Overview of Virgo on site computing infrastructure

Franco Carbognani – EGO

ET-PP WP8 Workshop| 30 June 2025

Overview

- Workflows
- Data Model
- Data flow schema
- Low-latency Analysis timescale
- The Challenge
- The Solutions
- Conclusions

Workflows

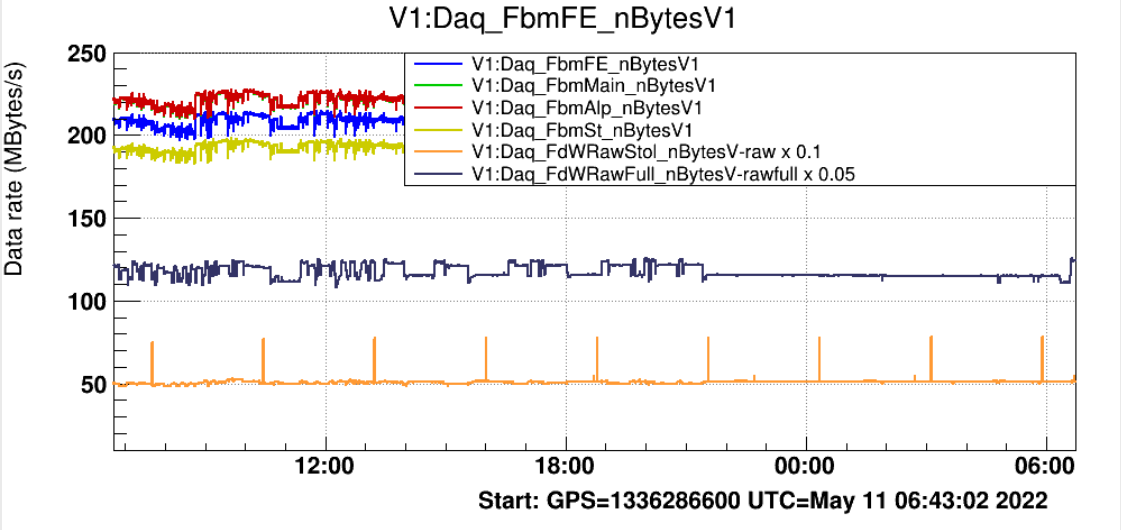
- Commissioning and Operation
- Detector Characterization (DetChar)
- Scientific Analysis (Low-latency)
- All workflows are driven by data from EGO's Cascina site

Data Model

- Data types: raw, reduced (rds), h(t), trend, minute-trend
- Frames format: 1-second shared memory to multi-second disk files
- Storage: circular buffers in Cascina; long-term archives in Tier-1 CCs

Raw Data, ~1.5 PB/yr:

