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
E-Mail: {meinhard.wolfmueller, stefanie.holzwarth, sarah.asam}@dlr.de

1. Main Project Requirements

The German Aerospace Center (DLR) has established the project TIMELINE in order to generate a scientifically sound, well calibrated and homogeneous 40-year time series of remote sensing-based global change relevant variables [1]. The map products are derived at 1 km resolution for Europe and North Africa from data of the Advanced Very High Resolution Radiometer (AVHRR) sensors on board of the National Oceanic and Atmospheric Administration (NOAA) satellites. Based on the raw HRPT (High Resolution Picture Transmission) AVHRR data and a thorough pre-processing including geometric corrections, data calibration and harmonization and the detection of data defects, level 1b (L1b) data are generated in a first step [2,3,4]. Based on the L1b data, water [5] and cloud [6] masks as well as a range of L2 thematic products are derived. Surface reflectance data are processed to L2 quality through atmospheric correction and a subsequent BRDF correction. While L1b and L2 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 resolution. The L3 data product suite consists of Normalized Difference Vegetation Index (NDVI), Snow Cover, Fire Hotspots and Burnt Area, Land Surface Temperature (LST), Sea Surface Temperature (SST), and different cloud properties [1] (Figure 1-1). This unique multi-decadal time series allows the investigation of long-term impacts of climate change on our environment.

A general re-processing of all L1b, L2 and L3 products in a unified, automated and flexible way has been provided. Data management ensures an efficient and user-friendly retrieval of localized time series data stacks. The time series products shall be visualized on-the-fly. Therefore, the overarching goals for the data base and the processing within the TIMELINE project are consistency, reproducibility, transparency, reliability and the generic approach.

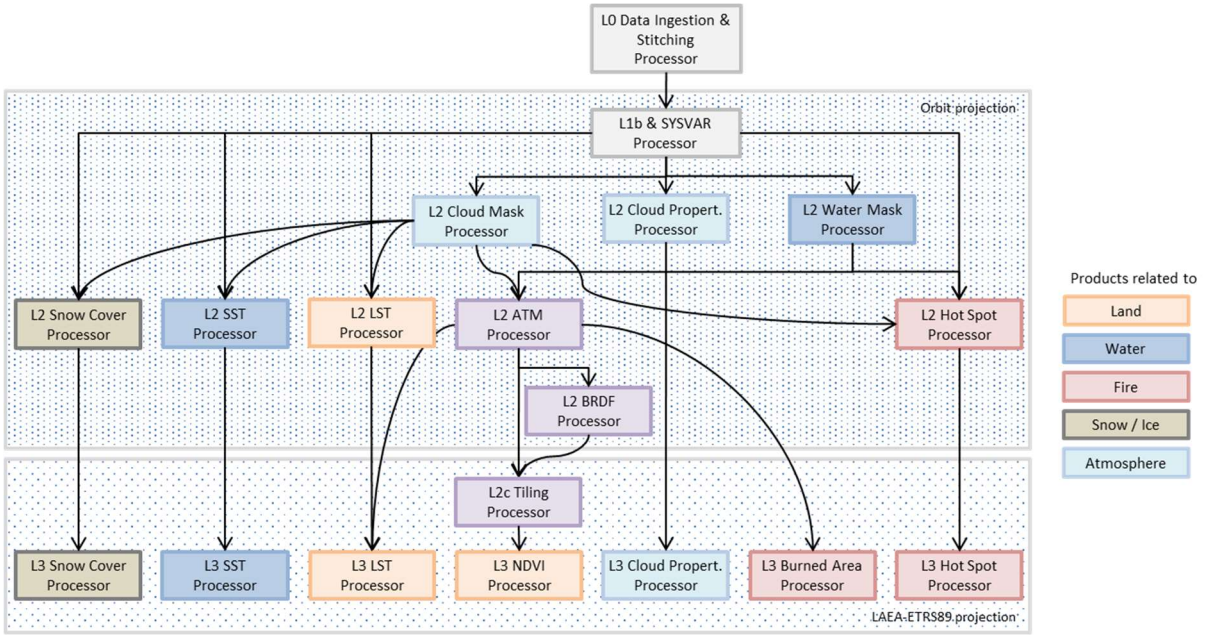

Figure 1-1 TIMELINE Products/Variables, Processors, Processing Levels and Product Dependencies

1.1 Main Scientific Aspects

The challenges to produce 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 (see Figure 1-2) are manifold [1]. A comprehensive effort is necessary to develop consistent, reproducible, reliable and generic variables enabling the detection of geoscientific phenomena and trends. Specifically, the spatial and temporal consistency of reflectance and thermal information are a prerequisite for unbiased time series analysis. Hence, 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. A further challenge for the generation of such a comprehensive product suite is on the one hand the development and management of state-of-the-art AVHRR-adapted thematic product algorithms, a thorough tracking of errors, uncertainties and quality information, as well as a proper metadata management, and on the other hand a careful reprocessing and version handling. Another important aspect is the generic concept of all processors, which allows e.g. for the currently ongoing integration of AVHRR data from EUMETSAT MetOP-A, -B and -C satellites.

