Postgraduate Seminar Series

Venue: Graduate Seminar Room
Date & Time: February 28, 2026 at 12:30 PM

Speaker Information

Md. Shariar Kabir (Std No. 0419052047) is a part-time M.Sc. student in the Department of Computer Science and Engineering, BUET. He completed his BSc in Computer Science and Engineering from BUET in 2019. His research interests include leveraging advanced machine learning (ML) techniques to solve real-world problems and refining modern NLP models to be safer and more effective for diverse populations. He is currently doing her postgraduate thesis under the supervision of Professor Muhammad Abdullah Adnan. He will be speaking about his ongoing research in this talk.

AgnoSVD: Dynamic Resource Allocation for Serverless Workloads using Collaborative Filtering

In serverless computing, determining the optimal resource configuration for each workload poses significant challenges, particularly due to the cloud provider's limited visibility into workload specifics and the non-linear relationship between resource allocation, execution time, and cost. Static default configurations lead to systematic inefficiencies: under-provisioned functions incur excessive execution times or timeouts, while over-provisioned functions waste resources and inflate billing costs. Existing approaches, such as exhaustive search (AWS Lambda Power Tuning), monitoring-based regression (Sizeless), and Bayesian optimization (COSE), each face important limitations: they are either too expensive to run per function, require invasive runtime instrumentation, or degrade when both CPU and memory must be jointly optimized. A further shared weakness is that each function is configured in isolation, ignoring the structural similarity across workloads that could allow knowledge transfer.

In this work, we present **AgnoSVD**, a system that predicts optimum resource configurations using Singular Value Decomposition (SVD)-based collaborative filtering. AgnoSVD constructs a sparse performance matrix where rows represent workloads and columns represent resource configurations. SVD factorizes this matrix into latent factor representations of workloads and configurations, enabling prediction at unobserved configurations via a dot product. Because the model learns entirely from observed execution times, it remains agnostic to the specific details of the functions and the resource configurations, with no need for source code inspection, runtime monitoring, or feature engineering. For a new function, only two reference profiling runs are required before the first recommendation, and the model reaches convergence within **two feedback iterations**. We evaluated two learning algorithms: Alternating Least Squares (ALS) and Stochastic Gradient Descent (SGD), with ALS achieving lower reconstruction error (RMSE 0.48% vs 4.7%) and faster training (0.1-0.2s vs 0.7-0.8s). To handle a growing training matrix, AgnoSVD incorporates variance-based and similarity-based *active learning* strategies for intelligent row selection. We tested the approach on Apache OpenWhisk and AWS Lambda using 99 functional workloads across diverse categories, including micro-benchmarks, image and video processing, ETL pipelines, and machine learning jobs, each evaluated at three input sizes. The model predicts near-optimal performance configurations with **90% probability** and cost configurations with **85% probability** at T=15%, achieving a **32.41% average cost reduction** and **5.18% average speedup** over the default configuration, outperforming Sizeless and other baselines.