Follow up on SciCatCON24

Highlighting discussions where DESY was involved.

Regina Hinzmann
2024-07-22

Overview

Reports.

- SciCatCON spilt in

1. **Talks**: release updates + facillity updates + PSI use cases + grants
2. **Discussions**: hands-on brainstorming sessions according to voted topics.

- Conclusion for DESY topics.

- Next SciCatCON

PSI
Welcome to PSI, the largest research institute for natural and engineering sciences in Switzerland.
MC

SciCatCON24 2.- 4. July at PSI

https://indico.psi.ch/event/15861/

Talks

Release status. Facility updates. PSI use cases.

Goals for jobs in v4.0

PSI

In the old backend, jobs were hardcoded, so not flexible to edit. Instead, we will now have a JSON that defines the job configuration:

- Configure jobs without code changes
- Refactor code to be modular and easy to add functionalities
- Enable the use of pre-built containers without modification
- Better testing, particularly integration tests
- (Phase 2) Move RabbitMQ/Kafka dependencies from the backend to independent plugins so facilities can only install what they use

Spencers slides see Day 1 *https://indico.psi.ch/event/15861/*

Roadmap

Status	Task
Done	Core Job implementation
Done	Permission Model
Done	Unit tests
Review	RabbitMQ action
In progress	Email action
In progress	URL action
Summer 2024	Early adopters: PSI, RFI, MAXIV
Fall 24	Merge to master

Talks

Release status. Facility updates. PSI use cases.

From several facility update’s/conversations I got the impression that indeed **many issues** or what they’re struggeling **overlap with ours**. Some examples:

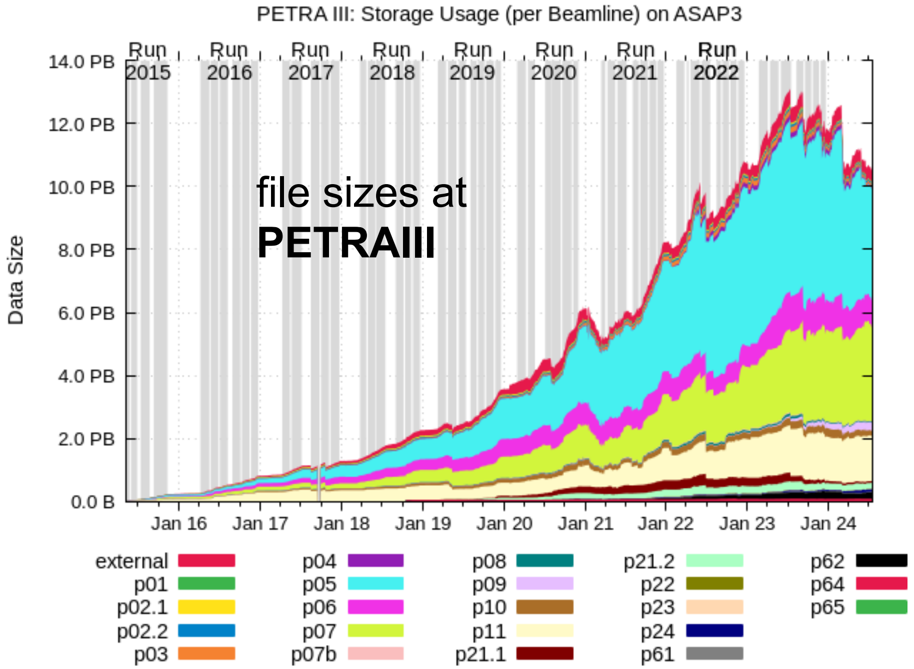
- Integrate scicat within their system in an efficient way to beamlines (e.g. MaxIV, SOLEIL)
- Ingestion and authentication (ESS): registered with ingestor account but want to add proposal.
- Many presentations (all?) use Kubernetes deployment.
- Spencers OpenEM project (goal: one scicat for all Swiss EM use cases) interests as at DESY: meta data validation; web-ingestor

From PSI use cases I learned

- Within the same user community of x-ray tomography scattering they’re **interested in** both, **raw and processed data**.
- Point raised: **scope of SciCat also reproducibility**? Becomes cheaper to redo an experiment than to save the actual data, their data estimate is ~26 PB/yr of SLS operation..