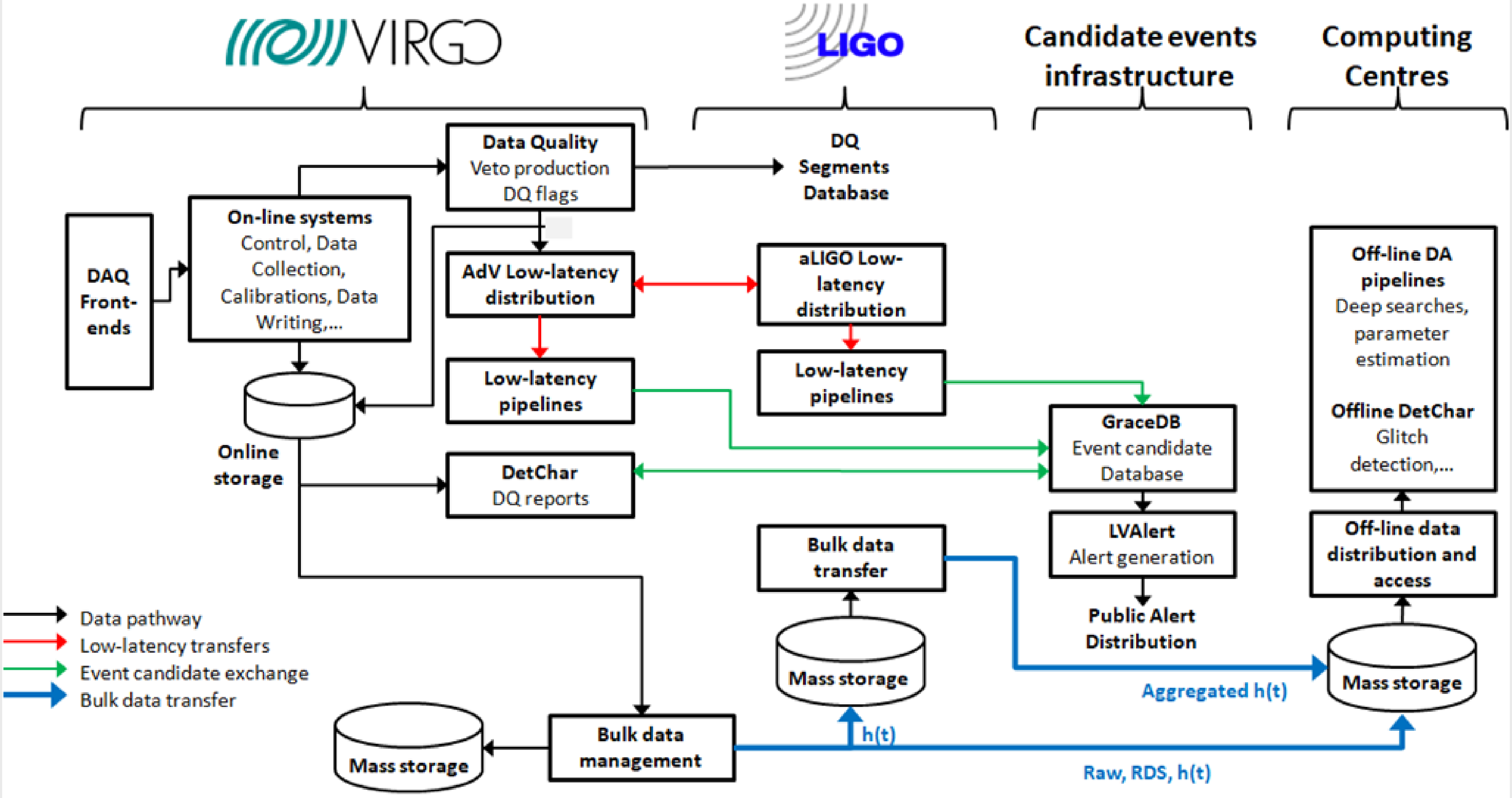
- **Full Bandwidth Raw**, not exported
- **Raw Data:** downsampled, include all auxiliary channels and calibrated/uncalibrated h(t) (~ 50MB/sec -> 4 TB/day -> 1.5 PB/yr), exported to Virgo Tier1 CC (CNAF, CC-IN2P3) for custodial storage during scientific runs
- **A few levels of reduced data sets** for various uses



Data for physics, ~ 5TB/yr/detector:

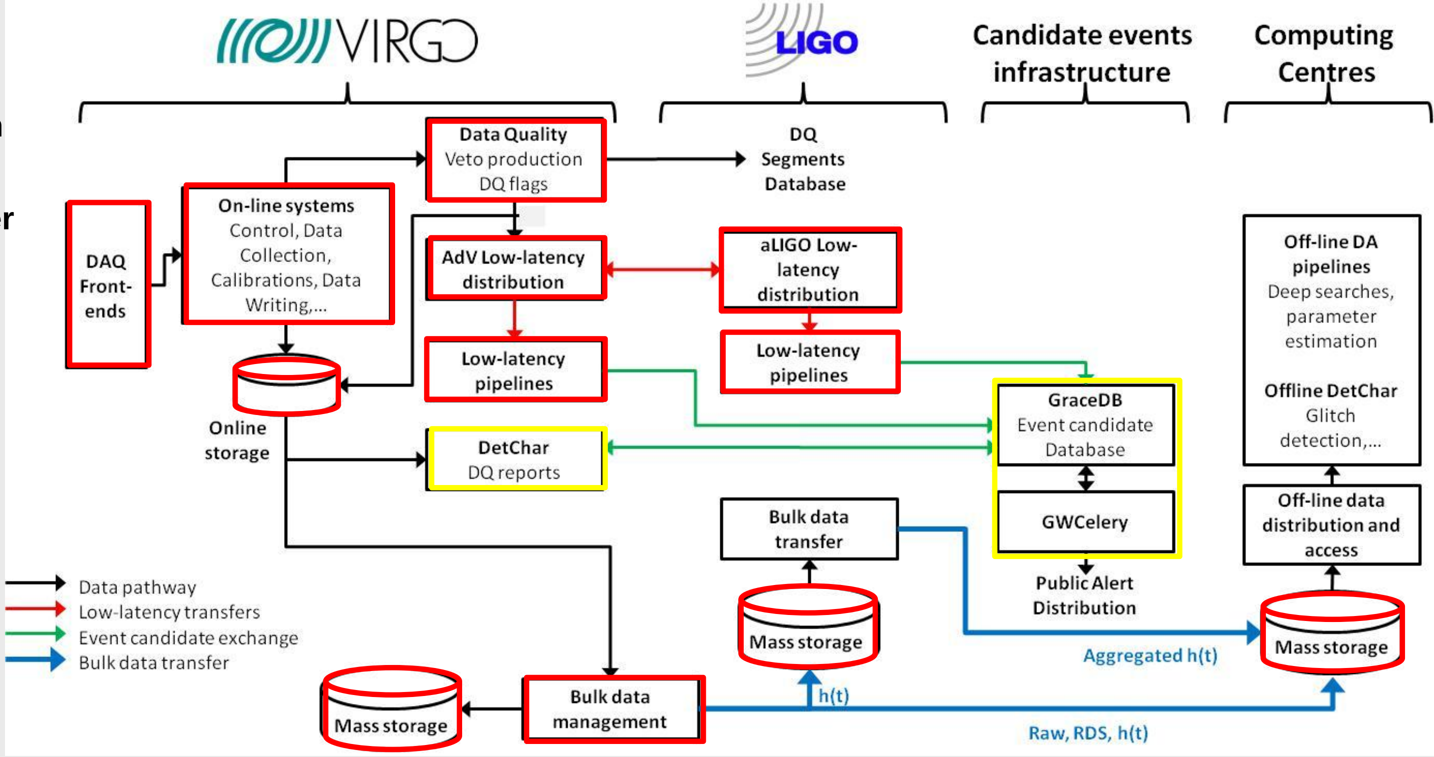
- **Virgo h(t): calibrated «strain» data**
  - sampled at 10 KHz, stored as ~1kSec frame files: aggregated h(t)
  - Includes state vector (data quality flags, vetoes,…)
  - Comes also in the “Analysis Ready” flavor
- **LIGO (and KAGRA) h(t)**
  - Transferred online to EGO for low-latency searches and made available to IGWN CCs for offline analysis

