INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
SHORT ABSTRACT OF THESIS

Name of the Student	:	**Abhijit Das**
Roll Number	:	**156101018**
Programme of Study	:	**Ph.D.**

Thesis Title: **Designing Data-Aware Network-on-Chip for Performance**

Name of Thesis Supervisor(s)	:	**Dr. John Jose**
Thesis Submitted to the Department/ Center	:	**Computer Science and Engineering**
Date of completion of Thesis Viva-Voce Exam	:	**12 October 2021**
Key words for description of Thesis Work	:	Network-on-Chip (NoC), Memory Hierarchy, Tiled Chip Multi-Processor (TCMP), Network Stall Time, Cache Miss Penalty, System Speedup

SHORT ABSTRACT

We are now in an era where data drives everything, and the demand for information processing is increasing exponentially. This increasing demand is driving a parallel increase in the number of processing cores in Tiled Chip Multi-Processors (TCMPs). It is indeed visible in the industry, with Intel Xeon Phi, AMD EPYC and Ampere Altra processors featuring up to 128 cores in their TCMPs. As the number of cores continues to increase in modern TCMPs, on-chip communication plays a pivotal role in determining the performance. Multi-hop packet-based Network-on-Chip (NoC) is a widely adopted communication infrastructure in modern TCMPs as it provides a high transfer bandwidth and is scalable. Data-driven applications expect minimum memory access latency, where NoC plays a significant role. In fact, it is reported that NoC is responsible for 60% to 75% of the miss latency experienced by the applications. Nevertheless, most of the existing memory access optimisations are NoC oblivious. These optimisations focus on the memory hierarchy and treat the underlying NoC like a black box. Since NoC plays a vital role in memory access latency, ignoring the nature of its infrastructure may severely impact the performance of applications in TCMPs.

This dissertation advocates for considering NoC and memory hierarchy together when designing techniques and proposing optimisations for TCMPs. The dissertation shows that designing data-aware NoC with the help of memory hierarchy improves resource utilisation, reduces memory access latency and improves system performance in TCMPs. Specifically, the architectures proposed in this dissertation are able to improve overall system performance by up to 19% with negligible storage, area and power overhead. While TCMPs continue to scale with the help of NoC, the proposed architectures can help with future design decisions.