



# *Multi-Cloud Performance and Security-driven Brokering for Bioinformatics Workflows*

**Saptarshi Debroy, PhD.**

**Assistant Professor in Computer Science, Hunter College of CUNY**  
**Doctoral Faculty, The Graduate Center of CUNY**



# Project Team Introduction

NSF Program: CC\* (Campus Cyberinfrastructure)

Program Area: OAC (Office of Advanced Cyberinfrastructure)



Award Number: 1827177

Project Title: **End-to-End Performance and Security Driven Federated Data-intensive Workflow Management**



**Prasad Calyam**  
Associate Professor  
University of Missouri



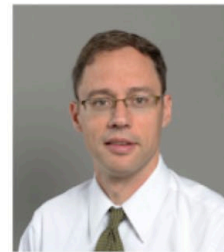
**Trupti Joshi**  
Assistant Professor  
University of Missouri



**Isa Jahnke**  
Associate Professor  
University of Missouri



**Saptarshi Debroy**  
Assistant Professor  
CUNY – Hunter College

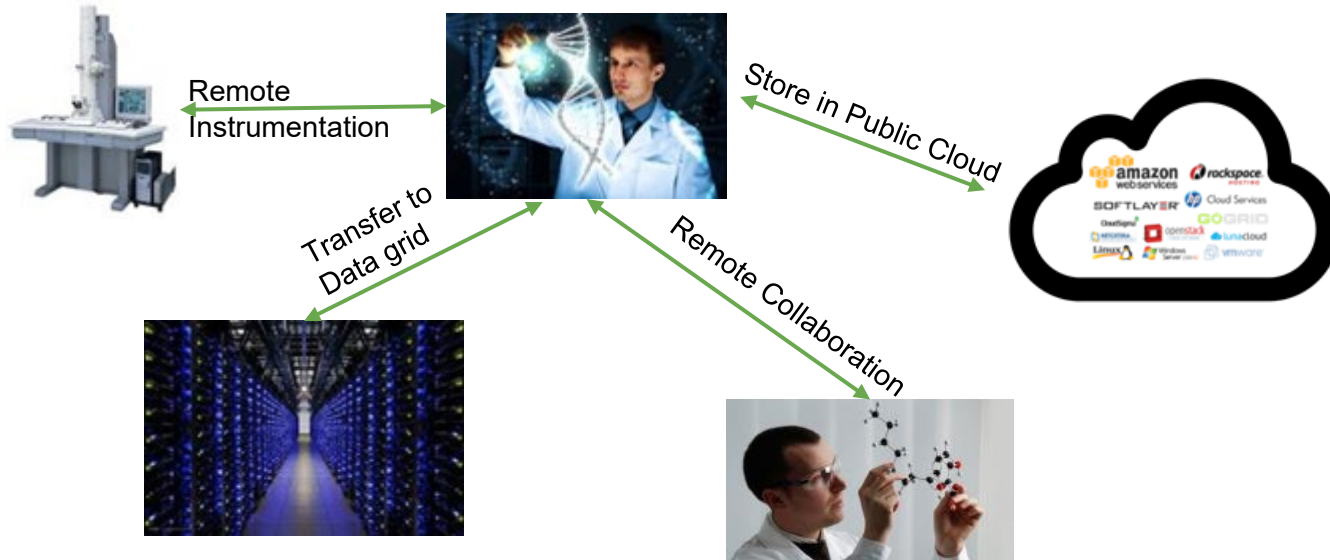


**Timothy Middelkoop**  
Director, Research Computing  
University of Missouri

## Graduate Students:

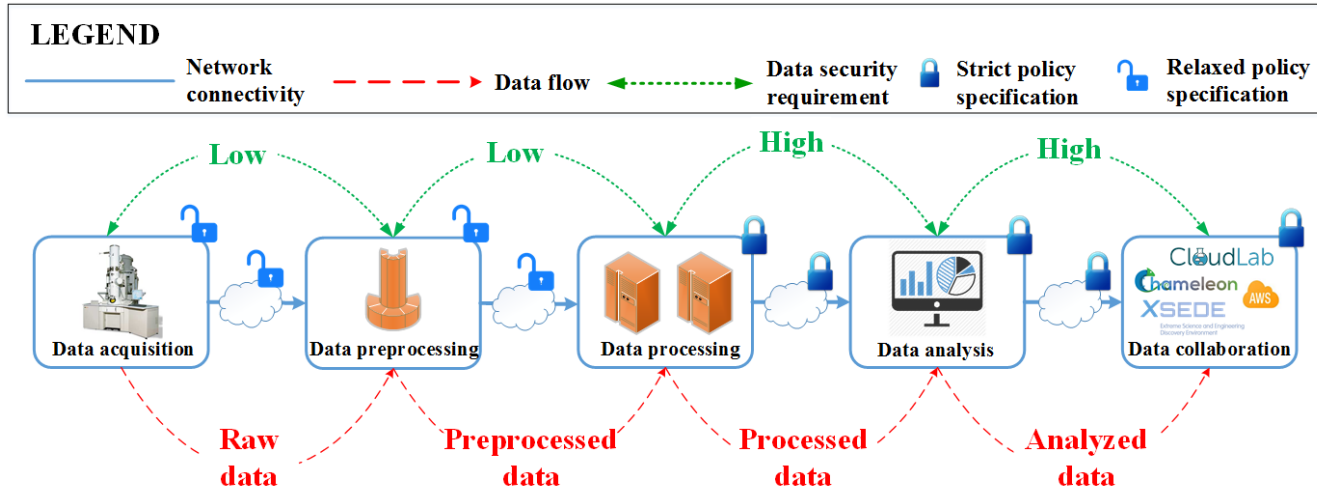
