VANDERBILT/JPL COLLABORATION ON A SCALABLE DATA PROCESSING PIPELINE FOR PROTEOMICS

Kirsten Anton, Luca Cinquini, Maureen Colbert, Dan Crichton (PI), Thomas Fuchs, Sean Kelly, Heather Kincaid, Ashish Mahabal, Chris Mattman, Rishi Verma, Paul Zimdars
California Institute of Technology & Jet Propulsion Laboratory (NASA)

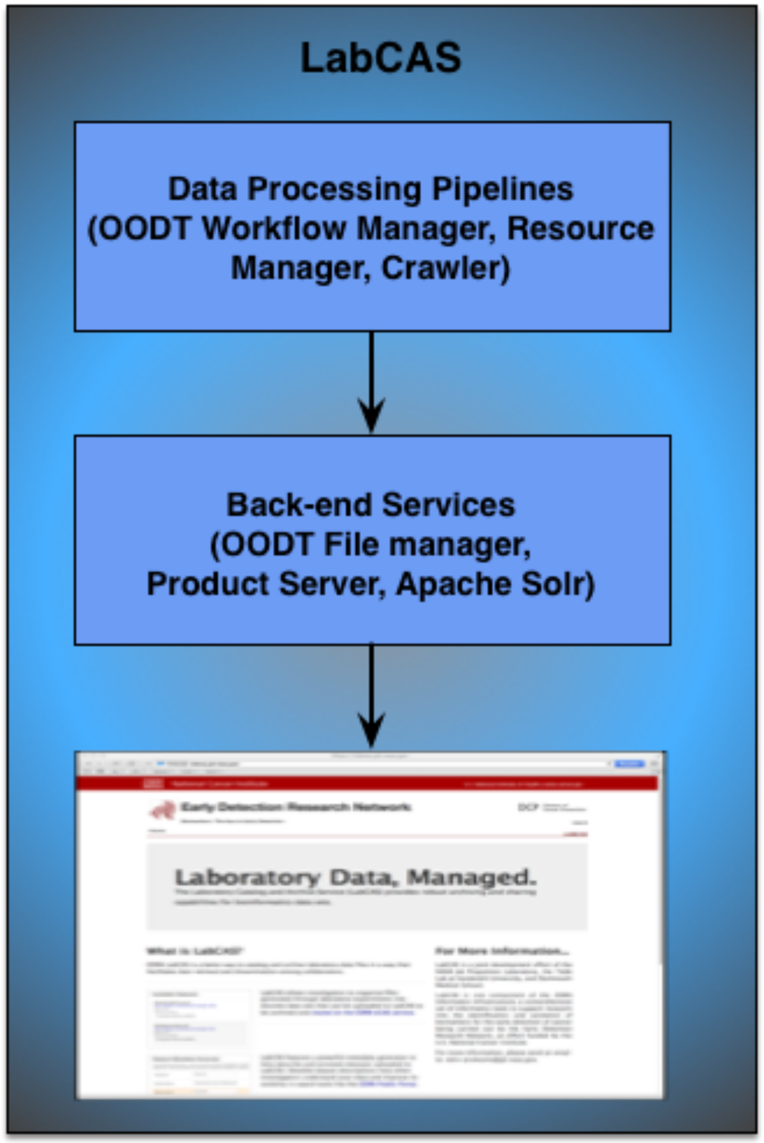
Matthew Chambers, David Tabb (PI)
Vanderbilt University

LabCAS: Laboratory Catalog and Archive Services

- LabCAS is an IT integrated environment for managing biomedical laboratory data: generation, publication, documentation, search, discovery and access
- Overall goal: support research for identification and validation of cancer biomarkers for early detection of cancer - focus on early sharing of results before publication
- Developed by JPL in collaboration with EDRN partners (Vanderbilt, Dartmouth, BU…)

Architecture Software Layers

- Data Processing Pipelines:
  - framework for processing laboratory data as integrated workflows, generate data products, and publish them to the archive
- Back-end Services:
  - components for publishing, searching and downloading data products
- Front-end Web Portal:
  - public site for authorized access of data products, as well as UI for execution of data processing pipelines


LabCAS

Data Processing Pipelines (OODT Workflow Manager, Resource Manager, Crawler)

