DYNAMO

Funded by the National
Science Foundation
Grant #1826997

Application Aware Software Defined Flows of Workflow Ensembles

George Papadimitriou, Eric Lyons, Cong Wang, Komal Thareja, Ryan Tanaka,
Paul Ruth, J. J. Villalobos, Ivan Rodero, Ewa Deelman, Michael Zink, Anirban Mandal

Innovating the Network for Data-Intensive Science Workshop
Supercomputing 20 - November 12th, 2020

renci

USC Viterbi
School of Engineering
Information Sciences Institute

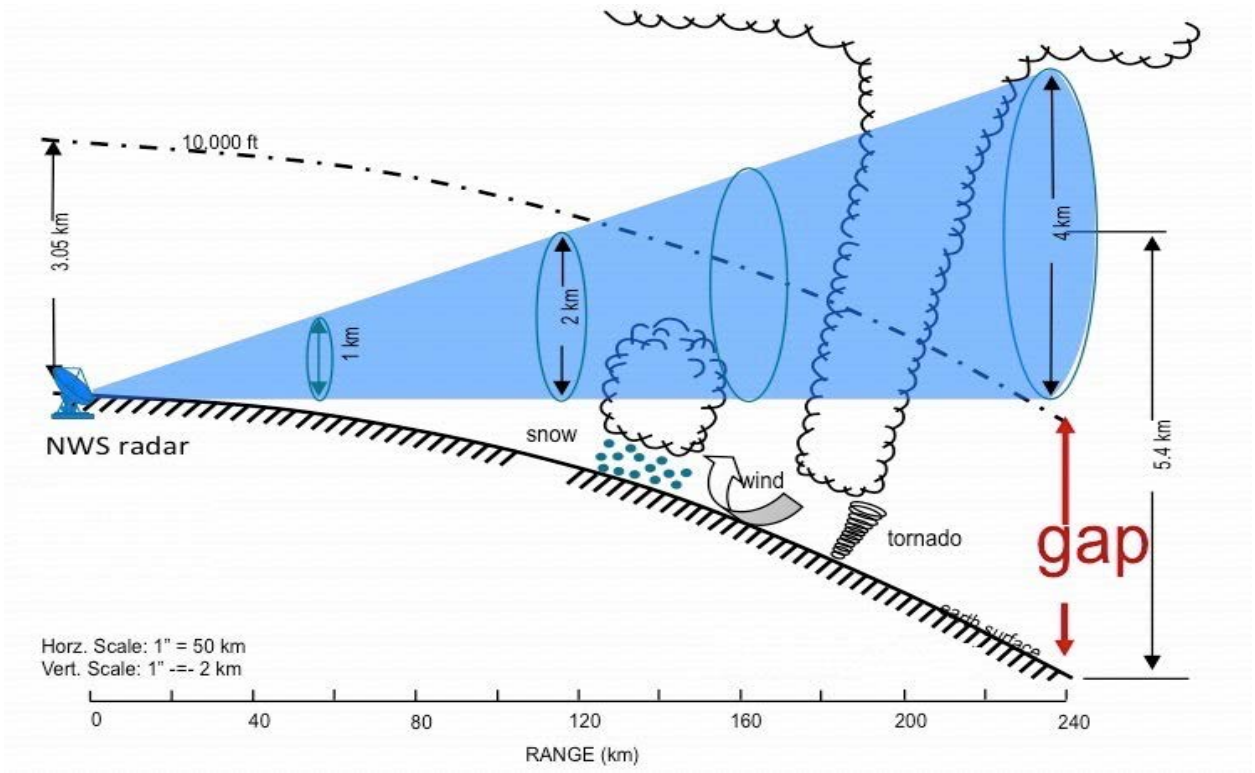
UMass Amherst

RUTGERS

Motivation
Characteristics of Today's Computational Science

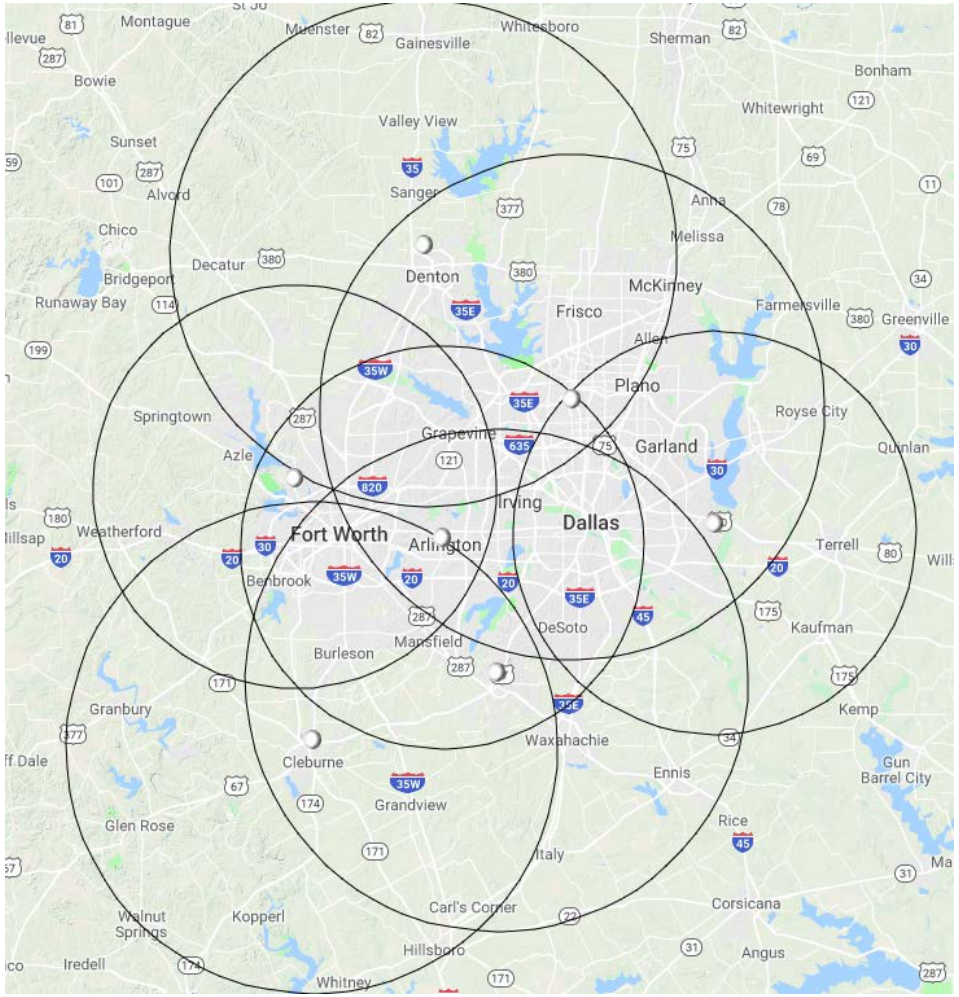
- Data transfer and compute intensive
- Complex workflows
- Distributed data repositories
- Highly distributed compute locations
- Major challenge: Integration of cyber infrastructure to science workflows

