

HSF Conditions Database: from Concepts to Implementation

Ruslan Mashinistov of behalf of HSF CDB team, experiments and contributors

Brookhaven National Laboratory (US)

ePIC S&C meeting 2026

September 16 , 2026

    @BrookhavenLab

Introduction

Conditions Data

Conditions data is any additional data needed to process event data

WHAT IS IT ?	EXAMPLES	VALIDITY RANGE
• Non-event parameters that change over time • Needed for calibration, alignment, and reconstruction	• Calibration constants, alignment geometry • Beam parameters, magnetic field maps • Bad channel maps	• Each conditions data file (payload) is valid for a specific range of runs • Ensures correct conditions are applied to each event

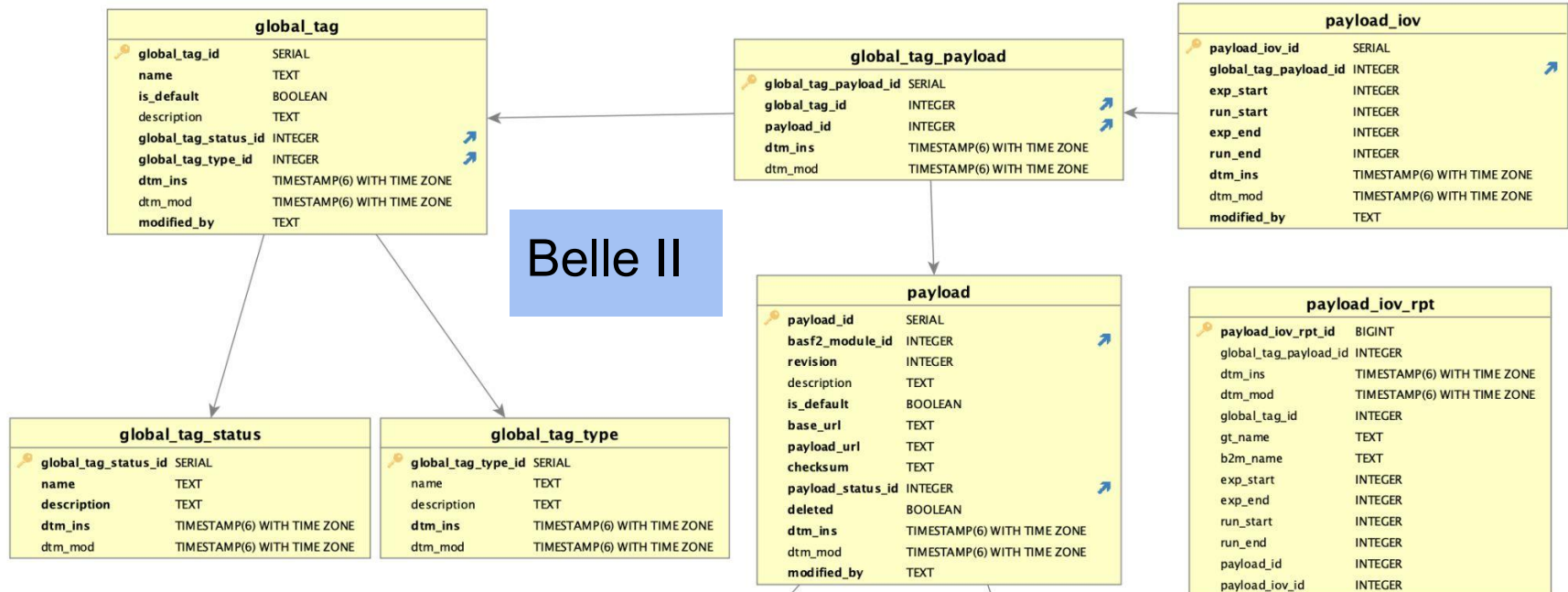
Key concept — Interval of Validity (IOV): Each conditions payload is valid for a specific range of runs/events. For example, a calorimeter calibration may be valid for **Run 1000 – Run 1500**, then **Run 1300 – Run 1500** replaced by an updated set.

Why it's deceptively hard

- Challenges
 - Bursty kHz-scale access from distributed computing
 - Reproducibility and strict versioning
 - Decades-long lifetime of the data
- Evolution
 - Initially, everything was handled at the DB level (CERN:COOL/CORAL->Oracle)
 - Payloads were then decoupled from metadata and stored independently
 - The CDB service evolved into a REST API
 - But first, let's take a quick detour into the database schema