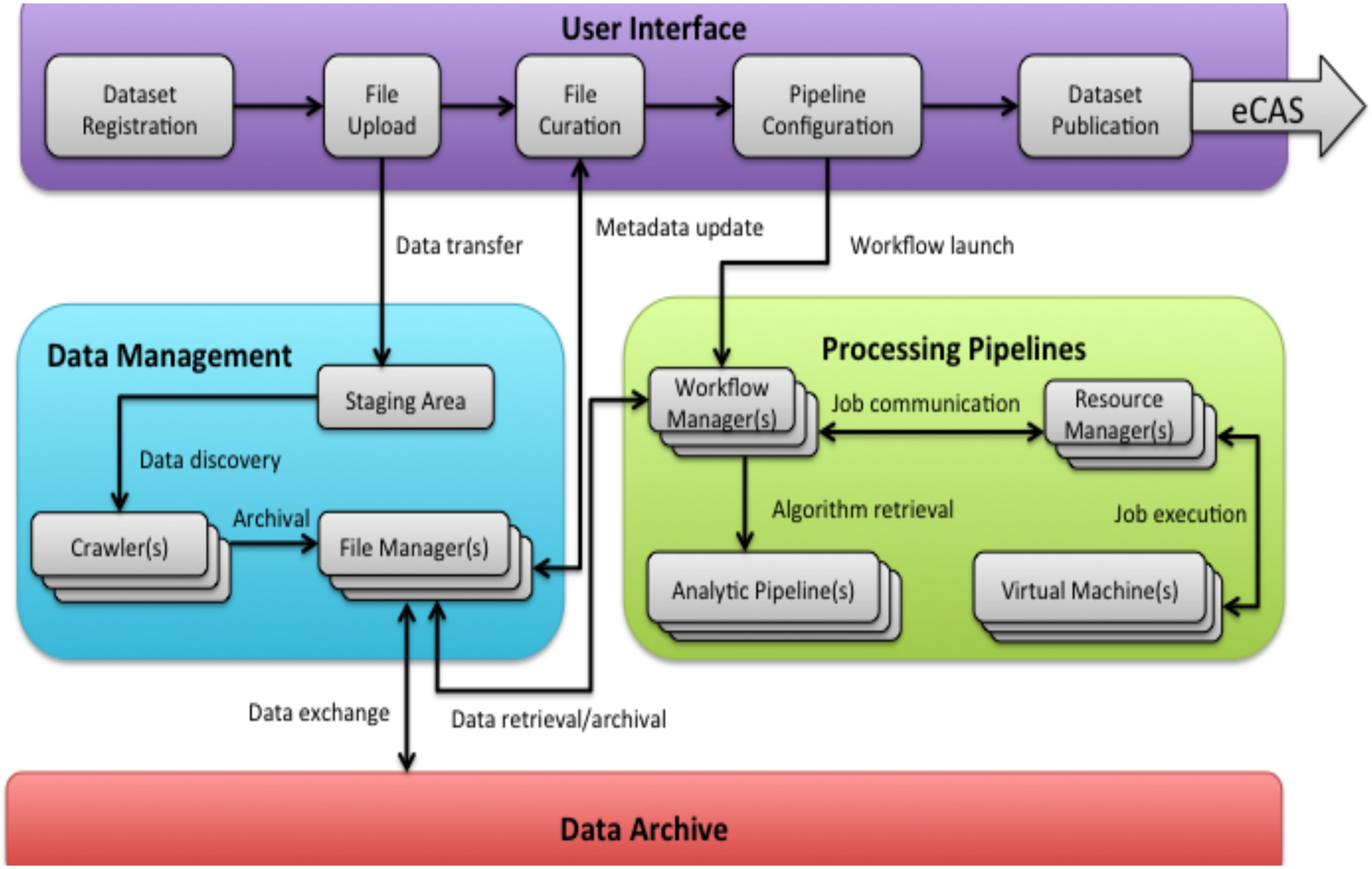
Back-end Services (OODT File manager, Product Server, Apache Solr)

LabCAS: Data Processing Pipelines

Our group is working at establishing a state-of-the-art computing environment at JPL for execution of biology data processing pipelines for the EDRN program.

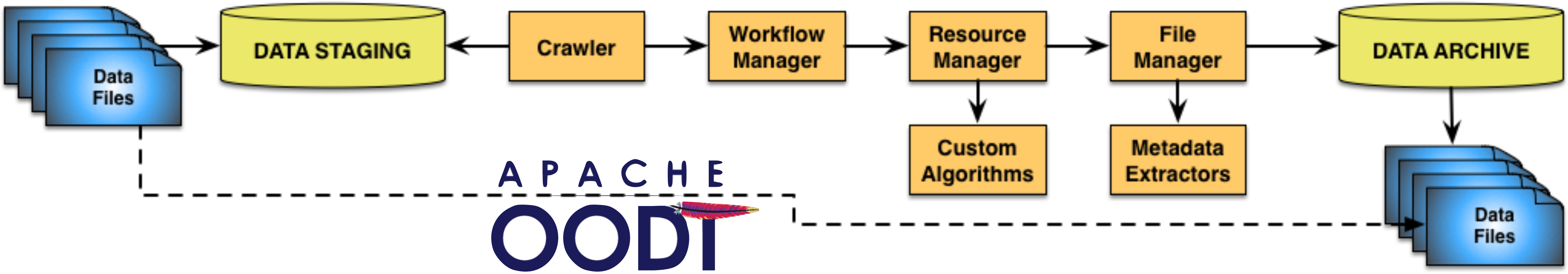
Goals

- Performance: easy configuration, distributed processing, support multiple pipelines
- Reproducibility: capture detailed metadata about workflow execution ("provenance") so that other investigators can reproduce and validate the results
- Sharing: publish data and metadata to the web portal for authorized access by other investigators and the general public



