NASA Information And Data System (NAIADS) for Earth Science Data Fusion and Analytics

NASA ESTO AIST-14-0014 Project

Constantine Lukashin (NASA LaRC)

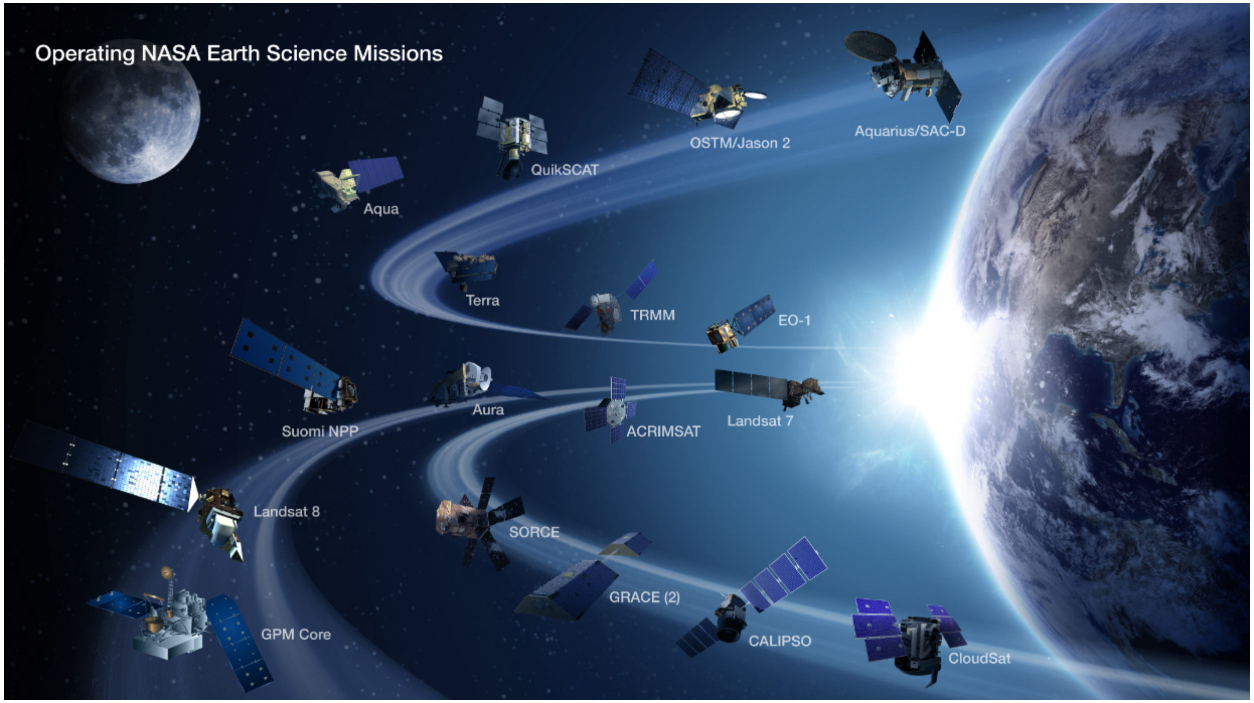
C. Roithmayr (NASA LaRC), V. Gyurjyan (DOE Jefferson Lab)

A. Bartle (Mechdyne Co)

A. Vakhnin (SSAI)

S. Mancilla, R. Oyarzun (UTFSM, Chile / DOE Jefferson Lab)

Earth Climate and Weather Systems



- Geostationary missions are not shown (15 internationally);
- Next 10 years: JPSS-1/2/3, GOES-R, PACE, TEMPO, CLARREO;
- New Decadal Survey and Venture Class missions;
- Missions last longer, sensors become more complex;
- The Climate Model outputs oversize the observations (currently 10s of PB);
- The OSSEs are becoming high priority (large data volume).
- *The next DS push for the Integrated Earth Observing System !*
- *Future constellation approach – data fusion is inevitable !*

Earth Observations and Climate Model Data

1. Observations:

- Expected volume of *used data products*: exceed 100 Petabytes in 10-15 years.
- Data is *distributed* nationally (NASA, NOAA), and internationally (ESA, etc.).
- NASA observational data *format is standard* HDF (Hierarchical Data Format)
- NOAA observational data *format is standard* NetCDF (Network Common Data Form)
- Observations from different sensors are *NOT synched / merged !*
- Each sensor/mission has a *separate data product line !*

2. Climate Models Output:

- Expected volume of CM outputs: may exceed Exabyte in 10-15 years.
- Data is *distributed* nationally (NASA, NOAA, DoE), and internationally (UK).
- Multiple Climate & Earth System models (30 – 50).
- Perturbed Physics Ensemble (PPE) for a single model: about 5 Petabytes.
- Climate Model output *format is standard* NetCDF

3. Relevant Data Centers:

- Observations: Langley and Goddard (NASA), NCDC (NOAA), LPDAAC (USGS), NSIDC.
- Climate/Weather Models: Goddard, GISS and Ames (NASA), NCAR (NSF), LLNL and LBNL (DOE), etc., etc., etc.

Earth Science Data Fusion (science requirement): to Maximize Information Content and Science Output

1. National Strategy:

- OSTP's Earth Civil Observation from Space (2013): Integrated Portfolio Management
- NASA Strategic Space Technology Investment Plan (2013) and Budget Memorandum (2014)
- NASA Strategy Plan 2014: Strategic Objective 2.2

2. Relevancy: Outstanding Science Output

- CERES, Earth's Radiation Budget: multi-sensor data fusion, up to 16 (NASA LaRC)
- CERES/MISR/MODIS data fusion: multi-sensor calibration validation on-orbit (NASA LaRC)
- CALIPSO, CloudSat, CERES, MODIS (A-Train) data fusion – Level-2 information (NASA LaRC)
- Required for future missions: CERES/RBI, TEMPO, CLARREO, ACE, and GEO-CAPE
- Required for future *satellite constellations* (baseline or small sats)

3. Major Challenges:

- IT infrastructure is not optimized for the task: data is distributed, slow connections, security…
- Traditional PGE-based workflow is I/O bound (job-per-file-per-CPU);
- Traditional configurations are network bound for heavy data handling;
- Time and resource intensive, the costs are very high.
- Opportunities are not realized: e.g. MISR/MODIS, A-train Level-1, etc.