The Data Flow Schema

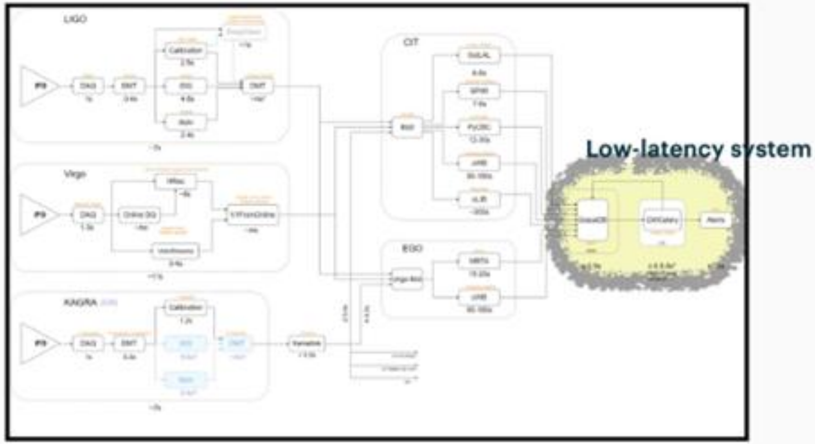


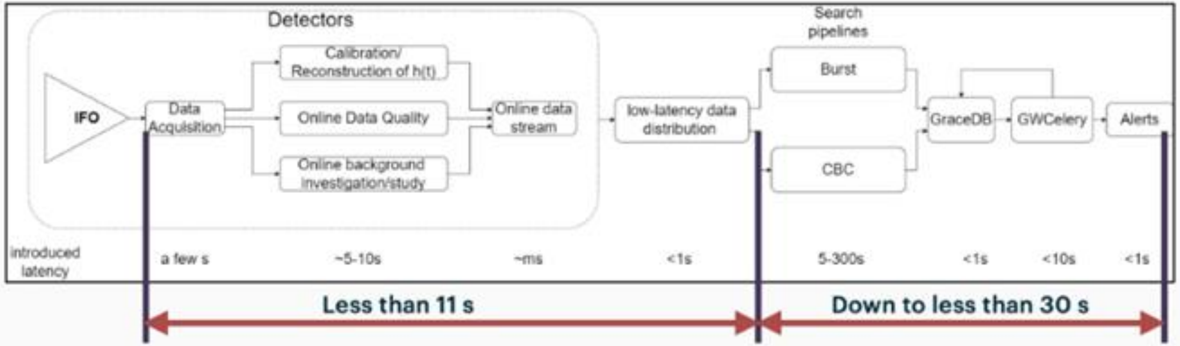
Data Flow in action

1. The signal arrives
2. Data composed into frames
3. Calibration of the data
4. Veto, DQ flags production
5. h(t) transfer
6. Low-latency matched-filter pipelines
7. Upload to GraceDB
8. Data written into on-line storage
9. Low-latency data quality
10. Low-latency sky localization
11. GCN Circular sent out
12. Data written into Cascina Mass Storage
13. Data transfer toward CCs



- We are running extensive tests (already started - up to engineering runs) from data acquisitions (synthetic) to alert generation, and we are monitoring latency.
- We have the signal ready to be analyzed online in less than 11 seconds from the arrival of the (GW) signal at the detectors.
- That makes pre-merger alerts possible (with negative latency) and the first preliminary alerts in less than a minute.
- The study will also allow us to test the effectiveness of the online pipeline to detect and assess the properties of the signal.