OODT: Object Oriented Data Technology

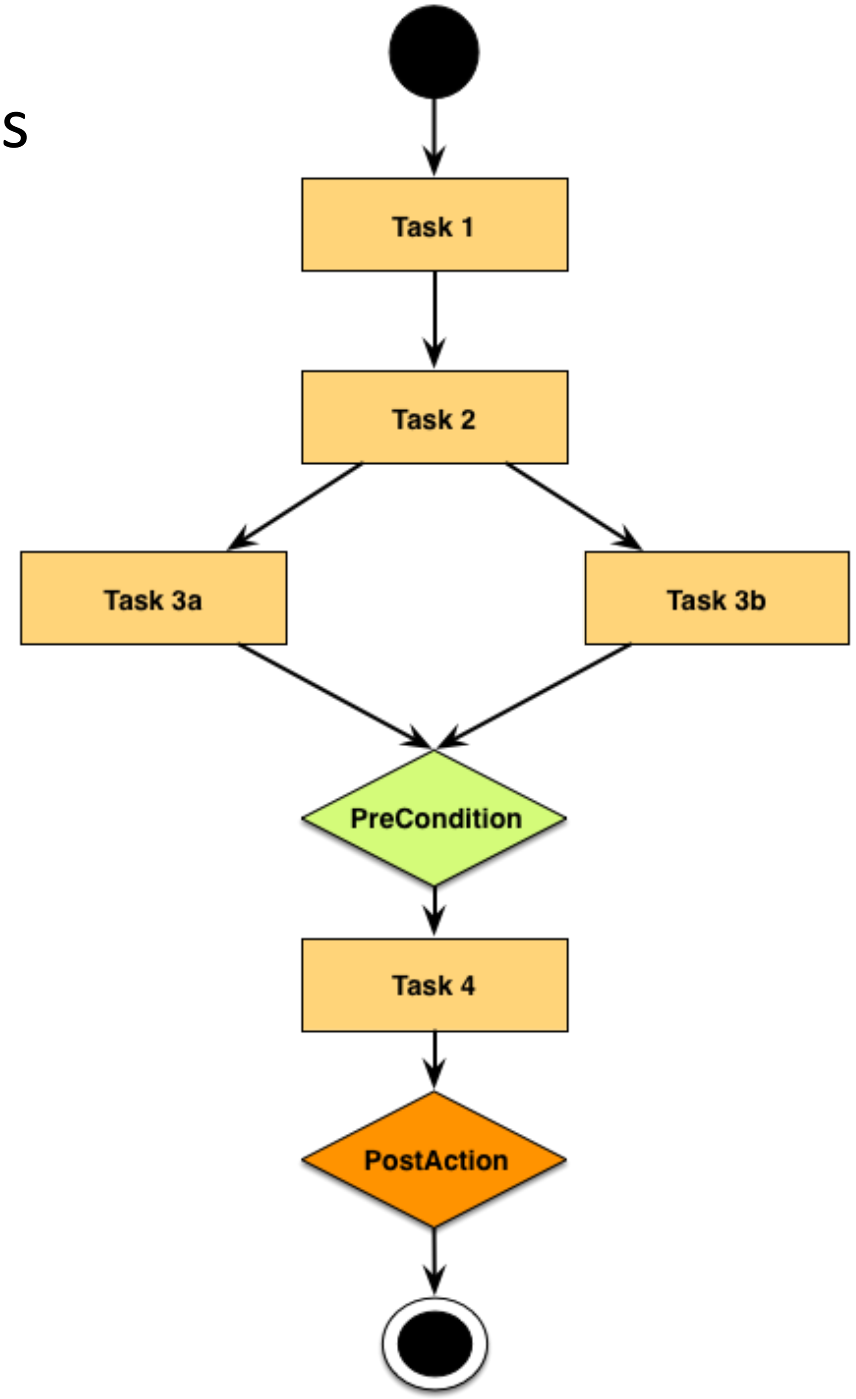
- Apache OODT is an Open Source framework for management, processing, discovery and access of scientific data
- Modular architecture allows to instantiate and combine different components to realize the most appropriate architecture for a specific data processing environment
  - File Manager: data access server and metadata catalog. May be backed up by Apache Solr - web-enabled, high performance search engine
  - Workflow Manager: general workflow engine for execution of pipelines composed of sequential or parallel tasks
  - Resource Manager: allocates computing resources for task execution
  - Crawler: service for monitoring data spaces and to trigger ingestion of files into the File Manager or submission of jobs to Workflow Manager



OODT Workflows Features

- Each Workflow Manager server can be configured to execute one or more workflows
- A client starts a workflow by sending an “event” with optional configuration metadata
- Each workflow is composed of an arbitrary number of tasks (sequential or parallel)
- For scalability, tasks can be run on the local Resource Manager, or sent to Resource Managers on other servers
- Optional pre-conditions cause task execution to wait until they are satisfied
- Optional post-actions trigger operations when a workflow terminates (on success or failure)
  - Example: start crawling for products to ingest
- Workflow products are categorized according to custom types:
  - Specific metadata elements
  - Specific archive location and versioner
  - Specific metadata parsers

APACHE OODT



LabCAS Pipelines Collaborations

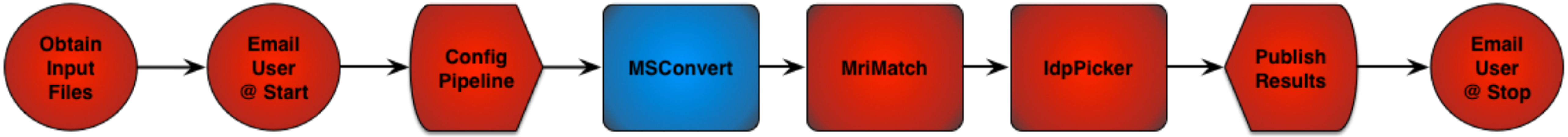
JPL is collaborating with several EDRN partners to enable their data processing pipelines to be executed within the LabCAS environment:

- First, understand and run the different stages of a pipeline as standalone tasks
- Then, instrument all stages as single runnable OODT workflow

- Vanderbilt University (PI: Dave Tabb)
  - Proteomics
- Boston University (PI: Marc Lenburg)
  - Biomarker Discovery
  - RNA Sequencing
  - Microarray pre-processing
- University of Washington (PI: Alvin Liu)
  - Microarray data
- Memorial Sloan Kettering Cancer Center (PI: Gunnar Ratsch)
  - Genomics
- Cedar-Sinai Medical Center (PI: Beatrice Knudsen)
  - Pathology
- Cedar-Sinai Medical Center (PI: Michael Freeman)
  - Proteomics