NAIADS Framework Requirements

- **IO Optimizer:** Data Event Builder in off-line software before any scaling !
- **Networking:** Ability for local and distributed data-Event building.
  Distributed software approach – move code to data, not otherwise.
- **New workflow / scaling:** Complete in-memory data-Event streaming, massive scaling;
- **Multi-lingual:** Support new and heritage codes, optimal language for given service.
- **Flexibility:** Re-configurable to multiple applications (fusion, processing, data mining, etc.).
  Service Oriented Architecture (SOA) approach.
- **Adaptability:** supporting various hardware configurations (cluster, cloud, servers, etc.),
  and various file systems: NFS, GPFS, Apache Hadoop / Spark, etc.
- **Portability:** Support various OS platforms (Unix, Linux , MacOS, …)
- **Standards:** I/O, transient data, and metadata (NetCDF, HDF)
- **Provenance:** track metadata at any stage of processing.
- **Traceability:** end-to-end test cases, extensive and transparent code documentation.
- **Modern good practices:** from management to coding, maintenance, persistence.

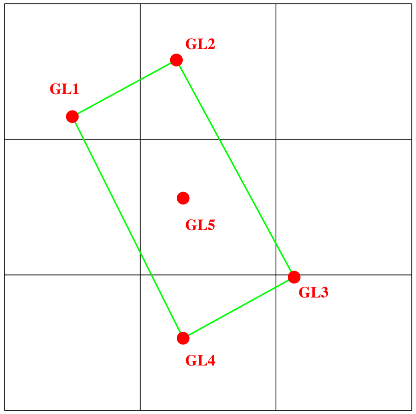
NAIADS Data Fusion: SCIAMACHY/MODIS/ECMWF

OBJECTIVES:

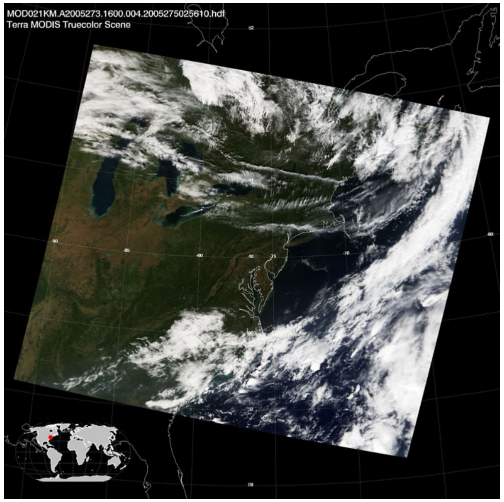
- To demonstrate NAIADS approach and full functionality using existing data;
- To benchmark NAIADS performance;
- Available data: 9 years of near-coincident measurements of from SCIAMACHY and MODIS;
- Create new fused SCIAMACHY/MODIS/ECMWF data product (requested by a number of projects).

SCIAMACHY Level-1 Data:

- Spectral measurement for every footprint: 30 km x 230 km;
- Swath 950 km (4 footprints) from 10 AM Sun-synch orbit.



SCIAMACHY FOOTPRINT 30 x 230 km

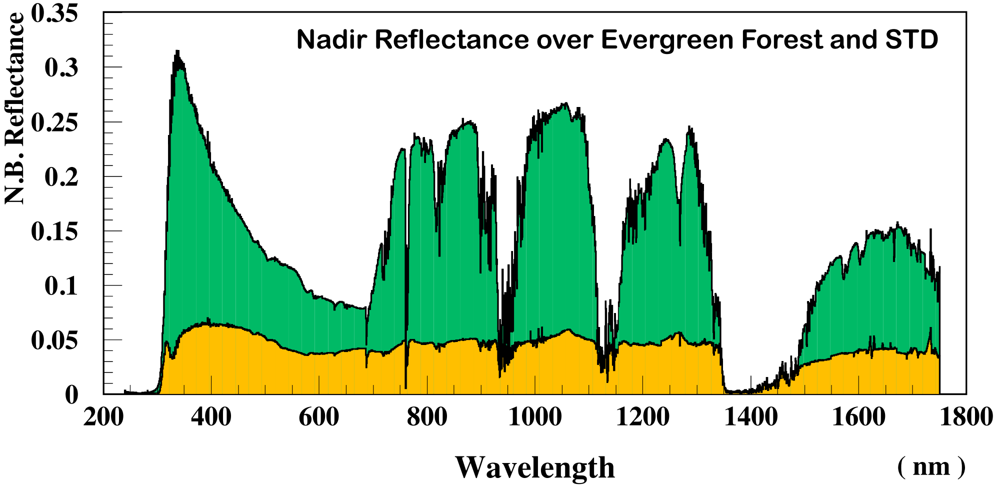


ECMWF Data (re-analysis):

Gridded (0.125°);
6 weather parameters;
Map every 6 hours;

MODIS/Terra Level-2 Data:

- Level-2 Cloud and Aerosol Data
- Spatial scale: 1 / 5 km and 10 km spatial;
- Swath 2300 km (global coverage daily);
- 10:30 AM Sun-synch orbit.



NAIADS: NASA LaRC and DOE JLab Collaboration

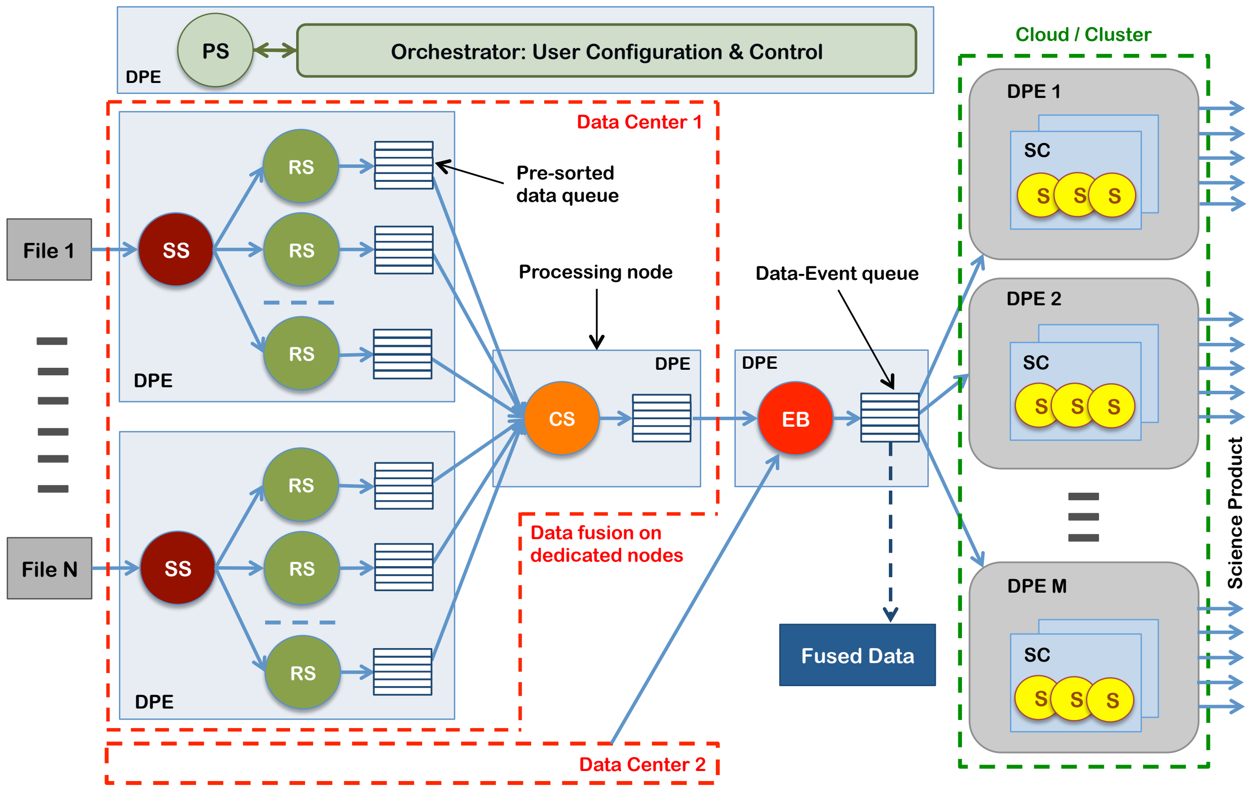
1. xMsg and CLARA Framework (DOE Jefferson Lab):

