Optimized Data Access from and to a Long-term Archive for the Processing of Time Series

M. Wolfmüller, S. Holzwarth, S. Asam, S. Kiemle, D. Krause, A. Scherbachenko

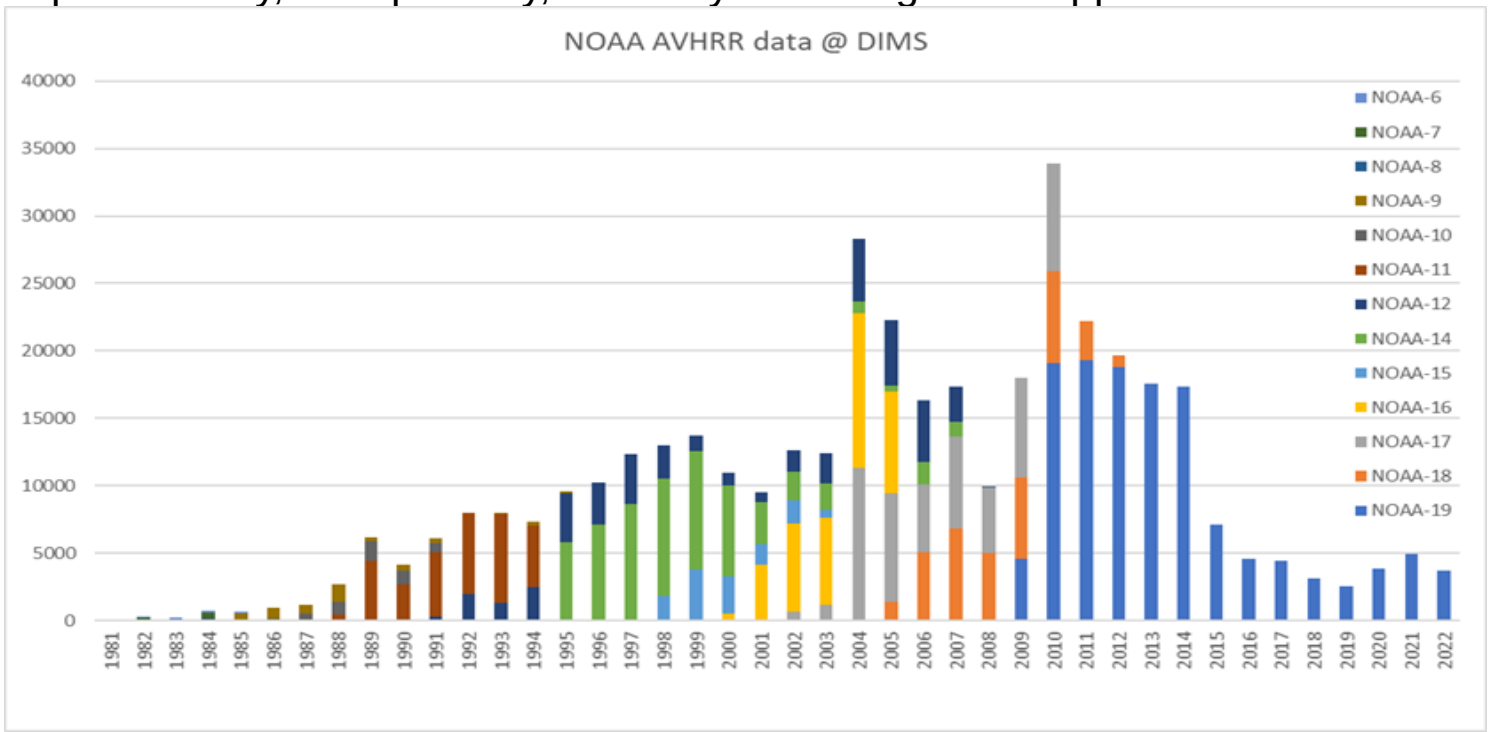
German Aerospace Center (DLR), German Remote Sensing Data Center (DFD),
Oberpfaffenhofen, D-82234 Weßling, Germany

PV2023 CERN
Geneva, Switzerland, May 2-4, 2023

Context and Main Aspects and Requirements for Data Access

- The production of a well calibrated and harmonized 40-year-long time series based on the AVHRR series of instruments (three versions of instruments) operated on 14 different NOAA platforms
- sensor degradation, different spectral responses and different radiometric drifts of the AVHRR instruments as well as orbit drifts of the NOAA satellites, have to be corrected
- overarching goals for the data base and the processing within the TIMELINE project are consistency, reproducibility, transparency, reliability and the generic approach.

