Network-aware Data Management Workshop

http://ndm-meeting.org

Location*: **Hilton 410**

November 16, 2015 Monday *2:00pm - 5:30pm*

2:00 - 2:05 Opening Remarks

2:05 - 2:35 Keynote Speech
Network Integration with Workload Management - the PanDA Example
Kaushik De (University of Texas at Arlington)

2:35 - 3:00
Managing Scientific Data with Named Data Networking (NDN)
Susmit Shannigrahi (Colorado State University)

3:00 - 3:30 Coffee Break

3:30 - 3:55
A Multi-domain SDN for Dynamic Layer-2 Path Service
Fatma AL-Ali (University of Virginia)

3:55 - 4:20
Design and Implementation of Control Sequence Generator for SDN-enhanced MPI
Baatarsuren Munkhdorj (Osaka University)

4:20 - 4:45
Approximate Causal Consistency for Partially Replicated Geo-Replicated Cloud Storage
Tayuan Hsu (University of Illinois at Chicago)

4:45 - 5:10
Hysteresis-based Optimization of Data Transfer Throughput
Md S Q Zulkar Nine (University at Buffalo (SUNY)

5:10-5:30 Closing Remarks

Network-aware Data Management

Monday, November **16**, 2015

(2:00 pm - 5:30 pm)
location: ***Hilton 410***

The 5th International Workshop on
Network-aware Data Management
in cooperation with ACM SIGHPC, in conjunction with SC'15:
IEEE/ACM International Conference for High Performance Computing,
Networking, Storage and Analysis.
Austin Convention Center
Austin, TX, USA

http://ndm-meeting.org

Keynote:

Title: ***Network Integration with Workload Management - the PanDA Example***
Kaushik De (University of Texas at Arlington)

Abstract: PanDA is the workload management system used by thousands of physics in the ATLAS experiment at the Large Hadron Collider. PanDA manages the execution of millions of user jobs per day at hundreds of sites worldwide. While PanDA was originally designed a decade ago to manage workloads on CPU's and Storage, over the past few years the role of networking has proven to be equally important. We will present results and discuss future plans for active network integration in PanDA.

Bio: *Kaushik De is a Professor of Physics and Director of the Center of Excellence in High Energy Physics at the University of Texas at Arlington. He started and co-led the PanDA software project from its conception more than a decade ago. He is the founding Director of the SouthWest Tier 2 Computing Center for ATLAS. He is currently deputy project leader for ATLAS Software and Computing in the U.S.*

Managing Scientific Data with Named Data Networking (NDN)

Chengyu Fan, Susmit Shannigrahi, Steve Dibendetto, Catherine Olschanowsky, Christos Papadopoulos, Harvey Newman, Edmund Yeh, Jean-Roch Vlimant, Azher Amin, Dorian Kcira, Iosif Legrand, Ramiro Voicu, David Randall, Kelley Wittmeyer, Mark Branson and Don Dazlich
(Colorado State University, Northeastern University, Caltech, CERN)

Abstract: Many scientific domains, such as climate science and High Energy Physics (HEP), have data management requirements that are not well supported by the IP network architecture. Named Data Networking (NDN) is a new network architecture whose service model is better aligned with the needs of data-oriented applications. NDN provides features such as best-location retrieval, caching, load sharing, and transparent failover that would otherwise be painstakingly (re-)implemented by each application using point-to-point semantics in an IP network. We present the first scientific data management application designed and implemented on top of NDN. We use this application to manage climate and HEP data over a dedicated, high-performance, testbed. Our application has two main components: a UI for dataset discovery queries and a federation of synchronized name catalogs. We show how NDN primitives can be used to implement common data management operations such as publishing, search, efficient retrieval, and publication access control.

