"Research-Defined Networks" for Big Data Science Applications

Prasad Calyam, Ph.D., Assistant Professor; **Saptarshi Debroy**, Ph.D., Postdoctoral Fellow
INSTITUTION: Department of Computer Science, University of Missouri-Columbia
TOPIC: DOE/SC science mission network and/or transport research layer challenge direction that might yield a significant leap forward

With the growing trend of large collaborative partnerships involving researchers, expensive scientific instruments and high performance computing centers, experiments and simulations produce peta-bytes (and more) of diverse data streams viz., Big Data, that is likely to be shared and analyzed by scientists in multi-disciplinary areas. To meet the data movement and processing needs, there is a growing trend amongst researchers within Big Data fields to frequently access remote specialized resources and communicate with collaborators using high-speed overlay networks. These networks use shared underlying components, but allow end-to-end circuit provisioning with bandwidth reservations. Furthermore, in cases where researchers have sporadic/bursty resource demands on short-to-medium timescales, they are looking to federate local resources with 'on-demand' remote resources to form 'hybrid clouds', versus just relying on expensive over-provisioning of local resources.

Current best practice is popularly referred to as building Science DMZs (De-militarized Zones) that have facilitated "Big Data Highways" that use cutting-edge technologies, such as software-defined networking, end-to-end performance monitoring, and network-virtualization. However, a truly '*Research-defined Network*' (RDN), built to control and troubleshoot all the networking aspects of data-intensive science applications is still to be realized by network researchers and service providers. Such RDNs need to support the full life-cycle of campus Big Data, from creation to computation and consumption.

Through RDNs, Big Data researchers should be able to dictate policies, security features, and QoS guarantees specific to their applications. Making data-intensive research a driving force behind campus Big Data network design will enable the network designers to better address open issues, such as: (a) assurance of satisfactory user QoE when simultaneously scheduling multiple science Big Data applications, (b) standardizing performance engineering techniques and protocols for easy use and wide-adoption, and (c) selectively replicating time sensitive or mission critical data across multiple platforms for reliability purposes and prudently selecting replication sites to avoid end-to-end performance bottlenecks.

Building such RDNs is a first step to federate different 'Big Data Highways' to create a 'Big Data Interstate system' where different campus Big Data network infrastructures seamlessly come together. Creating such federations should be aimed towards faster sharing of research data, enhancing cross campus research collaboration, and quicker troubleshooting of network performance bottlenecks. Although early efforts are taking shape in creating such multi-campus Science DMZ federations, there exists a number of open challenges in realizing such collaborations: (a) co-ordination of federated resources with adherence to policies of multiple-domains, (b) enforcing federated and transparent access control mechanisms over local autonomy to facilitate broader sharing, (c) building secured middlegrounds for performance visibility and network control across Science DMZ domains, and (d) creating social platforms or extending existing platforms for scientific collaborations, where Big Data researchers, network designers, and policymakers belonging to the same community can mingle, share data and expertise, collaborate, and create new policies and rules for the federation. Solving these open issues are fundamental in the future explorations that will lead to an *Internet that is personalized based on research needs of Big Data science applications* to technologically advance our society.