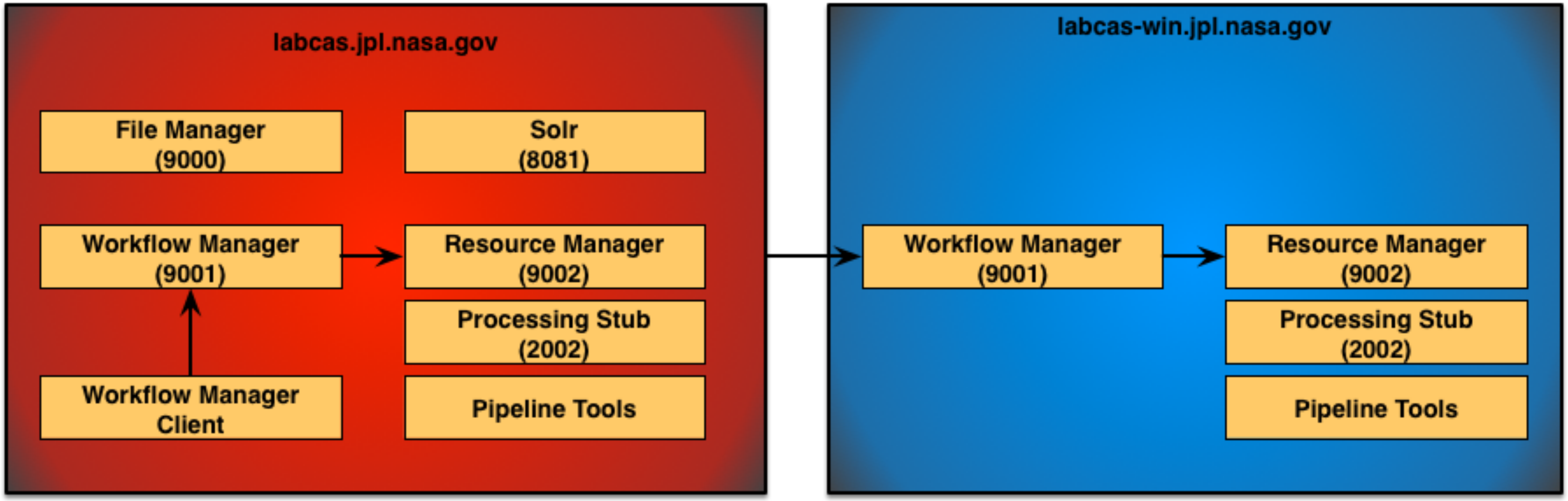
Proteomics Pipeline

- Data processing pipeline for identification and analysis of protein cancer biomarkers in body fluids
- Developed by David Tabb's group at Vanderbilt, part of CPTAC (Clinical Proteomic Tumor Analysis Consortium) activities
- Composed of several programs (part of ProteoWizard suite of open source tools):
  - MSConvert (pre-processing): conversion of RAW files to mzML format
  - MyriMatch (database search): searches sample data for peptides (.pepXML)
  - IdpPicker (filtering: IdPQonvert+IdPAssemble+IdPQuery): generates protein identification reports
- Instrumented as single OODT workflow executing a sequence of 13 tasks
- First test case at JPL consisted in running the pipeline on a medium-size pancreatic dataset composed of 675 input files of approx. 200MB each (total: 1.35TB)