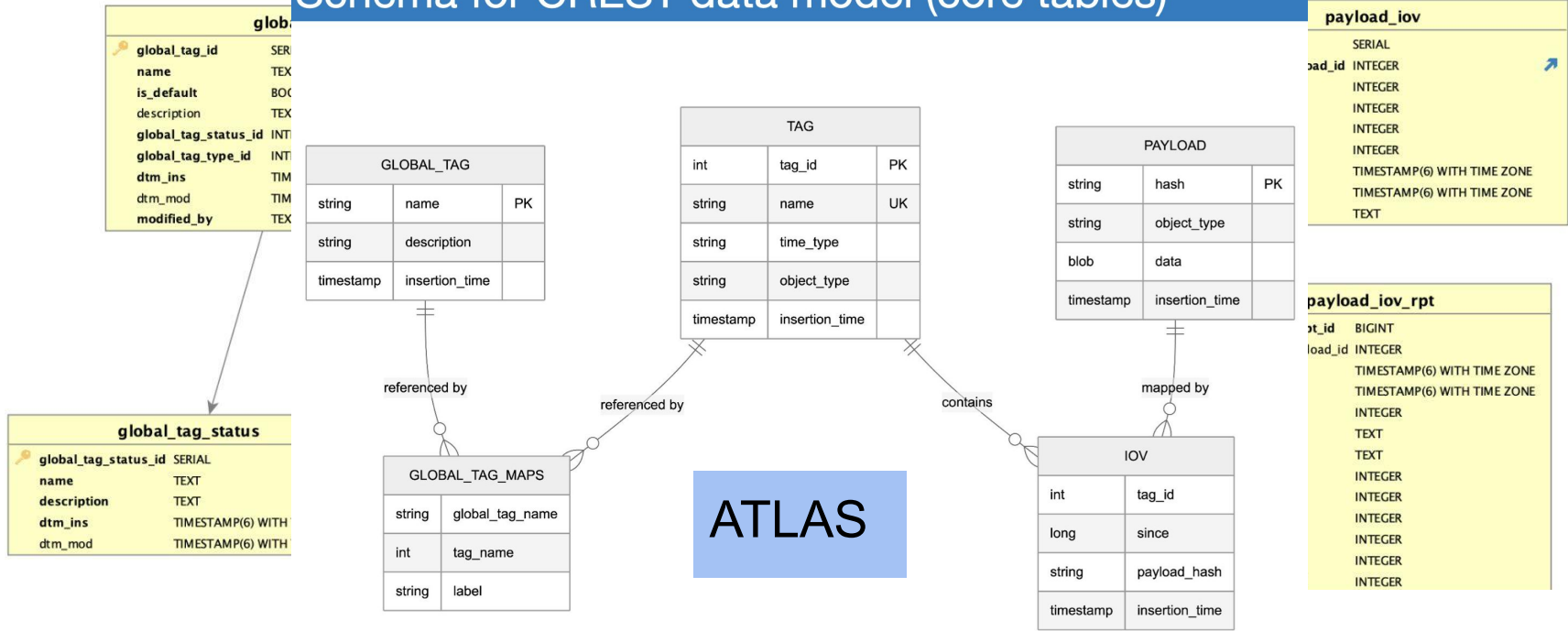
Database schema

DB schema



DB schema

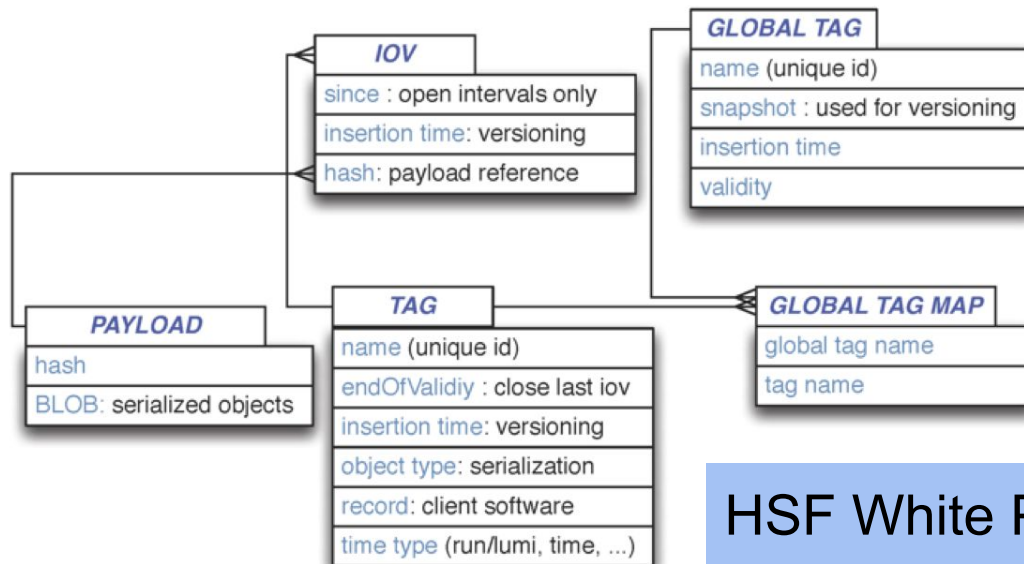
Schema for CREST data model (core tables)



DB schema

Schema for CREST data model (core tables)

[The Community White Paper](#) on **Conditions Data** is where **ATLAS** and **CMS** experts, under the HSF, distilled two decades of operational experience into shared best practices: separate payloads from metadata, generalize time as intervals of validity, ensure reproducibility with versioned global tags.



HSF White Paper

DB schema

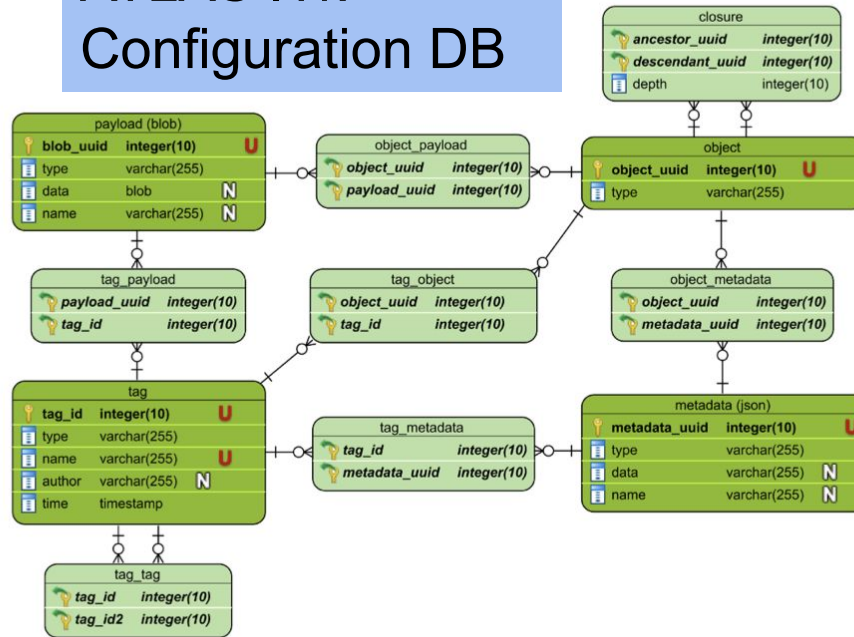
Schema for CREST

glob

Configuration data is
conditions data too

Detector configuration is
versioned,
time-dependent,
non-event data that must
be resolved consistently
for a run.

ATLAS ITk
Configuration DB



H TIME ZONE
H TIME ZONE

TH TIME ZONE
TH TIME ZONE

Database architecture

Edit permissions for a Global Tag: unlocked, locked, or frozen

Global Tag Status



Category of conditions data (e.g. CalorimeterCalib")

Payload Type



Global Tag



Payload List



Payload IOV



Payload Data

Consistent, named snapshot of all conditions

Collection of IOVs for one data type within a Global Tag

Conditions record with a validity range (start/end)

Actual file stored on external storage

- **Reproducibility** — “frozen” Global Tags guarantee identical conditions for reprocessing
- **Versioning** — clone & modify Global Tags for new processing campaigns
- **Non-overlapping IOVs** — exactly one valid payload per type at any point in time
- **Payload decoupling** — metadata (DB) is decoupled from payload files (storage)

Read optimization

Optimized SQL Query

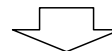
```
SELECT pi.payload_url, pi.major_iov, pi.minor_iov, pt.name, ...
FROM "PayloadList" pl
JOIN "GlobalTag" gt ON pl.global_tag_id = gt.id AND gt.name = %(my_gt)s
JOIN LATERAL (
    SELECT payload_url, major_iov, minor_iov, ...
    FROM "PayloadIOV" pi
    WHERE pi.payload_list_id = pl.id
        AND pi.comb_iov <= CAST(%(my_major_iov)s + CAST(%(my_minor_iov)s AS
DECIMAL(19,0)) / 10E18 AS DECIMAL(38,19))
    ORDER BY pi.comb_iov DESC
    LIMIT 1
) pi ON true
JOIN "PayloadType" pt ON pl.payload_type_id = pt.id;
```

For each PayloadList (Type)



Get Payloads descending
ordered by combined IOV

Limit return to 1 line - latest
Payload for a given IOVs



And then append the results
of each subquery to create
the final output

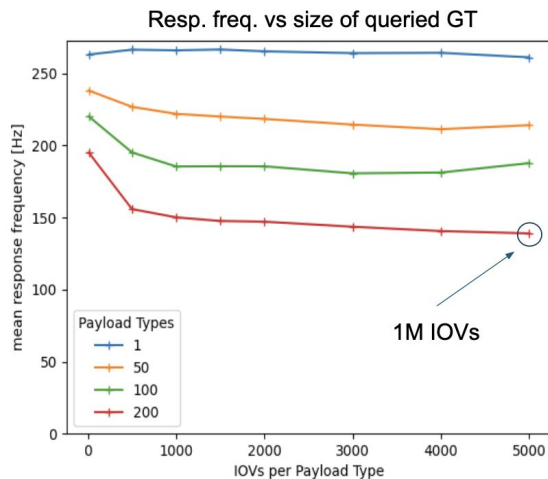
- Without LATERAL, each sub-SELECT is evaluated independently
- Covering index on Payload table including combined IOV and reference to the PayloadList
- AI-assisted PostgreSQL optimization (see Sarthak's slides)