Number of used L0 AVHRR Scenes per NOAA Sensor

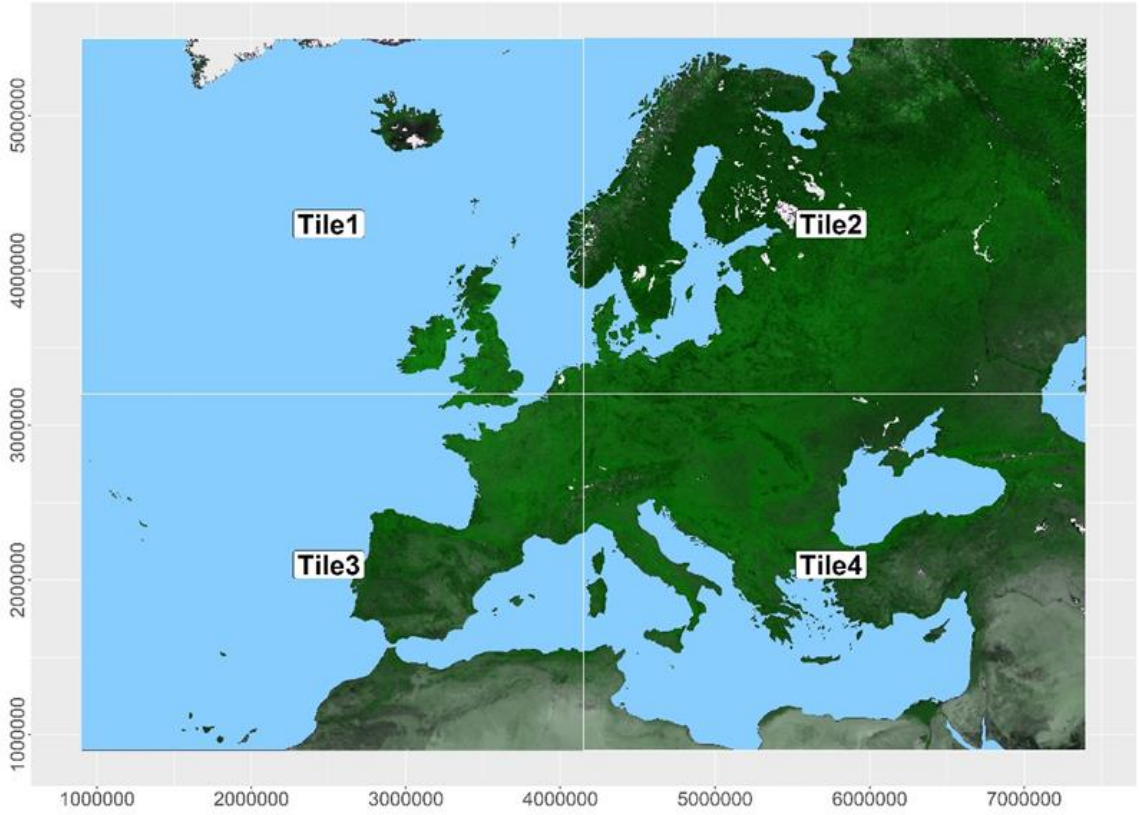


DLR

TIMELINE

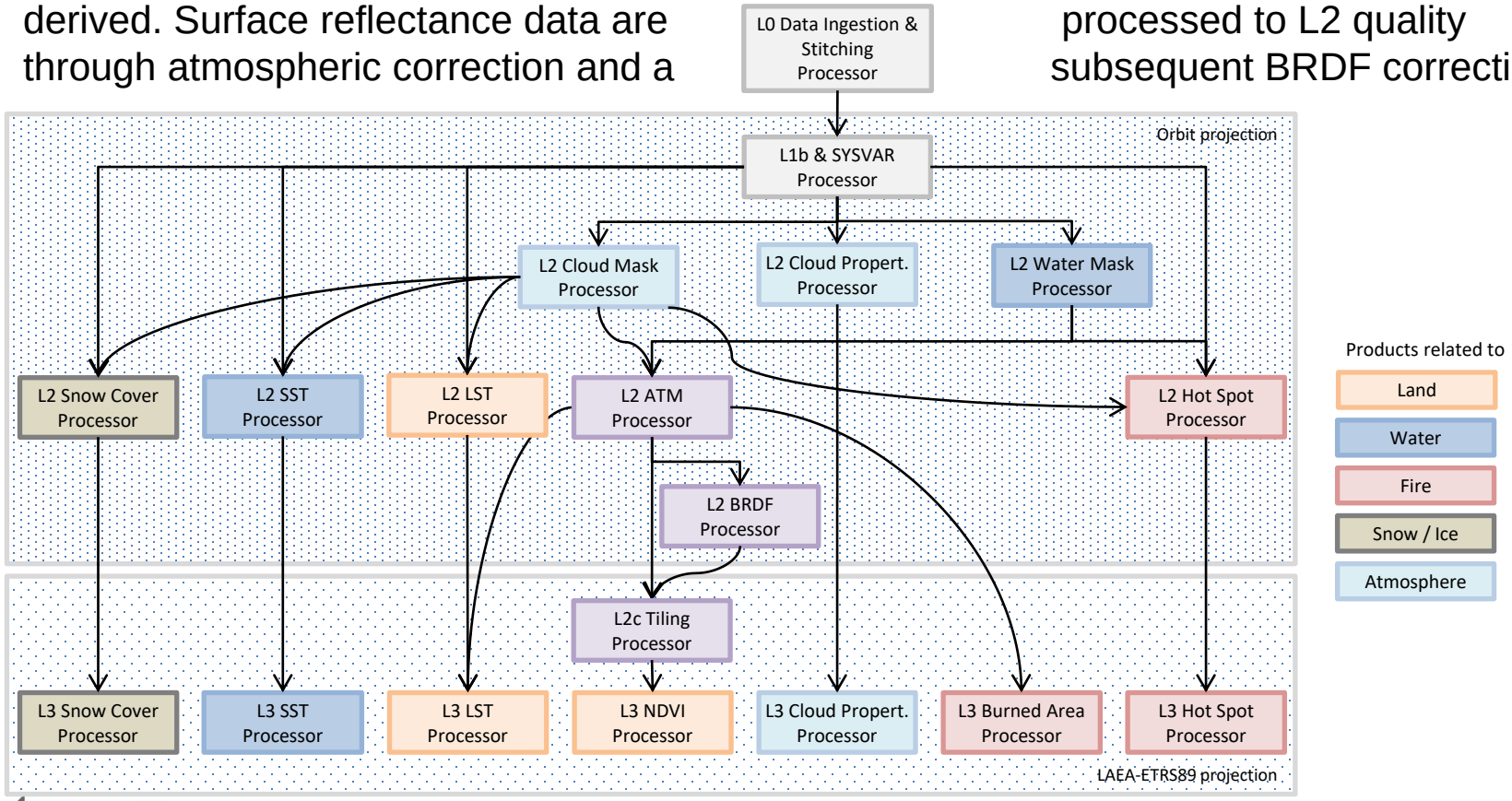
Context and Main Aspects and Requirements for Data Access

- The map products are derived at 1 km resolution for Europe and North Africa
- L1b and L2 products are scene-based data in orbit-geometry,
- L2c and L3 data are projected to LAEA-ETRS89 and gridded, in daily, 10-day and monthly temporal
- TIMELINE Area is overlaid with four L2c / L3 Tiles
- a careful reprocessing and version handling has been established

