

ASAPO: A high-speed streaming framework to support an automated data-processing pipeline.

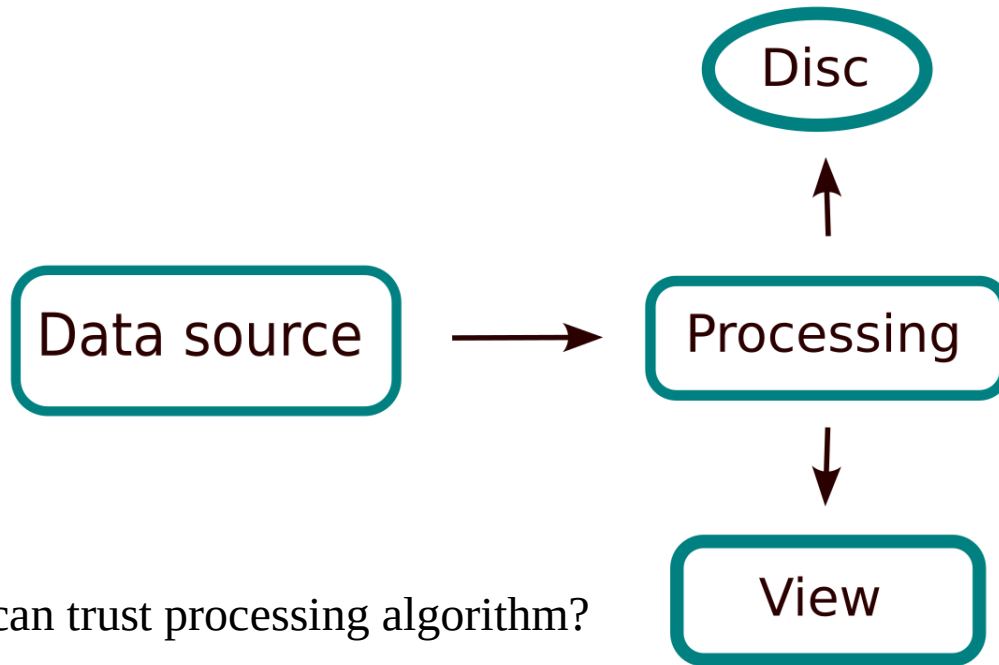
Mikhail Karnevskiy

DESY IT

NOBUGS 2024

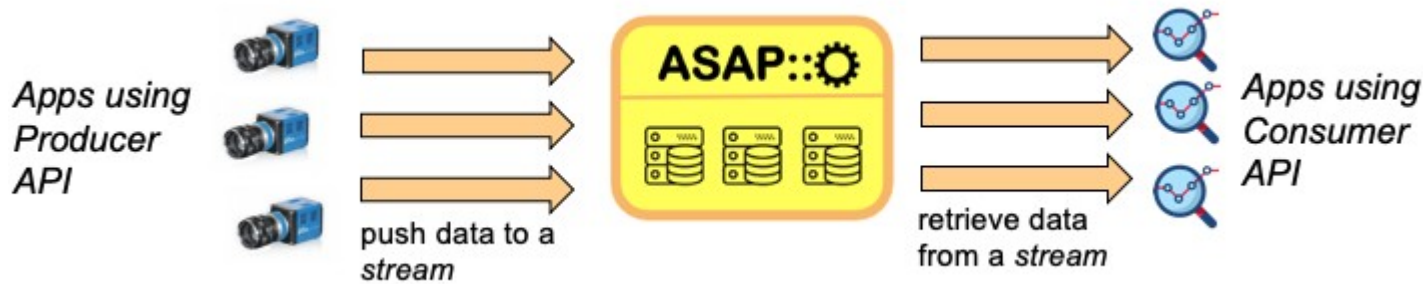
25 Sept, 2024

Motivation



- How we can trust processing algorithm?
 - Re-process?
- How easy it can be modified?
- How robust is data flow?