Motivation
CASA: Collaborative Adaptive Sensing of the Atmosphere

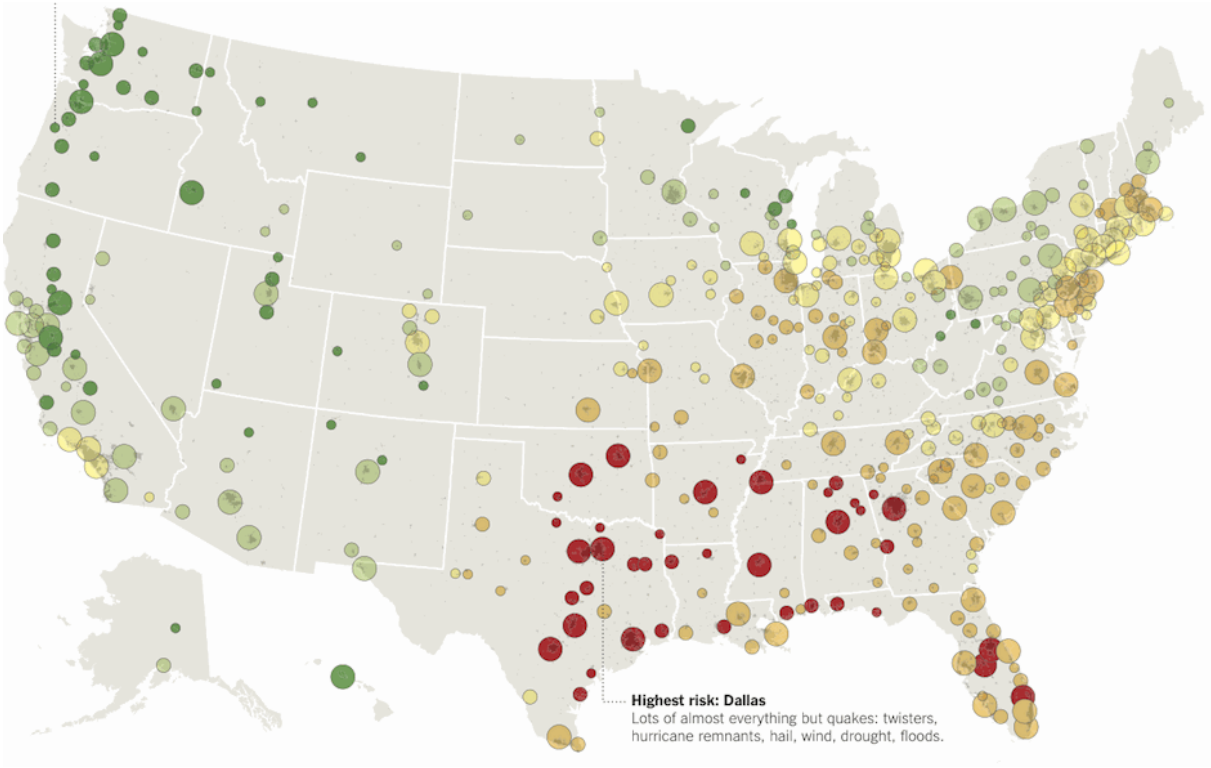


- Traditional Next Generation Weather Radars (NEXRAD)
  - High power, long range
  - Limited ability to observe the lower part of the atmosphere because of the Earth's curvature
- CASA
  - Network of short range Doppler radars
  - Adjustable sensing modes in response to quick weather changes
  - Suitable for near-ground weather events: tornado, hail, high winds

Motivation
CASA DFW Network



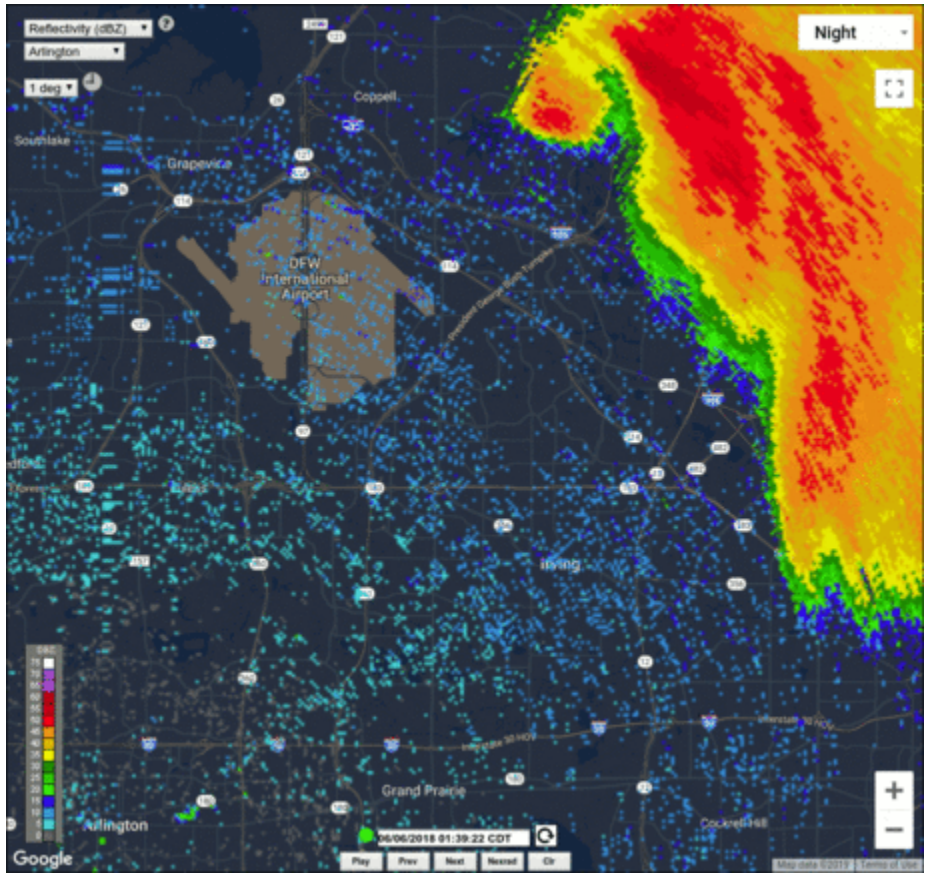
- > 7M people, >100K businesses, >1500 Corporate HQs



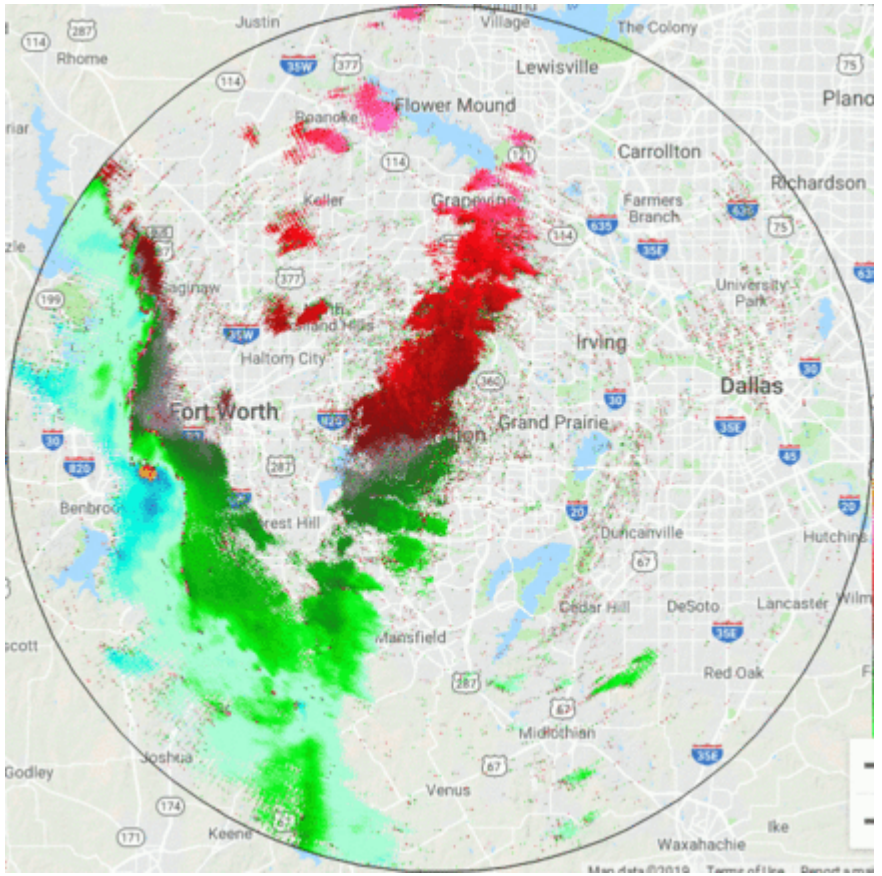
Motivation
CASA Data Rates

- CASA exposes complex network and compute requirements
  - ~100 Mbps per radar raw data, processed locally
  - ~10 Mbps per radar down-sampled moment data
  - ~1 Mbps combined radar product and image data
- Data transfer to DFW Radar Operations Center at NOAA Southern Region Headquarters (SRH)
- Data transfer to Univ. Of North Texas for data collection and further processing