- **Minh Nguyen (CUNY)**
- **Xiaojie Zhang (CUNY)**
- **Ashish Pandey (MU)**
- **Soumya Purohit (MU)**
- **Ramya Bhamidipati (MU)**
- **Mauro Lemus (MU)**

# Data Intensive Applications Today



- Science data-intensive applications require **on-demand** resources
- Motivates adoption of **hybrid cloud (private/public/community)** architectures
- Programmable technologies (e.g., SDN, OpenFlow, OpenStack) and **federated resources** (e.g., CyVerse, NSF Cloud, AWS) make such collaboration possible

# Security concerns for cross-domain resource use



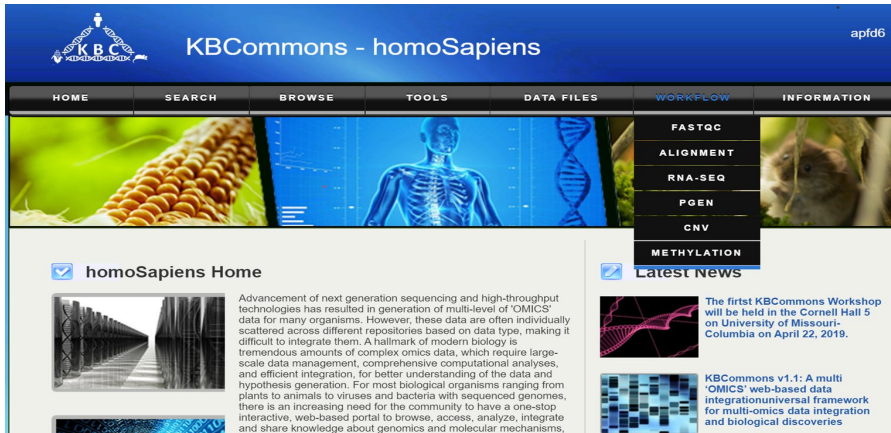
- Data has different life-cycle stages with varying security requirements
- Serious security concerns for data leaving campus or if resources are used across domains

Example of distrust at local institutions:

*“Division of IT strongly advises ... to discourage if not prohibit ... from using public cloud services for University related activities...”*