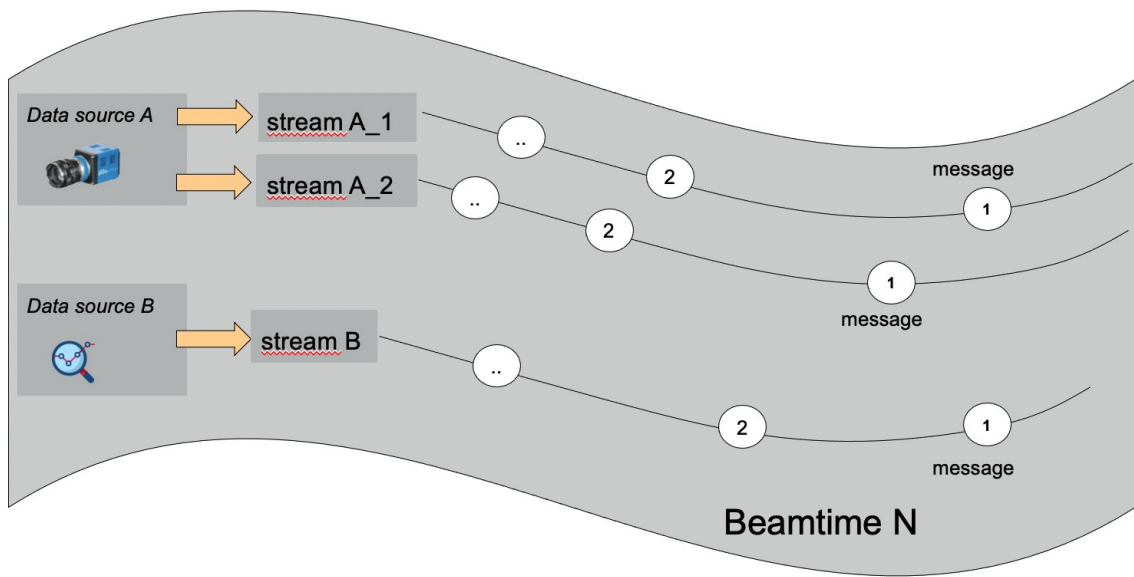
ASAPO introduction



- High-bandwidth communication between state-of-the-art detectors, the storage system, and independent analysis processes across DESY facility
- In-memory data transfer with optional caching on disc
 - Large, scalable in-memory cache
 - Saving data to disc as service
 - Deliver data from disc
- Easy to use C++/Python interfaces:
 - Producer: sends data to ASAPO
 - Consumer: get data from ASAPO

Data in ASAPO

- Messages are uniquely identified by beamtime, data-source name, stream name, and index
- Data-source name and stream names are arbitrary strings defined on client side
- Messages are indexes from 1 to N.
- Each message contains a binary data blob and a JSON metadata.
- Data is stored in memory-cache and on disk (optionally), metadata is stored in database
- Several data-sources can be combined in asapo to a dataset