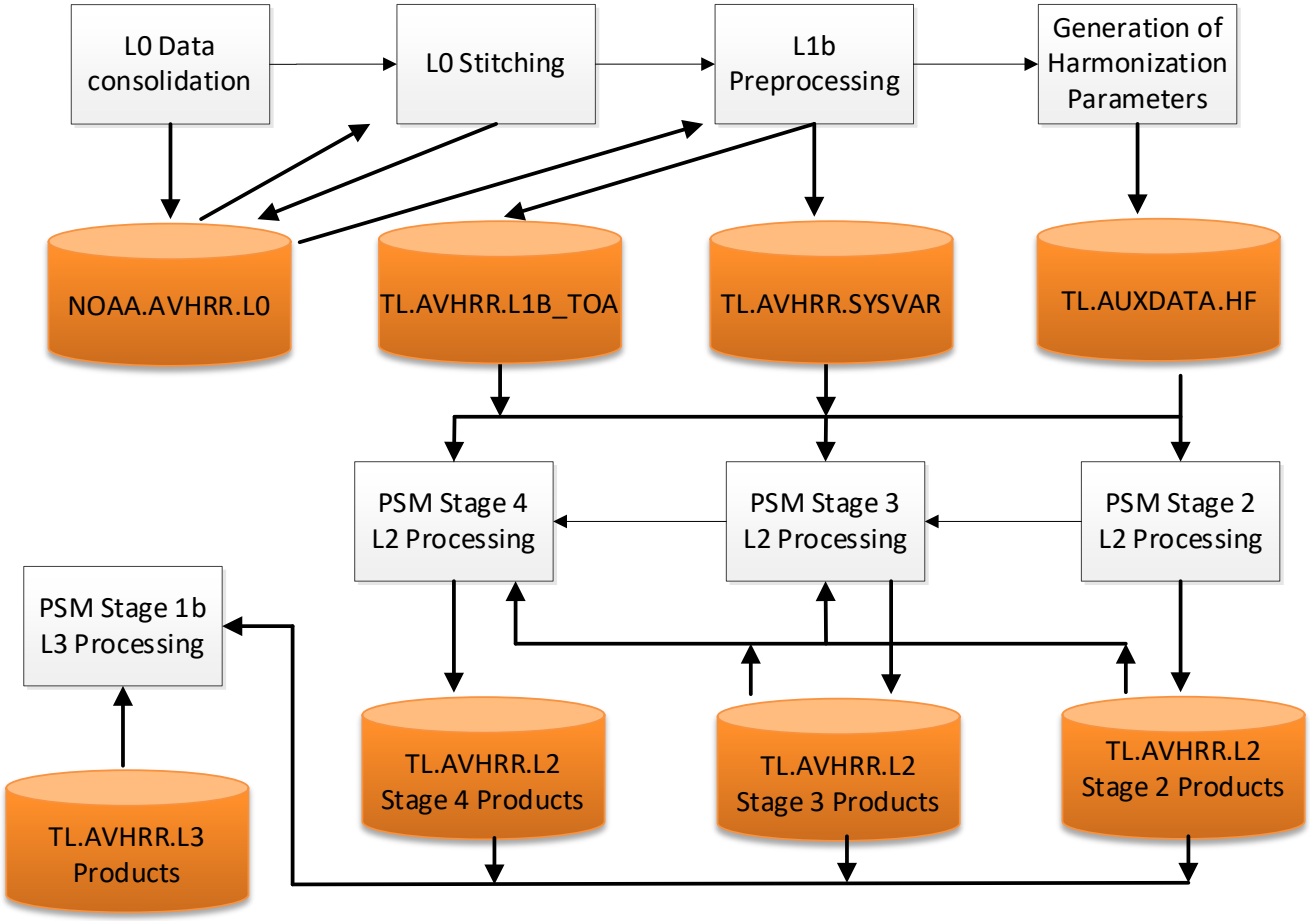


Context and Main Aspects and Requirements for Data Access

- Generation of consistent, reproducible, reliable and global change relevant variables for level L1b, L2 and L3
- Consideration of the product dependencies when using the different processors
- Based on the L1b data, water and cloud masks and a range of L2 thematic products are derived. Surface reflectance data are processed to L2 quality through atmospheric correction and a subsequent BRDF correction.



Main Project Scenarios and Production Sequence from L0 to L3 Products



Main Project Scenarios

THE TIMELINE PROCESSING SYSTEM HAS TO SUPPORT THE FOLLOWING PROCESSING SCENARIOS

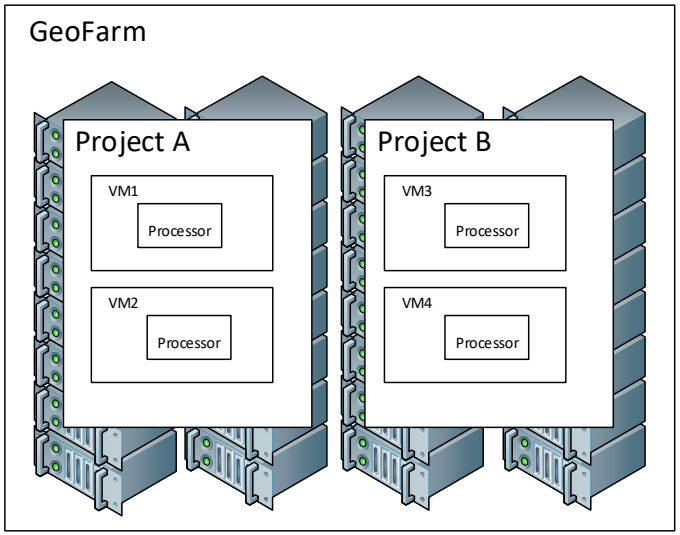
- L0 Scenarios
  - L0 data consolidation
  - L0 stitching
- single scene processing scenarios
  - the L1b preprocessing (TL.AVHRR.L1b_TOA, TL.AVHRR.SYSVAR),
  - the L2 processing of several L2 products (TL.AVHRR.L2_CM, TL.AVHRR.L2_CP, TL.AVHRR.L2_WM, TL.AVHRR.L2_ATM, TL.AVHRR.L2_BRDF, TL.AVHRR.L2_SC, TL.AVHRR.L2_LST, TL.AVHRR.L2_SST, TL.AVHRR.L2_HS)
- Gridding scenarios
  - the L2c processing of several L2c products (TL.AVHRR.L2c_ATM, TL.AVHRR.L2C_LST, TL.AVHRR.L2C_SST)
  - the final L3 processing of several L3 products (TL.AVHRR.L3_CP, TL.AVHRR.L3_LST, TL.AVHRR.L3_SST, TL.AVHRR.L3_SC, TL.AVHRR.L3_NDVI, TL.AVHRR.L3_BA, TL.AVHRR.L3_HS)
- While L0 data are unreferenced orbit segments before stitching, L1b and L2 are scene-based data in orbit-geometry, and L2c and L3 data are projected and gridded composites, with one file per tile. The study area is divided into four separate tiles, which are distinguished in the file name through the abbreviations "t01"-"t04“.