Single radar Reflectivity

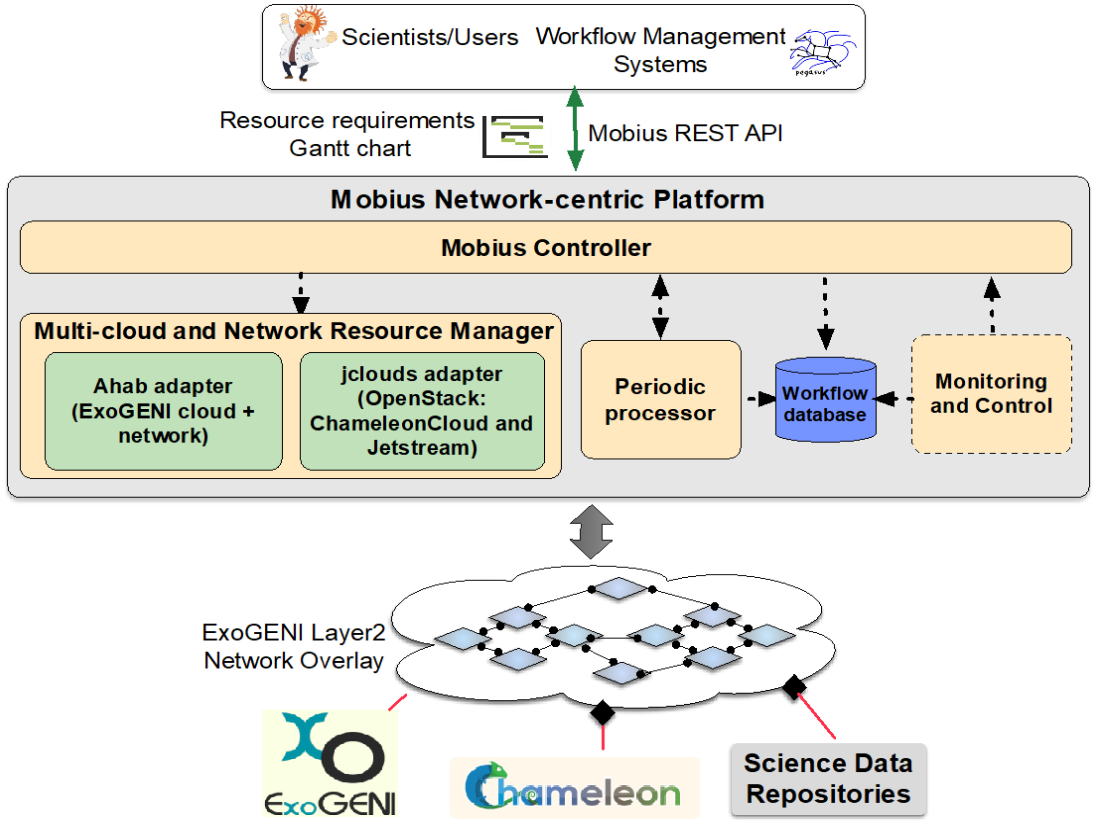


Single radar Velocity



DYNAMO

DyNamo
Workflow Provisioning and Orchestration



- Developed novel algorithms and mechanisms to optimize data flows across different national CI
- Mobius
  - A network-centric platform
  - Provides an abstraction layer over different resources
  - Achieves dynamic resource provisioning across ExoGENI, Chameleon, XSEDE and more.
- Data-aware workflow scheduling using Pegasus Workflow Management System
- Deployed solutions for usecases in observational science communities, such as CASA.

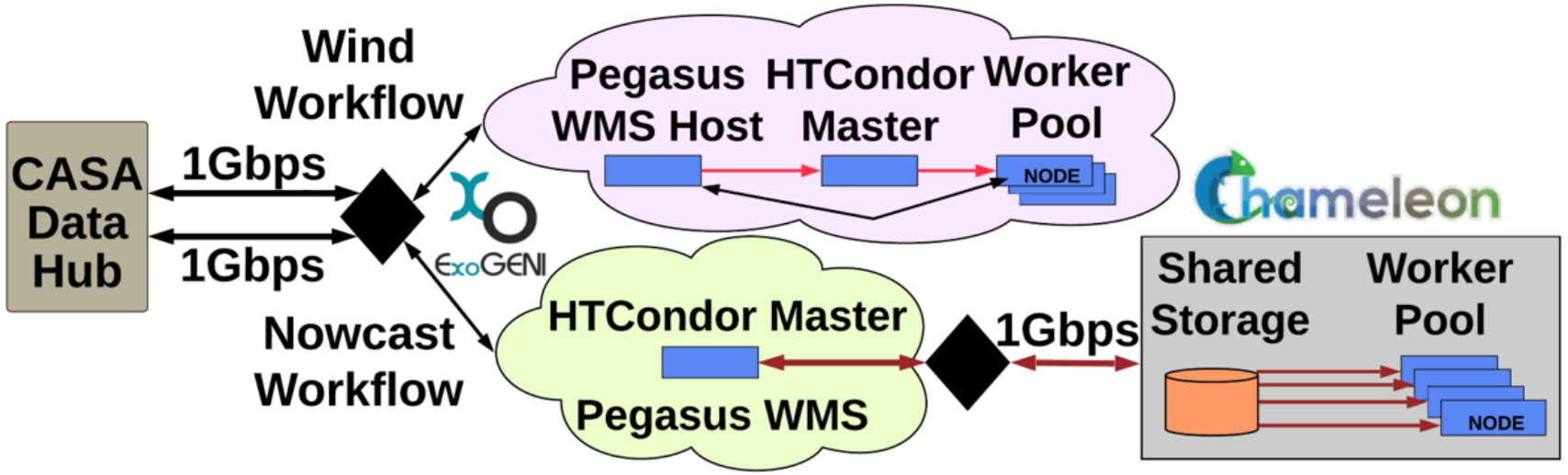
E. Lyons, G. Papadimitriou, C. Wang, K. Thareja, P. Ruth, J. J. Villalobos, I. Rodero, E. Deelman, M. Zink, and A. Mandal, "Toward a Dynamic Network-centric Distributed Cloud Platform for Scientific Workflows: A Case Study for Adaptive Weather Sensing," in *15th International Conference on eScience (eScience)*, 2019, p. 67–76.

renci

USC Viterbi School of Engineering Information Sciences Institute

UMass Amherst

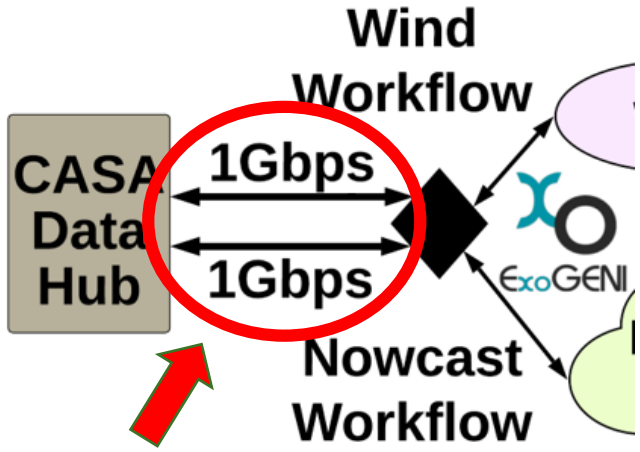
RUTGERS



- High speed data movement via ExoGENI's dedicated layer-2 overlay networks
- Compute and storage resources on both ExoGENI and Chameleon clouds
- Dynamic resource provisioning on ExoGENI and Chameleon clouds
- Workflow instrumentation with Pegasus WMS and HTCondor