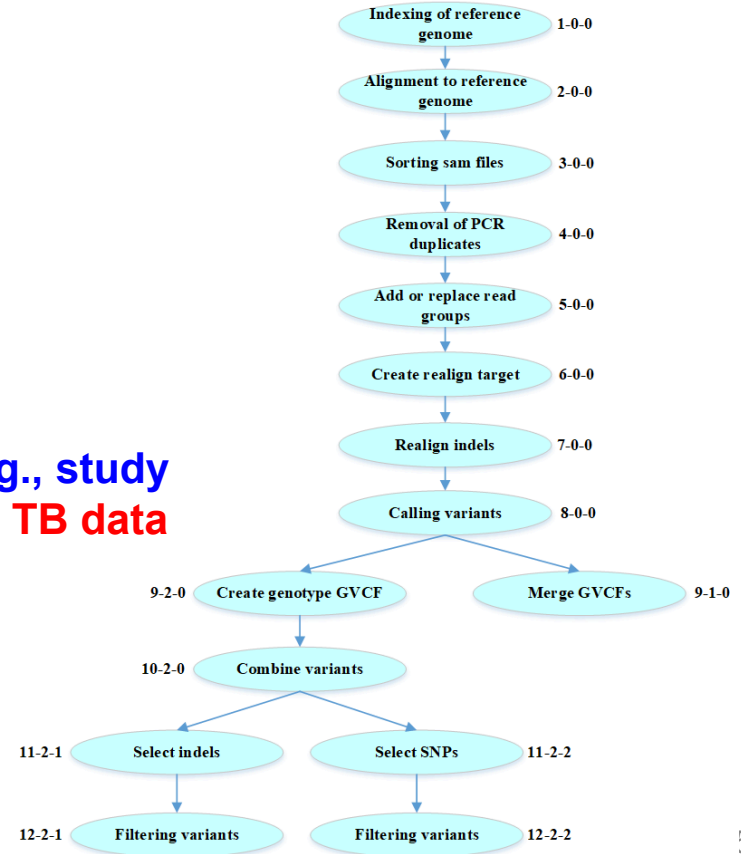
-- Division of IT, University of Missouri

# KBCommons Science Gateway and SoyKB



Science gateway portal hosting bioinformatics workflows e.g., study of genetic mutations in plants and organisms (involves TB data sets)

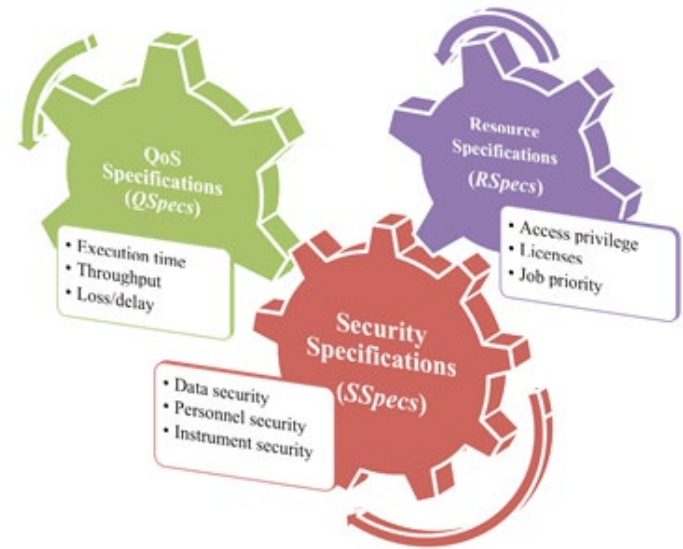
- The multi-step workflows provide biological users with an avenue to analyze their datasets.
- Support needed to accommodate security levels to handle protected genomics data



