Performance and Security Challenges in Science Workflows

Dipak Ghosal
dghosal@ucdavis,.edu
University of California
Davis, CA

CCS CONCEPTS

• **Networks** → **Network architectures**; **Network design principles**; *Network protocol design*; *Transport protocols*; *Network resources allocation*; *Network performance analysis*; *Network security*; • **Security and privacy** → *Denial-of-service attacks*; • **Computing methodologies** → *Reinforcement learning*;

KEYWORDS

science workflows, software defined networking, network telemetry,deadline driven data transfers, security, data transfer nodes, machine learning

ACM Reference Format:
Dipak Ghosal. 2019. Performance and Security Challenges in Science Workflows. In *Systems and Network Telemetry and Analytics (SNTA'19), June 25, 2019, Phoenix, AZ, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3322798.3329260

ABSTRACT

Scientific workflows are complex, often generating large amounts of data that need to be processed in multiple stages. The data often generated at remote locations must be transferred from the source and between the distributed HPC nodes interconnected by high-speed networks that carry other background traffic. Increasingly, many of these scientific workflows require processing to be completed within a deadline, which, in turn, imposes deadline on the network data transfer. A recent example of a deadline-driven workflow occurred when LIGO and Virgo detectors observed a gravitational wave signal associated with the merger of two neutron stars. The merger, known as a kilonova, occurred in a galaxy 130 million light-years from Earth in the southern constellation of Hydra. The data from this initial observation had to be processed in a timely manner and sent to astronomers around the world so that they could aim their instruments to the right section of the sky to image the source of the signal.

Complex workflows with deadline driven transfers of large data sets is also pertinent in enterprise networks. More and more the underlying network is a software defined network that supports fine grain real-time network telemetry. Deadline-aware data transfer requests are made to a network controller that may accept the request if it can meet the deadline. The network controller schedules the flows by setting maximum and minimum pacing rates of the deadline flows, metering the background traffic at the ingress routers, and appropriately routing the flows. The goal of the scheduling algorithm is to maximize the number of flows that meet the deadline while maximizing the network utilization.

To help facilitate the movement of large science data, many organizations utilize data transfer nodes (DTNs). DTNs help maximize data transfer efficiency particularly when there is a bandwidth mismatch between the core network and the network to which the receiving end-system is connected. An important aspect of a DTN based network architecture is the security issue. This is addressed by the Science DMZ model which attempts to guarantee reliable and high performance data transfers. DTNs can become a critical point of failure in a Science DMZ if performance becomes compromised due to security attacks. As protecting the performance of the DTN is critical in guaranteeing the performance of the science workflows, the Science DMZ avoids avoids sending data through standard firewalls as these can inject significant delay and processing overheads. Science DMZs rely on anomaly detection systems and Access Control Lists (ACLs) which maybe vulnerable to external denial-of-service attacks as well as various insider attacks. Typically, network intrusion detection systems (NIDS) use network metrics derived from the network packet stream to identify anomalies and flag attacks. In science workflows with large data transfers this is a challenging problem. A potential approach is to use system performance metrics of the DTN to identify external and insider attacks.

In this talk we will discuss the performance and security challenges of science workflows. With regards to the deadline driven data transfers, we will discuss the challenges in designing the network controller. We will discuss the need for a two-level autonomous control system. At the network level, the goal is to develop a network controller that can leverage fine grain network telemetry data and information of deadline flows to schedule the flows by providing strict traffic pacing limits. For the end-system, the challenge is to develop a control theoretic approach that will follow the pacing directives from the network controller. We will discuss the applicability of machine learning approaches to implement the network controller. We will discuss the need for strict predictability of the completion time of the data transfers and how formal-method based approaches aided by network telemetry data can be leveraged to achieve that.

With respect to the security issues we will consider various types external denial-of-service attacks as well as insider attacks such as as data exfiltration and data corruption. We will discuss how system performance metrics can be used to identify standard DoS attack such as a TCP-SYN flood attack directed at a DTNs. We will discuss how anomaly detection system based on Hierarchical Temporal Memory (HTM) can used to detect detecting performance anomalies in the DTN caused by external and insider attacks.

SNTA'19, June 25, 2019, Phoenix, AZ, USA

ACM ISBN 978-1-4503-6761-5/19/06.
https://doi.org/10.1145/3322798.3329260