- Publish/Subscribe messaging middleware (data streaming);
- Based on the ZeroMQ socket library: messaging (C++);
- Multi-lingual binding (Python, Java, C++);
- Service Oriented Architecture framework for data processing applications;
- CLARA takes care of multi-threading and scaling;
- System real time monitoring services.
- Documentation: *https://claraweb.jlab.org*

2. NAIADS software (NASA LaRC):

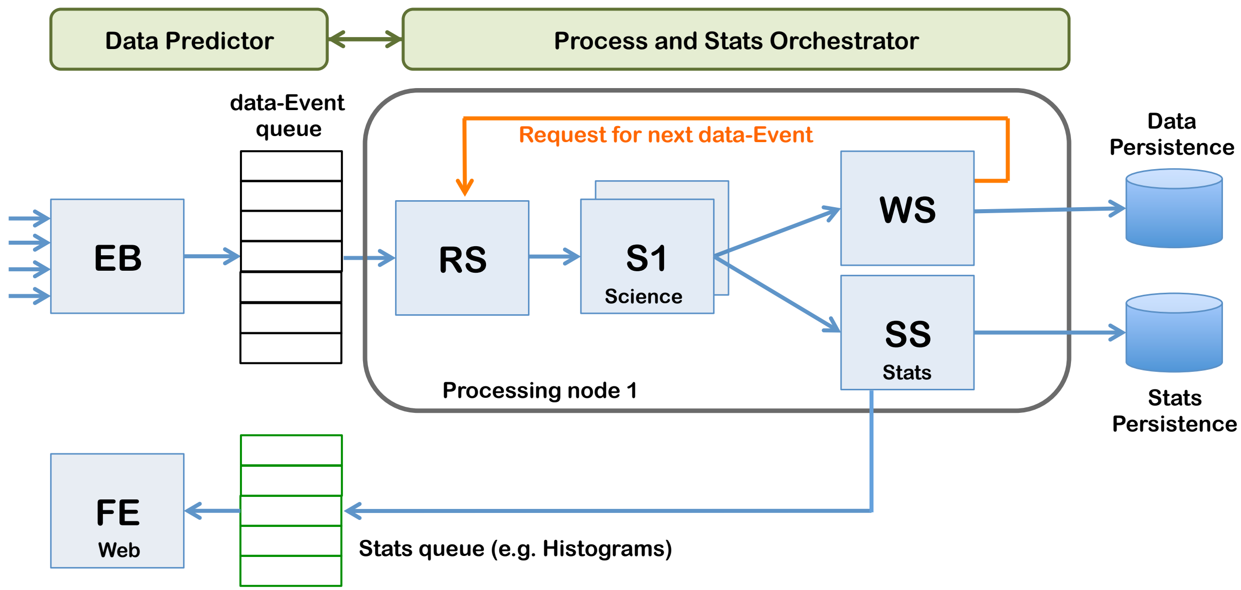
- NetCDF streaming (Python, Java, C++).
- Data fusion algorithms: spatial convolution, temporal interpolation (Java)
- Statistics Services (Java)
- Workflow integration (Python, Java, C++).
- NAIADS Wiki and Git Repository on EARTHDATA: *https://wiki.earthdata.nasa.gov/display/NAIADS/*

NAIADS/CLARA Architecture



- SOA CLARA/xMsg for flexibile and multi-lingual applications.
- PS: multi-sensor coincident data predictor service.
- SS: into-memory fast data staging service (multi-file).
- RS: parallel from-memory data reader service (pre-sorting).
- CS: data concentrator service in a data center (IO/network optimization).
- EB: complete data-Event Builder (adaptation to algorithm).
- Scaling: data-Event streaming to Cloud with minimized IO.

NAIADS Workflow Example: shown for a Single Node



- xMsg (ZeroMQ) messaging.
- EB: complete data-Event Builder.
- RS: parallel data reader service.
- S1: science algorithm service.
- WS: data persistence service (writer).
- SS: statistics service (e.g. histograms).
- FE: front end user web service.

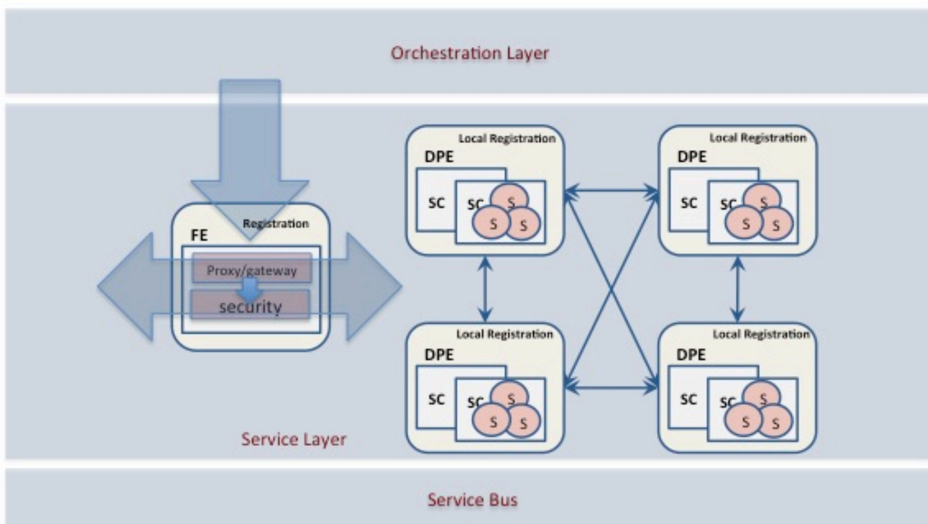
Scaling: Automatic multi-treading in nodes.