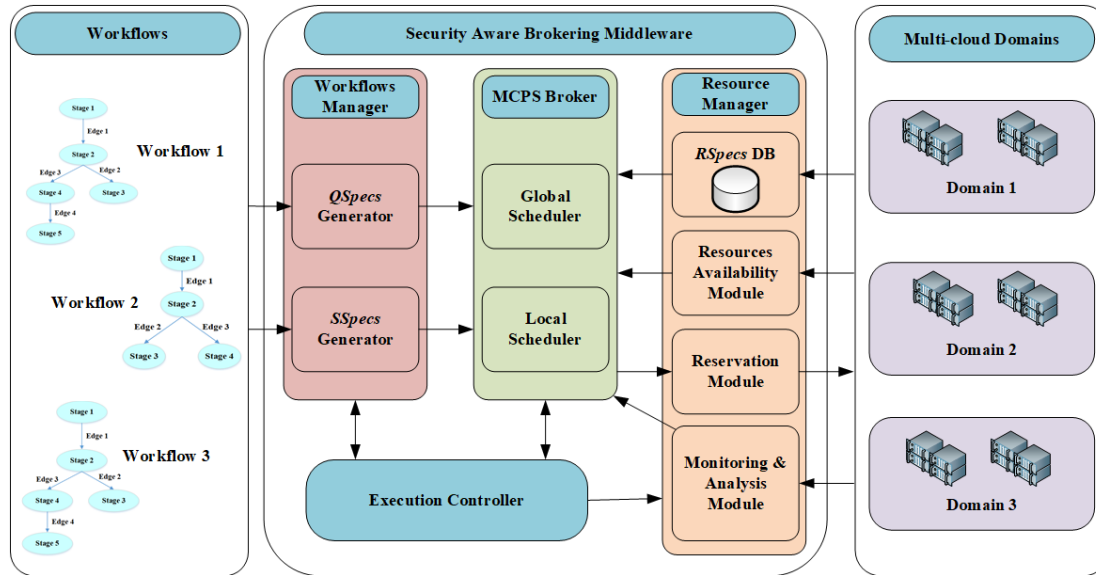
# Security & Performance Inter-Conflicts Problem

We characterize the inter-conflicts problem as:

- “Friction” among user requirements and domain/ data source policies
- Strict security requirements of data custodians adherence may restrict performance
- Institutional resource policies may not satisfy security requirements



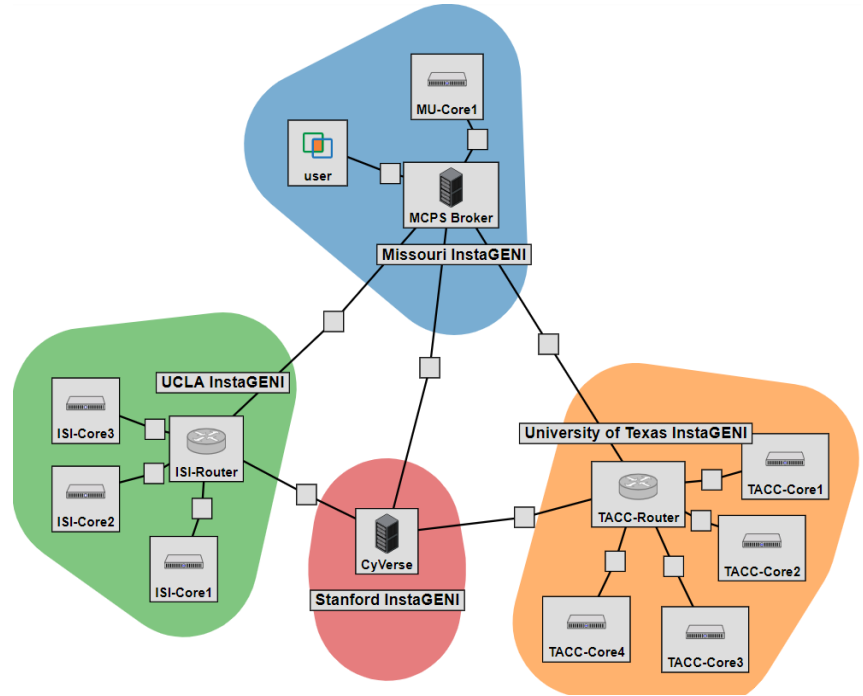
# Performance and Security driven Resource Brokering



- Global scheduling algorithm allocates workflow DAG vertices to domains with *security and policy* satisfaction.
- Local scheduling algorithm chooses optimal computing core within the chosen domain for *performance* satisfaction.

# GENI Implementation

- GENI infrastructure based testbed
  - Multi-cloud resource domains approximately based on the real computing centers used for SoyKB workflows
- Compute capability and network bandwidth mismatches to mimic real-life SoyKB implementation.
- Replicates security policies of TACC, ISI, and MU domains as well as dynamic resource utilization levels.