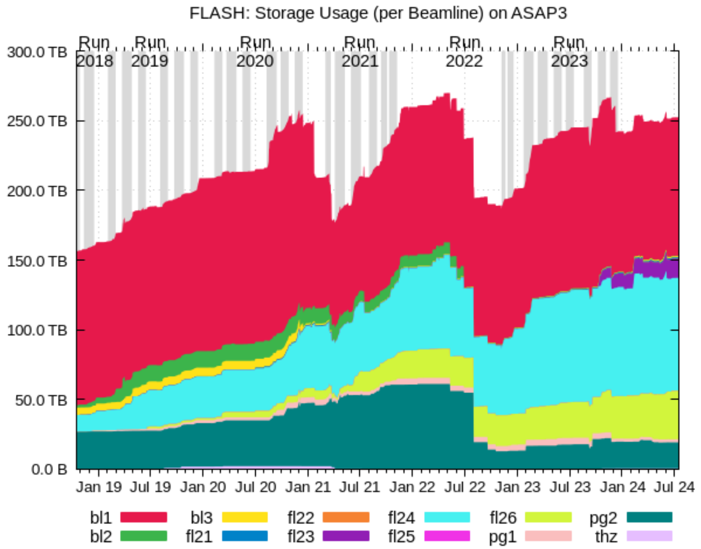
How much photon data is produced at DESY?

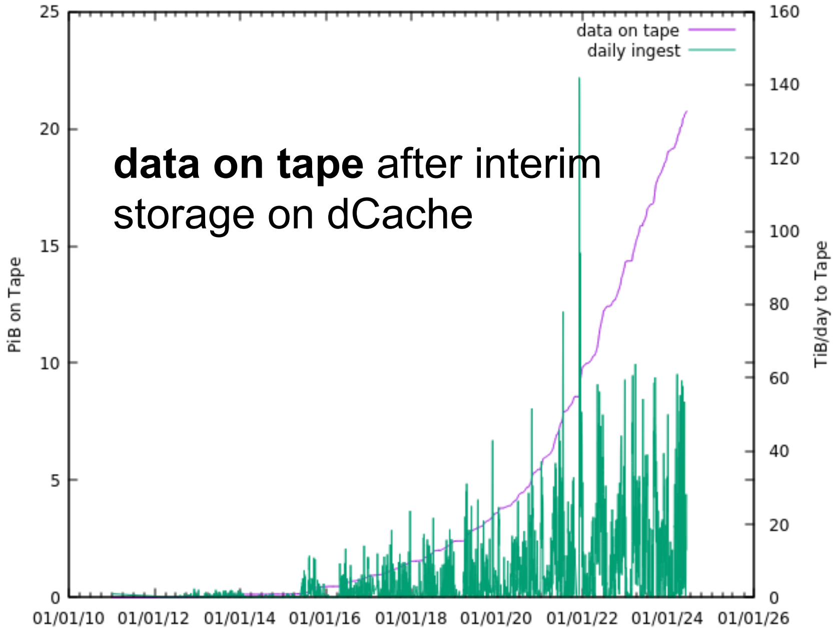
PetraIII and Flash



file sizes at **PETRAIII**

file sizes at **FLASH**





data on tape after interim storage on dCache

1. **ASAP3 (GPFS) storage** on limit is 15 PiB.
2. **Data on tape**: passed 20 PiB.

DESY FS and IT Plans for SciCat in upcoming year.

Make SciCat useful to the user

shown slide

All work targets concretely this goal: **Have SciCat run and operated in production**.

We'd like to have an idea how to setup **SciCat such that it becomes available at all DESY beamlines**.
Readiness for PETRA IV (DESYs next generation synchrotron).

- **This year's main goals**
  - An example: Make it easier to find datasets (currently use `grep` of part of filename strings in `ls` directory). Thanks to elastic search this feature makes our users happy. ✓
  - Setup of a DOI minting service for datasets just like PSI has done.

  Very helpful collaboration with PSI -- **Thank you, Carlo**! ✗

- **Next years goal**:
  - Now (2024) they are not yet connected amongst each other.
  - Set up of a performant, reliable system requires installation of monitoring tools *NOW*. Users do complain about *"slow down due to the currently massif metadatasets".* Need of quantified figures. Install monitoring tools.

  Infrastructure of performant and reliable systems (helm, Kubernetes, OpenStack + monitoring tools).

Summary of discussion points

~shown slide

How do other labs…

- Priorties for this and next year

1. Landing page prototyping
2. Github issues (most relevant maybe atomic patches)
3. Performance monitoring large scale setup.

Max answer regarding DataCollection: If the institute thinks and implements its use case, there there will be progress.

- …view some of our issues mentioned, how useful would they be for you? Eg.
  - update of proposal class,
  - introduction of DatasetCollection,
  - extend DOI fields according to DataCite
- Do we want a **PID/DOI also for proposals**?
  - Would it have an impact on dataset DOIs?
  - What is scientifically significant for having PIDs for proposals?
- Where do other labs see the role of a ***data curator***?
- We have a lot in the pipeline with very little expertise (and man power) yet.

Hands-on brainstorming sessions

2 rounds. 6 most wanted topics.