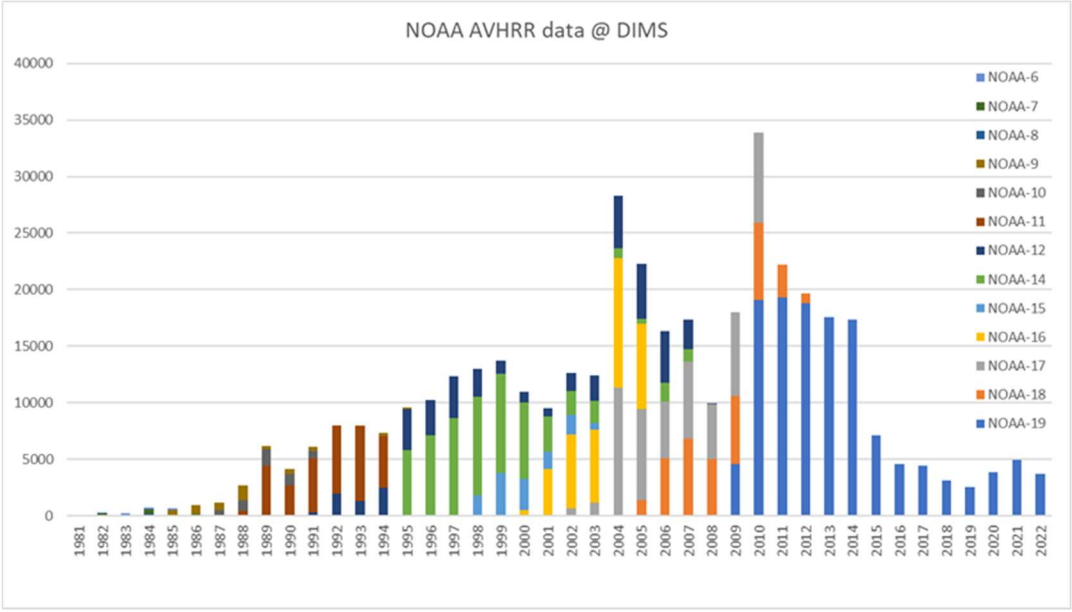

Figure 1-2 Overview of the number of L0 AVHRR scenes per NOAA sensor used for the TIMELINE project.

1.2 Main System Aspects

Figure 1-1 gives an overview over the products/variables, their processing level, the needed processors and their product dependencies. A TIMELINE Processing System has to be provided to support these processing scenarios using the processing platform GeoFarm and the DSDA long-term archive available at DLR DFD.

2. Project Scenarios

The processing of data sets is the essential functionality which allows creating the needed products. The need for reprocessing can evolve out of new calibration of data, increased performance of processing algorithms, or the addition of new data sources as input for the project.

2.1 Scenarios for Re-Processing Phase

The Re-Processing scenario is subdivided into the following production steps which are initiated, controlled and organized by the TIMELINE operator (Figure 2-1):

- L0 Scenarios
    - the L0 data consolidation,
    - the 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)