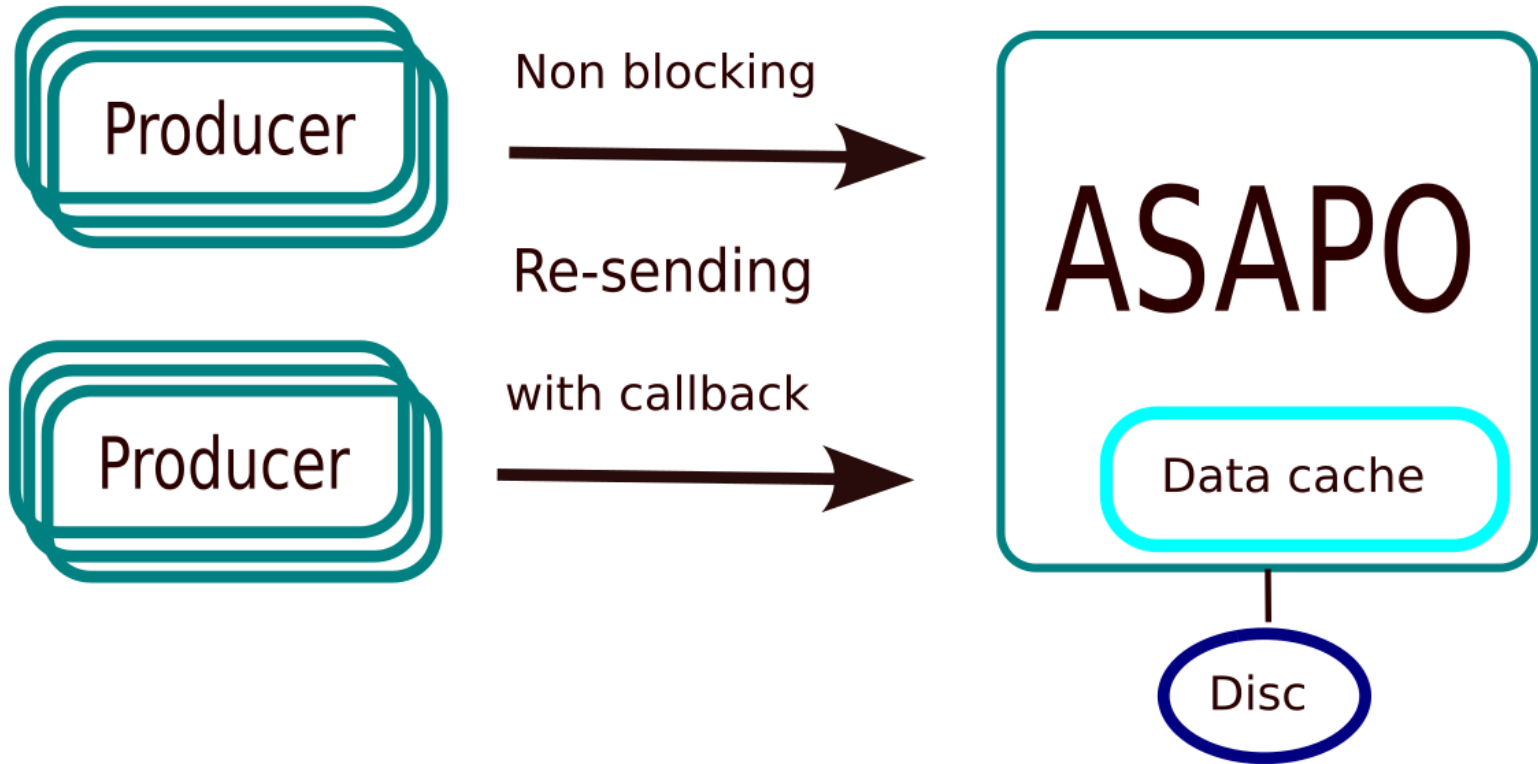
Message

Metadata