End-to-end latency study group (https://wiki.ligo.org/Operations/O4EndToEndLatency)

9

roberto.depietri@unipr.it

The (Online) Challenge

- The site computing infrastructure must guarantee:
  - The support for all identified Workflow
  - Adequate **storage and computing resources** for commissioning, detector characterization, low-latency searches
  - **Low-latency data distribution** among the different observatories and computing clusters for low-latency searches
  - **Bulk** (Aggregated h(t) and raw) **data transfer**
  - A **running environment** on dedicated resources as uniform as possible respect to the other sites
  - A robust support for **development and operation**

The (Common) Solutions

- Storage and computing resources
  - Dedicated or highly-prioritized resources
- Low-latency data distribution
  - Kafka
- Running environment
  - IGWN environments: Conda + CVMFS
- Aggregated h(t) data transfer
  - Rucio
- Raw data transfer
  - Legacy Solution (based on iRODS and WebDAV) -> Rucio
- A robust support for development and operation
  - GitLab

Storage and Computing Resources

Low-latency storage and computing mainly provided by observatory computing centers

- Low-latency alert infrastructure runs on dedicated resources with high priority to burst out into pool
- Search pipelines run on dedicated or highly-prioritised resources (preferably in an HTCondor-managed resource pool)
- Fast, direct access to small data files / shared memory
- Alternative HA deployment via Kubernetes is being tested on INFN-Cloud at CNAF

Storage and Computing Resources

Storage – Raw Data:

- Requirement from Operation/Commissioning moving from 1-year raw data buffer required on-site (~1.4 TB total) to a full (O4) run (~2.4 TB total)
- Going from 17 months buffer to a 25 months one
- Experience have shown that managing this storage as a pure circular buffer is the most maintainable solution
- Raw data retrieval from CC capability available but cumbersome.
- **This data buffer is truly essential for supporting the Operation/Commissioning workflow**

Data	Target buffer length in Cascina	2020 Buffer space in Cascina [TB]	2021 request [TB]	2025 request [TB]
Full Bandwidth raw	~ 1 week	40	**130**	
Raw data	1 year -> 25 months	~ 1000	**~ 1400 (400 TB increase)**	**~ 2400 (1000 TB increase)**
Omicron data /data/prod/omicron /data/dev/omicron	1 year	using shared areas under /data	**10 + 10 (20 TB increase)**	
AdV RDS	5 years	~50	**~160 (110 TB increase)**	
CBC low-latency (MBTA)	-	~120	**~180 (60 TB increase)**	

Storage and Computing Resources

CPU:

- Commissioning

Task	Processors	2021 request	2022 / O4 requests
Optical Simulations	20 CPU (cores)	**add two machines similar to ctrl26**	TBD
Targeted Studies: Scattered light	HTCondor farm (640 vcores)	**35 - 70 cores**	**35 - 70 cores**
Targeted Studies: Lock losses	HTCondor farm (640 vcores)	**4 - 8 cores**	**4 - 8 cores**

- Detector Characterization

Pipeline	Processors	2021 request	2022 / O4 requests
Omicron pipeline	60 CPU (cores)	**add one machine similar to olserver138**	TBD
DQR	HTCondor farm (640 vcores)	**Preserve capability to run DQRs with the same latency and efficiency than during O3**	**Preserve capability to run DQRs with the same latency and efficiency than during O3**

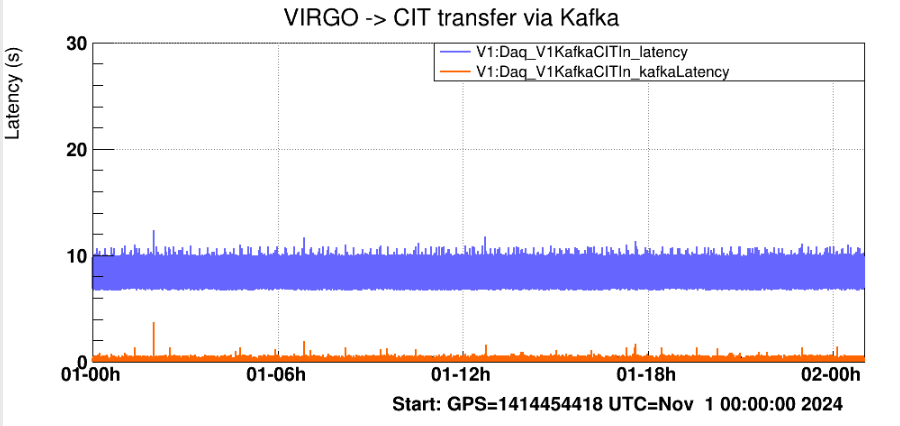
CPU:

- Low Latency Searches

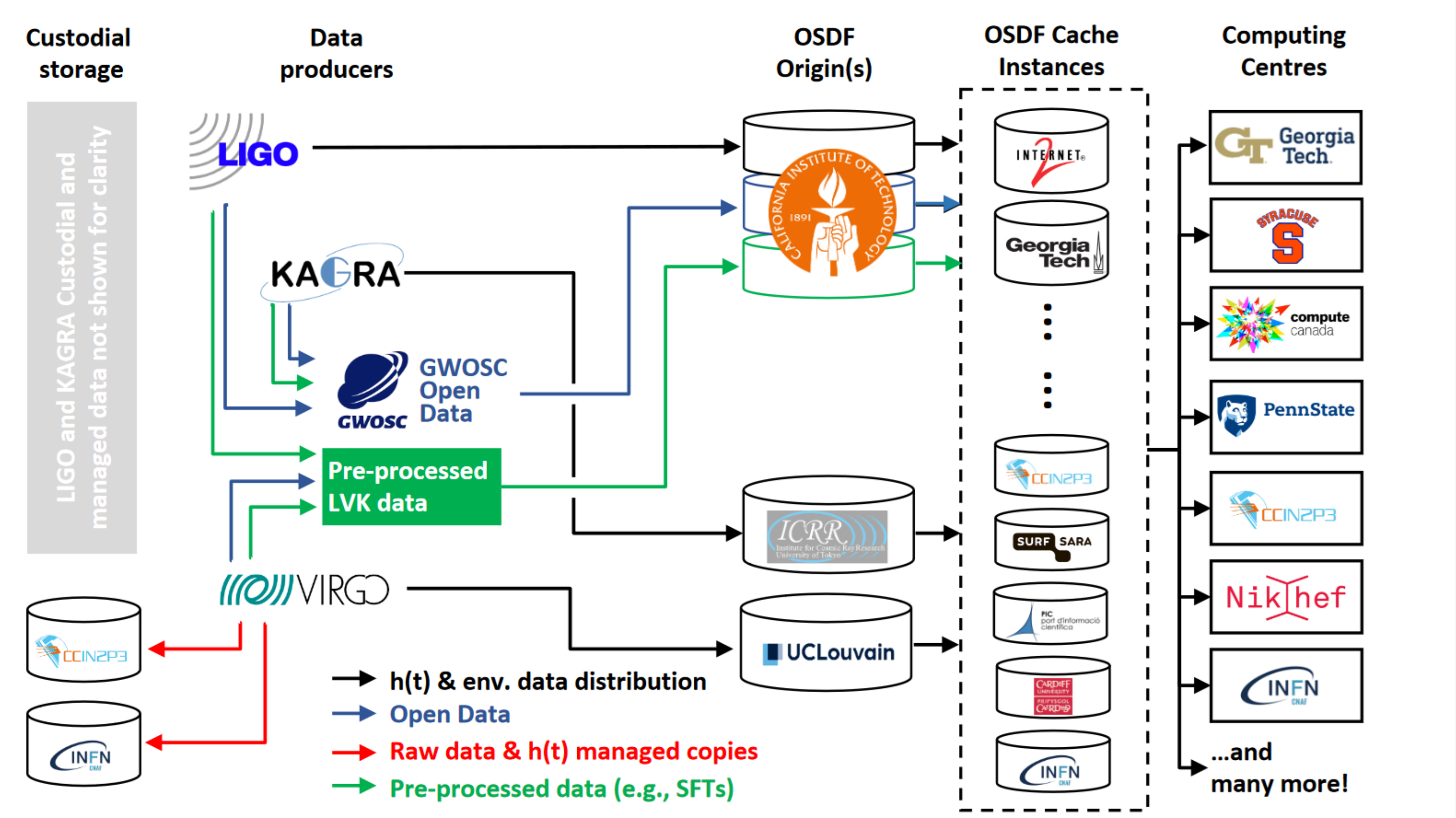
Pipeline	Processors	2021 request	2022 / O4 requests
CBC low-latency (MBTA)	880 vcores (9800 HS06)	same	**1760 vcores (19600 HS06) (100% increase)**
cWB low-latency	HTCondor farm (640 vcores)	**Preserve capability to run cWB with the same latency and efficiency than during O3**	**Preserve capability to run cWB with the same latency and efficiency than during O3**