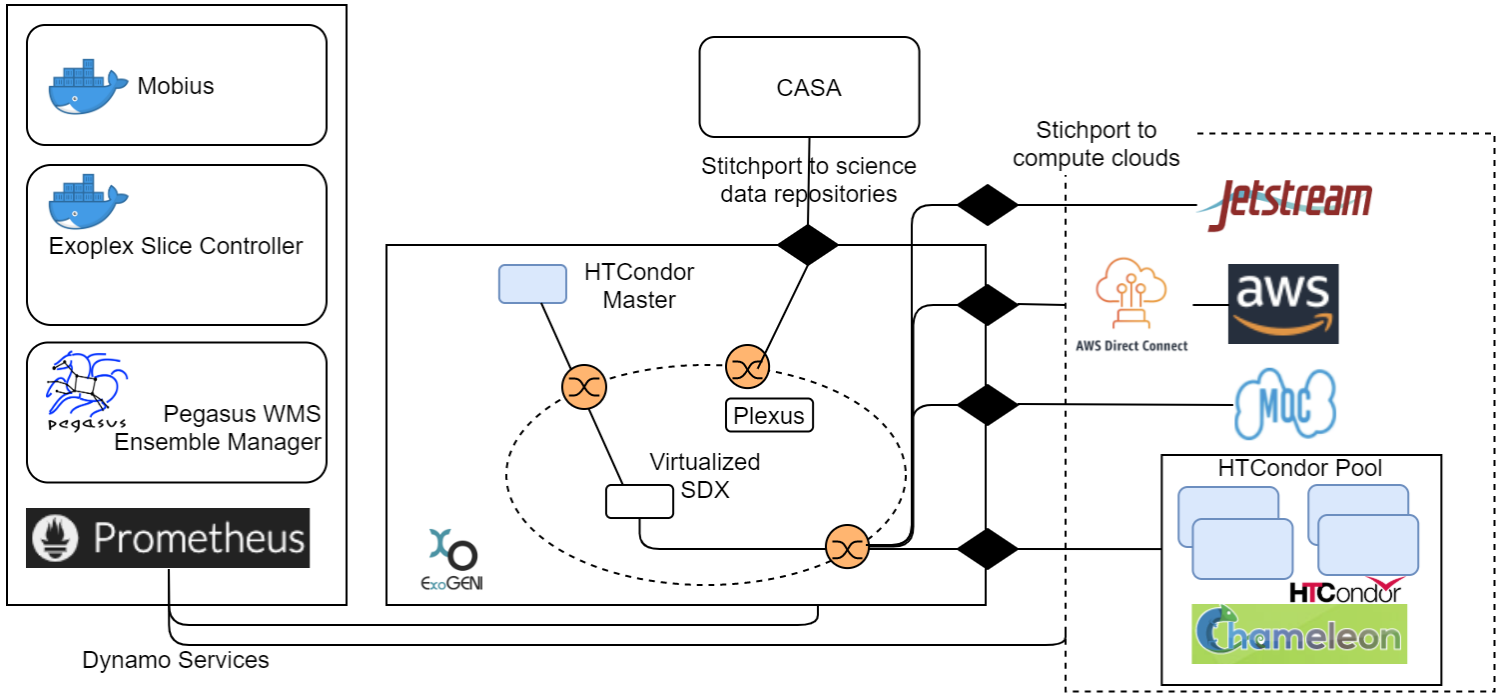
What's missing though?

- To provide isolation and performance guarantees each workflow should use its own dedicated layer2 connection
  - Sharing network resources fairly is not possible
- CASA's workflows have an event character, triggering new workflow executions as new input files arrive, creating workflow ensembles
  - External scripts are used to create and submit new workflow DAGs
  - No control over the workflow ensembles
- Monitoring the infrastructure in an easy and intuitive way is not available



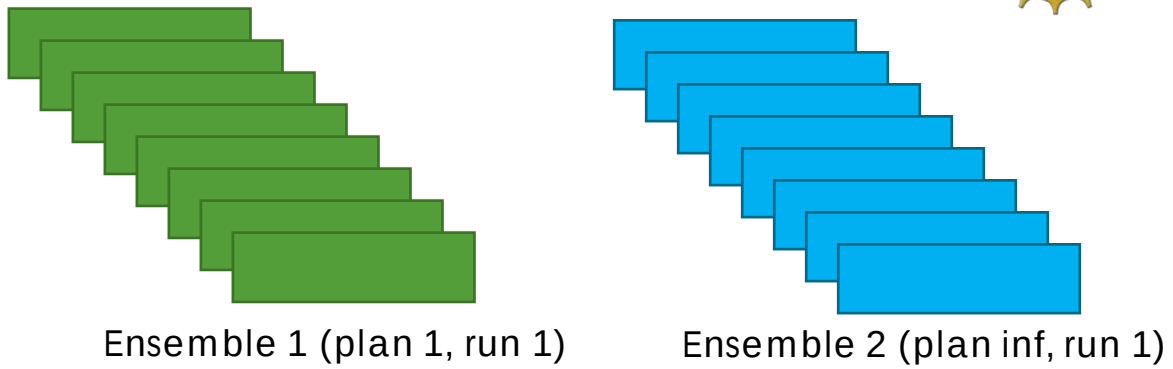
Approach

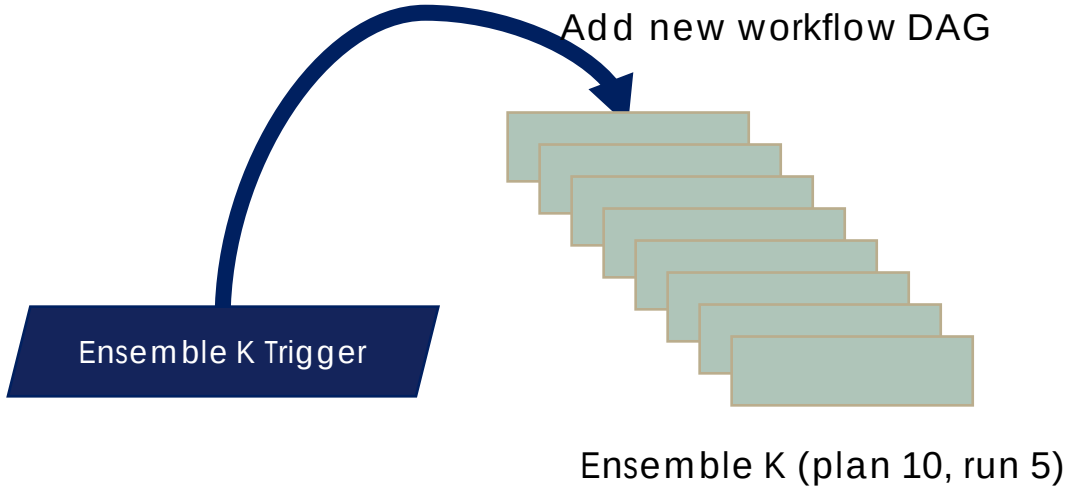
- We have extended the Pegasus Ensemble Manager to support the triggering functionality that CASA's workflows require
- We have enhanced the network capability of DyNamo with virtual software defined exchange functionality
- We have incorporated the Prometheus monitoring system into DyNamo provisioned resources