# User Interfaces

MCPS Broker User Interface

Not secure | pcvm4-1.instagenei.net.missouri.edu/ui1.html

## Welcome to MCPS Broker!

Please enter information here:

What is your workflow type? PGen

Comments:

Please upload your data:

Upload

Submit

MCPS Broker Admin Interface

Not secure | pcvm5-3.instagenei.net.missouri.edu/admin.html

## MCPS Broker Resource Dashboard

### Domain Statistics

Domain	Total No. of Cores	Available No. of Cores	Total Storage (MB)	Available Storage (MB)	Total Bandwidth (Mbps)	Available Bandwidth (Mbps)
MU	1	1	3200	2800	10	5
TACC	4	2	6400	5400	100	50
ISI	3	1	4800	3800	100	50

### Workflow Statistics

Workflow ID	Workflow Type	Execution Progress
1	PGen	97%
2	RNA-Seq	100%
3	PGen	2%

Refresh

MCPS Broker Admin Interface

Not secure | pcvm5-3.instagenei.net.missouri.edu/tacc-stats.html

## MCPS Broker Resource Dashboard

### TACC Domain Statistics

Job Order	Workflow Type	Workflow ID	Compute	Storage	Bandwidth	Execution Progress	SSpec Compliance
1	PGen	1	1	400	NA	100%	Yes
2	PGen	1	1	2100	NA	100%	Yes
3	PGen	1	1	2100	NA	100%	Yes
4	PGen	1	1	2100	NA	100%	Yes
5	PGen	1	1	2100	NA	100%	Yes
6	PGen	1	4	2000	NA	100%	Yes
7	RNA-Seq	2	3	1000	NA	100%	Yes
8	PGen	1	1	300	NA	100%	Yes
9	PGen	1	1	2000	NA	100%	Yes
10	PGen	1	3	1000	NA	100%	Yes
11	PGen	1	3	1000	NA	75%	Pending...

Back

MCPS Broker Admin Interface

Not secure | pcvm5-3.instagenei.net.missouri.edu/wid-1.html

## MCPS Broker Resource Dashboard

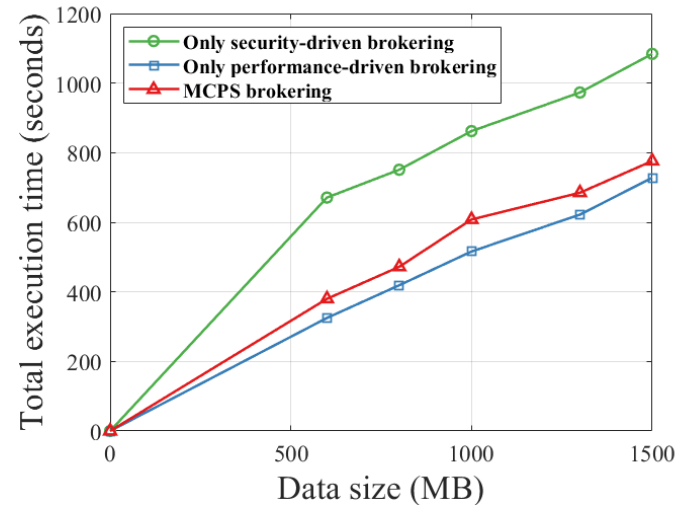
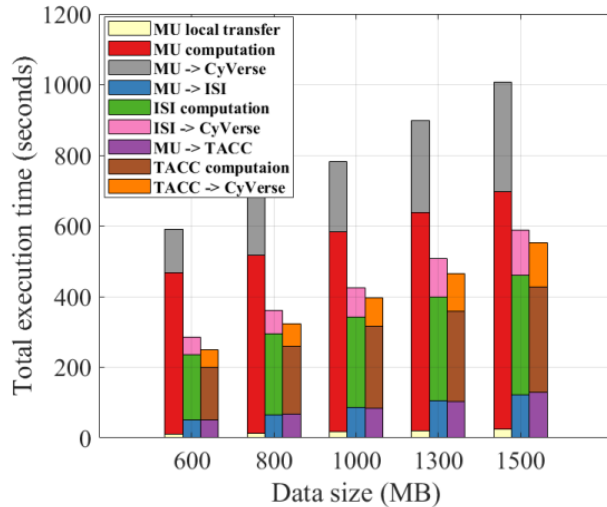
### Workflow ID: 1, type: PGen.

