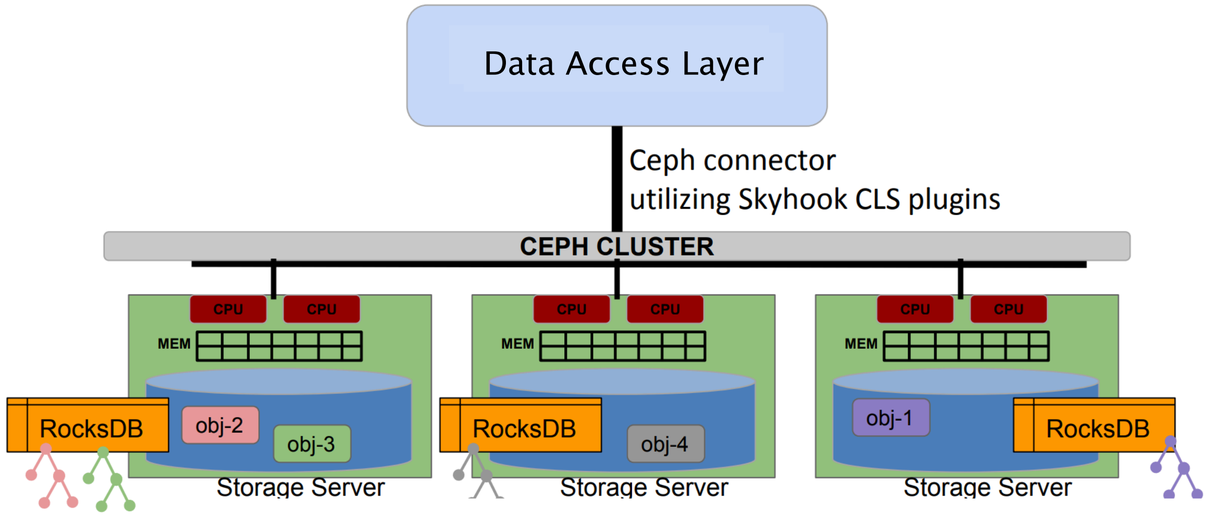
The Skyhook DM project has shown the ability to **ingest ROOT files** (particularly, CMS NanoAOD) and convert event data to the internal object-store format.

Ceph-side C++ plugins transition from on-disk format to desired memory format. For example, allows data management system to optimize disk layout without involving clients.

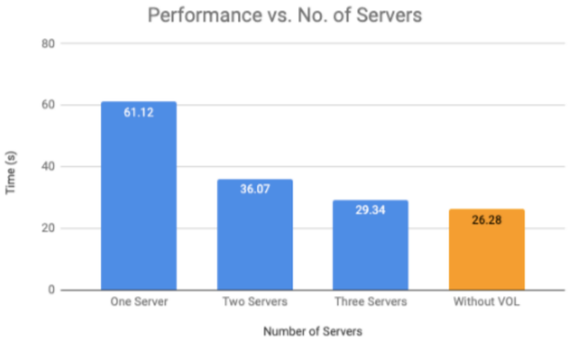
Uses Dask workers to distribute data to clients.

Data delivered as Arrow tables and (optionally) presented as dataframes.

Currently converting one of the Coffea demo notebooks to use SkyHook instead of downloading via HTTP.



Initial measurements of the overhead of the VOL layer in HDF5.



Upcoming Activities

Over the next few months, we plan to:

- Scale the TPC activities to cover production data and larger volumes.
- Complete our Data Challenge #1: last integrations between XCache, SkyHook, and ServiceX.
- Start scaling further the SkyHook & ServiceX components.
  - Work on user-ready UIs: right now, these are at best developer-friendly.
- Design a Data Challenge #2 to demonstrate scale (including IDDS to prepare data for ServiceX).
- Participate in the WLCG DOMA face-to-face at FNAL (prior to the Rucio Workshop).
  - Significant work to be done to write-up the DOMA pieces of the WLCG CTDR.

MORGRIDGE

INSTITUTE FOR RESEARCH

CORE COMPUTATION

morgridge.org

FEARLESS SCIENCE