prosEO – A Cloud-Native Processing Framework for Astronomy Data Pipelines

Thomas Bassler[2], Nicolás Corti Meneses[1], Peter Friedl[1], Anett Gidofalvy[1], Frederic Raison[1], Maximilian Schwinger[1]

1 - Deutsches Zentrums für Luft- und Raumfahrt (DLR), German Aerospace Center, Earth Observation Center | German Remote Sensing Data Center | International Ground Segment (IBS)
2 - Dr. Bassler & Co. Managementberatung GmbH

prosEO processing facility — DLR Earth Observation Center

prosEO is a **mission-agnostic, open-source, flexible, cloud-native processing** control system designed to handle the **end-to-end data processing** from raw ingestion to user-level products. Developed by the German Aerospace Center and partners for Earth Observation data and used across multiple missions, it can be advantageously ported to Astronomy.

Key Features

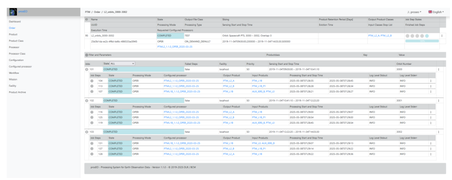
- **Multi-Mission Support**: process data from different satellite missions seamlessly
- **Cloud-Native & Modular**: scalability and easy integration into modern IT environments
- **Service-Oriented Architecture**: components are well-separated, allowing targeted customizations
- **Automatic Workflow Generation**: unlike traditional fixed input-output chains, prosEO can dynamically create processing workflows based on the requested output products
- **Metadata-Driven**: all production steps are tracked and queryable
- **Storage Agnostic**: can be integrated with cloud storage or on-premises data lakes
- **Interoperability**: achieving full interoperability with Copernicus / ESA

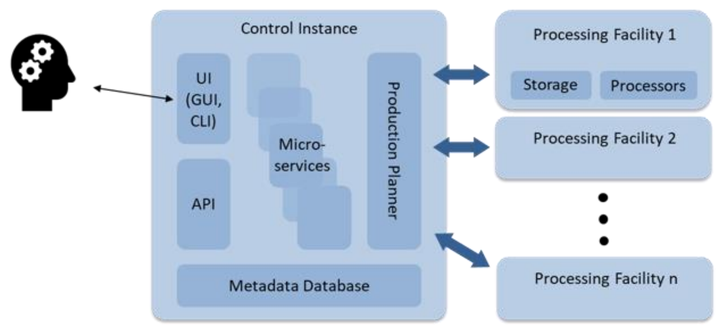
Future Plans

- Adapt prosEO to HPC
- Optimized choice of infrastructure, considering timeliness, cost, data availability and other constraints (Cloud, HPC)
- Take advantage of the Open Source Community leverage

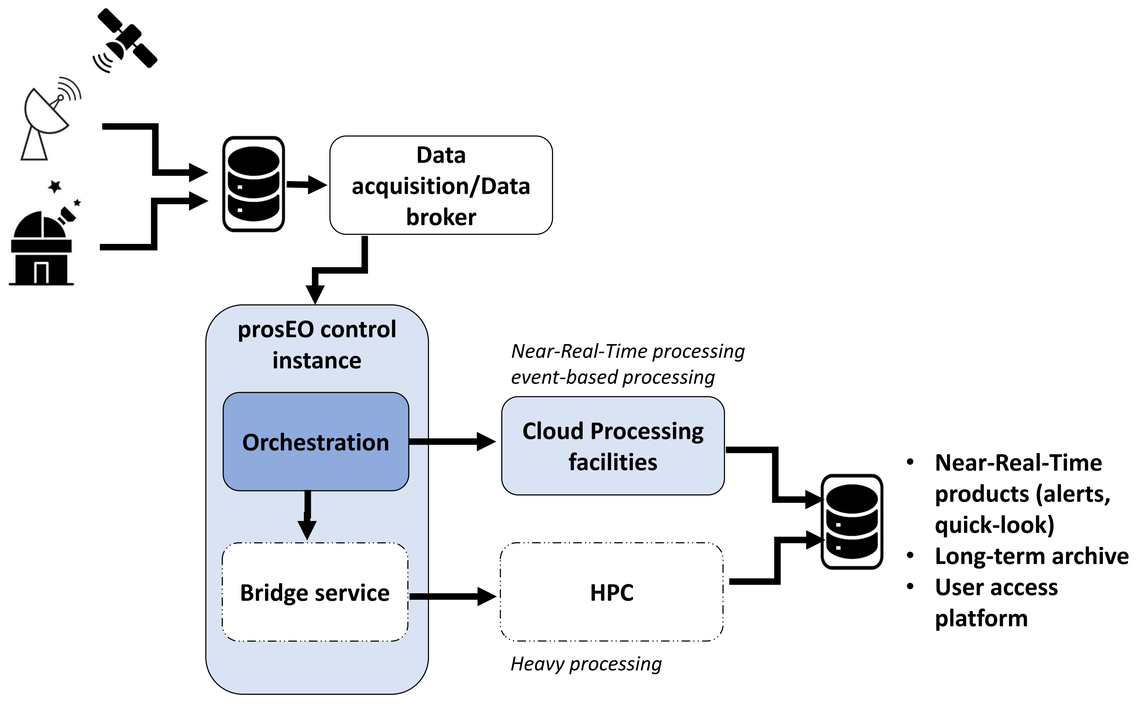
Current System Architecture & Core Components