Approach
Pegasus Ensemble Manager

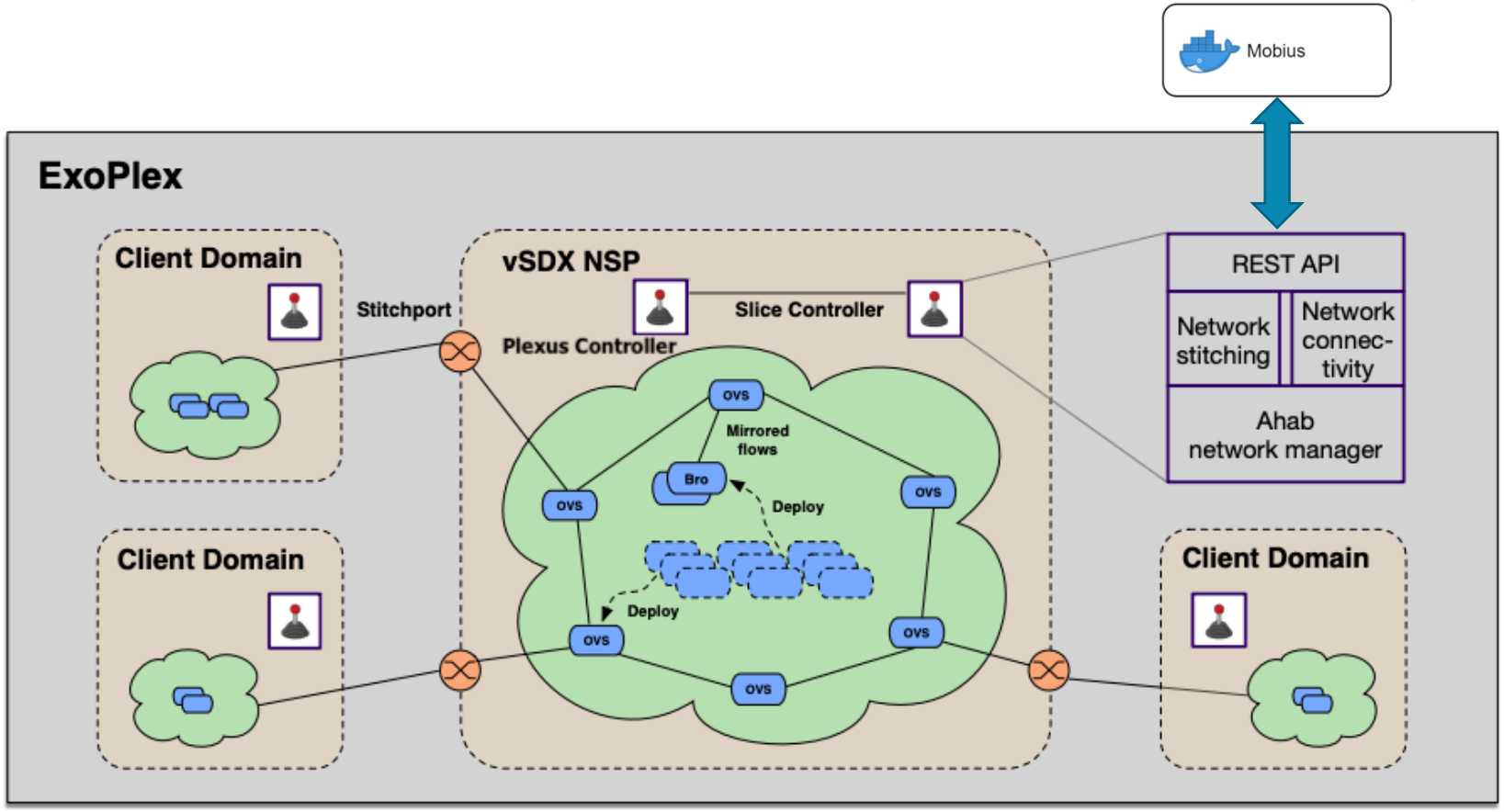
- Pegasus via its Ensemble Manager service can handle collections of related workflows
- The Pegasus Ensemble Manager (Pegasus-EM) controls
  - Number of workflows in an ensemble that can be planned concurrently
  - Number of workflows in an ensemble that can be executed concurrently
- Pegasus-EM supports trigger-based workflow creation and execution
  - It monitors for new files matching a given pattern
  - And generates new workflow DAGs based on user defined workflow generation scripts





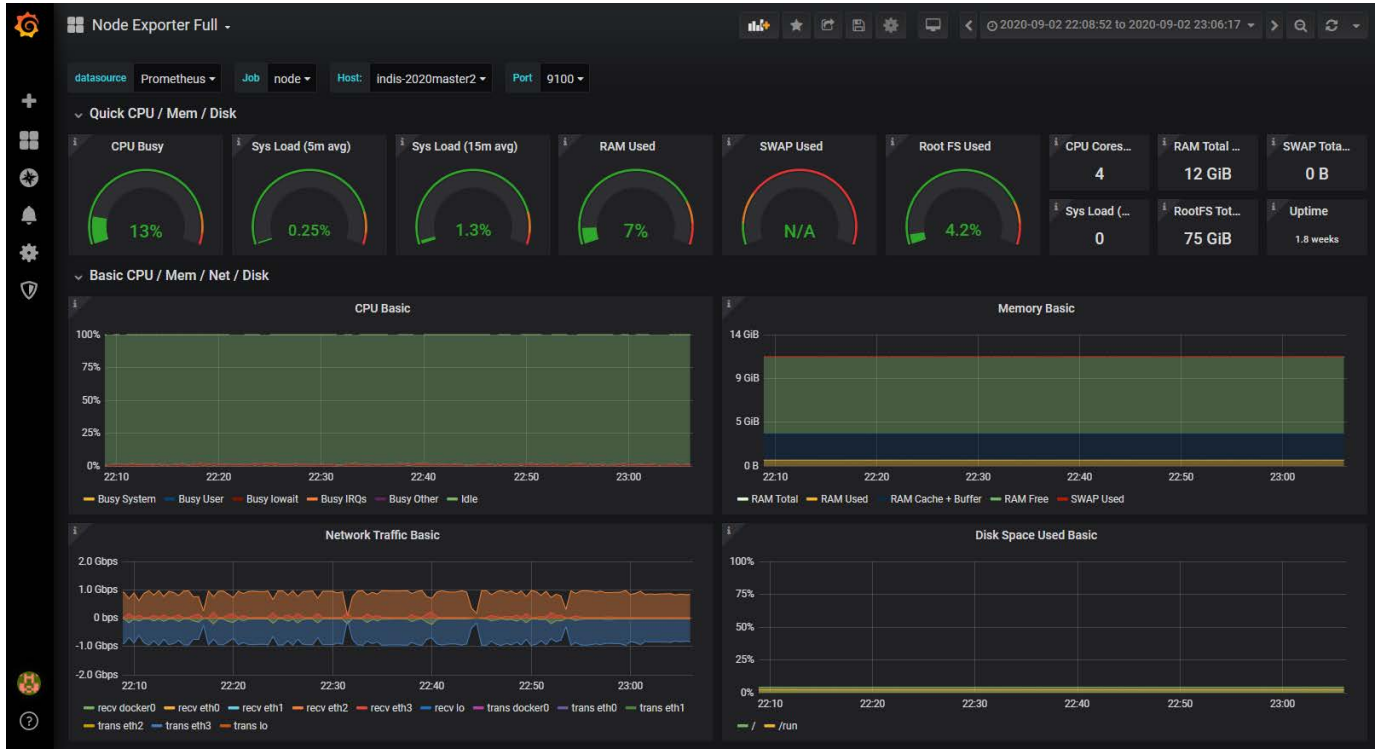
Approach
Virtual Software Defined Exchange

- The vSDX functionality is provided through the **ExoPlex** network architecture
- An elastic slice controller (**Plexus**) coordinates the dynamic network circuits, and the Bro security monitors
- Traffic flow and routing are managed by a variant of the Ryu SDN framework
- A REST API is exposed to allow interactions with the vSDX controller



