



**INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
SHORT ABSTRACT OF THESIS**

Name of the Student : **Dipika Deb**

Roll Number : **166101001**

Programme of Study : **Ph.D.**

Thesis Title: **Performance Enhancement of Tiled Multicore Processors using Prefetching and NoC Packet Compression**

Name of Thesis Supervisor(s) : **Dr. John Jose**

Thesis Submitted to the Department/ Center : **Department of Computer Science and Engineering**

Date of completion of Thesis Viva-Voce Exam : **17-05-2022**

Key words for description of Thesis Work : **Prefetching, Network on Chip, Compression, Packet, TCMP, Multicore**

SHORT ABSTRACT

The well-known memory wall problem is created because of the disparity between the processor speed and main memory speed, restricting a system from achieving the maximum performance benefit. In a multicore system, the performance is closely linked to how fast a cache miss is served, i.e., Average Memory Access Time (AMAT). However, in Tiled Chip MultiProcessors (TCMP), the inbuilt Network on Chip (NoC) plays an important role in determining AMAT. This is because the last level cache is shared and distributed among the tiles present in the system. In such systems, very often, the role of the underlying communication network gets unnoticed. Thus, cache misses experience additional delay apart from the conventional memory access latencies, which makes the block access time non-uniform. The additional delay is the network latency incurred to transfer the cache miss request and reply packet to the requesting tile. The non-uniform memory access latency across the tiles makes it unpredictable to estimate AMAT. Prefetching and NoC packet compression are the two techniques that can be used to reduce AMAT in TCMP. However, none of the existing techniques considers the on-chip communication overhead of TCMP. Considering the limitations of existing prefetching techniques, in this thesis, we propose efficient prefetching strategies that are aware of the underlying TCMP architecture. It identifies the false positive cases of prefetching that results in generating useless prefetch requests. These conditions prevail only on TCMP architectures due to its shared and distributed last level cache. The useless prefetch requests, thus generated, causes cache pollution which further results in generating unwanted NoC traffic. It further congests the network, thereby increasing the packet transfer rate in NoC. We notice that useless prefetch requests increases AMAT, hampering the system performance. We also cannot ignore the fact that cache pollution can be caused by useful prefetches by evicting important demand blocks from the cache. Hence, to reduce cache pollution we propose mechanisms for throttling useless prefetches and efficient strategies for placement of prefetch blocks. In order to reduce the on-chip communication latency, a novel packet compression technique is also proposed that

operates at a smaller granularity of data to achieve better compression ratio. Experimental analysis shows that the proposed prefetching and compression techniques perform better than the existing techniques. Thus, both the techniques combined reduces the on-chip communication cost that directly improves AMAT for TCMP architectures.

