**Title: Cost-effective and Locality-optimized MapReduce Clouds: Architecture and Techniques**

**Abstract:** Cloud computing and its pay-as-you-go cost structure have enabled infrastructure providers, platform providers and application service providers to offer computing services on demand and pay-per-use just like how we use utility today. This growing trend in cloud computing, combined with the demands for Big Data and Big Data analytics,  is driving the rapid evolution of datacenter technologies towards more cost-effective, consumer-driven and technology agnostic solutions. Cost effective resource optimization techniques that are highly effective and yet greatly agile are critical for both cloud providers and cloud consumers.

In this talk, I will present new techniques to address the challenges of locality-aware and cost-effective resource management and job scheduling for Hadoop-based Big Data processing in the Cloud.  First, I will present Cura, a utility-driven MapReduce Cloud resource planning and management system.  While existing MapReduce cloud services require users to select a number of complex cluster and job parameters and use those potentially sub-optimal per-job configurations, the Cura resource management achieves global resource optimization in the cloud by minimizing cost and maximizing resource utilization. Next, we will discuss the key features of Purlieus, a self-configurable locality-based data and virtual machine management framework that enables MapReduce jobs to access their data either locally or from close-by nodes in a datacenter including all input, output and intermediate data achieving significant improvements in job response time and throughput. In the third part of the talk, I will present our recent work, VNCache, an efficient solution for MapReduce analysis of cloud-archived data. VNcache dynamically integrates cloud-archived data into a virtual namespace at the enterprise Hadoop cluster and through a seamless data streaming and prefetching model, it enables Hadoop jobs to begin execution as soon as they are launched without requiring any apriori downloading. With VNcache's accurate pre-fetching and caching, jobs often run on a local cached copy of the data block significantly improving performance. Uniquely, VNcache is implemented with NO changes to the Hadoop application stack.