Main System Constraints and Requirements for the Processing System

MAIN SYSTEM CONSTRAINTS

- The TIMELINE products shall be provided as standardized NetCDF datasets.
- Usage of processing platform GeoFarm available at DLR DFD

.



GEOFARM

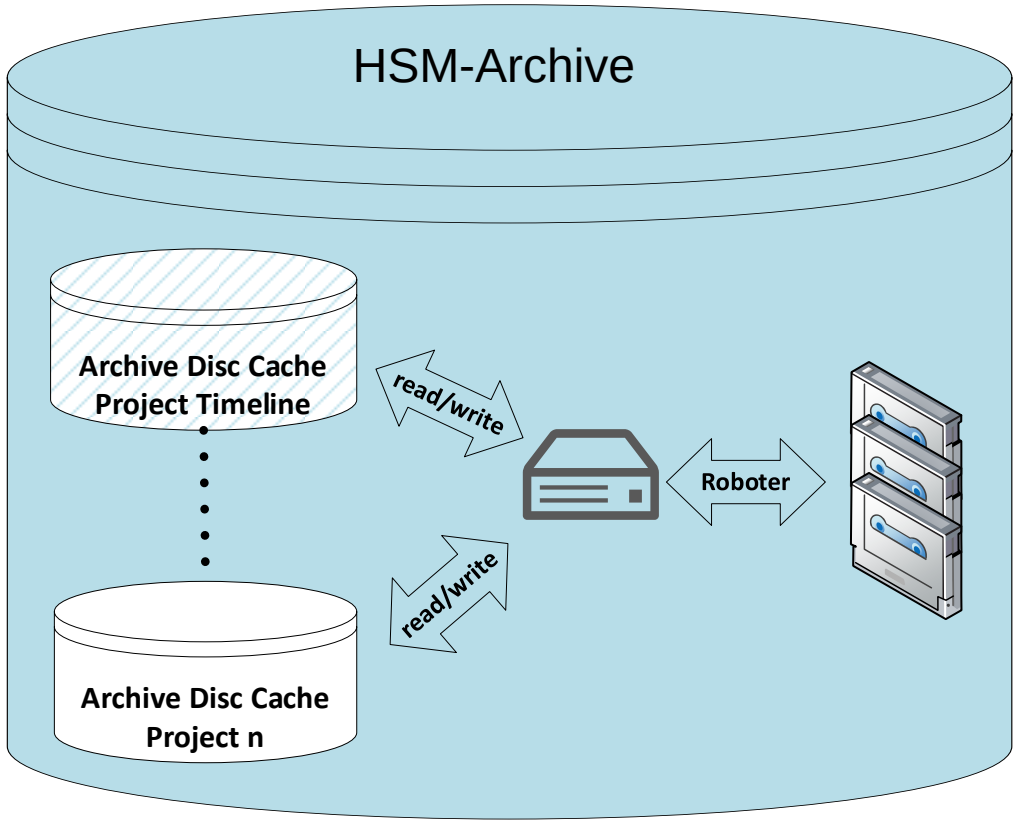
- homogenous virtualized environment (IaaS)
- provides new resources to a project without hardware reconfiguration
- is organized as a private cloud
- is based on DELL blade servers.

Main System Constraints and Requirements for the Processing System

Main System Constraints

- Usage of the DSDA Product Library and the long-term archive available at DLR DFD.

- Usage of available Tape drives
- Tapes are still a cost efficient possibility to archive products
- Are an efficient energetically method to make products available which are rarely used