Stages	Compute	Storage	Bandwidth	Domain	Execution Progress	SSpec Compliance
1. Indexing	1	400	NA	ISI	100%	Yes
2. Alignment	1	2100	NA	ISI	100%	Yes
3. Sorting sam files	1	2100	NA	ISI	100%	Yes
4. Removal of PCR	1	2100	NA	ISI	100%	Yes
5. Add/replace groups	1	2100	NA	ISI	100%	Yes
6. Create realign target	4	2000	NA	TACC	100%	Yes
7. Realign indels	1	1000	NA	ISI	100%	Yes
8. Call variants	1	300	NA	ISI	100%	Yes
9. Merge GVCFs	1	2000	NA	ISI	100%	Yes
10. Create GVCF	1	1000	NA	TACC	100%	Yes
11. Combine variants	1	1000	NA	ISI	100%	Yes
12. Select indels	3	1000	NA	ISI	100%	Yes
13. Select SNPs	3	1000	NA	TACC	100%	Yes
14. Filtering variants	3	1000	NA	ISI	75%	Pending...
15. Filtering variants	3	1000	NA	TACC	68%	Pending...

Back

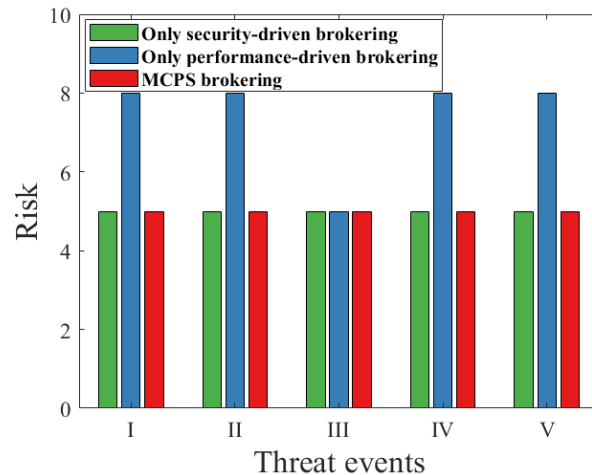
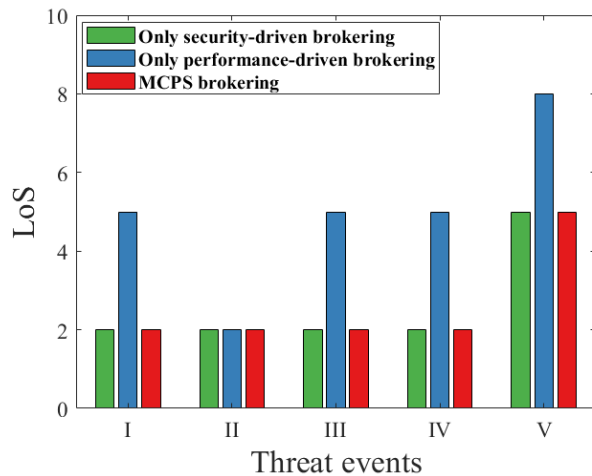
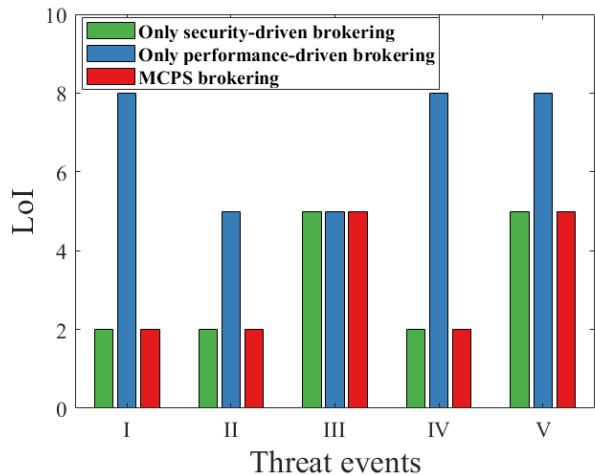
# Performance Evaluation

- Total workflow execution time comparison for PGen workflow
  - MCPS is **almost as good as** only performance-driven brokering (i.e., optimal)
- Total workflow execution time comparison for simpler RNA-Seq workflow
  - Remote computation is **better**



# Threat Analysis and Security Assessment

- The security compliance comparison for PGen
  - Likelihood of attack success (LoS) and overall Risk are similar to security-driven brokering (i.e., optimal)





<http://www.ontimeurb.net>