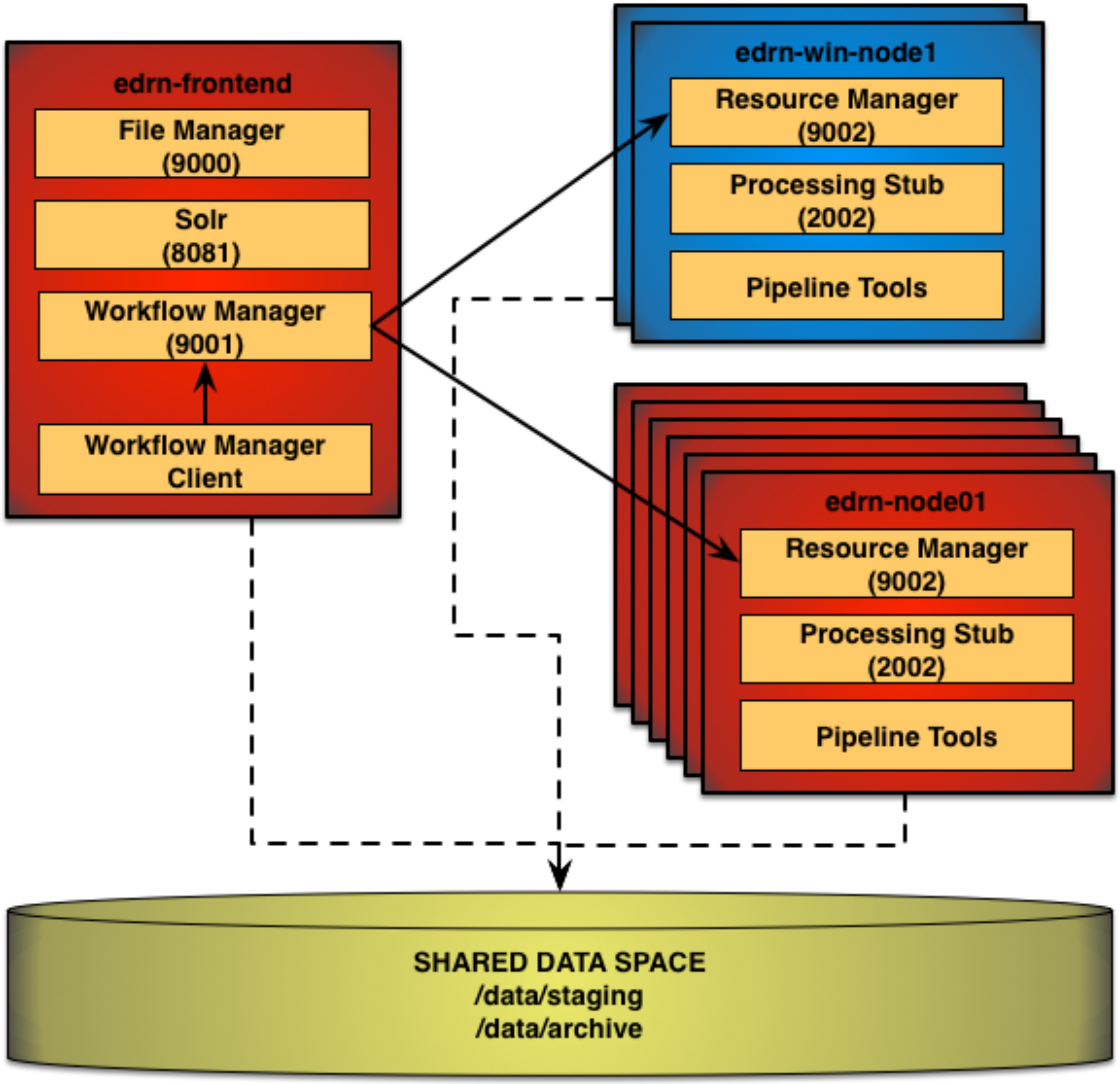
Proteomics Pipeline @ JPL : 1st attempt

- Use system of 1 Linux + 1 Windows servers (2 cores each)
- Each task setup to run over all files sequentially
- Tasks 1-7 (up to MSConvert) completed in about 1 week on Windows server
- Task 8 (MyriMatch) started on Linux server, stopped after estimating 40+ days before completion (approximately 1.5 hours for each file)



New EDRN Cluster Environment @ JPL

- System of 21 Virtual Machines (6 cores each)
  - 1 front-end server hosting common services
  - 18 Linux back-end processing nodes
  - 2 Windows back-end processing nodes
  - 10TB shared storage
- Automatic software replication from 1 processing node to all the others
- Execution of commands from front-end node to all the back-end nodes
  - Start/stop OODT services on all node simultaneously