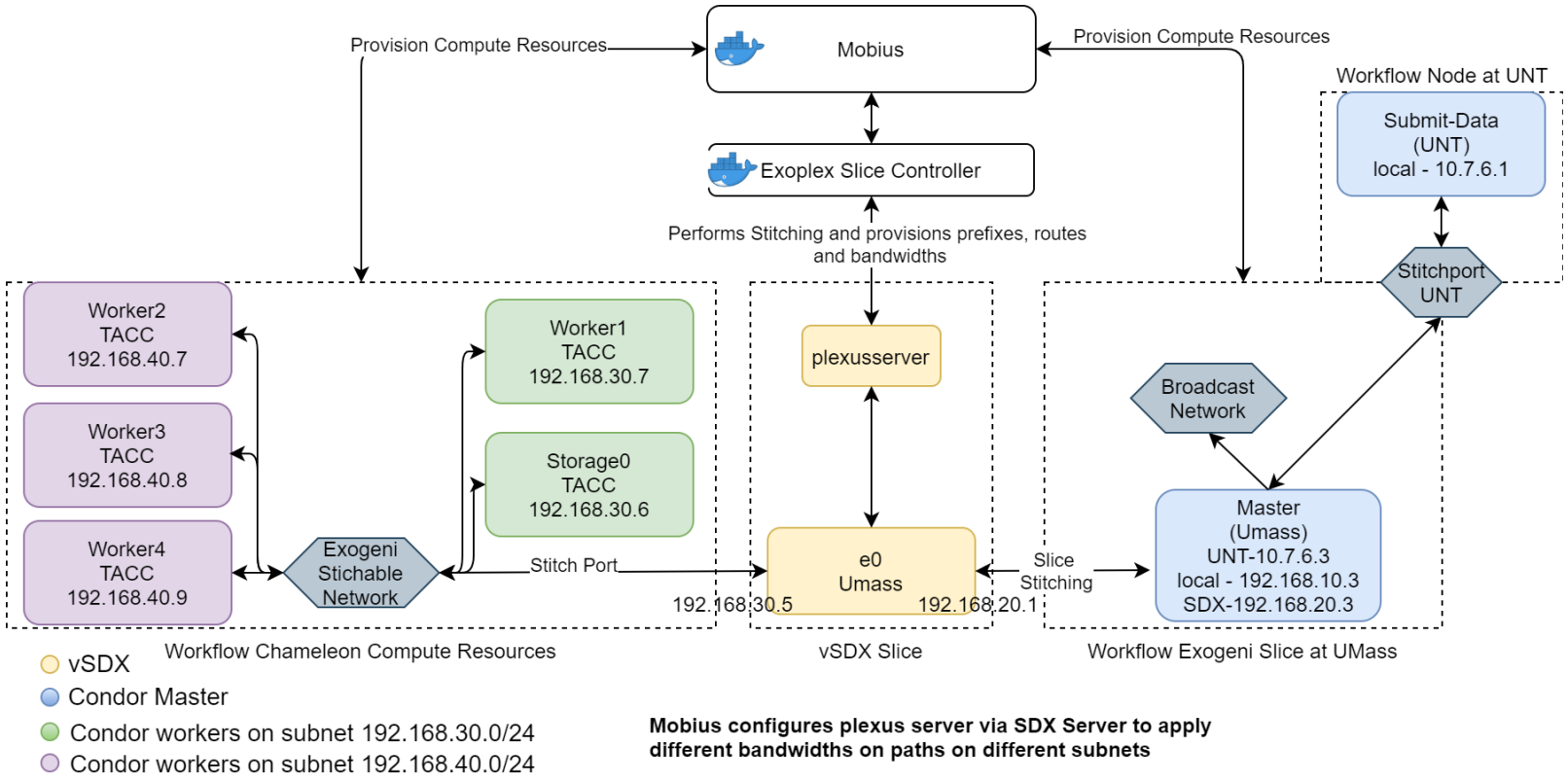
Approach
Prometheus Monitoring System

- Deployed automatically with any DyNamo provisioned resources
- A Grafana dashboard gives insight to each node's resources (cpu and ram utilization, disk IO, network IO, filesystem etc.)
- Provides a centralized view to all your project's resources

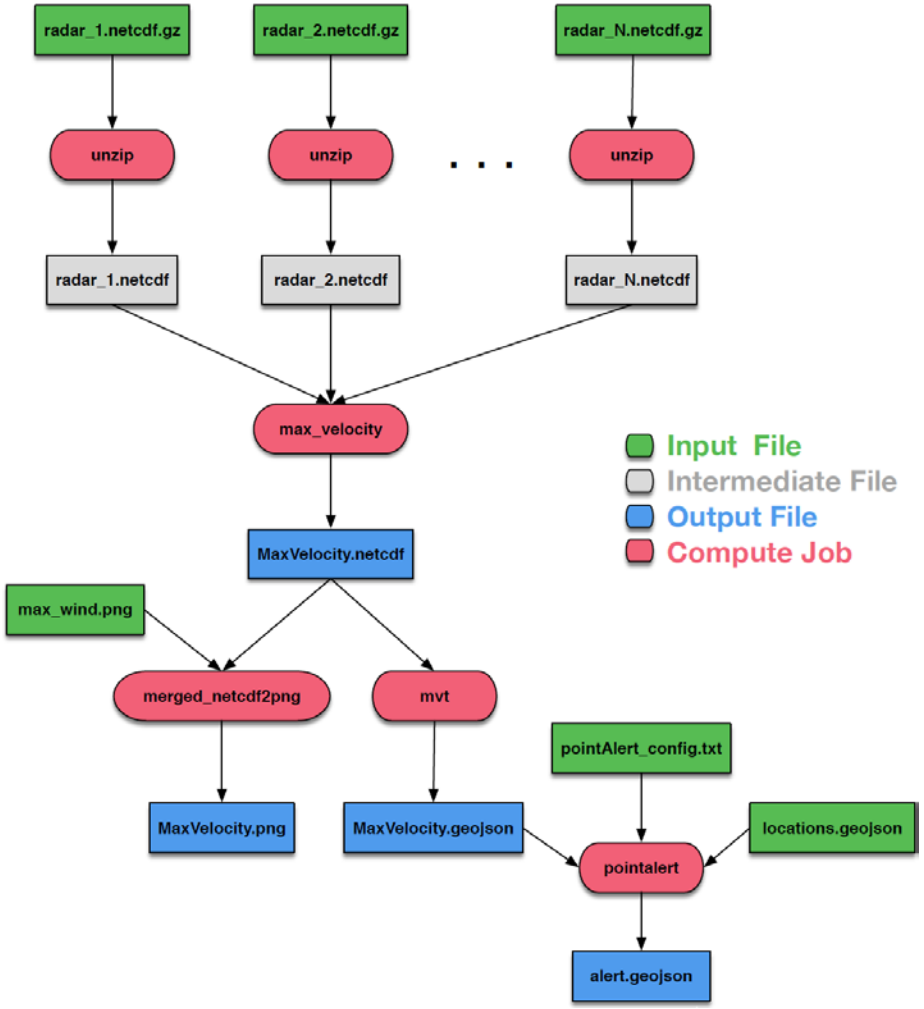


Evaluation Environment Deployment

- The Mobius controller and the Exoplex Slice controller were located at USC – ISI
- The submit node was located at UNT - running Pegasus Ensemble Manager
- Master node was located at ExoGENI UMass rack
- Workers were created at Chameleon TACC
  - 1 worker (48 compute slots) was allocated for the CASA Wind workflow
  - 3 workers (144 compute slots) were allocated for the CASA Nowcast workflow
- All connectivity was established via "stitching" and all links were 1Gbps