Low Latency Data Distribution

- Jointly developed IGWN Low Latency Data Distribution software (lldd) based on Kafka
- Running fairly smoothly so far for the overall network during O4
- Kafka is a modern high throughput stream processing software
  - Built in redundancy
  - Can survive if stream stops or Kafka broker goes down
  - Replication so no data loss from service downs
  - Highly scalable and reconfigurable
  - Can easily add additional observatories data



Data Transfer



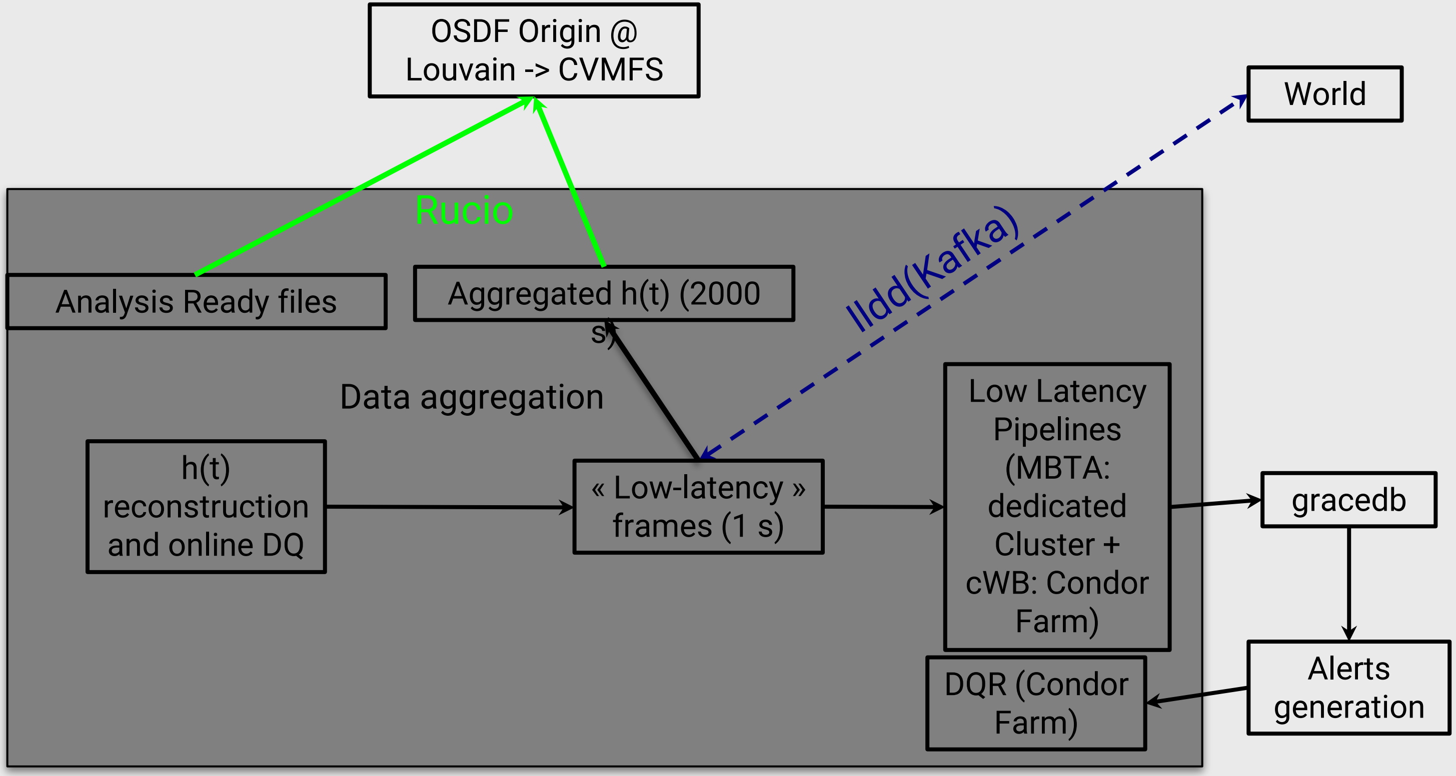
Aggregated h(t) Data Transfer

- Aggregated 2000 sec h(t) data (aka online files) have been flowing from Cascina to the OSDF/CVMFS Louvain origin using Rucio since before the start of O4
- Same mechanisms is used for the distribution of the Analysis Ready files wich comes in batches of variable sizes

Raw data transfer

- The iRODS + WebDAV Raw data transfer to CCIN2P3 and CNAF is going well and data is flowing into dedicated O4 directories
- Tests on Rucio for raw data transfer have started since a while
  - Infrastructure already setup
  - Data has been transferred from Cascina to Torino showing some performance issues which seems ortogonal to Rucio
  - A “near to production” storage area is being prepared at CNAF in order to disentagle Rucio tests from underlying hardware performances

The main components as running @ EGO site



Cascina Machines Clusters

The Cascina machines are grouped in clusters dedicated to different purposes:

- ctrlXX: dedicated to run only client applications for Detector Control or Low latency Analysis
- olserverXX and olnodeXX: dedicated to run server applications for Detector Control or Low latency Analysis, normally started from a Virgo Process Monitoring (VPM)
- farmnXX: usable for running tests application or general purpose computing (Matlab simulations, single run of detchar applications, etc.)

