Abstract: Cloud computing is rapidly transforming the way the IT industry operates, mainly due to its pay-as-you-use model of providing virtualized computing resources with on-demand scalability. Existing cloud service providers charge customers based upon the amount of resources used or reserved. However, there are no guarantees on the application level performance or quality-of-service that the given resources will provide. Performance interference between co-located applications in a multi-tenant Cloud environment is an important and challenging issue. Furthermore, the highly dynamic nature of Internet workloads, and the increasing complexity of cloud-hosted applications ranging from multi-tier architectures to Big Data Processing frameworks pose unique and significant challenges in managing application level performance. The situation is further complicated by the fact that cloud service providers need to control the power consumption in their data centers to avoid power capacity overload, to lower electricity costs and to reduce their carbon footprint. The complexity and scale of cloud systems make it increasingly difficult for administrators to manage them.

In this talk, I will introduce my work on developing middleware approaches to autonomic resource provisioning for performance control in Cloud environments. Firstly, I will present the design and implementation of a non-invasive mechanism to achieve performance isolation among co-located applications in virtualized servers. It mitigates the hard performance interference problem in multi-tenant Cloud environments in an energy efficient manner. Secondly, I will present my work on autonomic server provisioning for performance assurance of multi-tier Cloud applications. It involves the design of a model-independent and self-adaptive control system for robust performance guarantee in the face of highly dynamic workloads. Thirdly, I will present the design and implementation of an automation tool for resource allocation and configuration of Hadoop framework for cost-efficient big data processing in the Cloud. I will briefly present my works on coordinated control of application performance and power consumption in virtualized server clusters, and a power-aware consolidation and placement of scientific workloads in high performance computing GPU clusters. Finally, I will discuss my future steps focusing on predictable performance, energy-aware computing and reliability in Cloud and Big Data Processing.

Short Bio: Palden Lama is a PhD candidate in Computer Science at the University of Colorado, Colorado Springs. He received his B.Tech in Electronics and Communication Engineering from Indian Institute of Technology, Roorkee. His research interests lie at the synergistic intersection of Cloud Computing, Autonomic Computing, and Sustainable Computing. His work has led to 11 publications in top-ranked journals and premiere conferences in the field. Palden was awarded the Outstanding PhD Student of the Year in the Computer Science department in 2010.