INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
SHORT ABSTRACT OF THESIS

Name of the Student	:	**Amit Puri**
Roll Number	:	**186101102**
Programme of Study	:	**Ph.D.**

Thesis Title:
Design, Modeling and Optimization of Large-Scale Disaggregated Memory Systems

Name of Thesis Supervisor(s)	:	**Dr. John Jose, Prof. Tamarapalli Venkatesh**
Thesis Submitted to the Department/ Center	:	**Computer Science and Engineering**
Date of completion of Thesis Viva-Voce Exam	:	**29-Jan-2025**
Key words for description of Thesis Work	:	**Data Centers, Remote Memory, Disaggregated Systems, Large-Scale Systems**

SHORT ABSTRACT

This thesis discusses the challenges faced by traditional server systems in meeting the increasing memory requirements of modern server workloads, leading to the emergence of Disaggregated Memory Systems (DMS) as a solution. The traditional systems struggle with scaling up memory capacity, often resulting in under-utilization of onboard memory resources and inflexible hardware refresh cycles in data centers. DMS offer a more flexible approach by allowing memory to be attached as remote memory modes/pools connected to compute nodes through a high-speed coherent interconnect. This setup enables on-demand memory allocation, eliminates scalability and under-utilization issues, and facilitates independent upgrading of server memory resources, thereby reducing the total cost of ownership. However, the adoption of DMS introduces new system-level design and architecture challenges topped up by the absence of an architectural simulator for performance evaluation. Memory disaggregation increases the Average Memory Access Time (AMAT) due to the network interconnect between compute nodes and remote memory pools, potentially impacting application performance, especially in large-scale configurations. The thesis aims to address these challenges by proposing various system-level and architectural optimizations to reduce AMAT in DMS. It includes building an architectural simulator for performance evaluation, exploring hot-page migration techniques, studying memory bandwidth contention, and investigating network resource allocation to provide Quality of Service (QoS) to different applications. The experiments conducted with various workloads demonstrate that the proposed mechanisms can significantly enhance application performance in large-scale DMS deployments.