Initial performance testing

- Initial tests on OKD
 - Co-located client
 - Single-pod DB (with pooling)
 - No cache

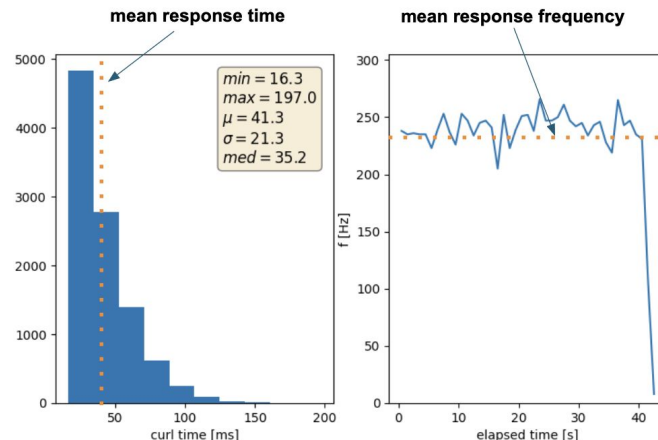
sPHENIX-driven scenarios

Scenario	Payload Types	Payload IOVs (per type)
tiny	10	100 (10)
tiny-moderate	10	2000 (200)
moderate	100	20000 (200)
heavy-usage	100	500000 (5000)
worst-case	200	5200000 (26000)



Achieved mean values:

- 40 ms response time
- 230 Hz response frequency



Write logic

Conditions Data – Use Cases

- HSF Conditions Database meeting: **use cases** <https://indico.cern.ch/event/1280790/>

- Most demanding use-case is

Fast-Processing. Goal:

- Publish data for analysis fast
- Maximize physics performance

Use case

Example

Online

- High Level Trigger

Reprocessing

- Run reco w/ improved calib.

Analysis

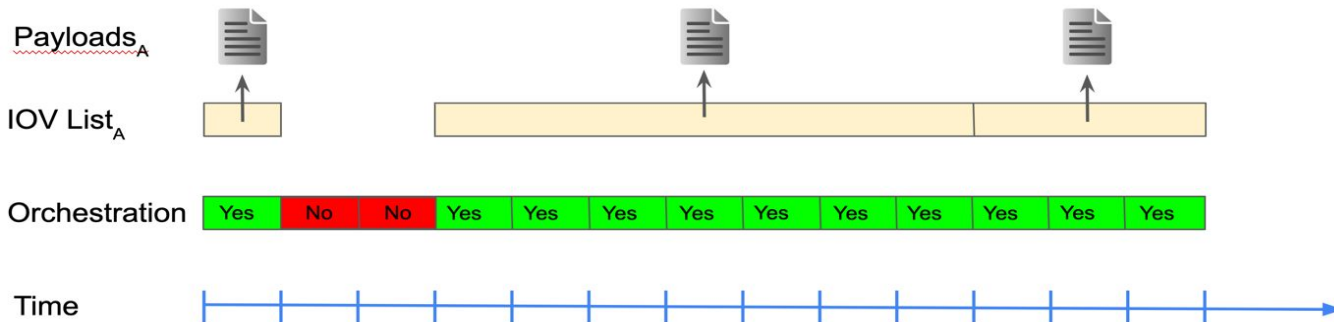
- High level physics analysis

Development

- Test new calib. within existing GT

★ Fast-processing

- Process data w/ just-in-time calib.



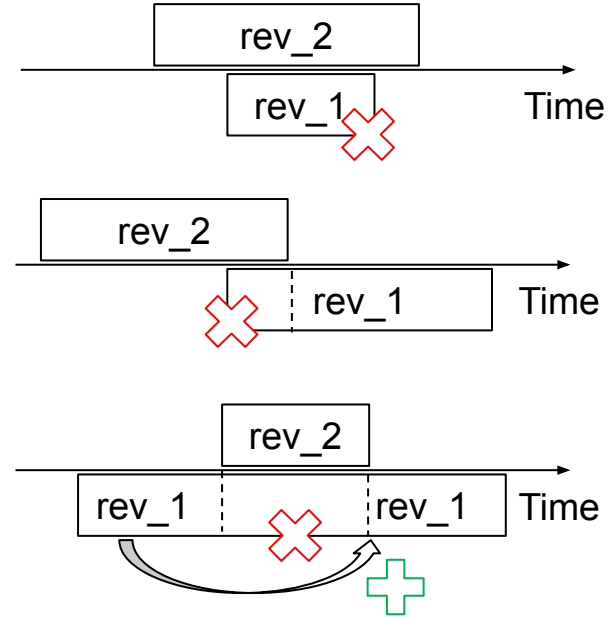
IOVs overlapping

- HSF CDB guaranties that only one (*or none) Payload per List is returned for the given IOV
- HSF CDB can resolve IOVs conflicts automatically

GT status

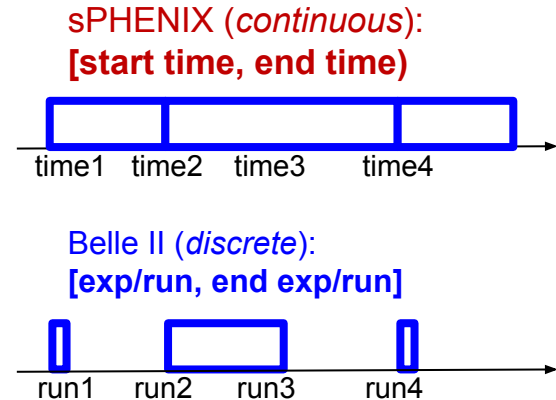
Locked
Unlocked
Frozen

* end IOV estimated by the client



IOVs in HSF vs Basf2

- IOV comprehension *plugin* (*Continuous/Discrete*)
 - Experiments can choose the desired behavior
- In HSF, the IOV is **exclusive** at the end
 - Requested by sPHENIX as they use timeframes
- Belle II's Basf2 expects that the interval is **inclusive**
 - Defined by exp/run pairs



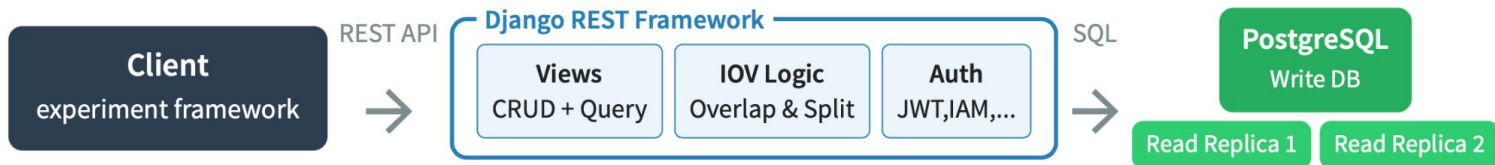
Development and Deployment

Important milestones

- Development started in September 2021
- In production for sPHENIX - January 2022
- Test instance for protoDune since January 2023
- CHEP2023 [presentation](#)
- [Presentation](#) at ePIC S&C WS at CERN in March 2024
- Belle II announced migration to HSF CDB in June 2024
- HSF reference implementation review slides [1](#), [2](#), [3](#) in July 2024
- Einstein Telescope chose HSF CDB - 2024
- PSQL cluster provided for sPHENIX - December 2024
- Belle II migrated to the OpenShift in June 2025
- CHEP2026 [presentation](#)
- HSF affiliated software [review](#) completed in September 2026
- Belle II estimates final migration to the HSF CDB in 2027