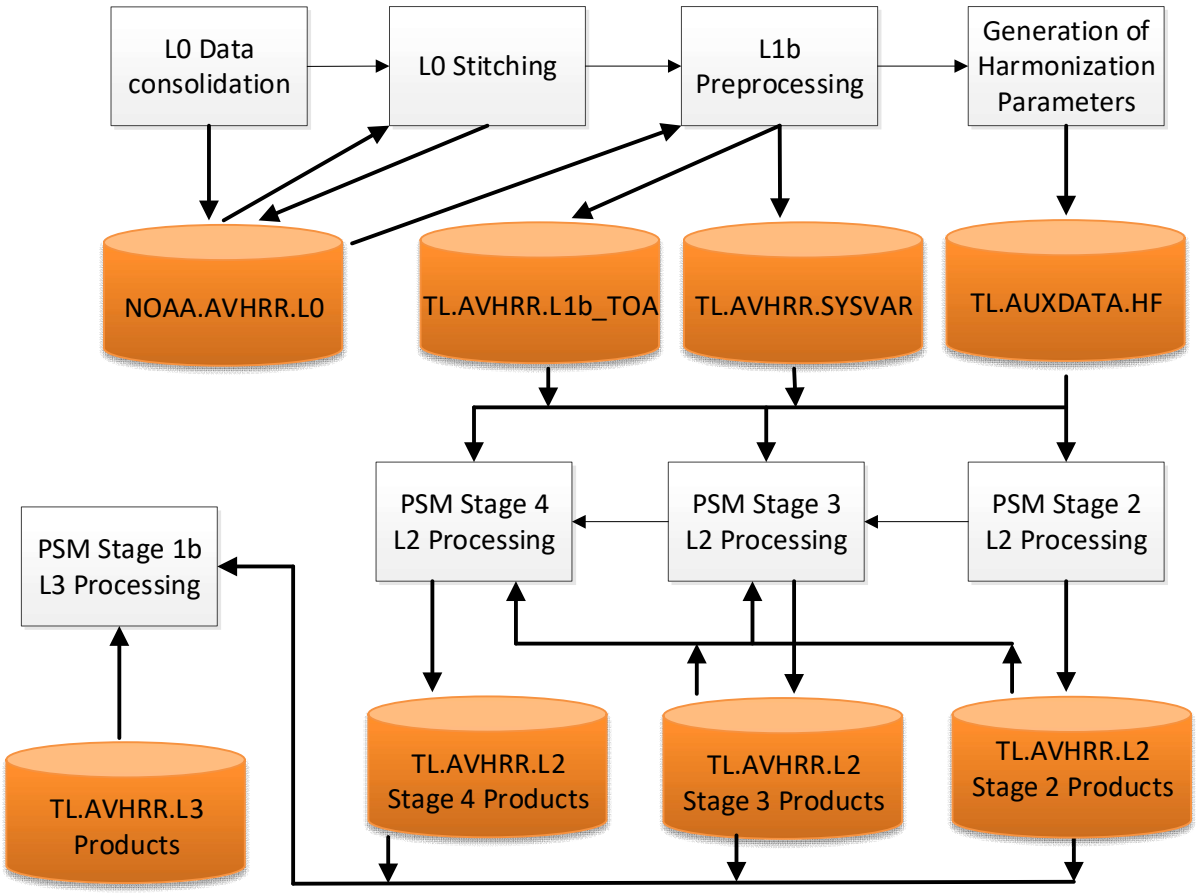

Figure 2-1 Production Sequence from L0 to L3 Products

3. Design

3.1 Product Design

The TIMELINE products are provided as standardized NetCDF datasets with an extensive set of global and variable attributes specifying format, content, creator source and history information, just to name a few. Versioning distinguishes between product and processor versions for each individual product. In addition, each map product has an associated quicklook file. All products follow a predefined file name convention.

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" (Figure 3-1).