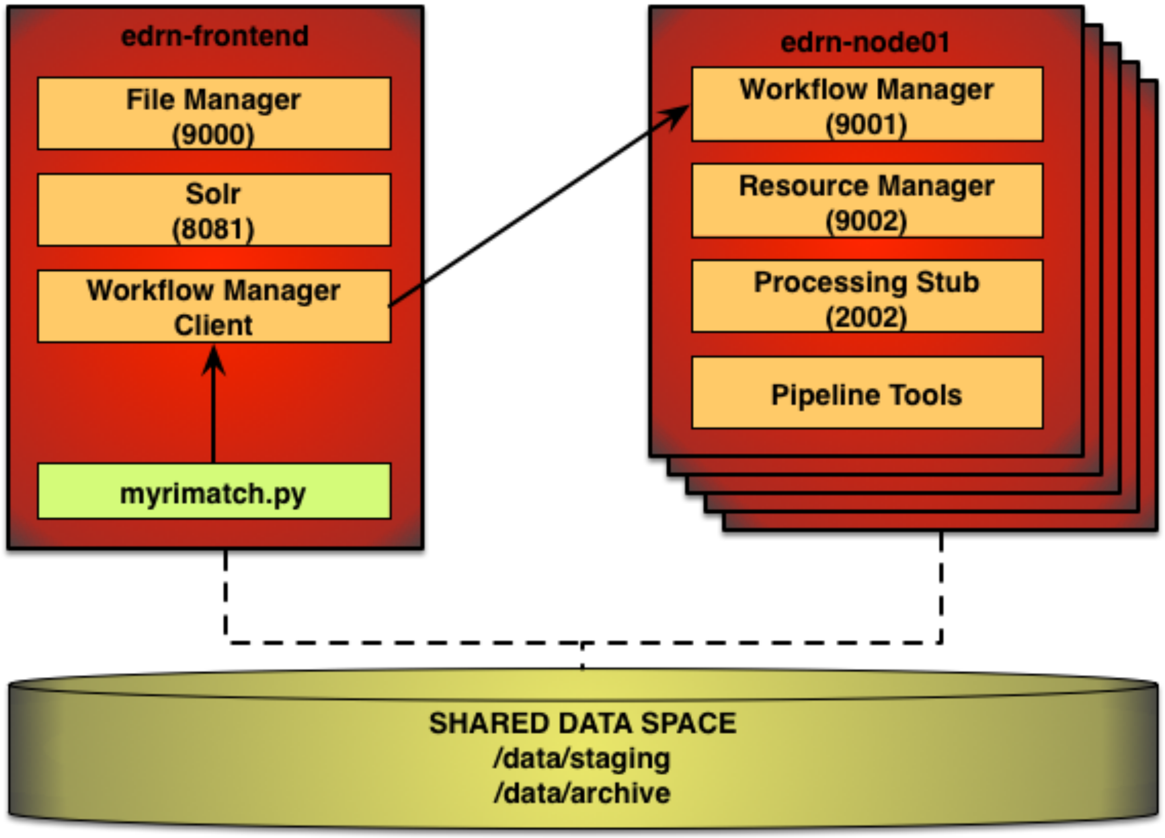


Proteomics Pipeline @ JPL : 2nd attempt

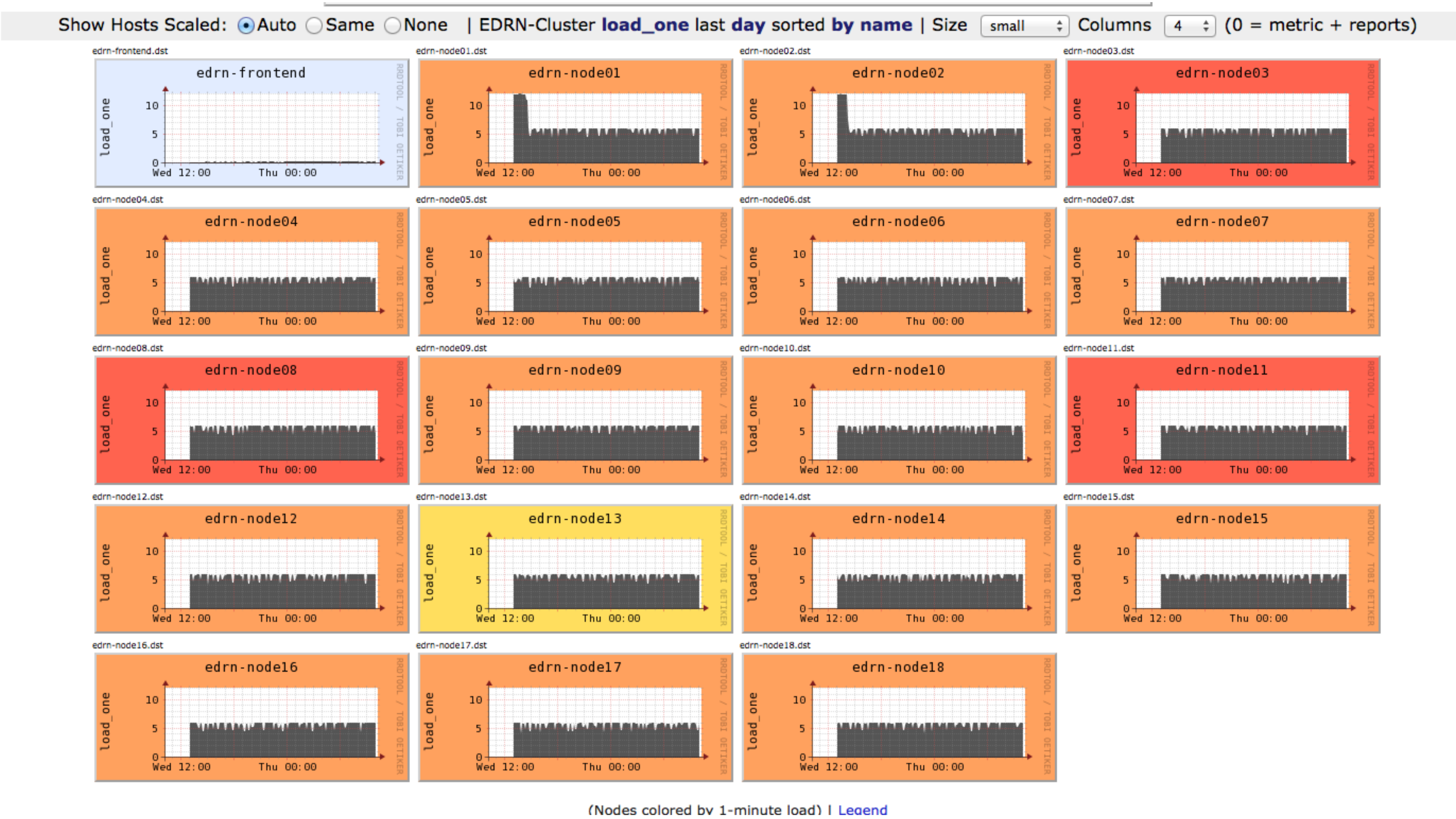
- Use new JPL/EDRN cluster
- MyriMatch step was identified as bottle-neck in Proteomics pipeline
- Because MyriMatch only processes 1 mzML file at a time, this step can be fully parallelized!
- Using mzML files already produced by the old pipeline, sub-workflow was setup to run MyriMatch in parallel over all available Linux nodes

Results

- Single MyriMatch job completes in approx. 30min (instead of 90+ min)
- Full MyriMatch sub-workflow completed in approx. 1 day (for all 675 files)!
- Performance improvements:
  - Factor of 3 from using more powerful hardware
  - Factor of 18 from parallelizing a single step onto multiple processing nodes