1. visualisation plugins, web-ingestor/validation, pyScicat or SDK and openAPI development
2. documentation, group/user mgt system, FE UX review

other discussed topics

- mono repository: no obvious disadvantage compared to many (immediate and obvious) benefits
- ..

Procedure:

After identifying 6 topics, the groups brainstormed/identified the toDo’s necessary to solve it. This took two sessions…

Group: pySciCat turned into SDK development

What is an SDK and why is this so important?

Goal: autogenerate *software development kits* (SDKs) for different languages using openAPI generators (language priorities 1. ts, 2. py, 3. other: go, java)

What is an SDK and why do we want it? source is this.

- *Is a set of tools written in one or more programming languages that make it easier to use a service or API.*
- *An SDK is a great abstraction layer.* It is written in the language the developer uses, and provides language idiomatic ways to call the API, managing things like HTTP calls, the mapping of JSON to objects, authentication, retries, etc.
- *A good SDK drastically reduces the code you have to write.*

Procedure for SDK dev in SciCat:

see `https://github.com/orgs/SciCatProject/projects/16`

Groups: meta data validation and visualisation

What is an SDK and why is this so important?

Spencer and Linus and others had a good conversation, we potentially hear about it later. Linus, thus DESY, committed to contributions by the work of Anjali A and potentially Julia K (CAU).

Groups: meta data validation and visualisation

Commitment to contributions by Linus (FS-EC) and Igor (FS-SC)

meta data validation

- led by Spencer B

visualization

- led by ? (Ian B or Laura S)

General Updates

- Since 2024-07-17 we run at P08 with the new way providing IDs:

  old: beamtimeID

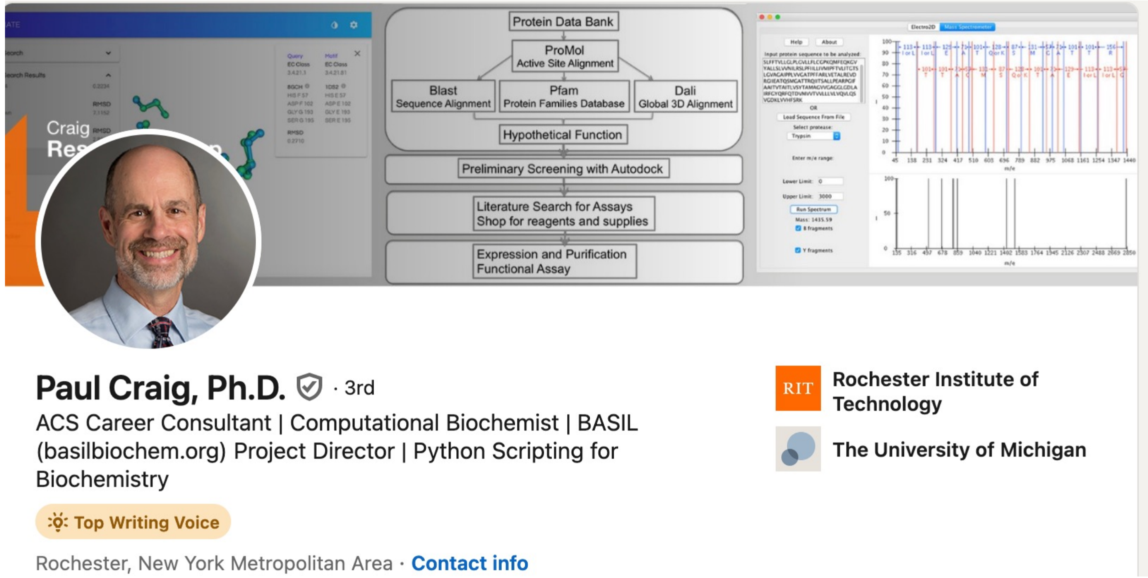
  new: DOOR_proposalID.beamtimeID

I’d like to extend it as default at other beamlines, especially FLASH, as well

– yes?

SciScat survey

- users wanted who have an opinion on SciCat usage. Please come forward! Deadline: Mid September 2024.


Craig
Res
Protein Data Bank
ProMol
Active Site Alignment
Blast
Sequence Alignment
Pfam
Protein Families Database
Dali
Global 3D Alignment
Hypothetical Function
Preliminary Screening with Autodock
Literature Search for Assays
Shop for reagents and supplies
Expression and Purification
Functional Assay
RIT

Paul Craig, Ph.D. · 3rd

ACS Career Consultant | Computational Biochemist | BASIL (basilbiochem.org) Project Director | Python Scripting for Biochemistry

Top Writing Voice

Rochester, New York Metropolitan Area · Contact info

Rochester Institute of Technology

The University of Michigan