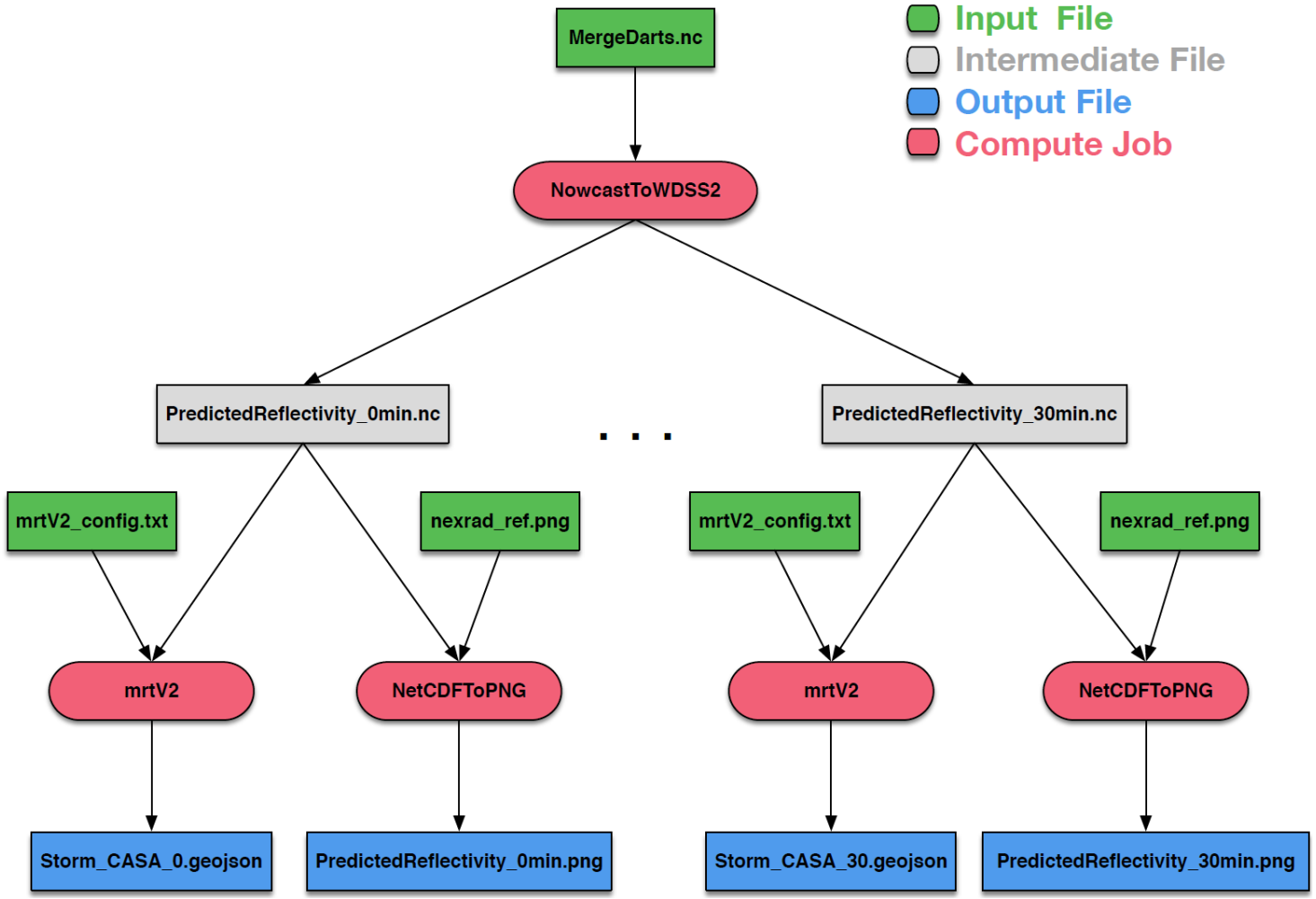


Evaluation
Pegasus Workflow - Wind



- The Wind workflow has a variable size that depends on the number of input files
  - All input radar moment data files are unzipped
  - The rest of the 4 tasks
    - compute maximum observed wind speed
    - create contours with velocity thresholds and images
    - notify points of interests
- The final 4 compute tasks are executed within a Singularity container
  - Consistent environment across execution sites
  - File size: 163 MB
- Testcase Data:
  - 30 minutes of real weather data
  - File size: 12 MB
  - Total size: 6 GB
- Pegasus Ensemble:
  - Execute one workflow every minute

Evaluation
Pegasus Workflow – Nowcast



- Each Nowcast workflow has 63 compute tasks.
  - 1st task splits the input data to 31 individual grids
  - 62 individual tasks compute reflectivity and contour images
- All tasks are executed within a Singularity container
  - Consistent environment across execution sites
  - File size: 153 MB
- Testcase Data:
  - 30 minutes of real weather data
  - File size: 9.6 MB
  - Total size: 287 MB
- Pegasus Ensemble:
  - Execute one workflow every minute

Evaluation
Network Needs – Wind Workflow Ensemble

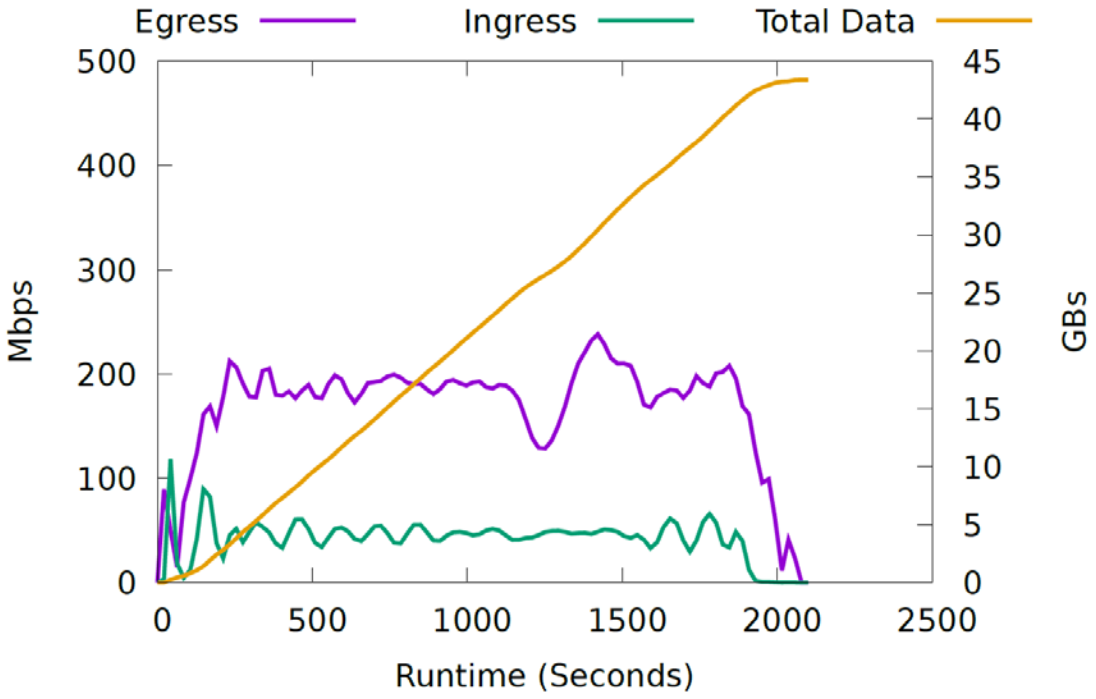


Fig. 8. Wind Ensemble - Network Utilization.

- For the given dataset, the Wind workflow ensemble transfers a total of ~44GB.
- Total runtime for the ensemble is ~2100 secs.
- Average egress bandwidth usage is ~200Mbps
- Peak egress bandwidth usage is ~250Mbps
- *Wind workflow ensemble doesn't congest the network*

Evaluation
Network Needs – Nowcast Workflow Ensemble

- For the given dataset, the Nowcast workflow ensemble transfers a total of ~280GB.
- Total runtime for the ensemble is ~3200 secs.
- Average egress bandwidth usage is ~900Mbps
- Peak egress bandwidth usage is ~960Mbps
- *Nowcast workflow ensemble congests the network, affecting transfers sharing the same resources.*

Fig. 9. Nowcast Ensemble - Network Utilization.

Evaluation
Workflow Makespans

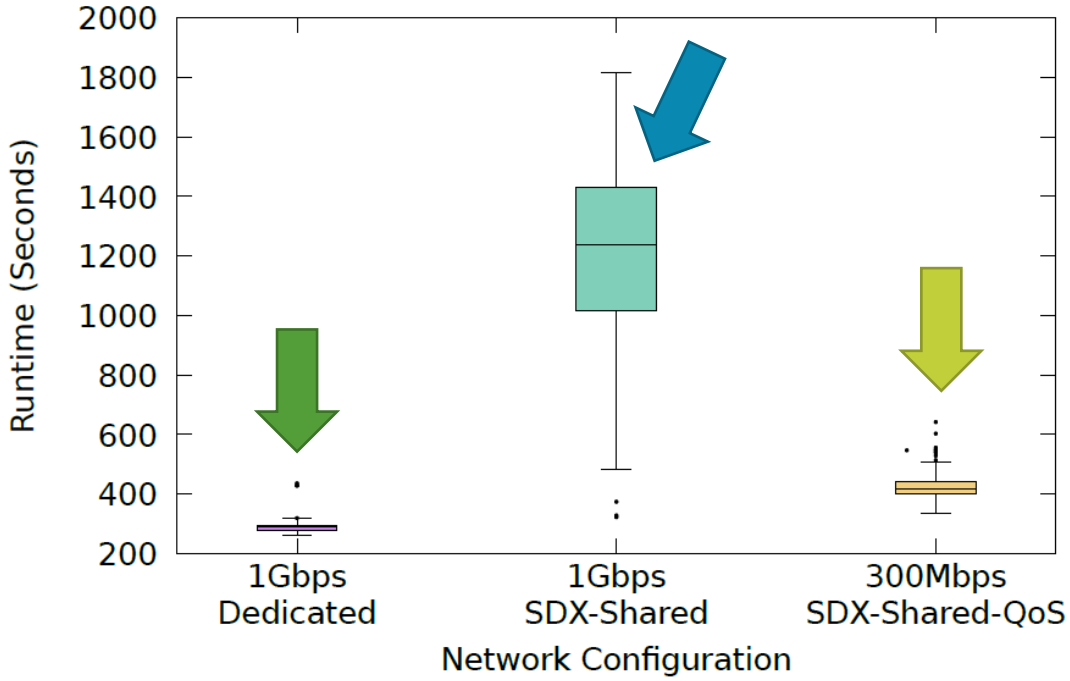


Fig. 10. Wind Ensemble Workflow Makespans.