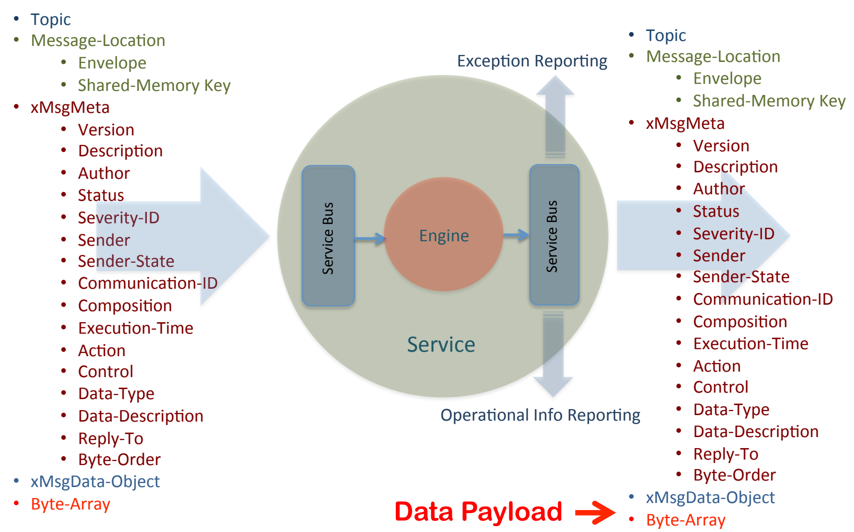
CLARA Framework

Integration with ZeroMQ (iMatix):

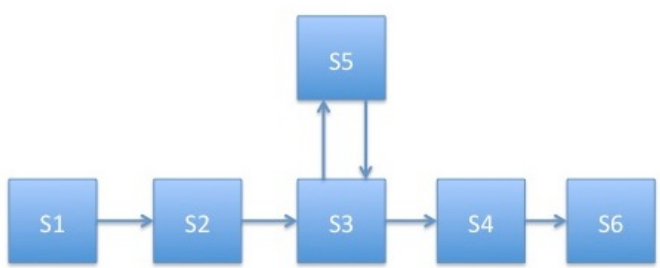
- ZeroMQ is messaging socket library;
- Solves the challenge of creating large multicore applications
- Pub/Sub (asynch) patterns;
- Very fast (tests at CERN, 2011).

Transient Data Envelop and data payload via NetCDF Streaming: Implemented in Java, Python, and C++



