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

PIs: Saptarshi Debroy, Prasad Calyam, Trupti Joshi
Affiliations: The City University of New York, University of Missouri-Columbia

Data-intensive science applications, such as bioinformatics often require specialized compute/networking/storage resources that are not always available locally on-site and need to use compute resources in remote cloud domains for processing which in tern requires high speed data transfer. Thus, researchers are increasingly adopting federated multi-cloud infrastructures (e.g., CyVerse, XSEDE) to support compute-intensive or data-intensive science collaborations. Adoption of such infrastructures are facilitated by Software-defined Networking (SDN) enabled campus Science DMZs for friction-less data movement and Federated Identity and Access Management (IAM) that enables campus researchers to reserve and seamlessly access local and remote cloud resources.

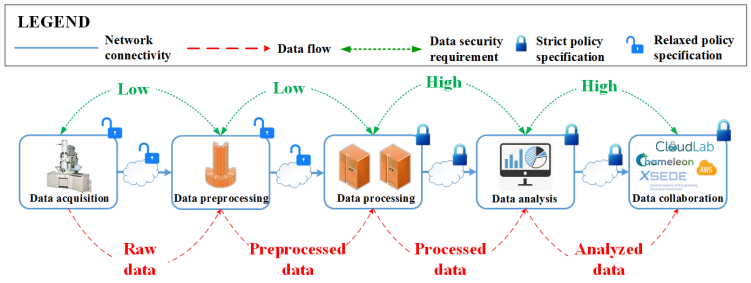


Figure 1: End-to-end lifecycle stages of a data-intensive application with dynamic security requirements using federated multi-cloud resources from domains with diverse resource policy specifications.

Allocation of such federated multi-cloud resources is typically based on applications' performance considerations (e.g., data throughput, execution time). However, such one-dimensional resource brokering fails to consider scenarios where applications' security requirements across different life-cycle stages (Low, Moderate, and High) contradict with remote domains' diverse security policies (ranging from very strict to very relaxed) as shown in Figure 1. It is a difficult proposition for users (especially when using complex workflows) to guess how to select options available within federated multi-cloud resources in a manner that overcomes bottlenecks such as resource capacity limitations, network bandwidth, security posture or cost factors at the various resource domains. Without a systematic framework and standardized tools, performance-security conflicts for applications and inefficient/expensive resource usage scenarios occur that are undesirable from a user perspective.

In our National Science Foundation (NSF) funded project (Grant no. OAC-1827177) project, we are investigating the need for security aware resource brokering over traditional one-dimensional resource allocation for multi-cloud data collaboration. We are addressing the lack of knowledge about: a) individual domain's security policies, b) how that translates to security assurance of the applications, and c) nature of performance and security trade-offs - that can cause performance-security conflicts for applications and inefficient/expensive resource usage (Figure 1).

In particular, we consider the implementations of high-throughput cloud-based bioinformatics data analysis workflows in the SoyKB science gateway developed for soybean and other related organisms. These workflows provide biological users with an avenue to analyze their in-house generated datasets (in terabytes per month) using multi-step workflows and conduct analysis in high performance computing environments that support the necessary security levels to handle Health Insurance Portability and Accountability Act (HIPPA) compliance. For example, a complex PGen workflow is used to efficiently facilitate analysis of large-scale next generation sequencing (NGS) data for genomic variations. Whereas a comparatively simpler RNA-Seq analysis workflow is used to perform quantization of gene expression from transcriptomics data and statistical analysis to discover differential expressed gen/isoform between experimental groups.

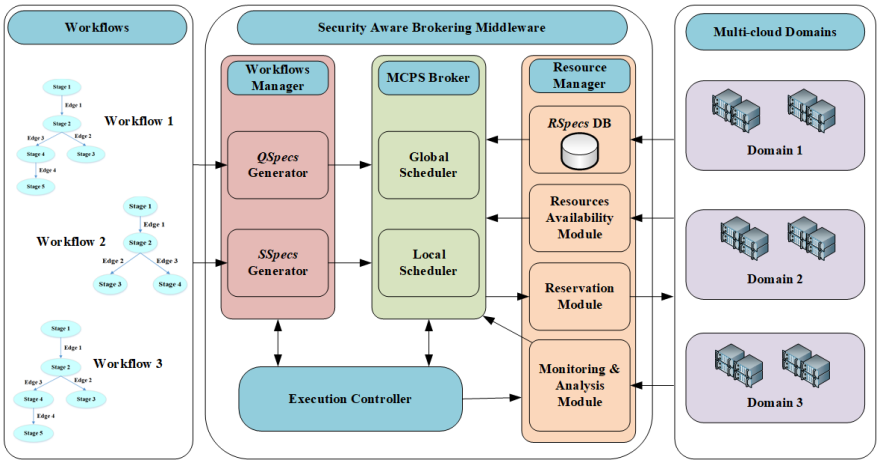
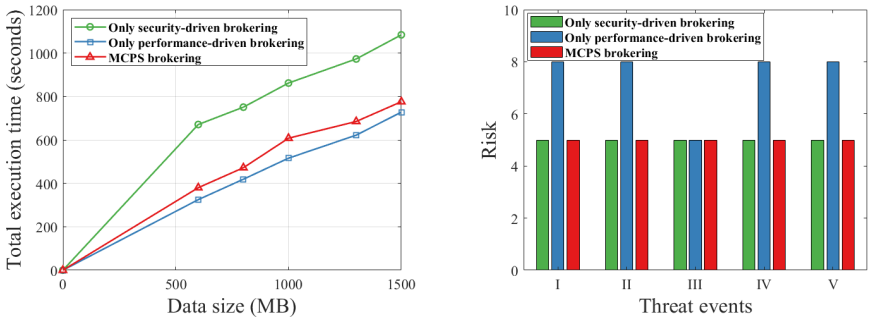


Figure 2: Security-aware resource brokering middleware services with MCPS Broker and its underlying components interaction

Currently, we are developing a security-aware resource brokering middleware framework, i.e., the MCPS (Multi-Cloud Performance and Security) Broker, within federated multi-cloud systems to allocate application resources by satisfying their performance and security requirements. The architecture of the MCPS Broker is demonstrated in Figure 2 workflow within SoyKB science gateway environment.



(a) Performance comparison (b) Security evaluation

Figure 3: Brokering scheme comparison

Our initial testbed implementation on GENI infrastructure is approximately based on the real computing centers used for SoyKB workflows: a local University of Missouri (MU) domain, as well as remote cloud domains, such as Texas Advanced Computing Center (TACC), and Information Sciences Institute (ISI). Here workflows are sent from the MU domain users through the MCPS Broker, which decides whether the workflows are processed locally at MU or remotely at TACC or ISI based on the global and local algorithm outcomes discussed with the results ultimately sent to CyVerse upon processing. Figure 3(a) shows that for different data sizes, MCPS brokering performs almost as good as 'only performance-driven brokering' in terms of choosing domains for processing that optimize total execution time. The security compliance comparison results are shown in Figure 3(b) that uses NIST based risk assessment method. The figure shows that the overall risk of different threat events (I to V) are similar for 'only security-driven brokering' and our proposed 'MCPS brokering' as these schemes almost always choose ISI or TACC over MU regardless of the formers' resource availability. *In the next phase, we are planning to scale up our implementation of MCPS Broker and case study evaluation with workflows on KBCommons. This will demonstrate the benefits of our proposed middleware in ensuring both performance optimization and security compliance for SoyKB and other workflows in KBCommons.*