Server implementation

REST API + PostgreSQL with optimized IOV lookups



- Django-based server: REST API with auth/authz and experiment-specific plugins
- Web UI in planning

- C++ client library + CL tools; client handles payload download/upload
- Helm chart for one-shot deployment on OpenShift or any Kubernetes platform

OpenShift / Kubernetes — Helm charts



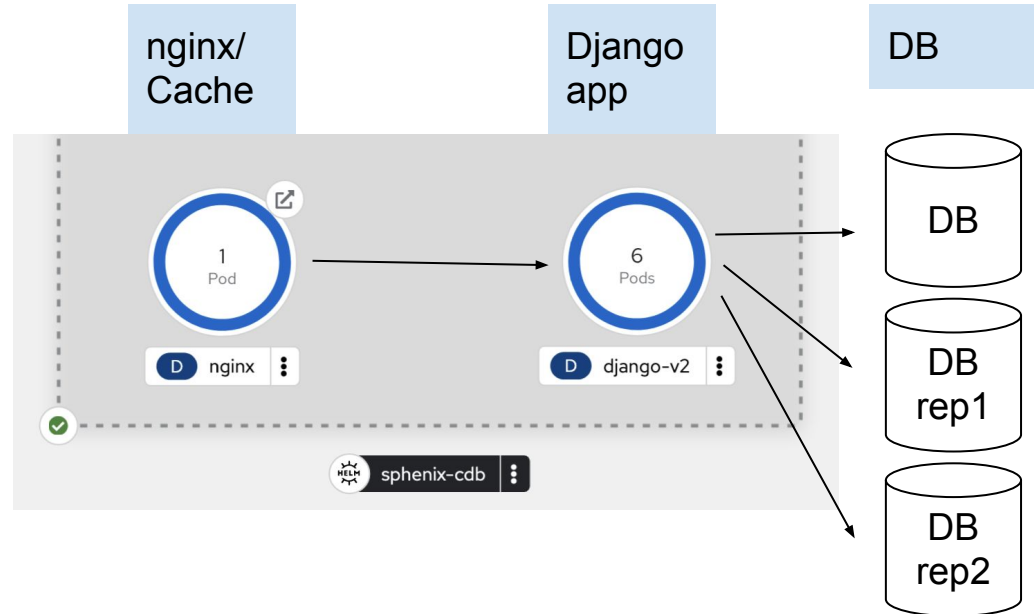
sPHENIX experiment on RHIC

sPHENIX Conditions Data Service

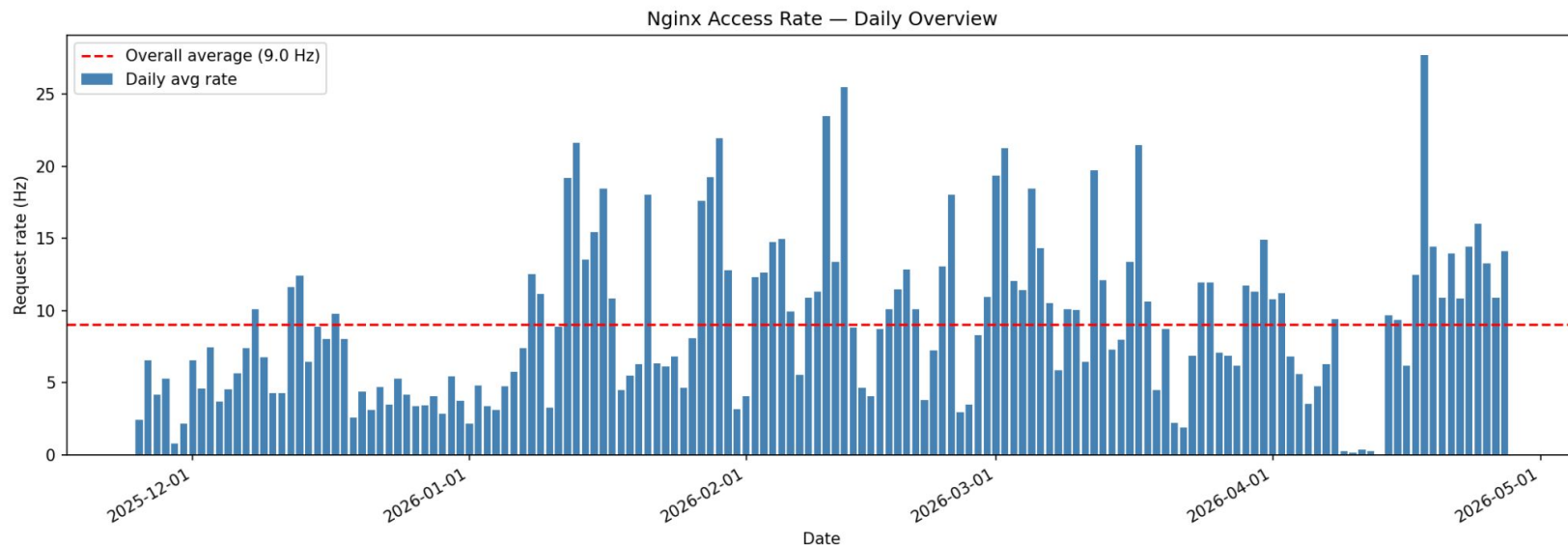
- Supports over 70,000 concurrent jobs on a farm running 132,000 logical cores
- Timeline
 - In production since January 2022
 - [HSF Conditions Database reference implementation](#) - CHEP2023
 - PSQL cluster provided for sPHENIX - December 2024
- Conditions data snapshot:
 - 133 global tags
 - Avg global tag size: 76 payload types, 19017 IOVs
 - Median global tag size: 76 payload types, 1881 IOVs
 - Global tag with most payload types: 163 types
 - Global tag with most IOVs: 133703 IOVs
 - In total, there are 204593 payload files totaling 26 GB in size, with a rough average of 127 KB/file