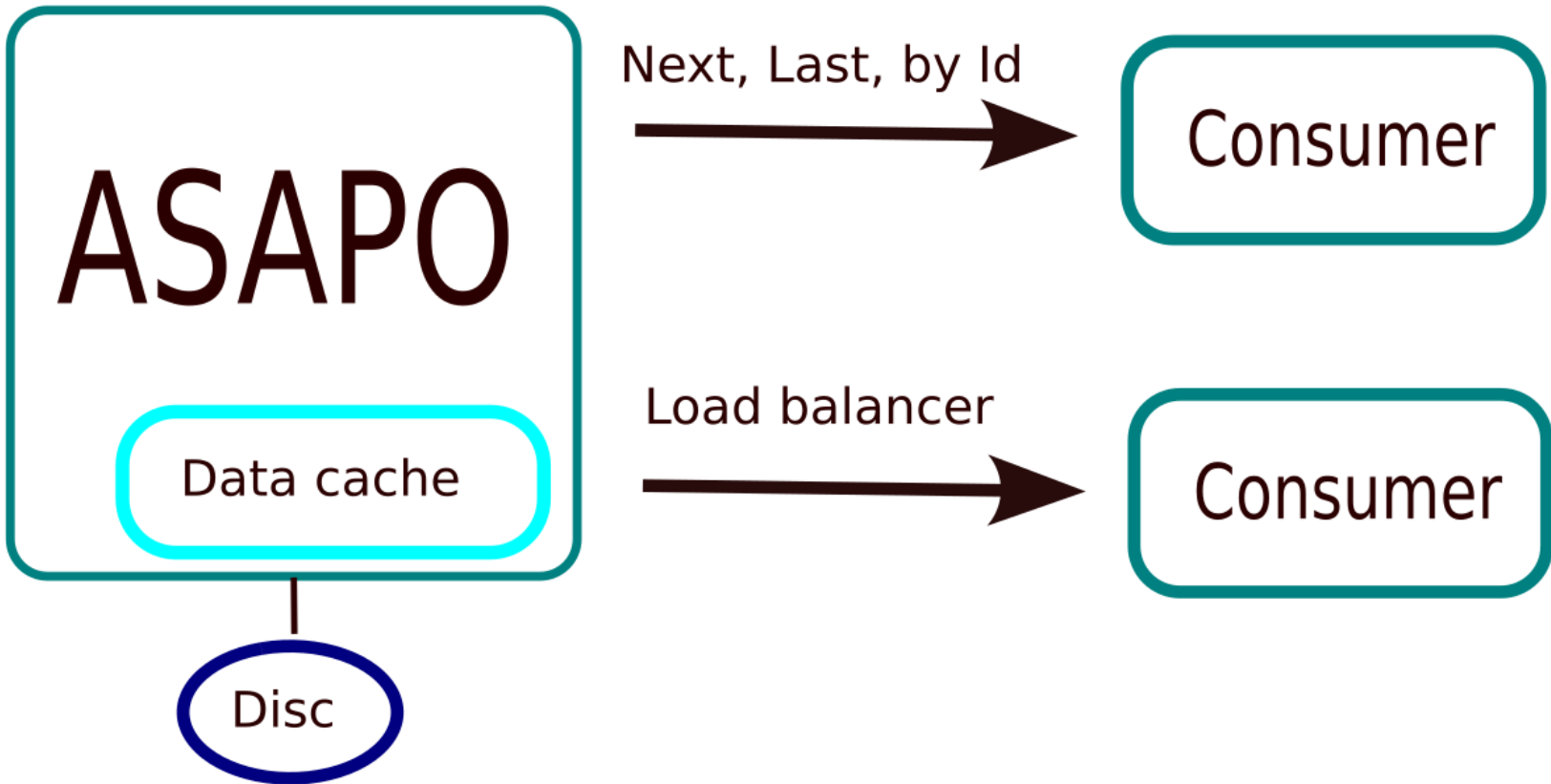
Id: 1
User meta (json string)
Other fields