A Multi-domain SDN for Dynamic Layer-2 Path Service

Scott Tepsuporn, Fatemah Alali, Malathi Veeraraghavan, Xiang Ji, Brian Cashman, A. J. Ragusa, Luke Fowler, Chin Guok, Tom Lehman and Xi Yang
(University of Virginia, Indiana University, ESnet, University of Maryland)

Abstract: This paper describes our experience in deploying a multi- domain Software-Defined Network (SDN) that supports dynamic Layer-2 (L2) path service, and offers insights gained from this experience. SDN controllers, capable of handling requests for advance-reservation and provisioning of rate- guaranteed L2 paths, were deployed in each domain. The experience demonstrated that this architecture can support global-scale multi-domain dynamic L2 path service. How- ever, to reach this scale, better tools are required for diagnostics of end-to-end L2 connectivity, and better error- reporting functionality is needed from the SDN controllers. As a use case for rate-guaranteed L2 path service, we experimented with high-speed large dataset transfers. We found that a combination of Circuit TCP (CTCP), in which the sending rate is held fixed, and a token bucket filter based rate shaper at the sending host, is best to achieve almost 0-loss, high-throughput transfers across L2 paths. Detailed studies were conducted to understand the impact of the rate-shaper and CTCP parameters to find the best settings.

Approximate Causal Consistency for Partially Replicated Geo-Replicated Cloud Storage

Ajay Kshemkalyani and Tayuan Hsu (University of Illinois at Chicago)

Abstract: In geo-replicated systems and the cloud, data replication provides fault tolerance and low latency. Causal consistency in such systems is an interesting consistency model. Most existing works assume the data is fully replicated because this greatly simplifies the design of the algorithms to implement causal consistency. Recently, we proposed causal consistency under partial replication because it reduces the number of messages used under a wide range of workloads. One drawback of partial replication is that its meta-data tends to be relatively large when the message size is small. In this paper, we propose approximate causal consistency whereby we can reduce the meta-data at the cost of some violations of causal consistency. The amount of violations can be made arbitrarily small by controlling a tunable parameter, that we call credits.

Design and Implementation of Control Sequence Generator for SDN-enhanced MPI

Baatarsuren Munkhdorj, Keichi Takahashi, Dashdavaa Dashdavaa, Yasuhiro Watashiba, Yoshiyuki Kido, Susumu Date and Shinji Shimojo (Osaka University, Japan)

Abstract: MPI (Message Passing Interface) offers a suite of APIs for inter-process communication among parallel processes. We have approached to the acceleration of MPI collective com- munication such as MPI_Bcast and MPI_Allreduce, taking advantage of network programmability brought by Software Defined Networking (SDN). The basic idea is to allow a SDN controller to dynamically control the packet flows generated by MPI collective communication based on the communication pattern and the underlying network conditions. Although our research have succeeded to accelerate an MPI collective communication in terms of execution time, the switching of network control functionality for MPI collective communication along MPI program execution have not been considered yet. This paper presents a mechanism that provides the control sequence for SDN controller to control packet flows based on the communication plan for the entire MPI application. The control sequence encloses a chronologically ordered list of the MPI collectives operated in the MPI application and the process-related information of each in the list. To verify if the SDN-enhanced MPI collectives can be used in combination with the proposed mechanism, the envisioned environment was prototyped. As a result, SDN-enhanced MPI collectives were able to be used in combination.

Hysteresis-based Optimization of Data Transfer Throughput

Md S Q Zulkar Nine, Kemal Guner and Tevfik Kosar
(University at Buffalo, SUNY)

Abstract: The achievable throughput for a data transfer can be deter- mined by a variety of factors such as network bandwidth, round trip time, background traffic, dataset size, and end- system configuration. For the best-effort optimization of the transfer throughput, three application-layer transfer parameters – pipelining, parallelism and concurrency – have been actively used in the literature. However, it is highly challenging to identify the best combination of these parameter settings for a specific data transfer request. In this paper, we analyze historical data consisting of 70 Million file transfers; apply data mining techniques to extract the hidden relations among the parameters and the optimal through- put; and propose a novel approach based on hysteresis to predict the optimal parameter settings.