sPHENIX CDB instance

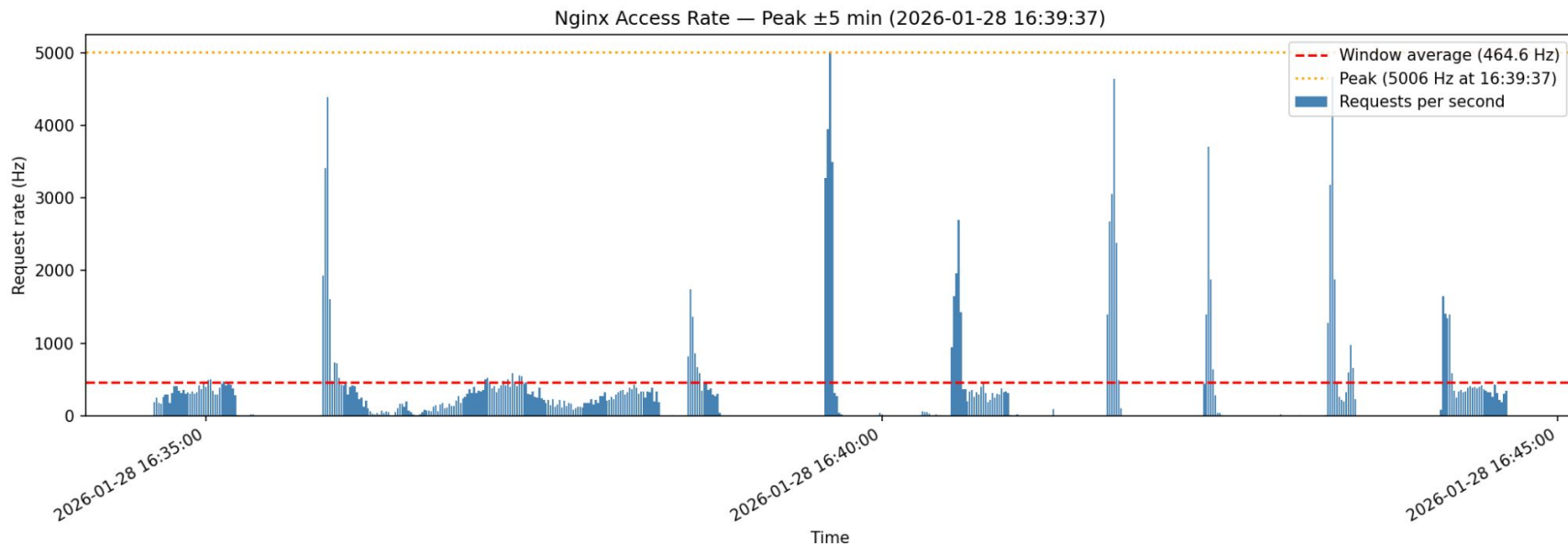
- Host: OpenShift@SCDF BNL
- Pods: 1 nginx, 6 Django - 4 threads each
- PostgreSQL
 - 3 VMs: 1 main, 2 replicas
 - CPU: 8 cores,
 - RAM: 32 GB
 - max_conn 5000
 - DB size: 3GB
- Payloads are handled by the client



sPHENIX conditions data daily average access rate



sPHENIX conditions data access pattern



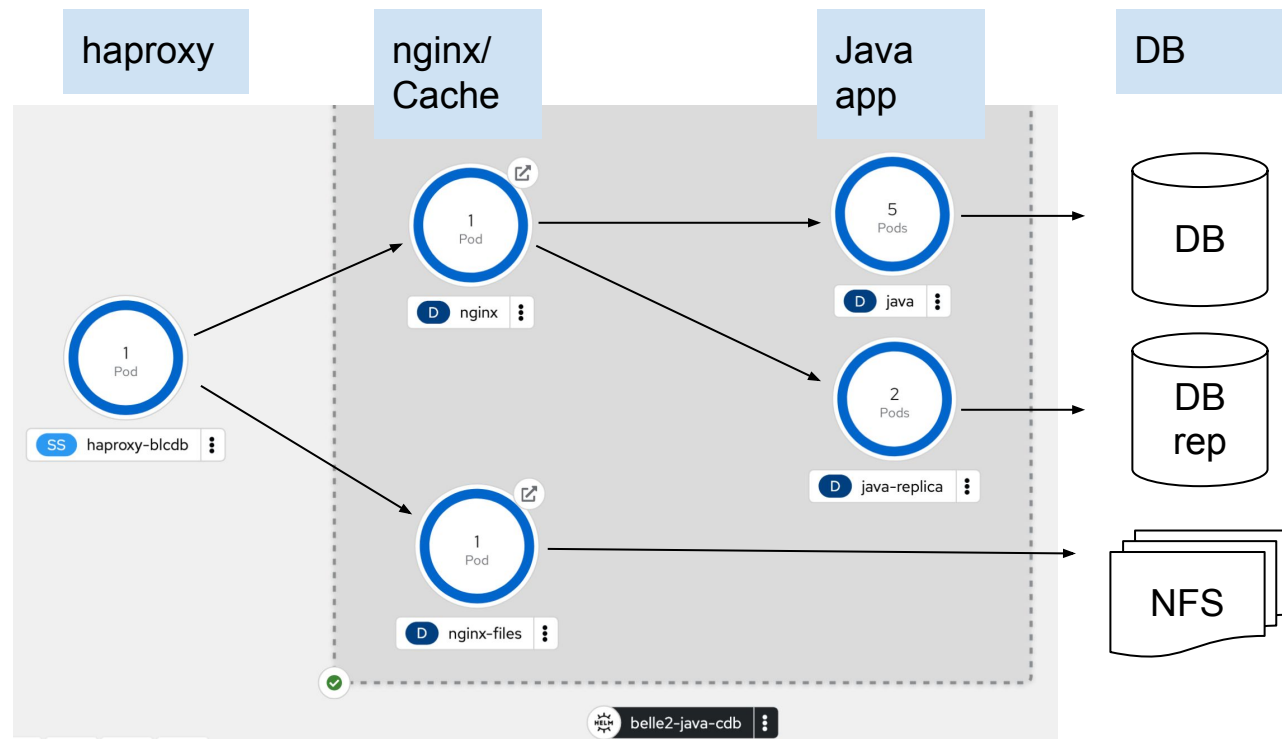
Belle II experiment on SuperKEKB

Belle II Conditions Data accounting

- Conditions data snapshot:
 - 3705 global tags
 - Avg global tag size: 86 payload types, 4273 IOVs
 - Median global tag size: 3 payload types, 12 IOVs
 - Global tag with most payload types: 2840 types
 - Global tag with most IOVs: 276124 IOVs
 - In total, there are 572048 payload files (in 10282 directories) totaling 27 GB in size, with a rough average of 47.2 KB/file