Main System Constraints and Requirements for the Processing System

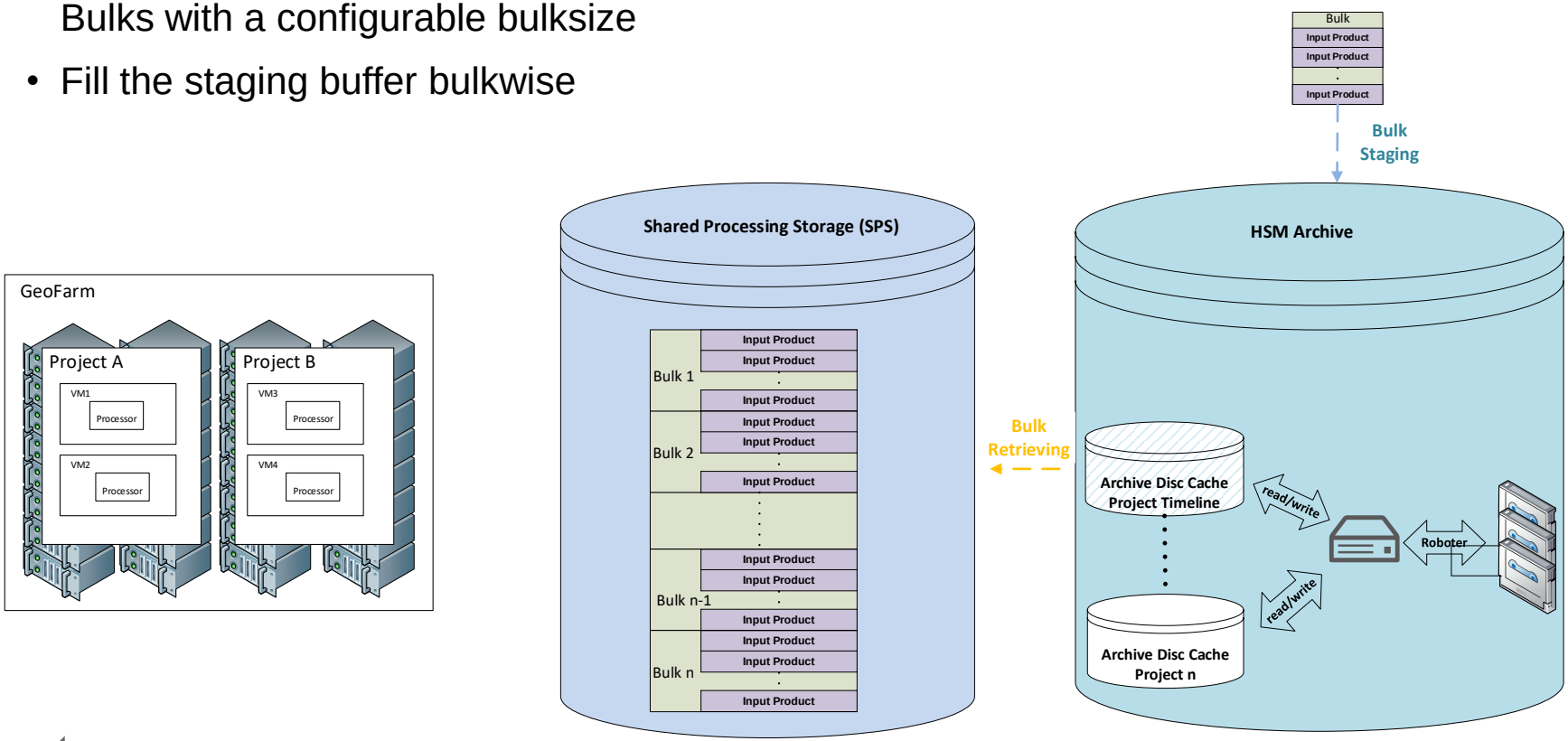
Main System Requirements

- a consistent version control on system and production levels
- the processing of several products within one processing run should be possible.
- the processing automatically follows a predefined product dependency of the requested output products
- the necessary input data should be selected generically and flexibly via the processing request.
- a high-performance processing in a sliding window directly from the archive should be possible.
- the workflow generation and processing should be highly automated.
- the needed processing power should be scalable

Efficient HSM Archive Access for GeoFarm

PRODUCT LIBRARY WITH HSM ARCHIVE

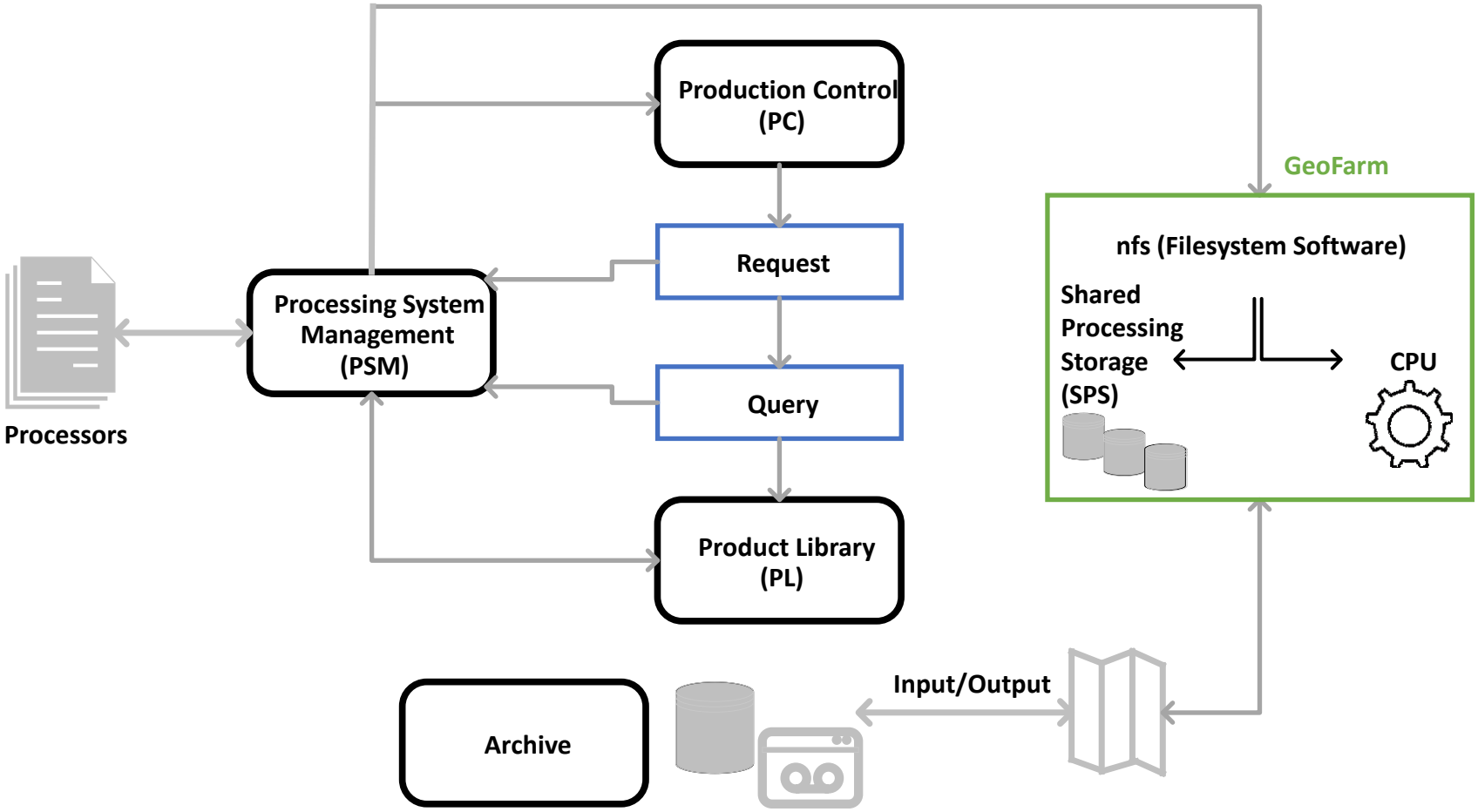
- Minimization of time consuming tape spooling
- Fill the staging buffer of the archive in order to keep the tape read access in streaming mode
- Subdivide the input data stream into Bulks with a configurable bulksize
- Fill the staging buffer bulkwise