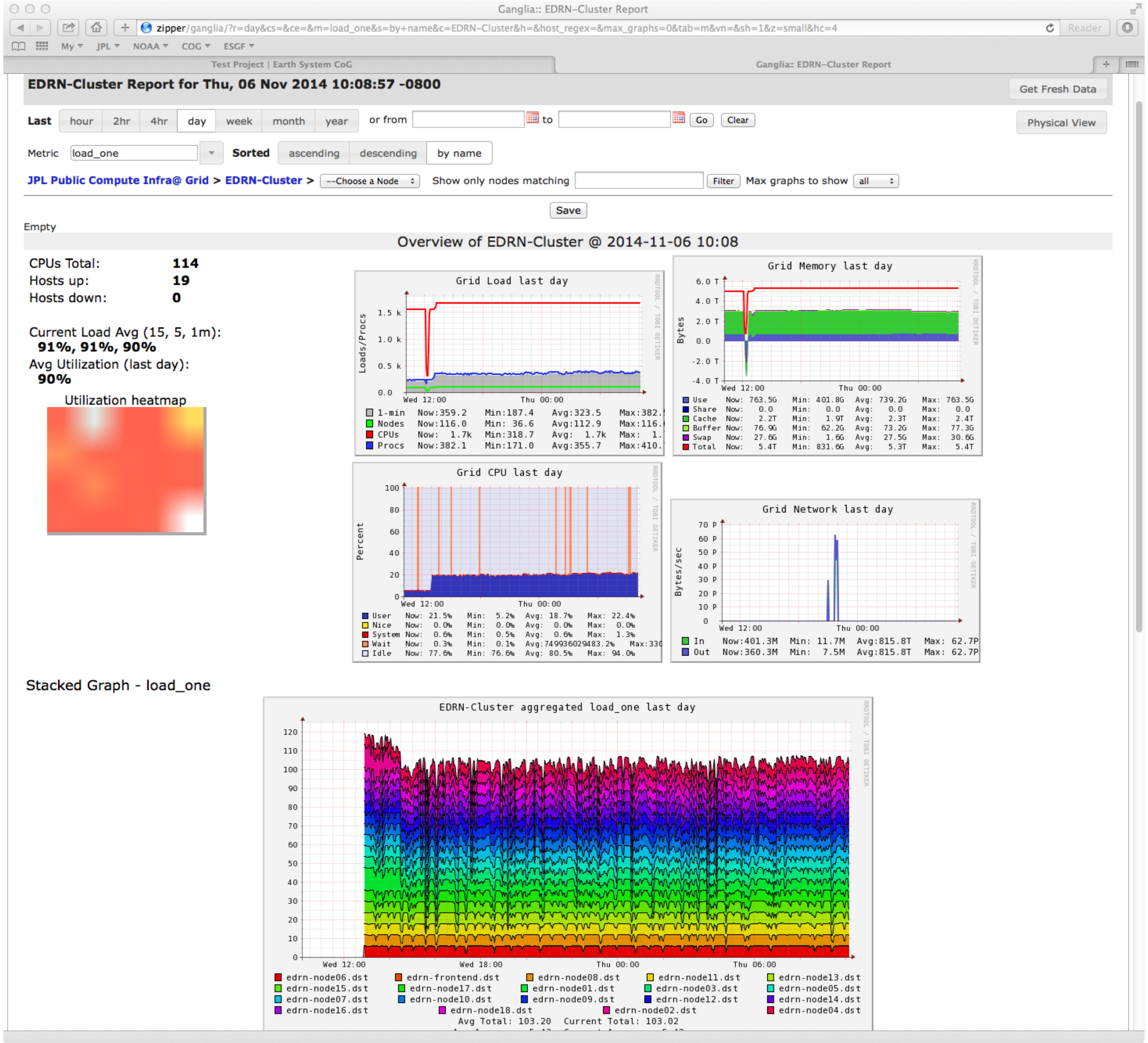
JPL/EDRN Cluster Usage (1 day)



Show Hosts Scaled: ● Auto ○ Same ○ None | EDRN-Cluster **load_one** last **day** sorted **by name** | Size small Columns 4 (0 = metric + reports)

(Nodes colored by 1-minute load) | Legend

JPL/EDRN Cluster Usage (1 day)


Ganglia:: EDRN-Cluster Report
zipper/ganglia/?r=day&cs=&ce=&m=load_one&s=by+name&c=EDRN-Cluster&h=&host_regex=&max_graphs=0&tab=m&vn=&sh=1&z=small&hc=4
Reader
My ▾ JPL ▾ NOAA ▾ COG ▾ ESGF ▾
Test Project | Earth System CoG
Ganglia:: EDRN-Cluster Report

EDRN-Cluster Report for Thu, 06 Nov 2014 10:08:57 -0800

Get Fresh Data
Last hour 2hr 4hr day week month year or from to Go Clear
Physical View
Metric load_one
Sorted ascending descending by name
JPL Public Compute Infra@ Grid > EDRN-Cluster > --Choose a Node
Show only nodes matching Filter Max graphs to show all
Save

Empty

Overview of EDRN-Cluster @ 2014-11-06 10:08

CPUs Total:	**114**
Hosts up:	**19**
Hosts down:	**0**

Current Load Avg (15, 5, 1m):
91%, 91%, 90%
Avg Utilization (last day):
90%

Utilization heatmap

Grid Load last day
Loads/Procs
1.5 k
1.0 k
0.5 k
0.0
Wed 12:00
Thu 00:00
1-min Now:359.2 Min:187.4 Avg:323.5 Max:382.
Nodes Now:116.0 Min: 36.6 Avg:112.9 Max:116.
CPUs Now: 1.7k Min:318.7 Avg: 1.7k Max: 1.
Procs Now:382.1 Min:171.0 Avg:355.7 Max:410.
Grid Memory last day
Bytes
6.0 T
4.0 T
2.0 T
0.0
-2.0 T
-4.0 T
Wed 12:00
Thu 00:00
Use Now: 763.5G Min: 401.8G Avg: 739.2G Max: 763.5G
Share Now: 0.0 Min: 0.0 Avg: 0.0 Max: 0.0
Cache Now: 2.2T Min: 1.9T Avg: 2.3T Max: 2.4T
Buffer Now: 76.9G Min: 62.2G Avg: 73.2G Max: 77.3G
Swap Now: 27.6G Min: 1.6G Avg: 27.5G Max: 30.6G
Total Now: 5.4T Min: 831.6G Avg: 5.3T Max: 5.4T
Grid CPU last day
Percent
100
80
60
40
20
0
Wed 12:00
Thu 00:00
User Now: 21.5% Min: 5.2% Avg: 18.7% Max: 22.4%
Nice Now: 0.0% Min: 0.0% Avg: 0.0% Max: 0.0%
System Now: 0.6% Min: 0.5% Avg: 0.6% Max: 1.3%
Wait Now: 0.3% Min: 0.1% Avg:74993602948 3.2% Max:330
Idle Now: 77.6% Min: 76.6% Avg: 80.5% Max: 94.0%
Grid Network last day
Bytes/sec
70 P
60 P
50 P
40 P
30 P
20 P
10 P
0
Wed 12:00
Thu 00:00
In Now:401.3M Min: 11.7M Avg:815.8T Max: 62.7P
Out Now:360.3M Min: 7.5M Avg:815.8T Max: 62.7P
RRDTOOL / TOBI OETIKER