Data blob

Ingest data to ASAPO

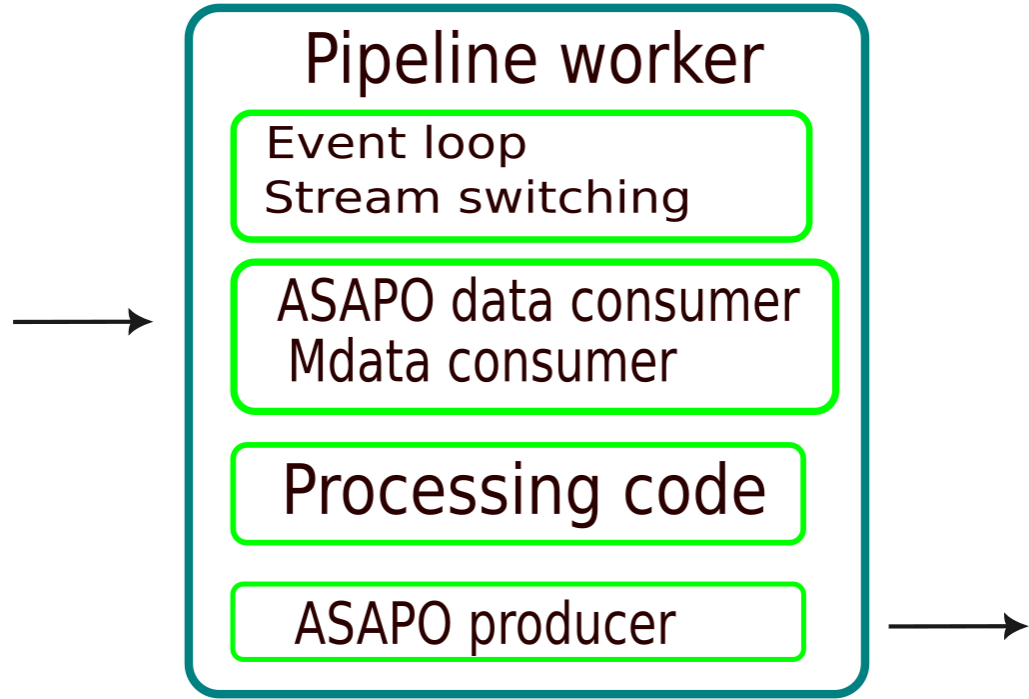
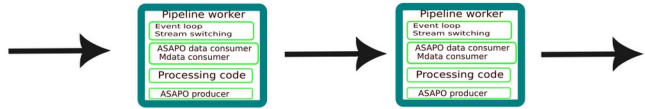


Retrieve data from ASAPO

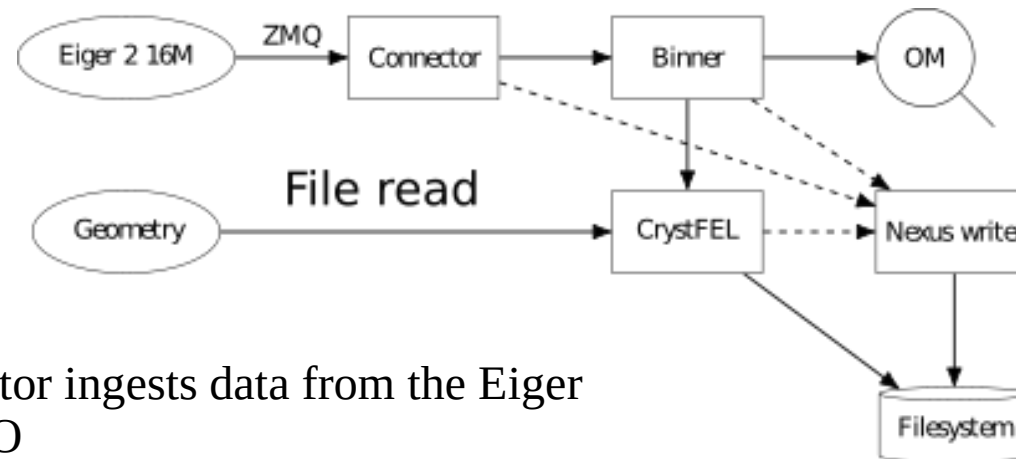


ASAPO-based pipeline

- Uses Python or Cpp ASAPO clients
- Data processing with a chain of workers
- Communication via ASAPO service
 - Workers does not know each other, but knows data-source to retrieve.