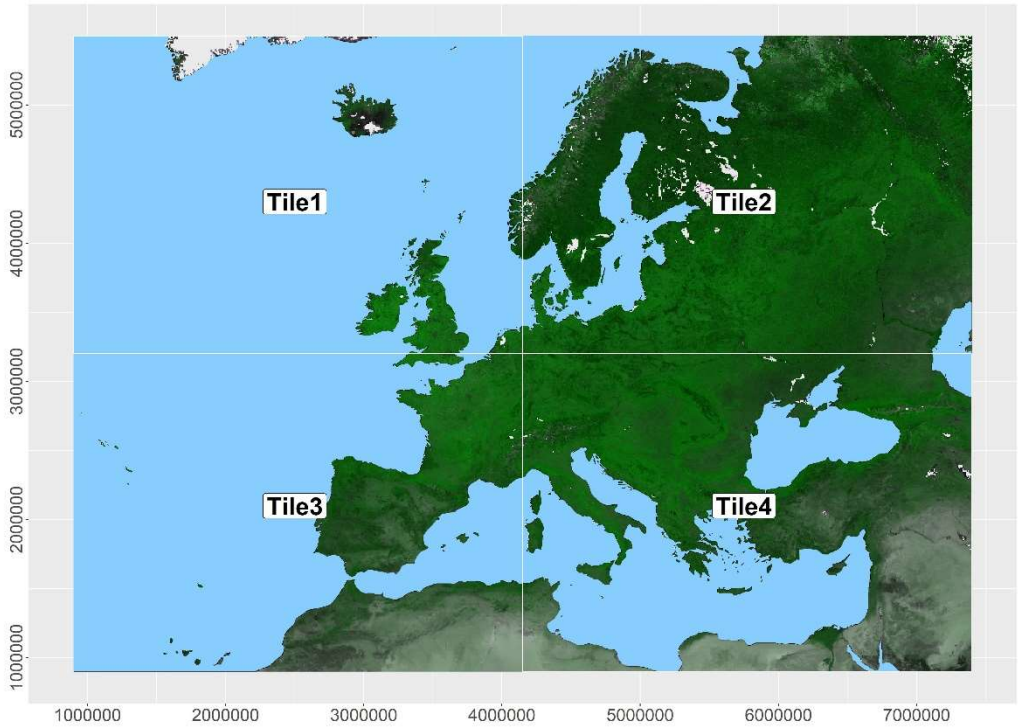

Figure 3-1 Extent of the TIMELINE area overlaid with the division of the four L2c / L3 tiles.

3.2 System Requirements

For the development of the TIMELINE processing system the following particular importance were pursued:

- 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 shown in Figure 1 1.
- 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.

The resulting processing system developed according to these requirements is able to execute very flexible processing requests using the following main input specifications:

- a list of product types that shall be produced in a predefined product version,
- a list of catalog queries defining the input products which are necessary for the entire processing,
- a metadata range (e.g. start time of the input products) which define how many input products shall be generated to the specified output products.

3.3 System Design and Architecture

Boundary conditions for the system design are the usage of the available "HSM Archive" and the infrastructure "GeoFarm" providing the processing resources and storage.