DLR

TIMELINE

System Design and Architecture - Timeline Processing System



Main Features of the Timeline Processing System

- Construction of a general re-processing system
- Buffered re-processing with sliding window directly from the archive
- Fully automated tape reload, product retrieval, product processing and archiving
- Workflow generation and distribution according to xml-based configuration
- Expandable for other sensors under the following boundary conditions
  - All necessary product types of the sensor are configured in the product library
  - Input and auxiliary products must to be selectable using product catalogue queries
  - The queries for the primary input products must have an attribute (e.g. time, orbit) defining an order
  - Processor calls must be scalable (node concept)

System Design and Architecture

TIMELINE PROCESSING SYSTEM

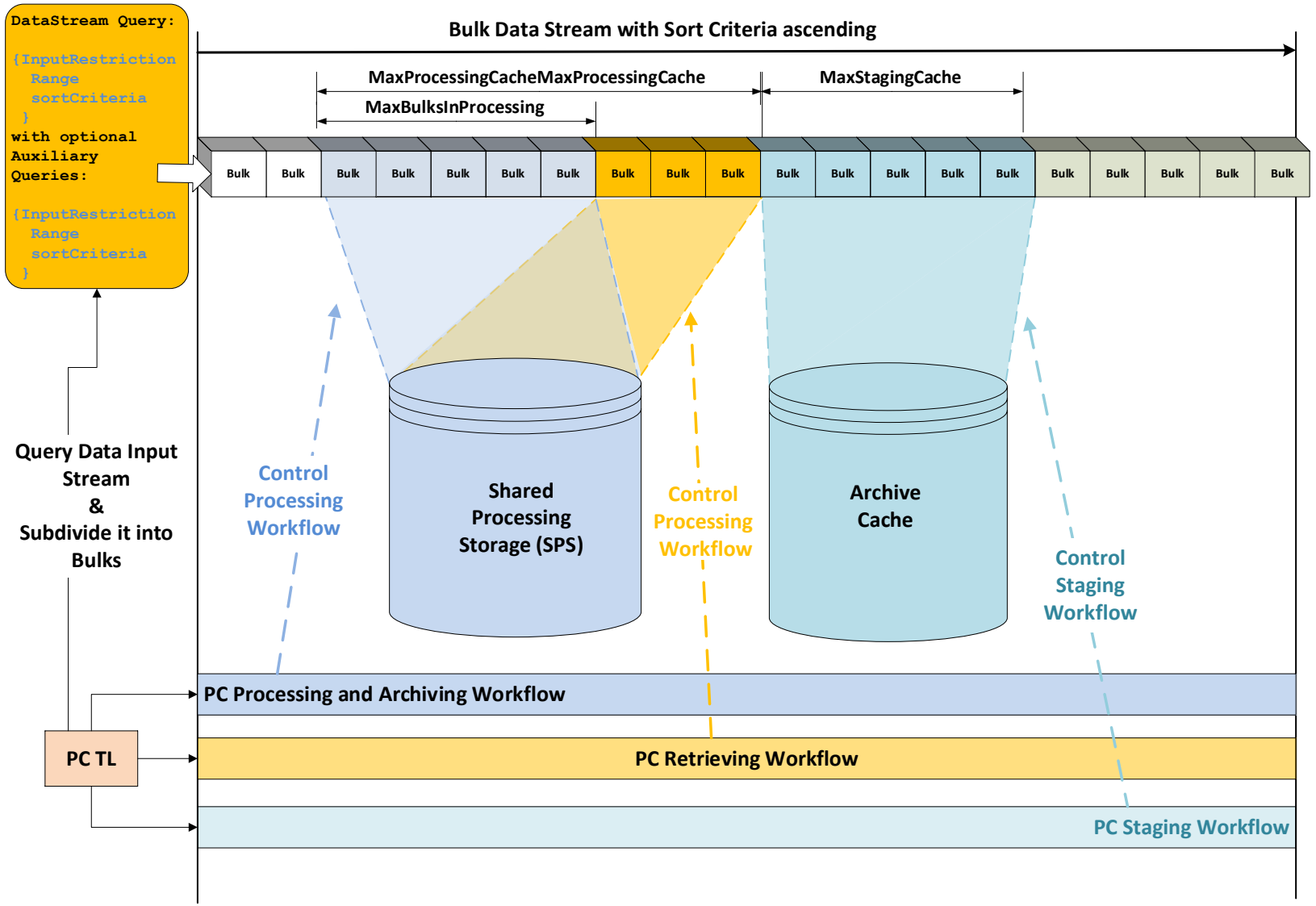
- The system is composed of the Product Library and Archive providing input data (NOAA AHVRR raw data as well as auxiliary data) and saving the generated outputs
- standalone scientific data processors which are embedded into a processing workflow by several Processing System Management components
- Production Control responsible for the overall Workflow and Bulk Generation and the Production Control
- Shared Processing Storage (SPS) provided by the GeoFarm

PRODUCTION CONCEPT

The starting point of all TIMELINE re-processing activities is a Re-Processing request containing at least the following parameters:

- Data Stream Query defining some input restrictions, an input range and a sort criteria for the input data stream form the Product Library/Archive
- Optional auxiliary queries for required auxiliary products
- A list of output product types that shall be generated including their intended product version
- and some workflow control parameter