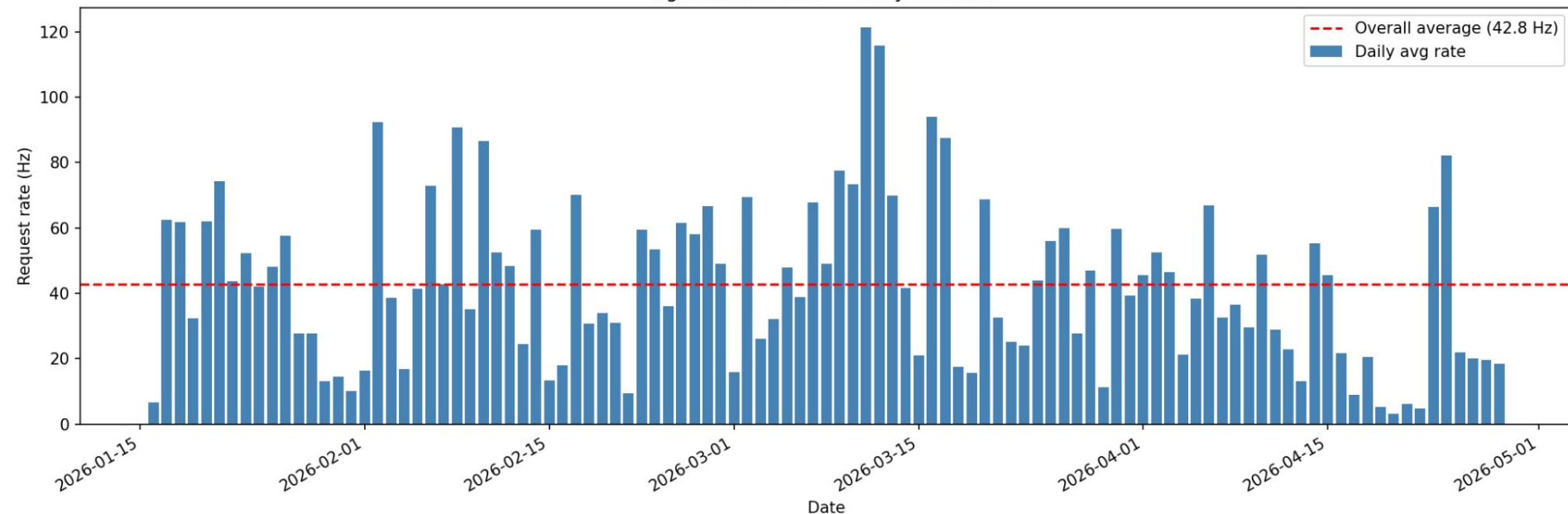
Belle II current Java instance

- Host:
OpenShift@SCDF BNL
- HAProxy pod exposes the instance externally
- 2 Nginx pods
 - Metadata Service
 - Http access to payloads
- 7 Java pods - 32+ threads
- PostgreSQL
 - VM in OpenShift
 - 32 CPU cores, 64G memory
- DB size 3.8 GB