3.3.1 HSM Archive

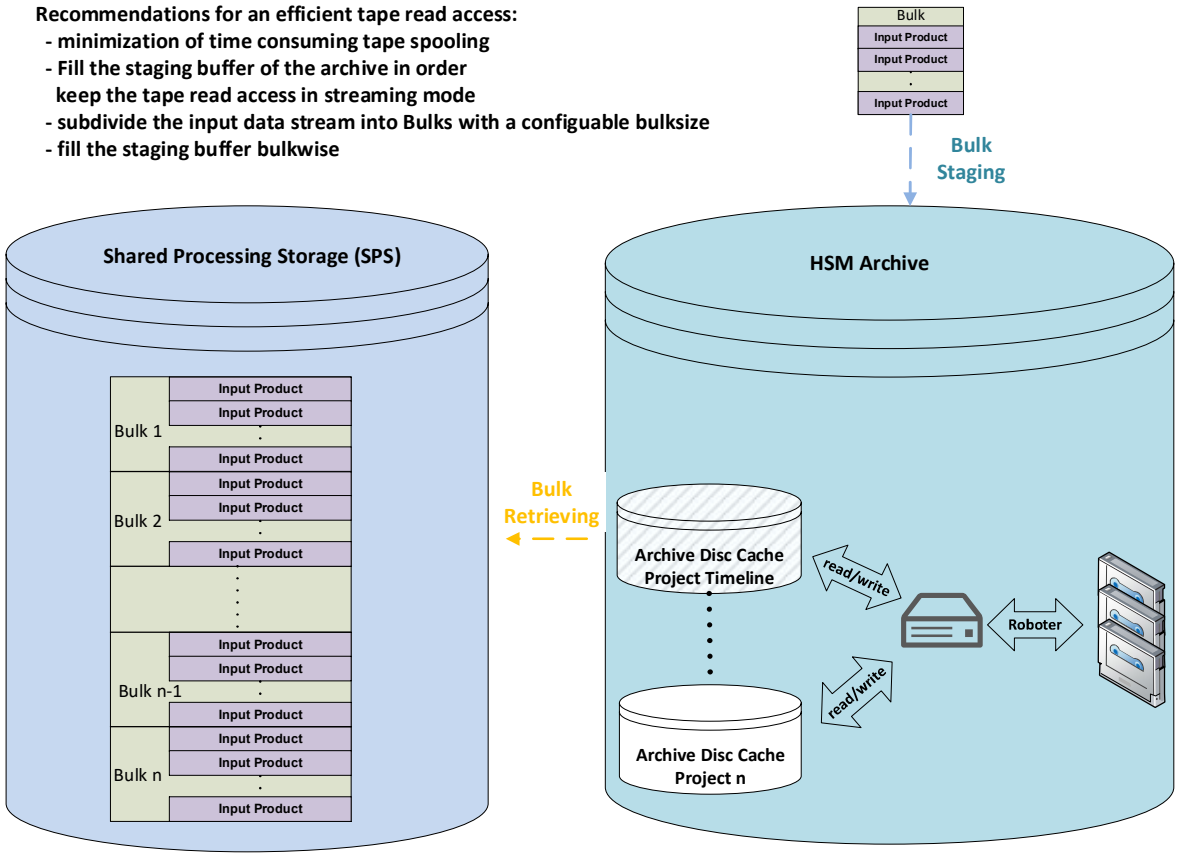

Figure 3-2 Efficient HSM Archive Access

3.3.2 Processing Infrastructure GeoFarm

EOC provides a homogenous virtualized environment (GeoFarm) which intends to provide IaaS (Infrastructure as a Service) to earth observation related projects. The environment provides new resources to a project without hardware reconfiguration. Tasks arising in this scope are split between the GeoFarm project and the projects requesting resources as shown in Figure 3-3.

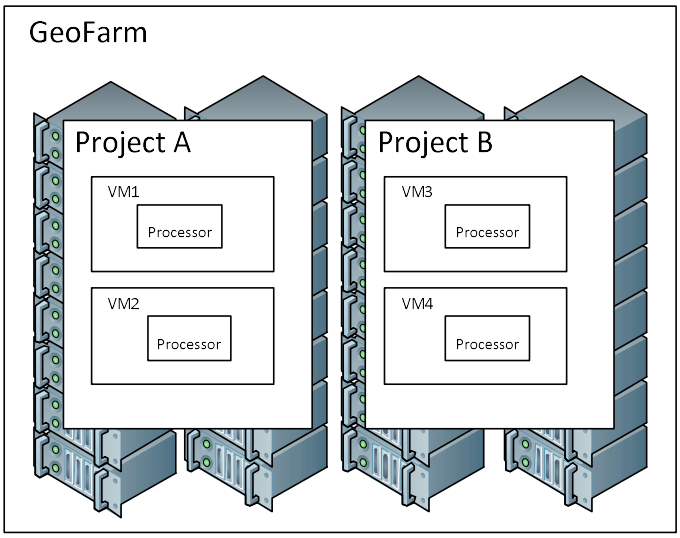

Figure 3-3 GeoForm Processing Environment

Hardware will be allocated to projects in the form of a virtual machine running in the virtualized environment. The virtual machine has to be configured such as the project needs and has to be provided by the project needing the resources.

The currently installed processing hardware is organized as a private cloud. It is based on DELL blade servers. There are several blade servers of the type Dell Power Edge 905 and Dell Power Edge 915M. On the M905 are installed AMD Opteron 8431 CPUs, on the M915 are AMD Opteron 6176 CPUs. There are always 4 CPUs per server. The actual processing environment has installed 672 cores, 3.3 TB RAM and 268 TB storage.

3.3.3 System Design and Architecture

The processing of all TIMELINE Re-Processing requests follows the scheme which is shown in Figure 3-4. 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.

