The School of Informatics, Computing and Engineering (SICE)

INTELLIGENT SYSTEMS ENGINEERING COLLOQUIUM SERIES

Rethinking System Design For Transient Cloud Computing

Prateek Sharma

Prateek Sharma is a Ph.D. candidate in the College of Information and Computer Sciences at the University of Massachusetts Amherst, where he is advised by Prashant Shenoy. His research interests span several areas of computer systems, including cloud computing, operating and distributed systems, and virtualization. His current research focuses on designing systems and abstractions that make large scale distributed computing platforms more efficient, as well enable applications to effectively harness their resources.

Abstract

Cloud computing platforms form the bedrock of today's computing ecosystem, and provide computing resources for applications in data science, scientific computing, and online web services. Today's cloud platforms run ever more complex applications with diverse requirements, resulting in new challenges in efficient use of cloud resources---both from an application and system design perspective. Increasingly, clouds and data centers are moving towards transient computing, a new model for resource allocation, that improves efficiency and reduces cost. However, transient servers can be unilaterally revoked by the cloud operator, and this uncertain availability results in loss of application state, application downtime, and performance degradation.

In this talk, I will present research challenges in the design of cloud systems and applications that arise due to transiency. First, I will describe a resource management technique, called server portfolios, that is inspired by financial portfolios, that enables distributed applications to effectively use low-cost cloud transient servers. Second, I will describe fault-tolerance techniques that can mitigate the performance degradation due server revocations for distributed data processing applications such as Spark. I will present the design of systems and abstractions that combine these two approaches, and show how a wide range of applications can use transient resources and reduce computing costs by up to 90%. Finally, I will discuss new directions in transient computing and edge cloud architectures that enable emerging applications such deep learning and Internet of Things, to effectively harness cloud resources.

TUESDAY, FEBRAURY 27, 2018
10:30 AM | LUDDY HALL, RM. 1106

SCHOOL OF
INFORMATICS, COMPUTING AND ENGINEERING