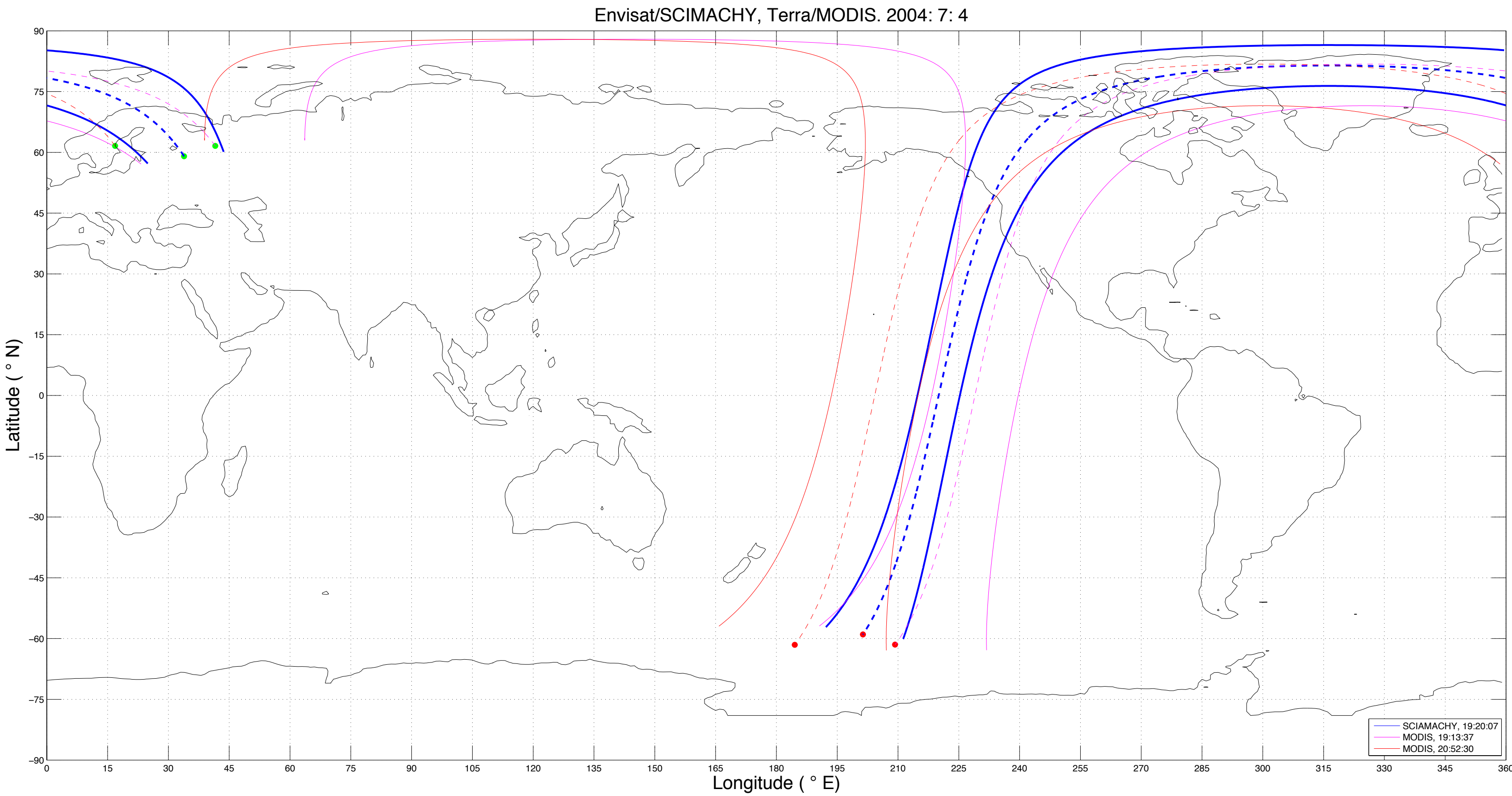


Composition Example: Data Calibration



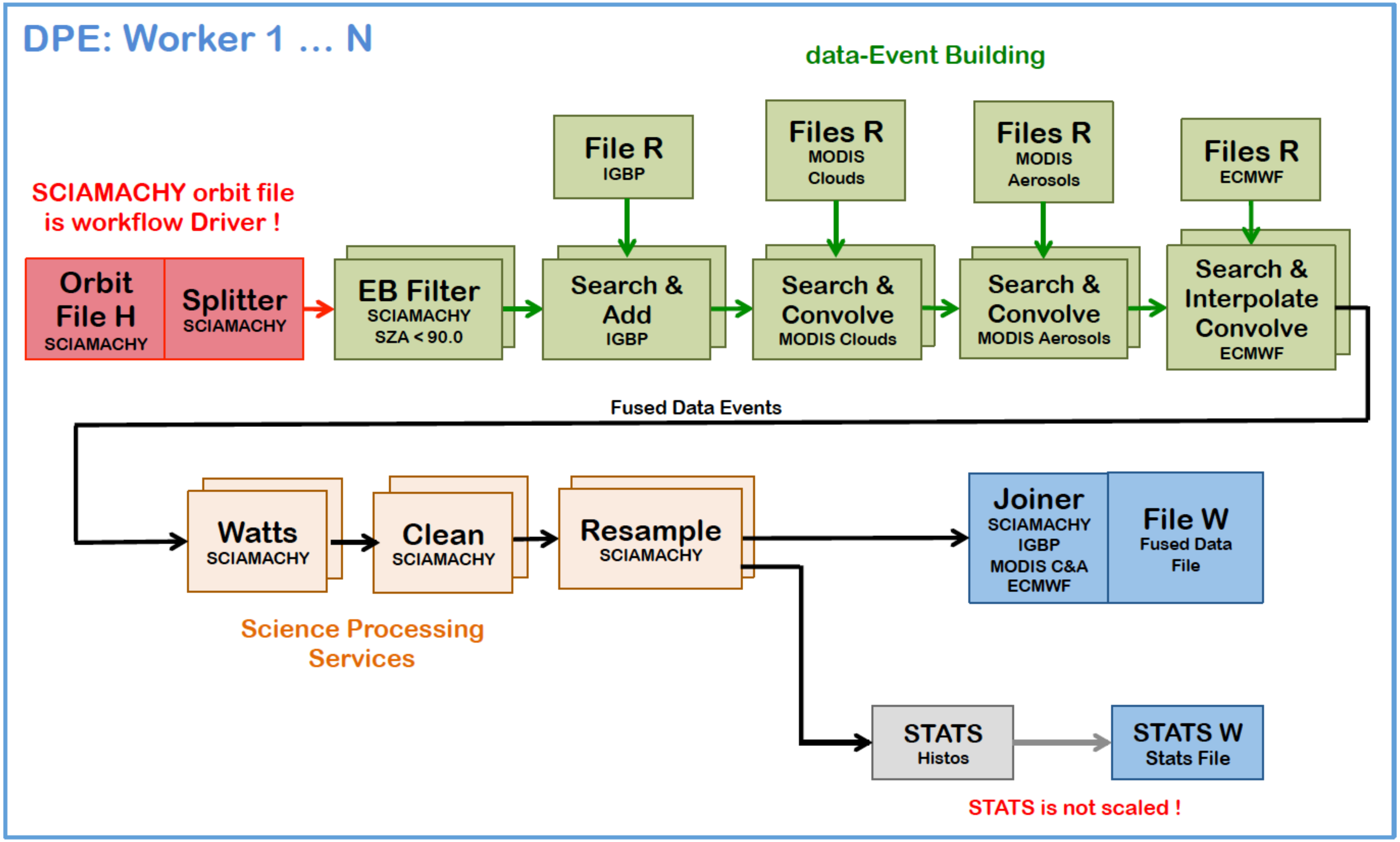

```
S1 + S2 + S3;
if ( S3 == "calibration") {
   S3 + S5;
} else {
   S3 + S4 + S6;
}
```

NAIADS: Orbital Modeling for Optimization



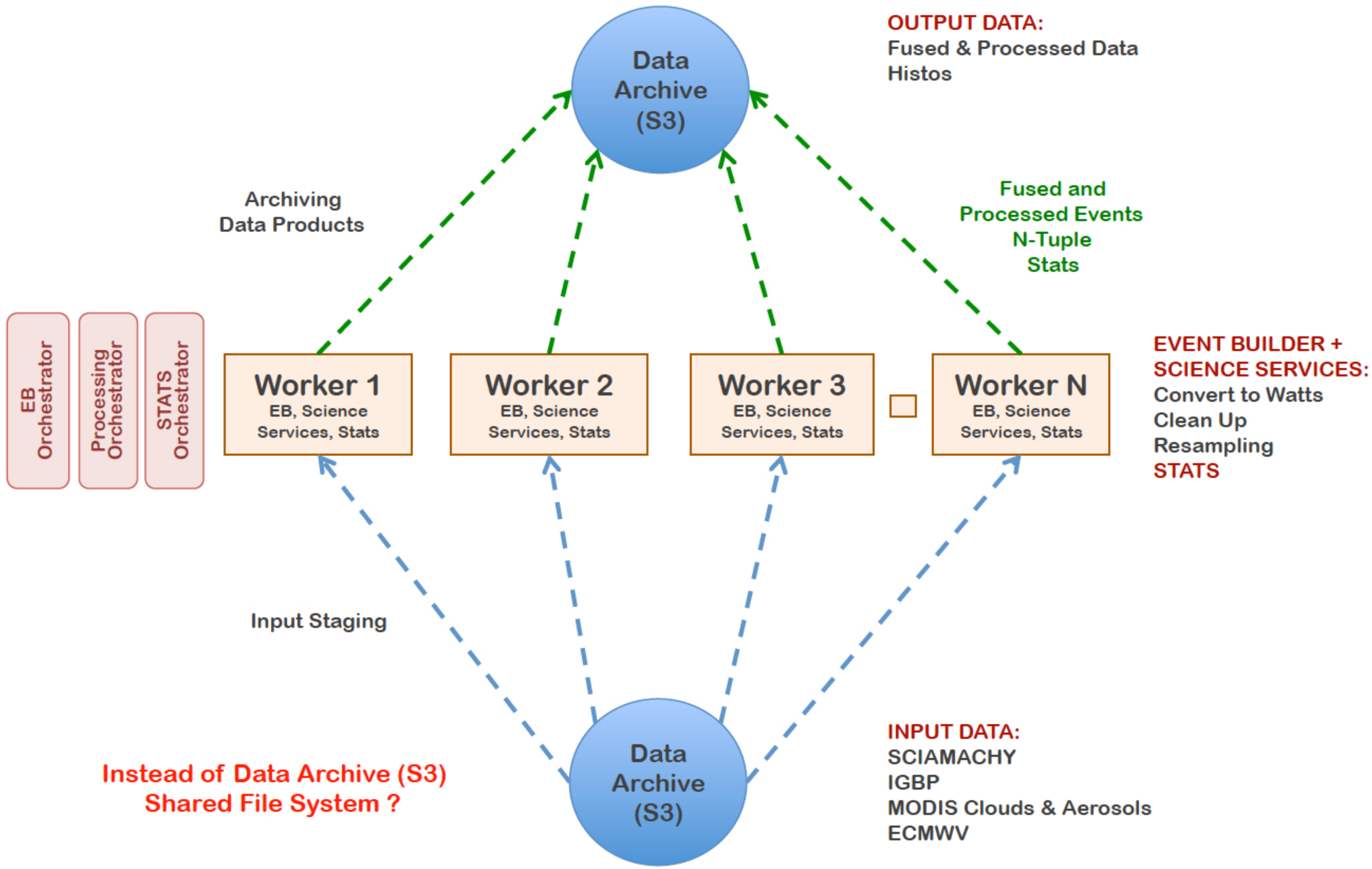
- Simplified General Perturbations (SGP4), is used together with daily, archived two-line element sets to propagate orbital ephemerides for the Envisat and Terra spacecraft.
- Typically, one of these MODIS measurement swaths precedes the SCIAMACHY swath in time, and the other MODIS swath has a later time.
- Distance of the eastward drift for each Envisat orbit is 48 km at the equator. The drift is directly attributable to the difference in orbital periods of the two spacecraft. The orbital period of Terra is approximately 1.7 minutes less than the period of Envisat.