- Mission Planning Interface
- Product Metadata Management
- Production Request System
- Processing Execution Engine





Global Architecture Scheme



In Usage

Mission	Operational scenario
INPULS	Systematic Production
Sentinel-5P	Bulk Reprocessing
Sentinel-1 A/B/C/D	On-demand Production
Sentinel-4 Commissioning	On-demand Production

Coming up

- CO2Image
- TanDEM-X / TerraSAR-X (Post Mission Reprocessing)

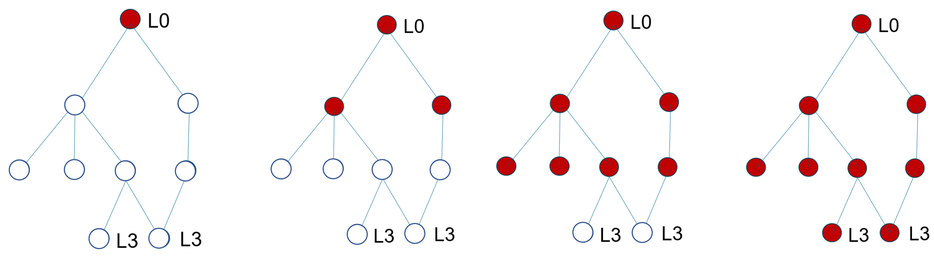
Highlights & Metrics

- Virtually unlimited horizontal scalability: 4.5 years of Sentinel-5P mission (1.6 PB) were reprocessed in 6 months with 4 virtual data centers and 125 worker nodes each
- Modular updates allowed for mission specific plug-ins with no impact on the prosEO core and other projects
- Enhanced traceability with metadata
- Integration into various infrastructures: HPC, Cloud

Processing Management

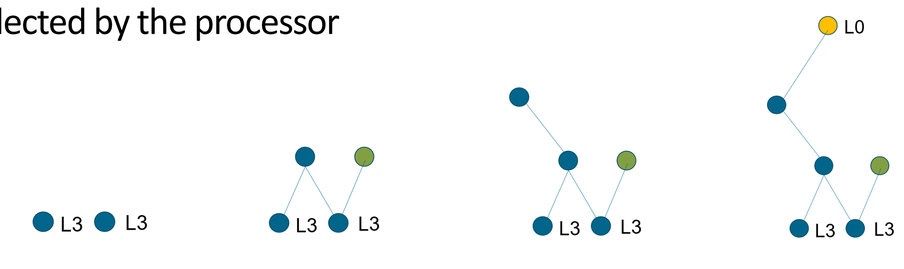
Workflow based : processing the data forward to the end products

At each processing step the processor waits for the input data and creates the next outputs. Step by step: L0 → L1 → L2 → L3

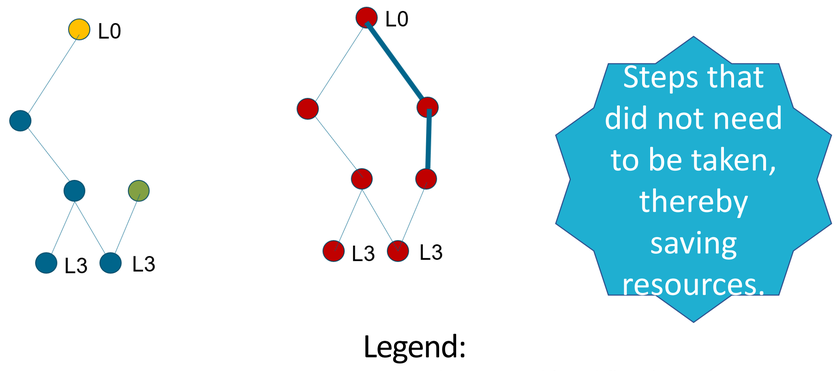


Dependency based: backwards from the end product

Each processing step is a standalone job in itself constructed on: the requested output, the processor version to be used, the input data selected by the processor



… From Effective to Efficient



Legend:
- Forward generated (workflow based)
- Archive
- Cache
- Backward generated (dependency based)

Technical stack:

docker · docker Compose · kubernetes · Spring Boot · RAML · Java · Maven · podman

prosEO processing facility on Github:

DLR Deutsches Zentrum für Luft- und Raumfahrt German Aerospace Center