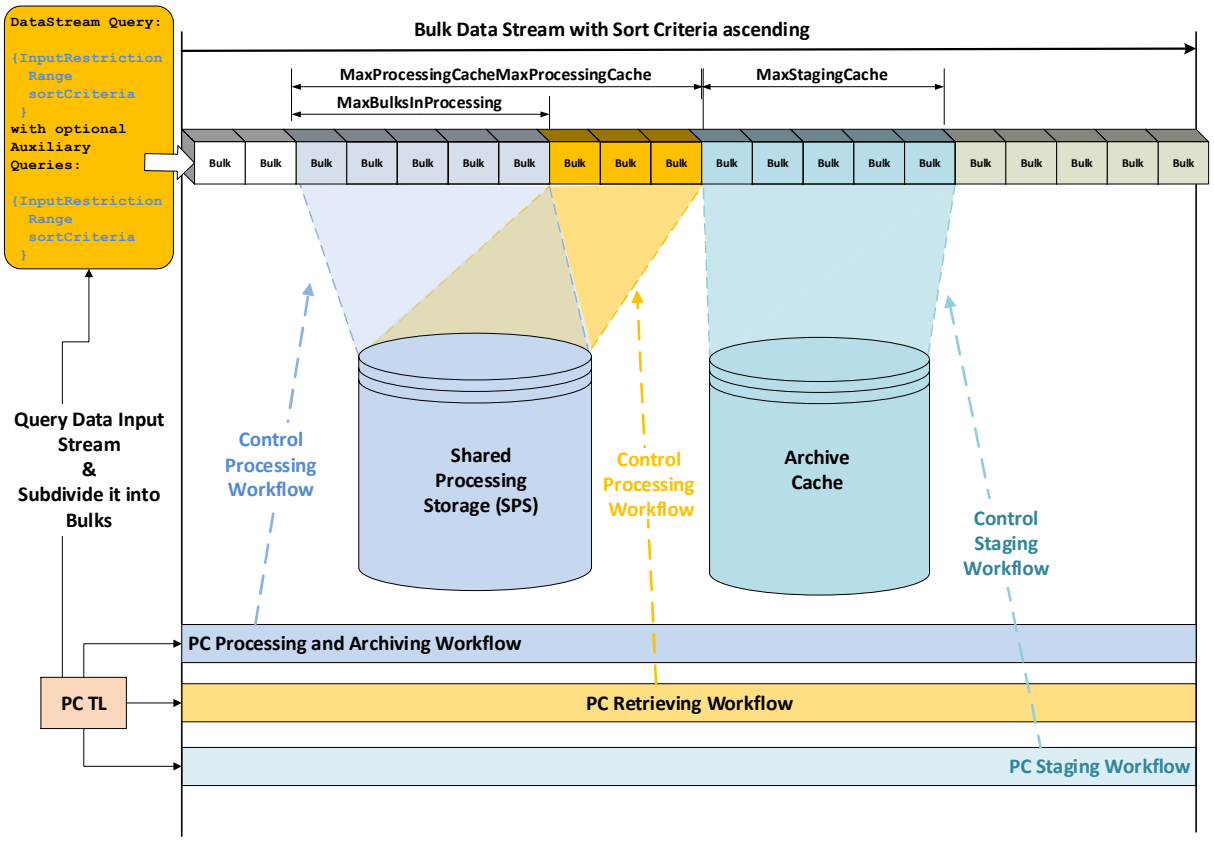

Figure 3-4 TIMELINE Bulk Data Stream Concept

Figure 3-5 shows a schematic overview of the processing and data management system to be built in the TIMELINE project. 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 Processing System Management and the Production Control components and the Shared Processing Storage (SPS) provided by the GeoFarm.