NAIADS Service Workflow

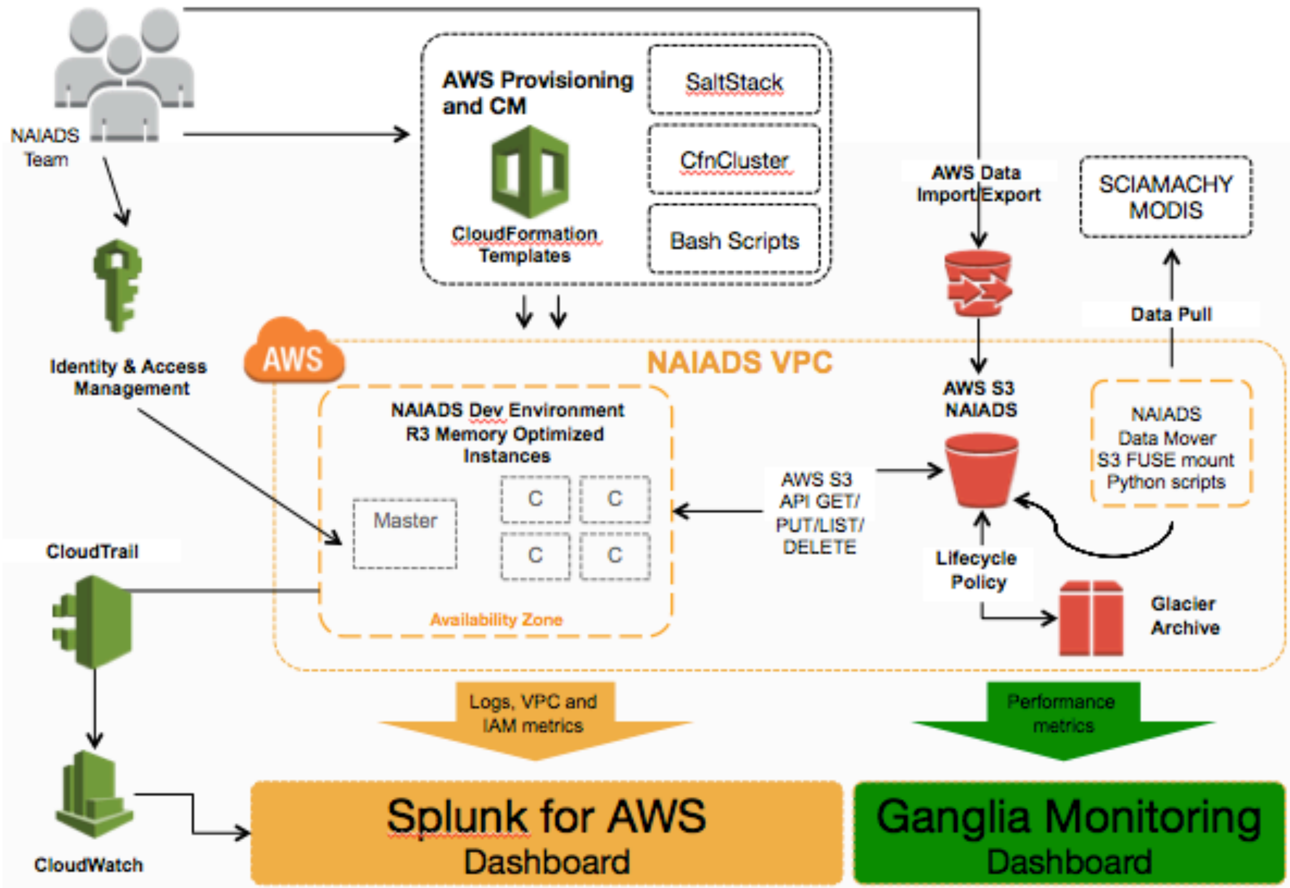


This workflow is implemented: in-memory processing on Event-by-Event basis.
Yes, one file can be processed in parallel by events ! (we have a demo)

NAIADS Current Architecture (AWS): In-the-node processing/scaling – an efficient approach for local processing system