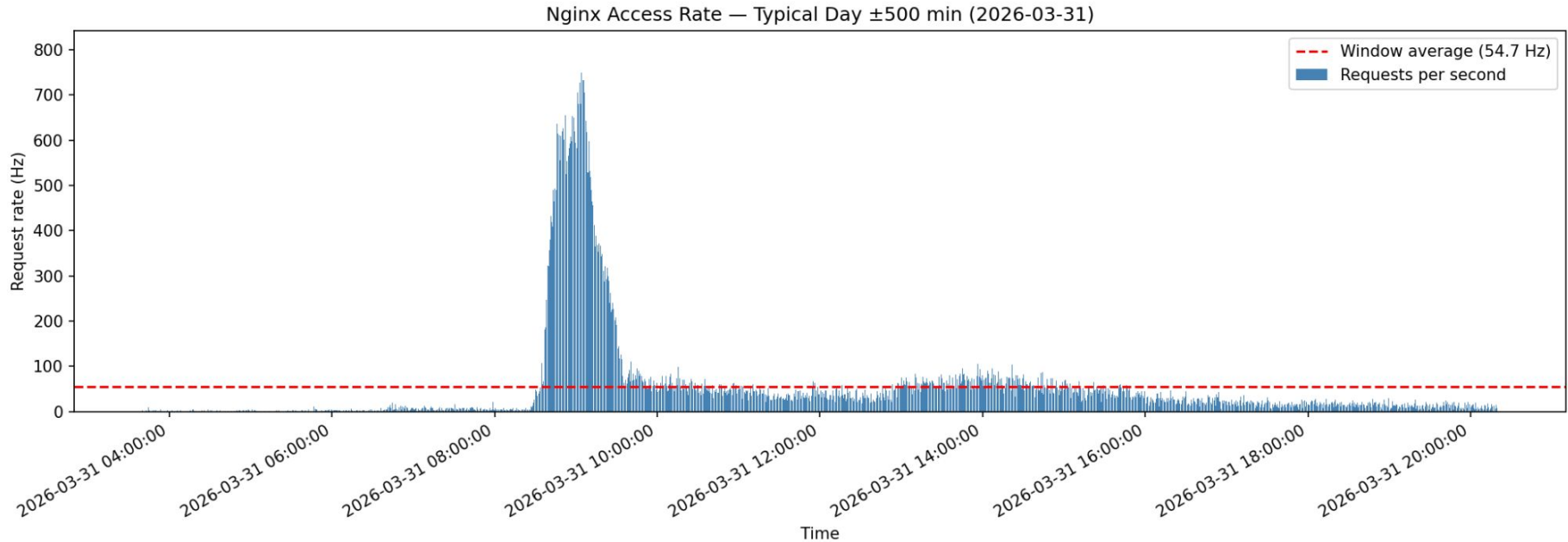


Belle II conditions data daily average access rate

Nginx Access Rate — Daily Overview

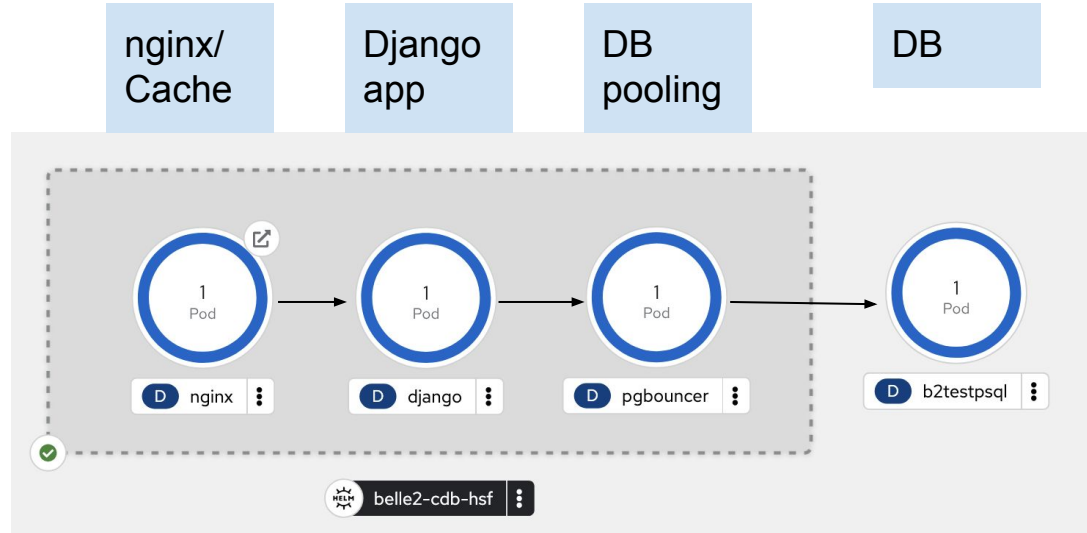


Belle II conditions data access pattern



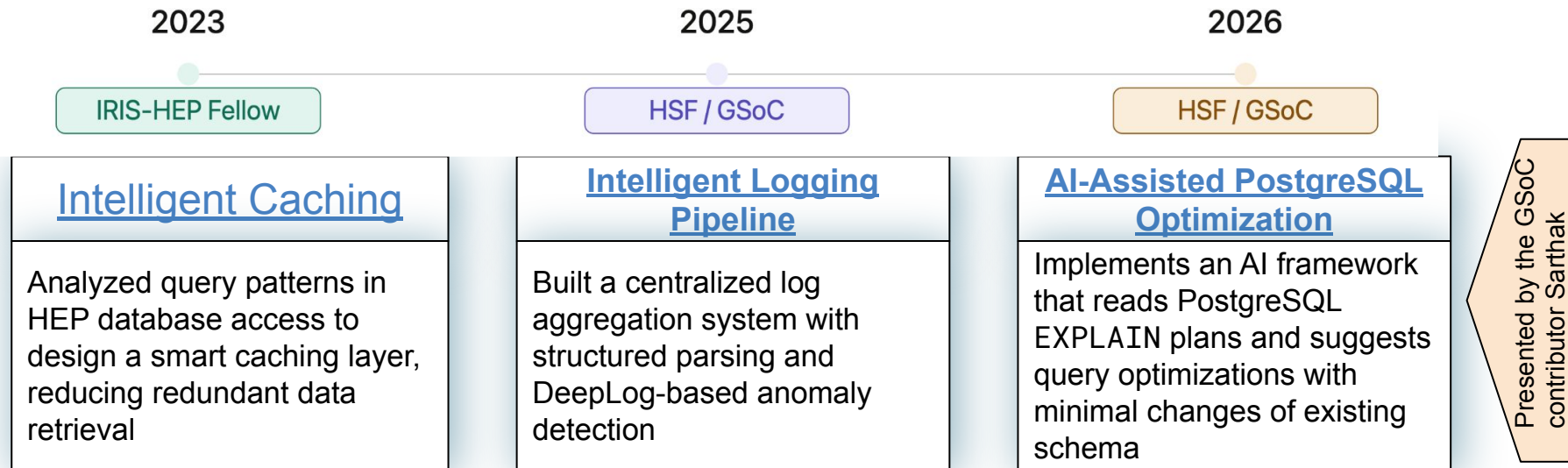
Belle2 HSF CDB test instance

- Test instance at OpenShift@SCDF BNL
- Minimal deployment - 4 pods
- Used for the conditions data migration and tests
- Will be gradually reconfigured for production



An ecosystem of projects around HSF CDB

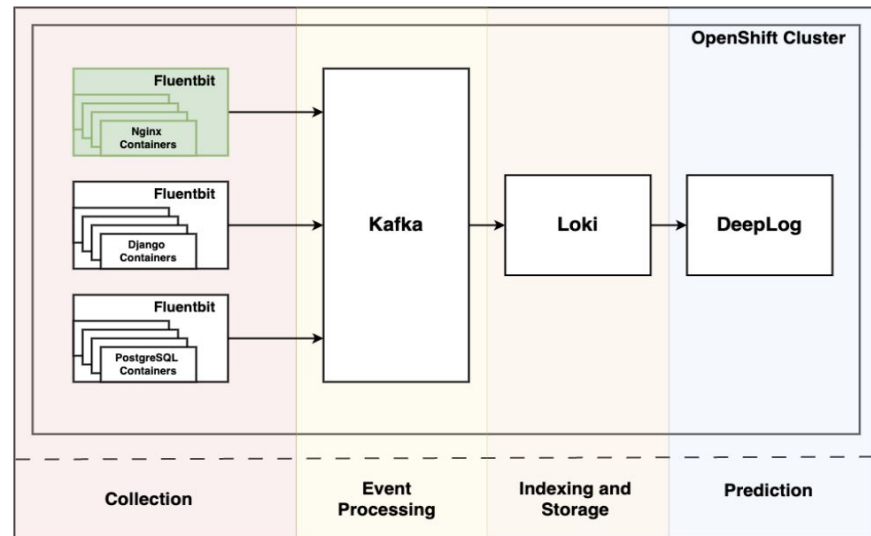
HSF/GSoC and IRIS-HEP projects



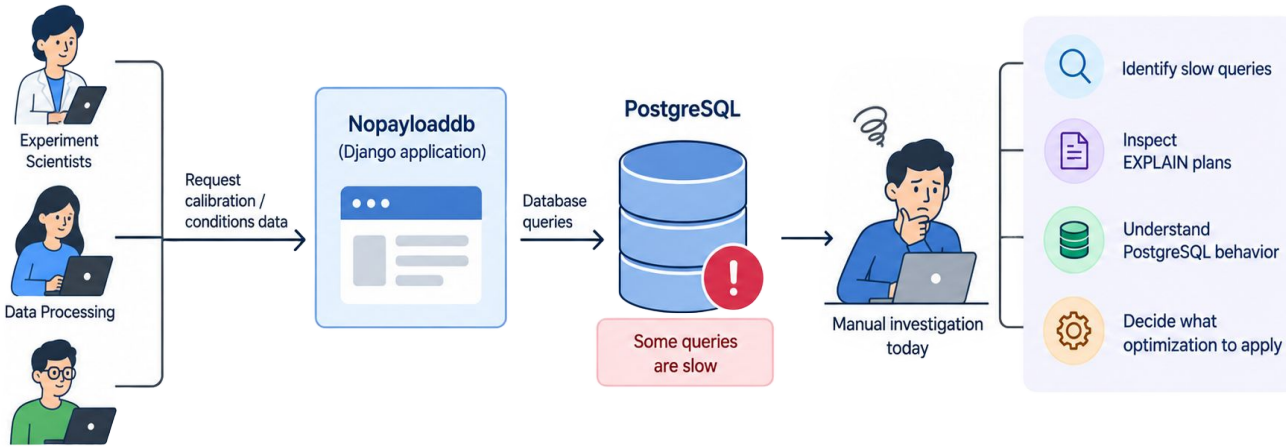
- **Applying data-driven and AI techniques** to make HEP computing infrastructure more efficient
- **Growing the next generation of HEP software engineers** through HSF and IRIS-HEP mentorship programs
- **Bridging physics experiments and modern software engineering**

Advanced logging

- We have submitted the project “Intelligent Log Analysis for the HSF Conditions Database” to the [Google Summer of Code 2025](#)
- Osama Tahir proposed and already prototyped AI-assisted logging system
- We’re adopting logging aggregation part