A more detailed description of data servers assignement at EGO can be found here

Cascina Storage Areas

Data Storage

- All intermediate and final data products (in particular large size files such as .gwf files) are supposed to be stored under the /data area. The use of home directories for storing intermediate large files (or very large number of files) must be avoided. The guidelines for the use of /data to achieve a proper on-line/out-of-line systems decoupling (See : [VIR-0394A-19]) must be followed

Domains:	on-line	out-of-line			
Top dir	/data/online	/data/prod	/data/dev	/data/archive	/data/web
Purpose	ITF control + online analysis processes/users (read/write, master copy)	general users/processes read	users/processes debugging, studies	old read-only data (only ones needed at Cascina, the rest => CCs)	for publishing ? (TBD)
Access	from online hosts only (olserverxx, HTCondor, ctrlxx)	mainly from farmnxx from online hosts: only "near-line mount" via filesystems capable of disconnected operation (i.e. webdav, no nfs)			
Retention	< 1month circular buffer	negotiated, circular buffer			
Volume size	fewGB – fewTB per task	< 50TB per task	~ 50TB per volume	~ 50TB per volume	
Availability / Maintenance actions	Highest: only selected bug fixes or feature / config updates	Normal: 1 day or 1 week downtime periods possible		unspecified	
Backup policy	snapshots only	disk-to-disk + snapshots	disk-to-disk + snapshots + tapes (2 full/year)	tapes (as needed)	

Async replica

Cascina Storage Areas

Online area

•*/data/online/* : contains the volumes for applications/tasks which have a critical on-line footprint for gw events triggering and follow-up

Volumes in the online area are typically small to make shorter the management/troubleshooting/recovery/maintenance operations and are accessible mainly by on-line machines (olserverxx , ctrlxx) in order to control the I/O load.

The data retention time for data in this area would be of the order of 10 days.

Out-of-line area (dubbed also "nearline" even if technically is not the right name)

•*/data/prod/* : contains live replicas of the on-line data volumes for general access

Volumes in the "prod" area can be larger respect to their online counterpart in order to keep a longer history for "active" data.

The delay of the appearance of the last produced data here from when they are written on the online area would be in the range from 0 to 5 minutes.

•*/data/dev* : contains the volumes for the, mainly interactive, development, testing and analysis

They can be seen as "home directories" dedicated to each task

•*/data/archive*: contains "cold" read-only data

Running Environment: IGWN Conda

Sustainable software development and distribution = **CVMFS** + **Conda**

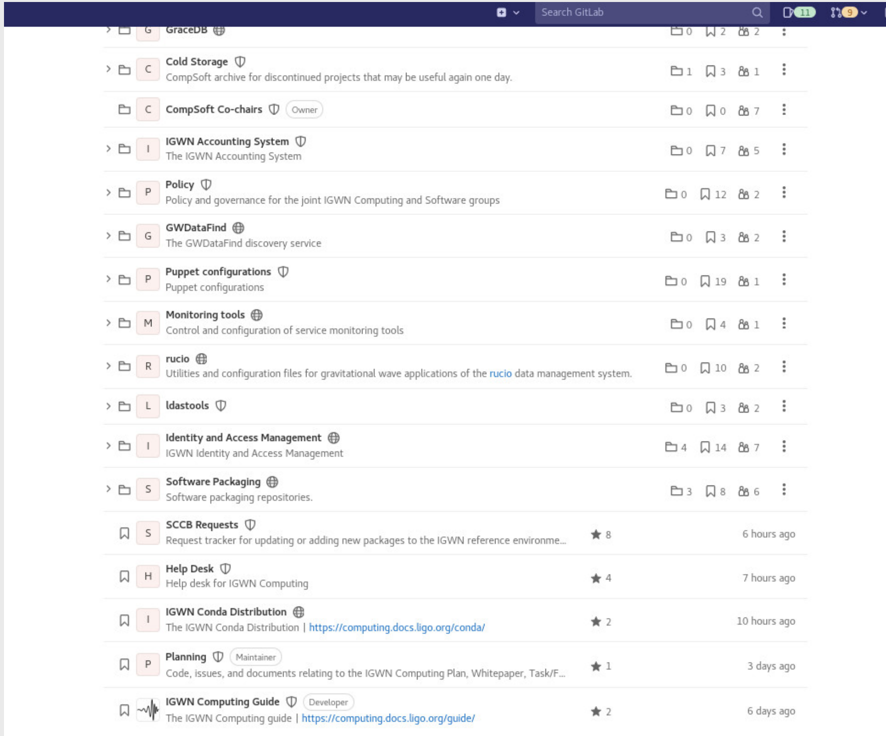
Conda environments hosted in CVMFS file system available at all IGWN sites and CC