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

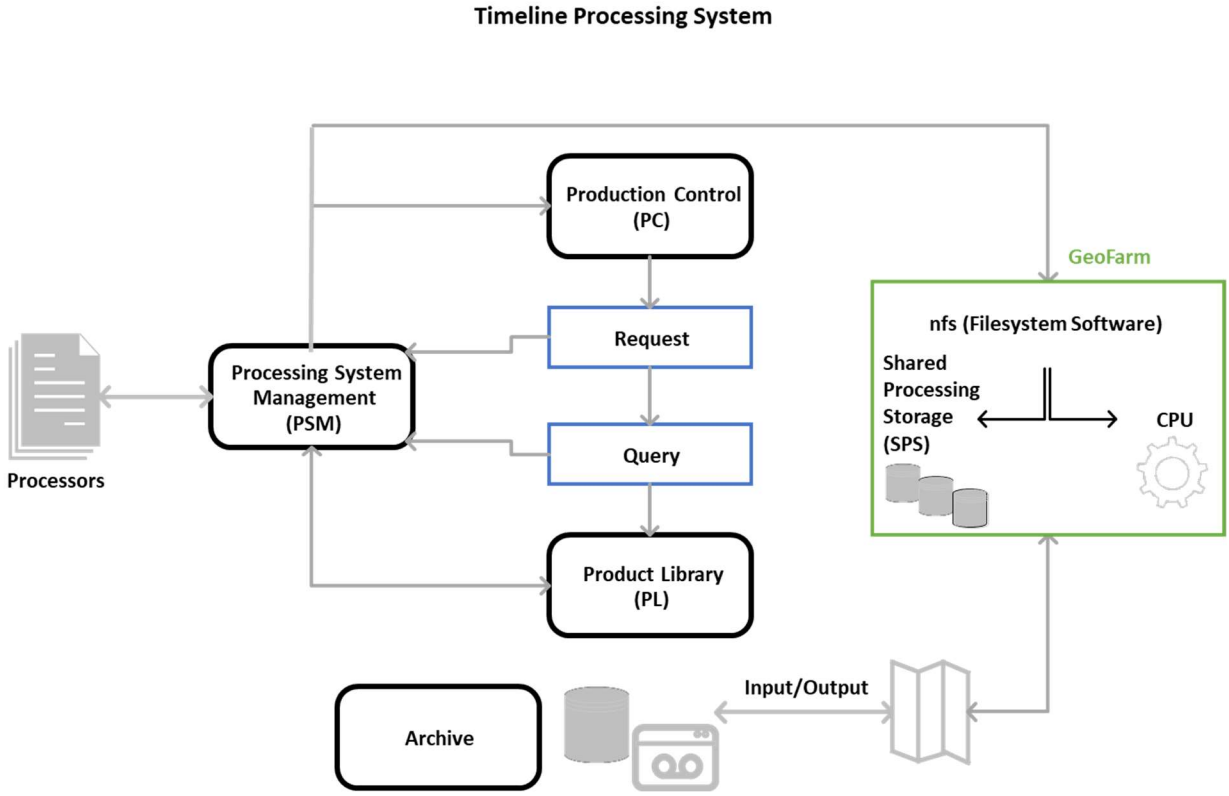

Figure 3-5 Components, Data Structures and Configurations of the TIMELINE Processing System

The main components of the system are:

- **Product Library (PL) and Archive**
  ensures a complete, consistent and version controlled product management of EO products providing search, staging, retrieval and archiving functions. An efficient staging of input product files from tape into the archive and processing disk cache has been provided decoupling the optimized archive tape access of thousands of files from their transfer through the local network onto the processing cache.
- **Production Control (PC)**
  is responsible to read all types of TIMELINE Re-processing requests and organizes and controls its necessary processing chains and workflows. It provides a continuous data stream directly from the tape Archive into the shared processing storage and organizes the complete processing with the help of the parallel staging, retrieval and processing workflows. Additionally an error handling and reporting of failed processing requests is maintained.
- **Processing System Management (PSM)**
  supports interfaces for product access, archiving and production requests as well as processor interfaces in order to schedule and organize processing steps and/or processing workflows invoking configured processors/algorithms. It handles the processing of the L1b, L2, L2c and L3 products using the available processing resources in the TIMELINE project.
- **Scientific Processors**
  the scientific processors responsible to generate the required output products.
- **Shared Processing Storage (SPS)**
  provides a shared disk space for the input and output products used by the involved Processing Systems and Processors

4. Configuration and Performance

4.1 Configuration

In order to ensure an optimized production workflow, the following caches and configuration parameters must be adjusted according to the specific production scenarios: size archive cache, size SPS (necessary areas for the input and output products), bulk size, MaxStagingRequests, MaxProcessingRequests and MaxProcessingNodes (Figure 4-1).