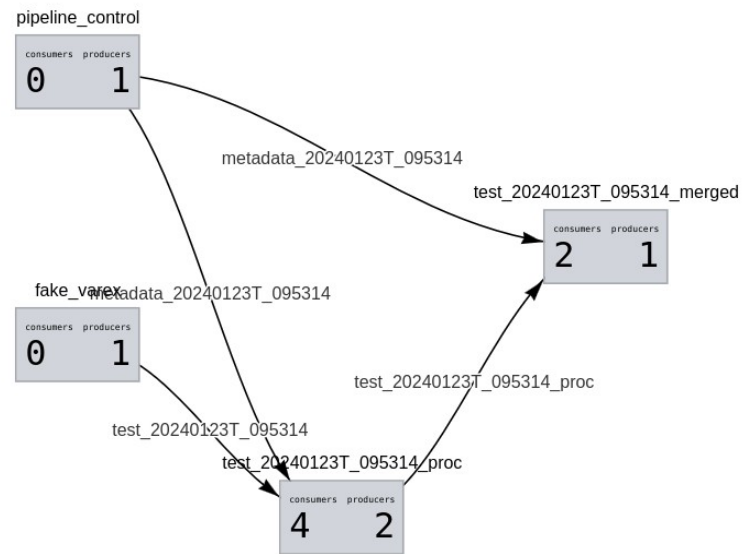
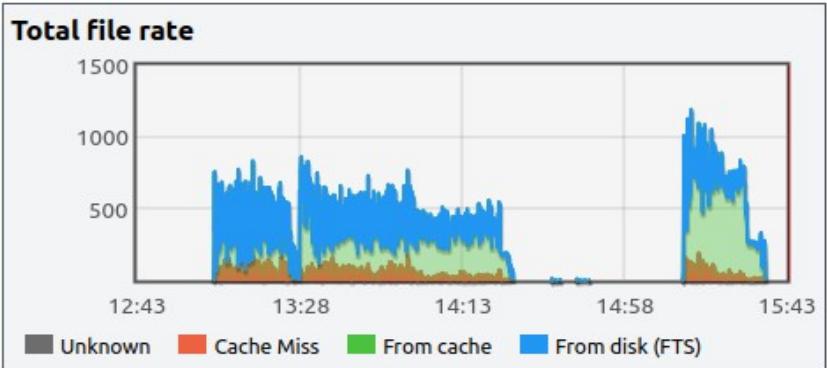
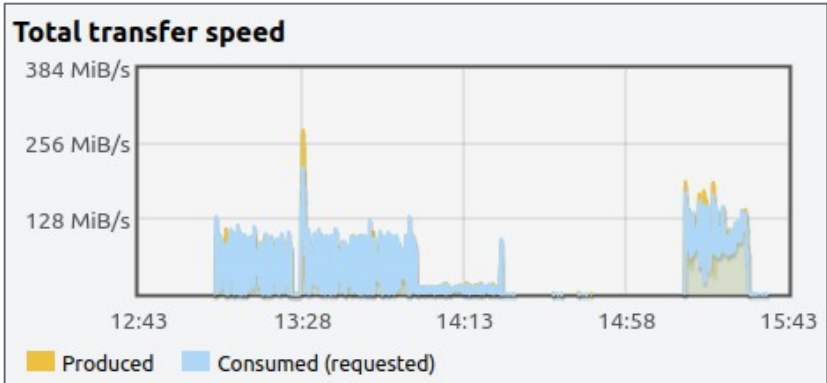
Pipeline example



- ASAPO-Eiger-Connector ingests data from the Eiger ZMQ stream to ASAPO
- (Optional) Binner reduces images resolution to speed up later processing steps
- CrystFEL for peak search, indexing, and integration
- OM (OnDA Monitor) for live visualization
- Nexus writer can write raw, binned, or filtered (hits only) images and metadata to disk, depending on which data source it is connected to
- Currently, geometry/analysis results are read/written by CrystFEL from/to disk directly

Monitoring

- Web service to visualize data-flow is ASAPO
 - Show message rate and file rate
 - Show delay in data processing
 - Visualize the pipeline topology
- Further development is required



Pipeline step	Delay time
Integration	2 secs
Writer	0.597 secs

Try ASAPO

- Git at DESY: <https://gitlab.desy.de/asapo>





- Pipy client packages. 



- Docs: <https://asapo.pages.desy.de/asapo/>

ASAPO standalone service:

- Single docker with all asapo services
- Monitoring via Grafana  
- Limited functionality (not scalable)
- Fully functional API



Summary

- Several data-processing pipelines based on ASAPO are used at different experiment-stations at DESY Petra-III facility.
- Reliable performance and support have been demonstrated. Over 100 Hz continuous message rate with multiple MB message size.
- Online data processing improves the performance of the beamline, simplifies data post-processing and enable data reduction.
- Asapo can be useful at other facilities.