Running Environment: IGWN Conda

- IGWN Conda Distribution provides pre-built, automatically-distributed environments of approved software
- available via CVMFS on any machine (no authentication required)
- can be replicated on any workstation
- Provide effective OS independence
- Leverage on CMake/Meson for software build. See: [CMT with Meson or CMake build delegation tutorial.]
- Provide a very effective solution for otherwise unmaintainable number of custom software builds
- On large part of IGWN sites (including Virgo Cascina) the IGWN environment can be activated manually or is by default at user login

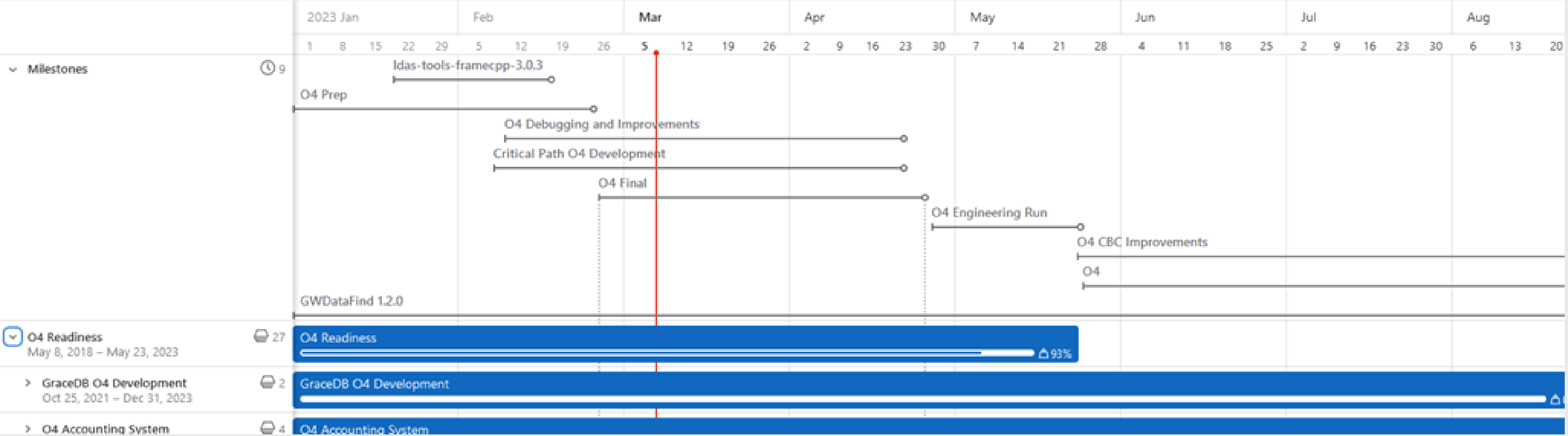
LIGO, Virgo and KAGRA Computing & Software teams increasingly working together using GitLab centered DevOps practices. IGWN Gitlab provides/supports:

- Code repository for IGWN Detectors Control, Data Analysis and Low-latency software
  - Enables easy collaboration on software
- Continuous integration (CI) capabilities to enable the automation of building, testing and deployment of code.
- Software Configuration Control Board (SCCB) activities
- IGWN Computing HelpDesk (via tickets)



Development and Operation

- Conda software distribution management
- IGWN Computing Guide (via GitLab Pages)
- IGWN Computing Planning (via the growing GitLab support for projects management)
    - Issues
    - Epics



Conclusions

- The transition from custom, incompatible LIGO, Virgo and KAGRA-specific infrastructures to a common multi-collaboration infrastructure, based on mainstream tools, is well ongoing and has payed already many dividends (e.g. Gitlab, Conda, Kafka, Rucio, ..) in particular at the Cascina site
- The upgrades implemented in time for the O4 Science Run are serving us well
- We are under pressure to reduce Cascina Computing spending footprint. The possibility to run the Low-latency Scientific Analysis elsewhere than Cascina, now fully enabled by the Kafka based Low Latency Data Distribution, is currently being investigated.