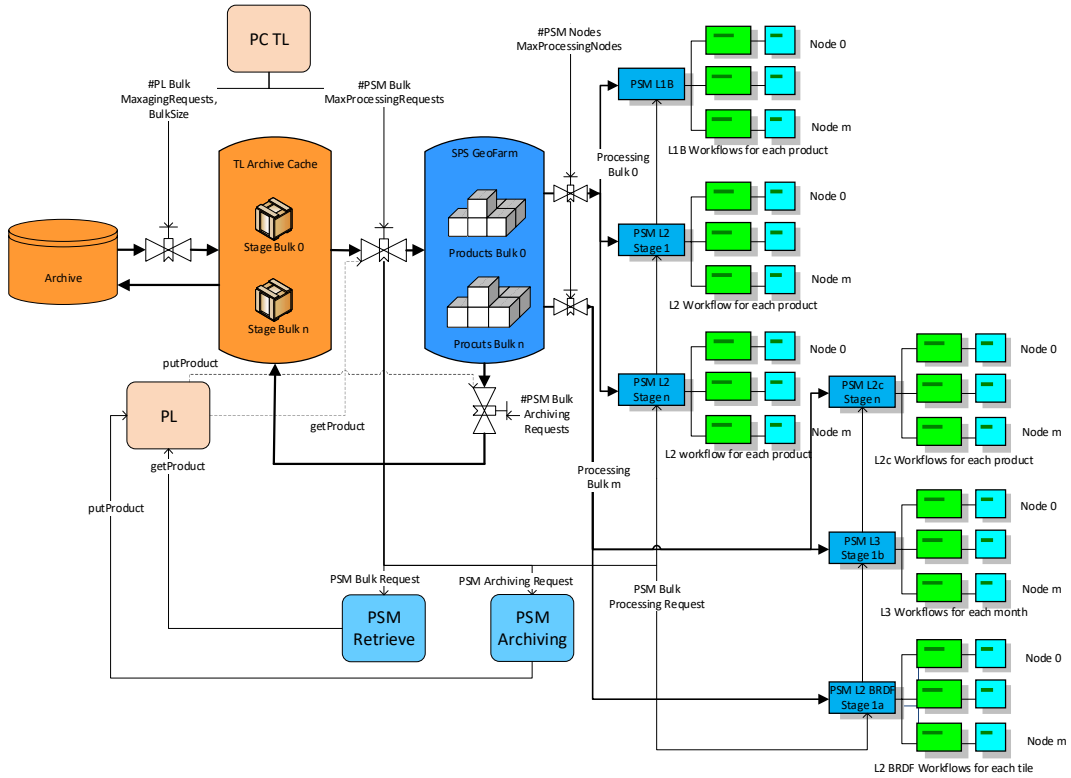

Figure 4-1 Control Parameters for Throughput

4.2 Performance

In most processing scenarios a configured bulk size of 100 input products per bulk and a staging buffer size of nearly six bulks were enough to keep the tapes in streaming mode and to supply the retrieving and processing workflows continuously with input data from the archive. With this configuration the input and the output for one bulk ranges between 125 GB and 380 GB. Therefore, the archive Cache has been configured between 2 and 3 TBytes. In the main processing scenarios the needed processing cache in the SPS varies between 4 TByte (L1B-Processing) and 16 TByte (L3-Processing) and the used processing nodes for parallel processing varies from 24 (L3-Processing) and up to 80 (L1b/L2-Processing).

In the TIMELINE project we have agreed upon that all developed scientific processors use only one core to process the inputs to its resulting outputs (e.g. single scenes or one L3 product for one month). The acceleration of the complete processing has then been achieved by configuration of the PSMs with several processing nodes. The resulting number of configured processing nodes depends on the required processing resources (processing time, RAM) of specific scientific processor.

In summary the TIMELINE Processing System has been generated 18 output product types with ca. 1,5 million of generated products with a volume of ca. 500TBytes. For this generation ca. 2,5 million of input products have to be got from the archive with a volume of ca. 1,5 PByte. For this processing result the system has been active for ca. 450 days.

5. References

1. Dech, S.; Holzwarth, S.; Asam, S.; Andresen, T.; Bachmann, M.; Boettcher, M.; Dietz, A.; Eisfelder, C.; Frey, C.; Gesell, G.; et al. Potential and Challenges of Harmonizing 40 Years of AVHRR Data: The TIMELINE Experience. Remote Sens. 2021, 13, 3618, doi:https://doi.org/10.3390/rs13183618.
2. Molch, K.; Leone, R.; Frey, C.; Wolfmüller, M.; Tungalagsaikhan, P. NOAA AVHRR Data Curation and Reprocessing - TIMELINE. In Proceedings of the Big Data from Space (BiDS' 2013), Frascati, Italy, 5–7 June 2013.
3. Bachmann, M.; Tungalagsaikhan, P.; Ruppert, T.; Dech, S. Calibration and Pre-processing of a Multi-decadal AVHRR Time Series. In Remote Sensing Time Series: Revealing Land Surface Dynamics, Kuenzer, C., Dech, S., Wagner, W., Eds.; Springer International Publishing: Cham, 2015; pp. 43-74.
4. Dietz, A.J.; Frey, C.M.; Ruppert, T.; Bachmann, M.; Kuenzer, C.; Dech, S. Automated Improvement of Geolocation Accuracy in AVHRR Data Using a Two-Step Chip Matching Approach—A Part of the TIMELINE Preprocessor. Remote Sens. 2017, 9, 303, doi:https://doi.org/10.3390/rs9040303.
5. Dietz, A.J.; Klein, I.; Gessner, U.; Frey, C.M.; Kuenzer, C.; Dech, S. Detection of Water Bodies from AVHRR Data—A TIMELINE Thematic Processor. Remote Sens. 2017, 9, 57, doi:https://doi.org/10.3390/rs9010057
6. Klüser, L.; Killius, N.; Gesell, G. APOLLO_NG - a probabilistic interpretation of the APOLLO legacy for AVHRR heritage channels. Atmos. Meas. Tech. 2015, 8, 4155-4170, doi:https://doi.org/10.5194/amt-8-4155-2015.