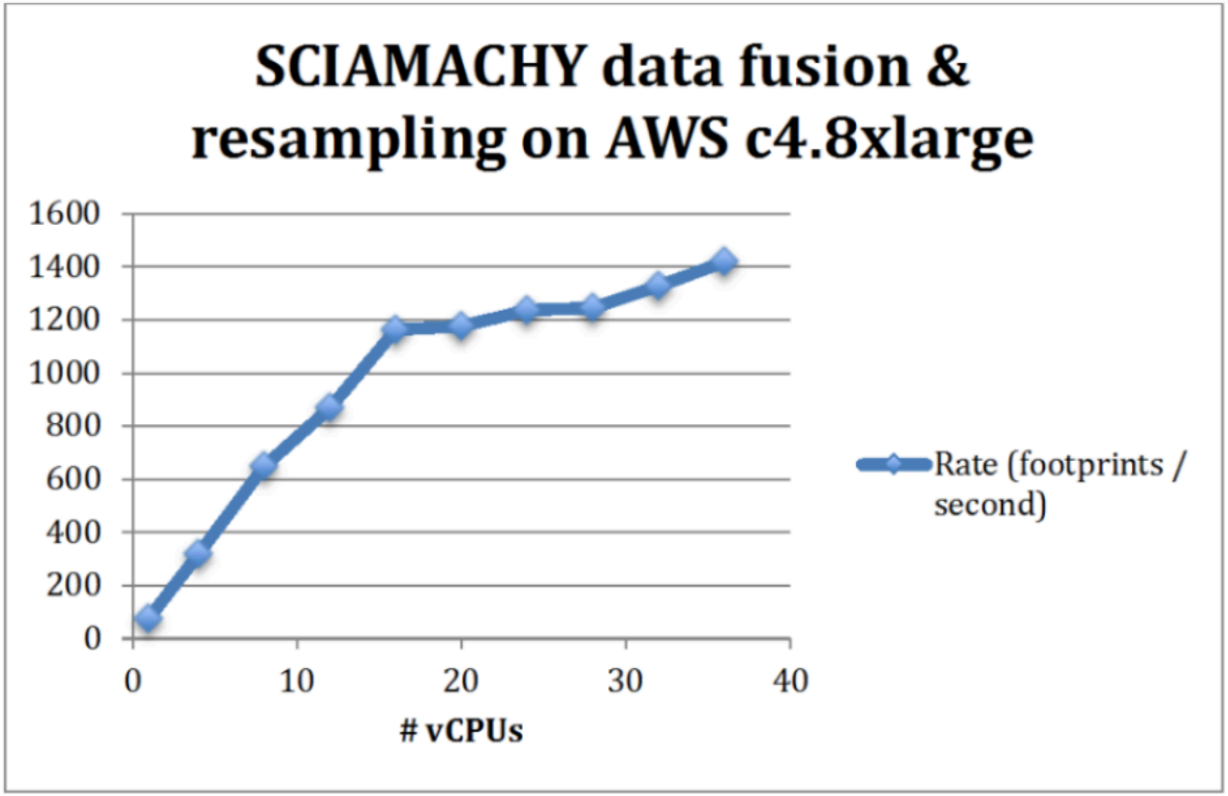
NAIADS: AWS Cloud Architecture



NAIADS Development and Automation and VPC Security and Monitoring: see the NAIADS Report for details.

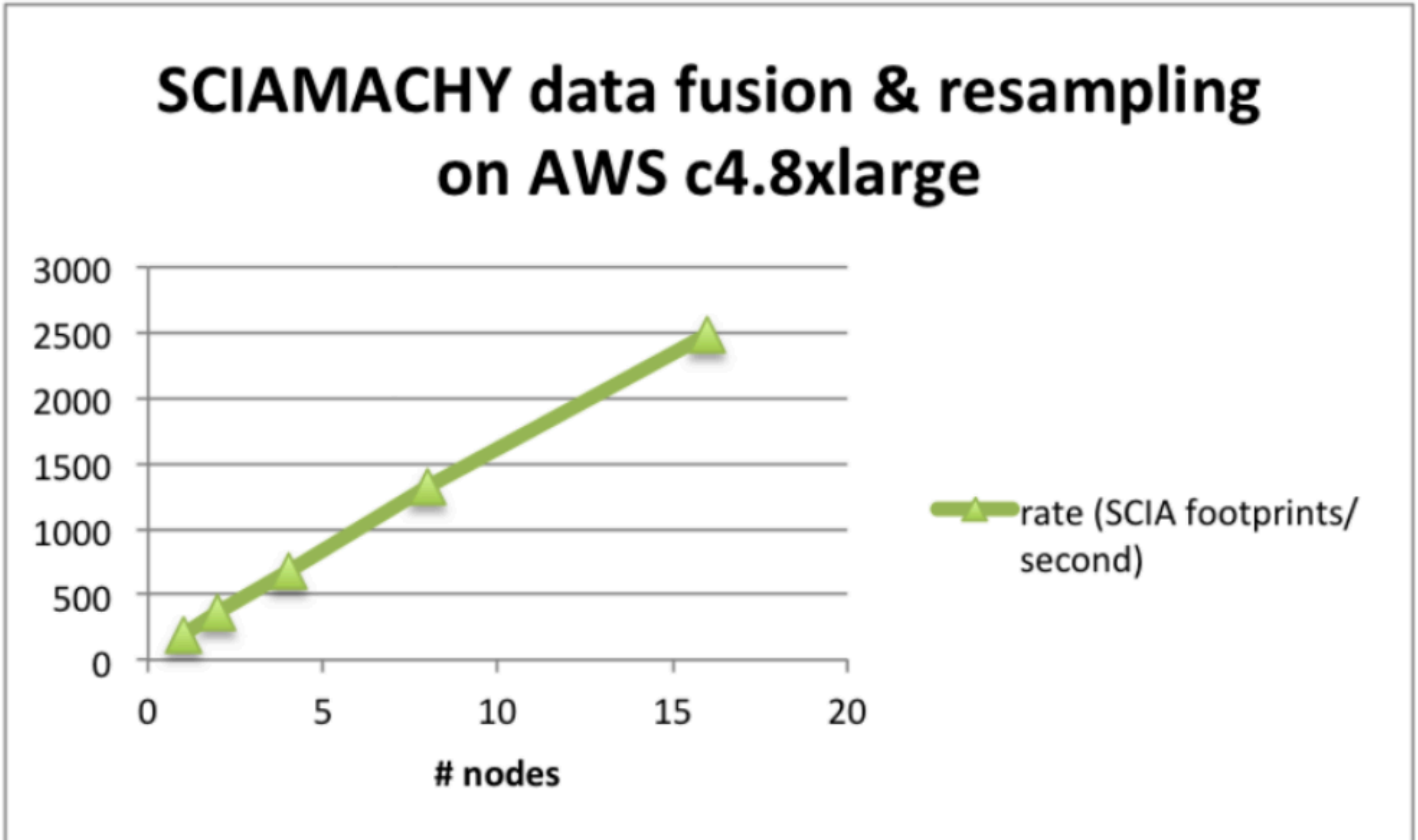
Multi-core Workflow Scalability

- AWS c4.8xlarge instances, 36 vCPUs ~= 18 physical cores
- Rate based on average workflow execution over 10 SCIAMACHY files
- Does not include service configuration & setup.
- Staging data from S3 is time consuming.