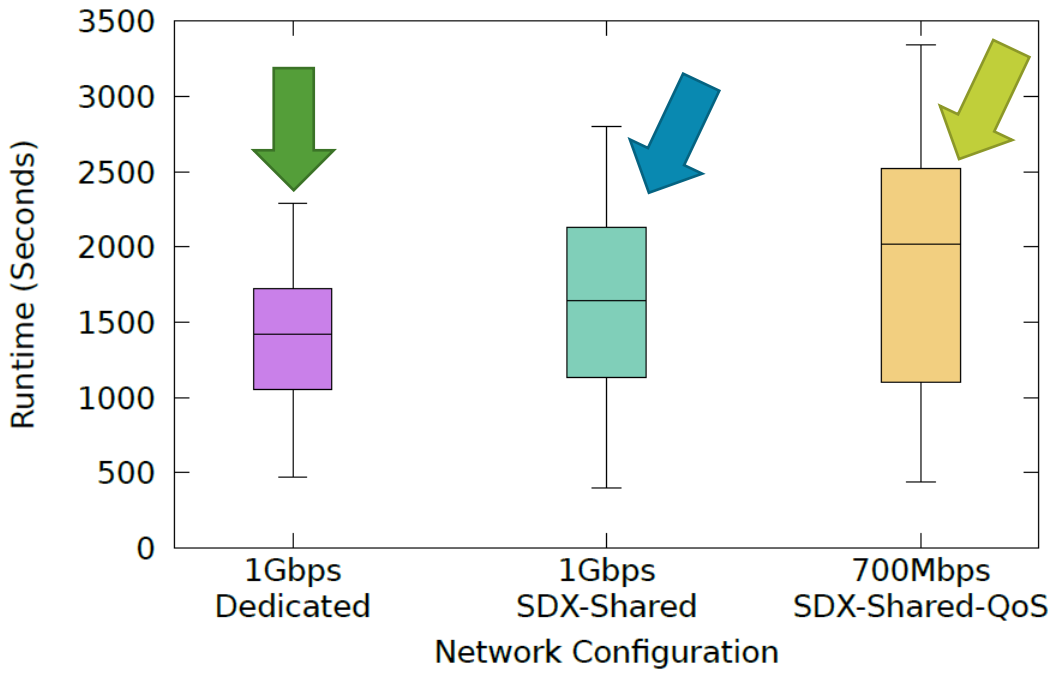


Fig. 11. Nowcast Ensemble - Workflow Makespans.

Statistics from 900 workflow submissions and over 240,000 file transfers generating more than 900GBs of network traffic.

Evaluation
Transfer Durations

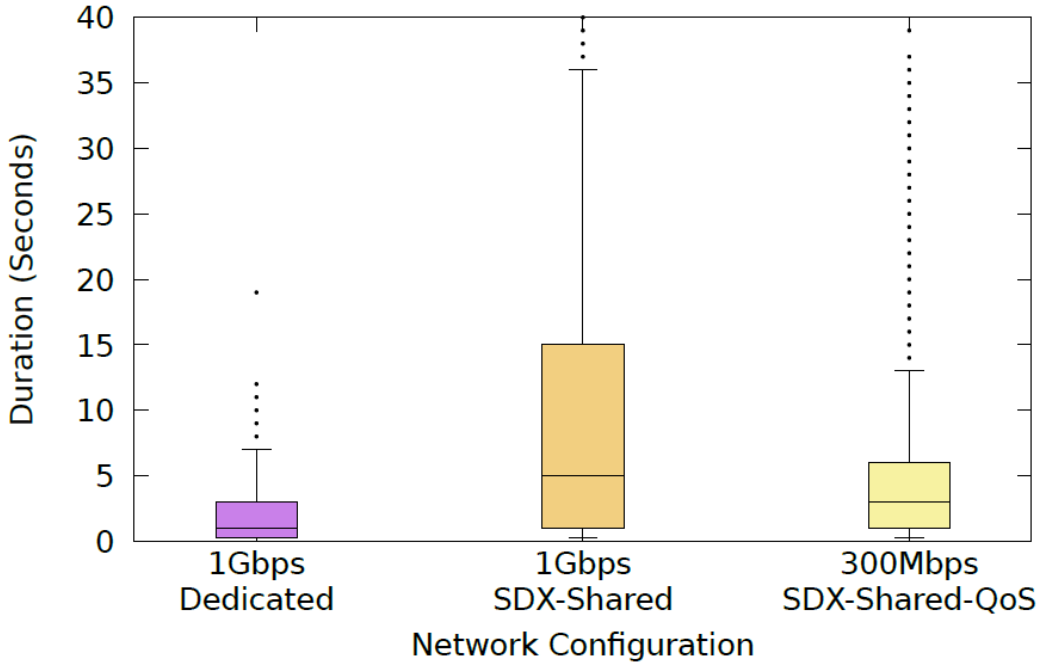


Fig. 12. Wind Ensemble - Workflow Data Transfer Durations.

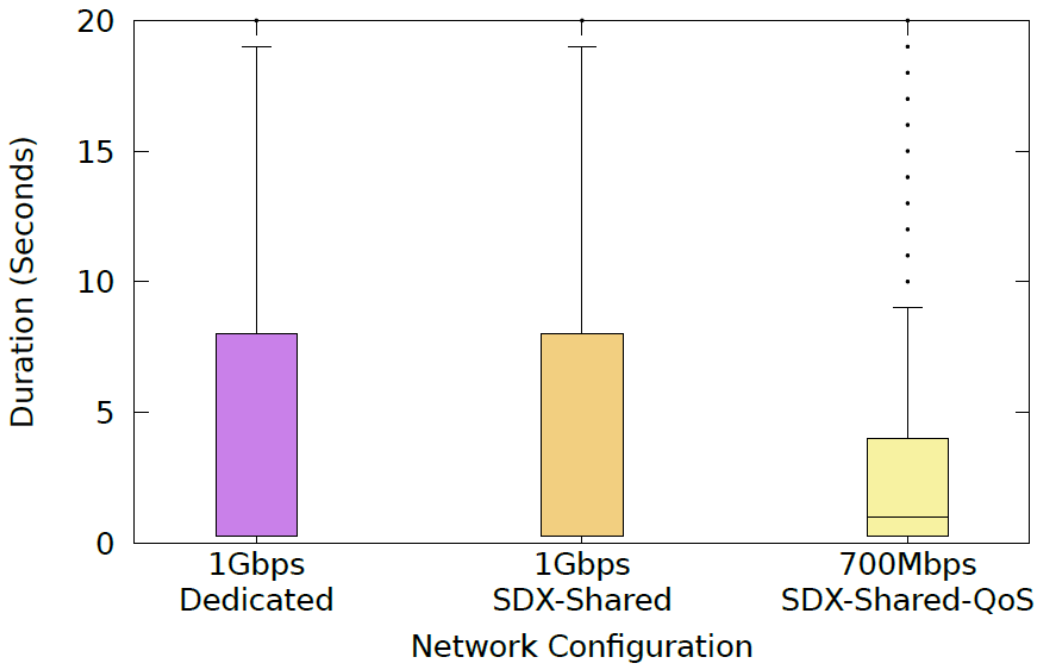


Fig. 13. Nowcast Ensemble - Data Transfer Durations.

Summary

- DyNamo: a multi-cloud platform with high-performance adaptive computing and networking support for science workflows
- DyNamo enables automation, dynamic infrastructure management
- Using the Pegasus Ensemble Manager, DyNamo provides fine control over workflow ensembles (e.g., CASA)
- The vSDX capabilities allow CASA scientists to use more efficiently their resources while maintaining the quality of service for their applications
- Prometheus monitoring offers a lightweight and intuitive way to access resource utilization metrics

DYNAMO

Funded by the National
Science Foundation
Grant #1826997

Thank you !

Questions ?

renci

USC Viterbi
School of Engineering
Information Sciences Institute

UMass Amherst

RUTGERS