**Title:** Adapting the Serverless Platform for Emerging Application Patterns

**Speaker:** Dr Vivek M. Bhasi

**Venue:** RM 101

**Date:** 6th Feb 2025

**Time:** 3PM - 4PM

**Abstract:**

Cloud computing initially emerged with the promise of providing high-performance computation, cost-efficiency, and elasticity. While these benefits have largely been realized, users continue to face the burden of managing numerous virtual resources, shouldering responsibilities such as autoscaling, redundancy, load balancing, and virtual machine migration. The complexity of these low-level administrative tasks inspired a shift in demand toward a more accessible path for developers for deploying new applications (apps). This demand gave rise to serverless computing, where users focus solely on coding their apps, often represented as Directed Acyclic Graphs (DAGs) of serverless functions. Cloud providers handle all operational tasks, including instance selection, autoscaling, and fault tolerance, enabling a streamlined development process. At the same time, serverless computing benefits providers also by allowing them to maximize system utilization through efficient scheduling of serverless functions on under-utilized resources.

This talk will focus on works where I have been the primary contributor, with a focus on my contributions to advancing serverless computing, particularly addressing its limitations in performance, resource usage, and cost optimization for emerging application patterns. My early research centered on provider-side resource management and performance optimization, with a focus on efficient serverless function container management and request scheduling. These efforts aimed to minimize function memory footprints while meeting Quality of Service (QoS) guarantees, for apps with unique characteristics (such as having dynamic DAG structures or input size–sensitive functions). Building on this foundation, my current work has expanded into more hardware-focused research, particularly in the emerging domain of Heterogeneous Serverless Computing. This includes designing GPU-enabled serverless frameworks that utilize intelligent time and spatial GPU sharing mechanisms to efficiently execute Machine Learning (ML) inference workloads.

**Brief bio:**

Vivek M. Bhasi is a postdoctoral scholar at Penn State University, where he also earned his Ph.D. in Computer Science and Engineering. His research centers on cloud computing, with a focus on serverless computing resource management, and heterogeneous hardware optimizations for real-time ML inference. He also explores systems for ML, leveraging model and hardware characteristics to optimize training and/or inference.