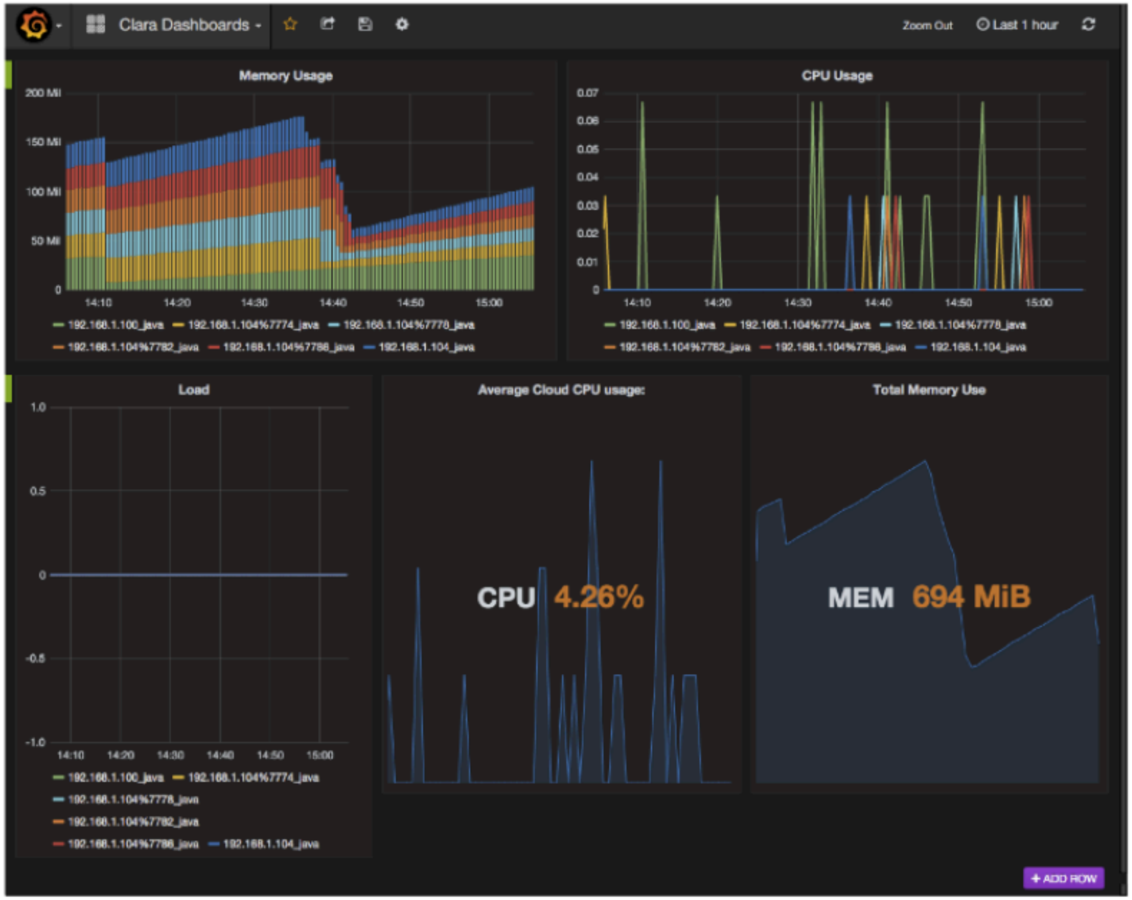


Horizontal Scaling

Linear scaling: improvement from previous tests



NAIADS/CLARA Web Dashboard



Grafana & InfluxDB: tools optimized for real-time process monitoring.

Diagnostic tool for monitoring Services and their performance.

Diagnostic tool for monitoring System performance.

NAIADS Statistics Services (Java):

REQUIREMENTS for Histograms (frequency) and Profiles (averages):

Binned Objects	Methods
H1D (frequency)	Addition/Subtraction; Normalization; Integration; Fitting
H2D (frequency)	Addition/Subtraction; Normalization; Integration; Projections; Fitting
P1D (averages)	Normalization; Combining (+/-): means, STD, counts; Fitting
P2D (averages)	Normalization; Combining (+/-): means, STD, counts; Fitting

KEY PRINCIPLE: Science language is math, focus on numeric results !

PROGRESS:

- Data interface with NetCDF streaming is implemented;
- Data gridding/averaging part is implemented;
- We use numerically stable algorithm for the statistics calculations;
- Methods: add, subtract, normalize, and projection of 2D to 1D are implemented;
- Exporting the data from stats structures into JSON, and using Electron (Javascript library) to make interactive visualization for the browser.

NAIADS Technology Infusion: SRB Heritage Codes

Expected Accomplishments: Demonstrate Data Production Improvement for NASA's Surface Radiation Budget (SRB) Project.

Requirements:

- SRB is planning to process 34+ years of data for scientific analysis and societal benefits;
- Multi-satellite fusion and increasing spatial resolution;
- Become "operational" for regular processing of new observations to lengthen record;

Benefits:

- Improved data production capability to enable efficient production (factor of 10 speed up at least);
- Support additional fusion data sets;
- Support higher data resolution;
- Support faster reprocessing with improved inputs/algorithms.

Modernize Original Codes for High Performance Computing:

Project will require restructuring and partial recoding that will increase modularization and scalability

Resulting codes adaptable to evolving cloud computing environment

NAIADS Reporting and Papers

- Presentation at the ESTF 2015, Pasadena, CA, June 2015 (C. Lukashin)
- Papers at the IEEE Big Data 2015 in Geoscience Workshop, Santa Clara, CA:
    1. *Component Based Dataflow Processing Framework (V. Gyurjyan, et al.)*
    2. *Earth Science Data Fusion with Event Building Approach (C. Lukashin, et al.)*
- Presentation and Demo at the ESTF 2016, Annapolis, MD, June 2016 (A. Bartle).
- ESIP Review, Fall 2016.
- In-depth seminar for the NAIADS and CLARA software design: January 2017
- Final report by the end of 2017.

NAIADS Approach and Expected Impacts:

1. **Novel Data Fusion Approach:**
   Optimization of the IO in large data applications & improving efficiency of process scaling. *Event Builder in offline software.*

2. **Multi-Lingual Solution:**
   Support for most used programming languages (Java, Python, C++), providing framework from new development and heritage codes. Provide choice for optimal implementation for given algorithm.

3. **Generalized Data Statistics:**
   Set of essential statistical analysis tools for data of multiple types: from low to high level data products, observations or models, space or ground experiments.

4. **Flexible Data-Streaming Framework Platform:**
   Workflow adaptability to multiple applications – data processing, mining, metadata generation, sub-setting, data analysis.

5. **Foundation and test bed for a large distributed community data system.**

IMPACT: These technologies has potential to improve efficiency and reduce costs of data handling within Earth Science community.