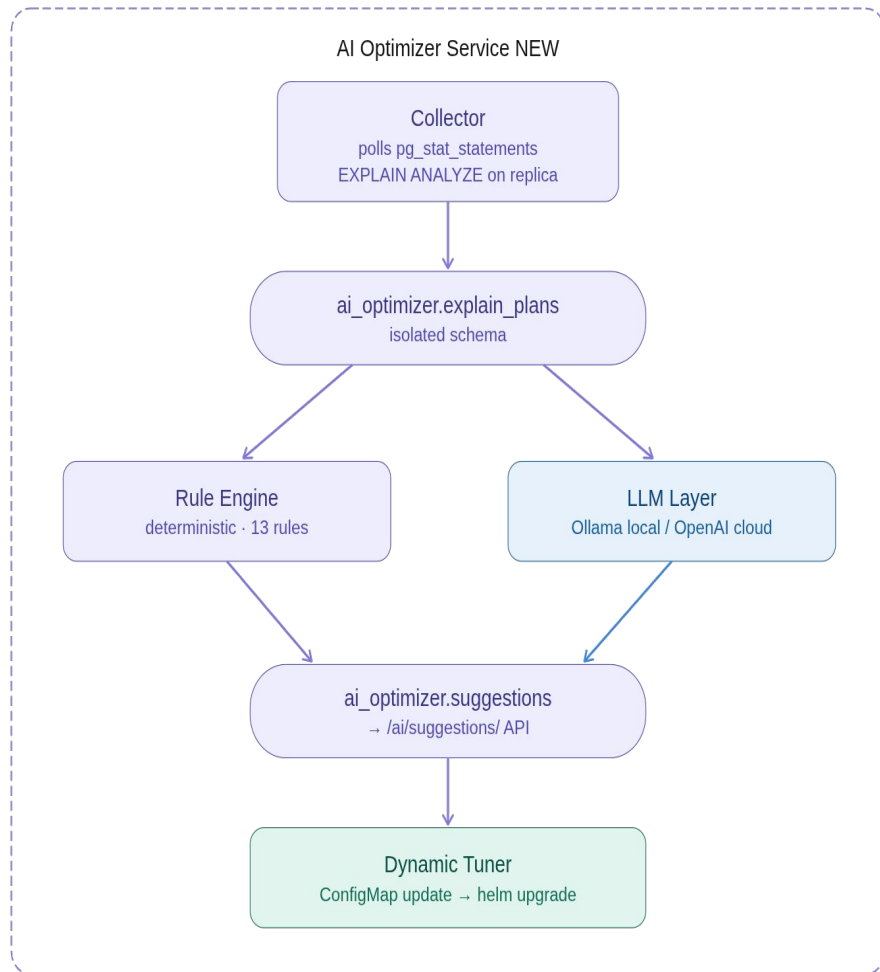
AI-assisted PostgreSQL optimization



Database queries become slower as the database and workload grow and as new workflows are added

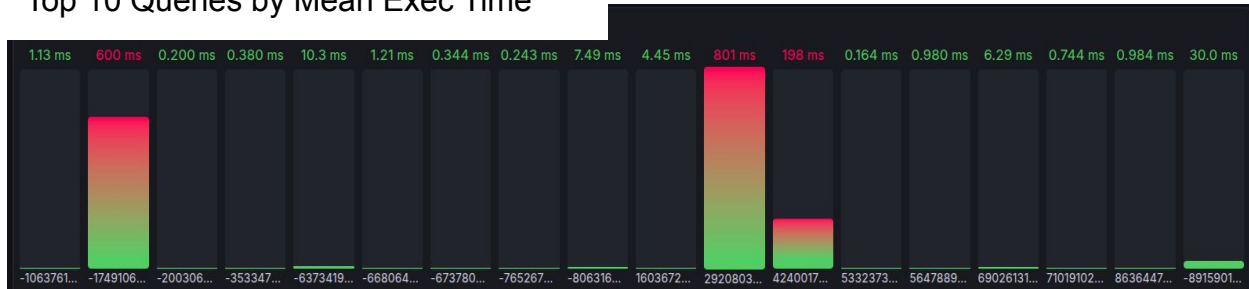
Architecture

- Rule Engine contains deterministic rules to detect common issues
 - Indexes
 - Nested loops
 - Vacuum delays etc.
- Optional LLM layer
- Dedicated API for structured optimization suggestions with administrator review and approval



Optimization in action

Top 10 Queries by Mean Exec Time



```
{  
  "id": 226,  
  "id": 226, "-8470970736517998187",  
  "queryid": "-8470970736517998187",  
  "rule_id": "R1",EX",  
  "category": "INDEX",  
  "priority": "HIGH",eq Scan detected on PayloadIOV (120000 rows across 8 loop(s)). Consider adding an index for this access path.",  
  "Seq Scan detected on PayloadIOV (120000 rows across 8 loop(s)). Consider adding an index for this access path.",  
}
```

List suggestions: http://URL/api/cdb_rest/ai/suggestions/?status=pending

Apply suggestion

```
sarthak@professor:~/Videos/nopayloaddb/nopayloaddb$ docker compose exec webapp python manage.py apply_approved_suggestions  
WARN[0000] /home/sarthak/Videos/nopayloaddb/nopayloaddb/docker-compose.yml: the attribute 'version' is obsolete, it will be ignored,  
applying suggestion id=229: CREATE INDEX CONCURRENTLY covering_idx ON "PayloadIOV" (payload_list_id, comb_iov DESC NULLS LAST);  
applied suggestion id=229  
apply approved_suggestions: 1/1 applied
```

Conclusion

- Status
 - Production since 2022 (sPHENIX): 70k+ concurrent jobs, 132k cores, peaks to 5000 Hz
 - Belle II migration underway — target 2027: about 36,000 concurrent jobs running across about 38 grid sites
 - Being evaluated by Einstein Telescope / ETAP
- Development
 - Web UI development
 - Kubernetes-native PostgreSQL cluster (GSoC 2026)
 - Anomaly detection and intelligent log analysis (GSoC 2025)

Web interface

CDBGlobal TagsPayload TypesPayload ListsIOVsCompare

Conditions Database

Browse and manage conditions data

Global Tags

Browse, search, and inspect all global tags with their statuses and payload list counts.

4062

Payload Types

View all registered payload type definitions used across the conditions database.

12153

Payload Lists

Explore payload lists and their associations with global tags and payload types.

338928

Payload IOVs

Browse all payload IOVs (intervals of validity) across the database.

8248561

GT Comparison

Compare two global tags side by side to see differences in their payload lists.

CDBGlobal TagsPayload TypesPayload ListsIOVsCompare

Global Tags

All registered global tags in the conditions database

Allfrozenlockedunlocked

Refresh

4062 items

NAME	STATUS	AUTHOR	PAYLOAD LISTS	IOVS	CREATED ▲	UPDATED
leps2013_InitialTest_GlobalTag	frozen	TEST USER	0	0	May 12, 2015 at 06:24 PM	Apr 24, 2018 at 10:10 AM
6_COPY-OF_leps2013_InitialTest_GlobalTag	frozen	TEST USER	0	0	May 12, 2015 at 06:24 PM	Aug 11, 2015 at 05:53 PM
234_COPY-OF_leps2013_InitialTest_GlobalTag	frozen	TEST USER	0	0	May 12, 2015 at 06:24 PM	Apr 24, 2018 at 12:23 PM
BelleI_GlobalTag_Tutorial	frozen	TEST USER	0	0	May 13, 2015 at 12:14 AM	Nov 2, 2015 at 05:03 AM
teddy_test	frozen	TEST USER	15	15	Jun 2, 2015 at 10:36 PM	Apr 12, 2018 at 06:48 AM
139_COPY-OF_teddy_test	frozen	TEST USER	0	0	Jun 2, 2015 at 10:36 PM	Apr 24, 2018 at 09:48 AM

 Brookhaven
National Laboratory

39

Thank you

Backup

Belle II Conditions Data service

- Belle II runs up to about 36,000 concurrent jobs across about 38 grid sites
- Timeline
 - Belle II uses B2S - Java implementation of the CDB
 - HSF CDB migration has been announced in June 2024
 - Belle II migrated to OpenShift in June 2025
 - Belle II estimates final migration to the HSF CDB in 2027

The original [Java CDB \(B2S\)](#) and its VM-based deployment on RHEL could not sustain request rates beyond 100-200 Hz