TIMELINE Bulk Data Stream Concept



DLR

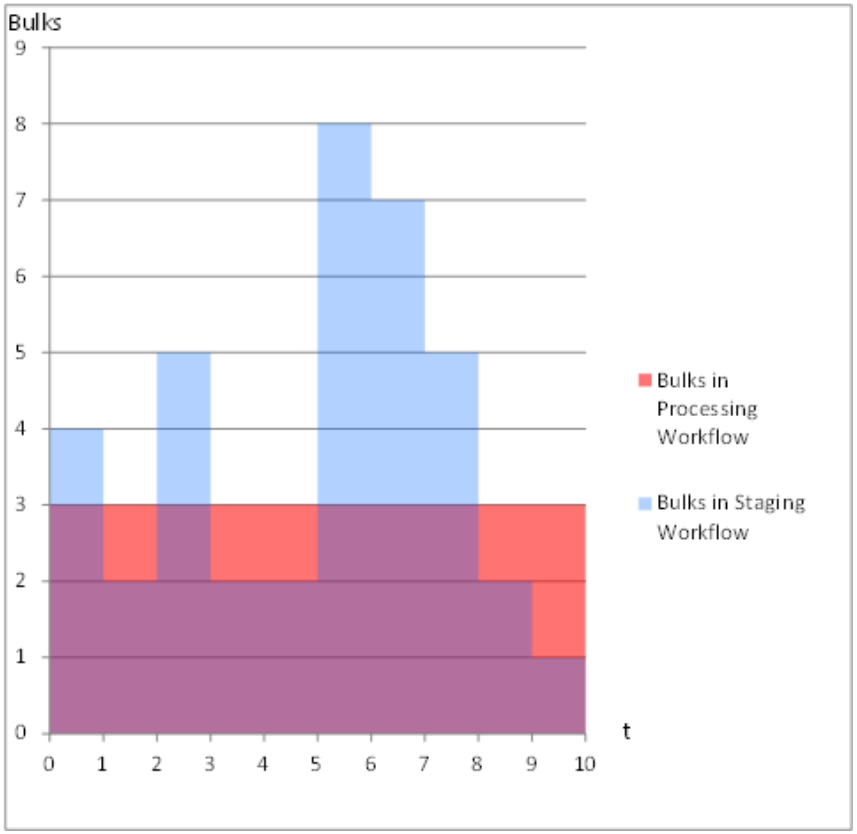
TIMELINE

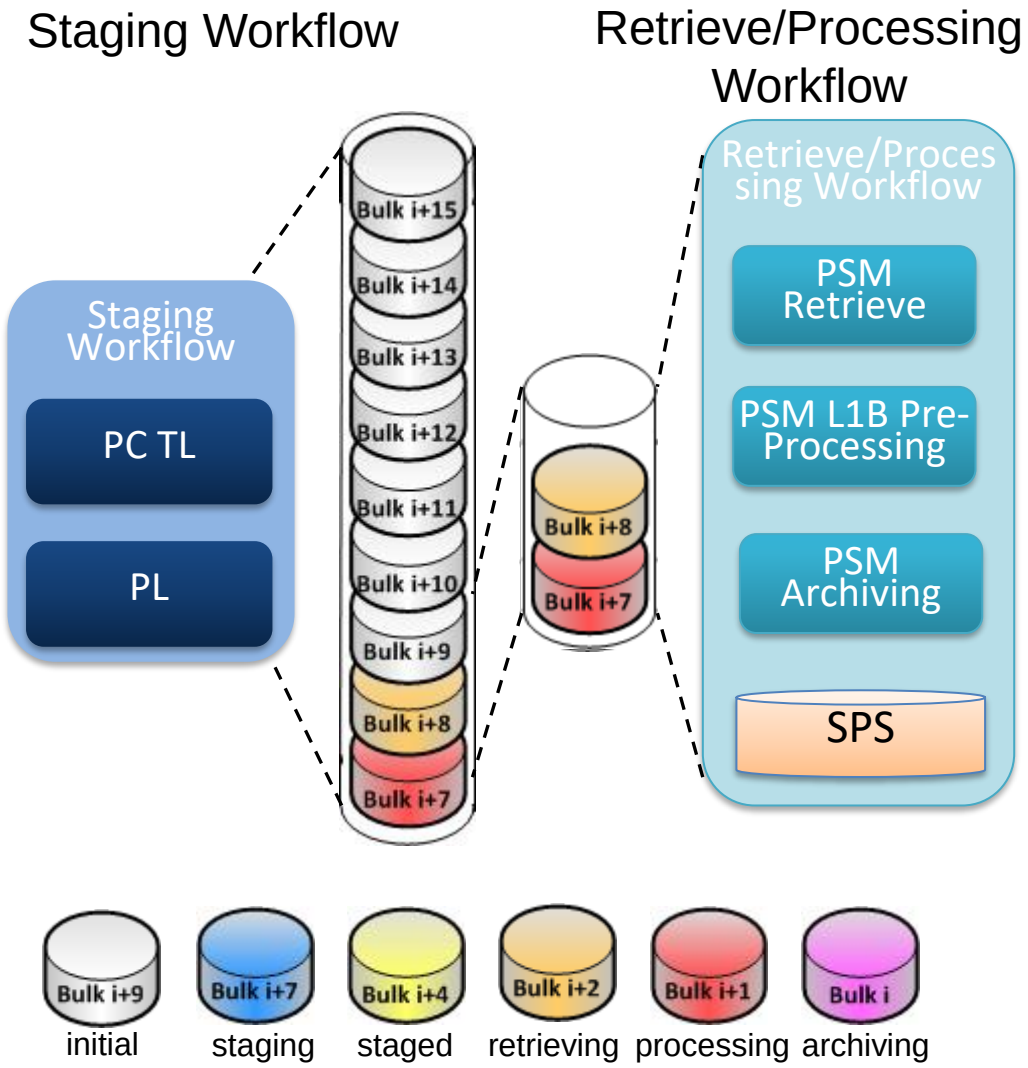
System Design and Architecture

DATA MANAGEMNET

- The processing of the three workflows is controlled and synchronized by the corresponding configuration parameters maxBulksForStaging, maxBulksForRetrieving, maxBulksForProcessing in combination of the bulk status.
- The size of these configuration parameters must be defined regarding the available resources within the Archive Cache and the SPS. PC TIMELINE tries to keep as much as possible number of bulks active within the three workflows.
- The staging workflow is able to stage new bulks from tape archive into the archive disc cache if the number of active bulks is lesser than the parameter maxBulksForStaging.
- The retrieving workflow is able to copy already staged bulks from archive disc cache into the SPS (processing disc cache) if the number of active bulks within the retrieve workflow is lesser than the parameter maxBulksForRetrieving.
- The processing workflow is able to process the input products of already retrieved bulks to the requested output products if the number active bulks within the processing workflow is lesser than the parameter maxBulksForProcessing.