Stacked Graph - load_one

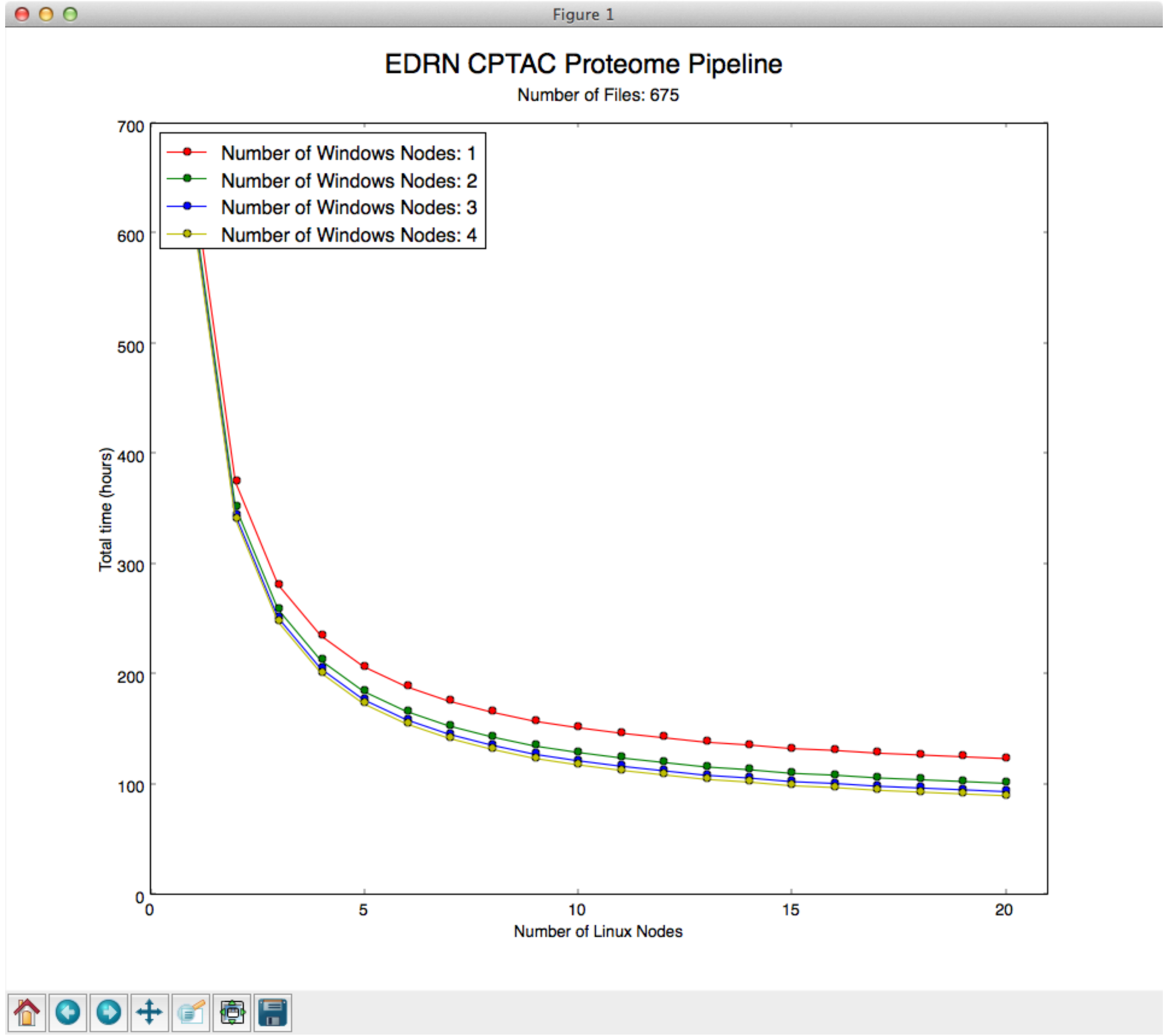
EDRN-Cluster aggregated load_one last day
120
110
100
90
80
70
60
50
40
30
20
10
0
Wed 12:00
Wed 18:00
Thu 00:00
Thu 06:00
edrn-node06.dst edrn-frontend.dst edrn-node08.dst edrn-node11.dst edrn-node13.dst
edrn-node15.dst edrn-node17.dst edrn-node01.dst edrn-node03.dst edrn-node05.dst
edrn-node07.dst edrn-node10.dst edrn-node09.dst edrn-node12.dst edrn-node14.dst
edrn-node16.dst edrn-node18.dst edrn-node02.dst edrn-node04.dst
Avg Total: 103.20 Current Total: 103.02

DAWN Simulation for Proteomics Pipeline

DAWN (Distributed Analytics, Workflows and Numeric): model to simulate and optimize Big Data computational pipelines, developed at JPL under the Data Science initiative

Analysis of Proteomics distributed workflow (using rough benchmarking):

- Clear reduction in overall elapsed time when MyriMatch step is executed on multiple nodes
- Efficiency gain levels off around 15 nodes
- Another gain is obtained by using 2 Windows nodes instead of 1
- Full pipeline should complete in 4-5 days



Future Work

Proteomics Pipeline

- Re-run and benchmark full pipeline on new hardware for pancreatic test case
  - Submitted "report" to Vanderbilt to fix IdpPicker bug on Linux OS
- Run pipeline on other datasets
- Finalize pipeline product metadata
- Publish generated products to LabCAS

LabCAS

- Total redesign of Front-end Web Portal
  - Streamlined uploads of data files
  - Flexible UI for execution of data processing pipelines
  - Enhanced searching of datasets (by keywords or facets)
  - Improved access to documentation
- Back-end Services and Pipelines
  - Ability to package specific workflows for deployment at other institutions
  - Develop test suite
  - Upgrade to latest OODT release