- Automatic, „Bulk“-controlled, data driven Processing

Processing Workflow – Implementation





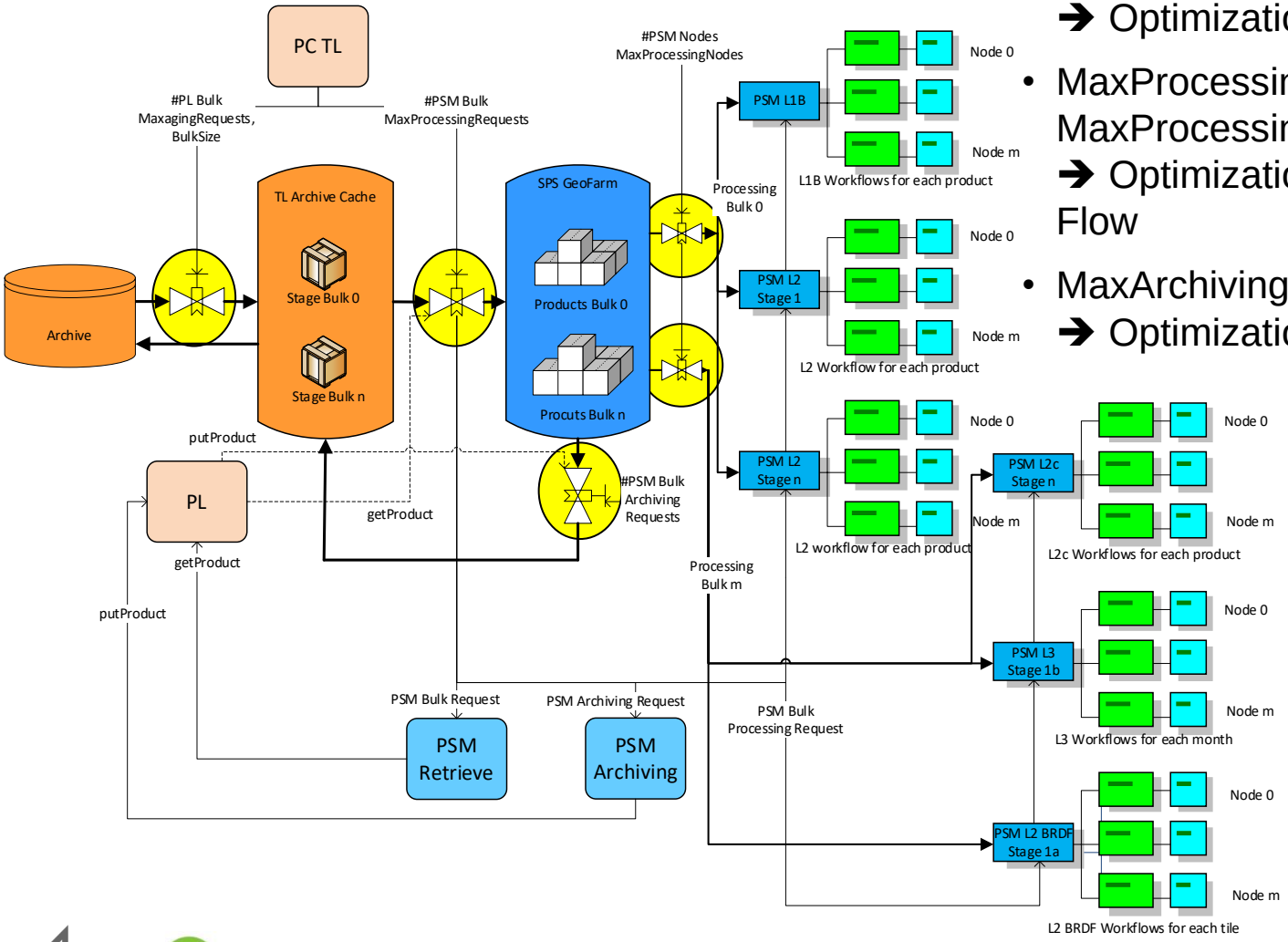
DLR

TIMELINE

Control Parameters for Throughput

Adjusting Screws

- MaxStagingRequests, MaxRetrievingRequests, BulkSize
  ➔ Optimization Archive Flow
- MaxProcessingRequests, MaxProcessingNodes
  ➔ Optimization of the Processing Flow
- MaxArchivingRequests
  ➔ Optimization of Archiving Flow



Processing System Performance and Preview

PERFORMANCE

- The production of time series for L1B, L2 and L3 products between 1981 and 2022 based on the AVHRR series of instruments
- Processing of ca. 150.000 L0-Szenen input products to 18 output produkt types
- Generation of ca. 1,5 Mio output products with a data volume of ca. 500TByte
- Workflows with approx. 30,000 input/output and a data volume of 1,5TByte per day were continuously achieved

Preview

- We are currently in the process of getting an overview of which parts of our IT systems are consuming which electricity.
- This is the base to calculate the power consumption for the data processing of a data set which are stored on different archive media (e.g. tapes, Fast-LTA (logical tapes), discs)