INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
SHORT ABSTRACT OF THESIS

Name of the Student	:	SHIRSHENDU DAS
Roll Number	:	10610112
Programme of Study	:	Ph.D.

Thesis Title: **Effective Utilisation of LLCs by Managing Associativity, Placement and Mapping**

Name of Thesis Supervisor(s)	:	Dr. Hemangee K. Kapoor
Thesis Submitted to the Department/ Center	:	CSE
Date of completion of Thesis Viva-Voce Exam	:	23-01-2016
Key words for description of Thesis Work	:	ChipMultiprocessor (CMP), Last Level Cache (LLC), NUCA.

SHORT ABSTRACT

Tiled based CMP (TCMP) has become the essential next generation scalable multicore architecture. The cores in TCMP commonly share a large sized Last Level Cache (LLC). NUCA is used in LLC to divide it into multiple banks such that each bank can be accessed independently. Static NUCA (SNUCA) has a fixed address mapping policy whereas dynamic NUCA (DNUCA) allows blocks to relocate nearer to the processing cores at runtime.

It has been observed that the LLC of the current TCMP architectures is not utilised properly. Better cache utilisation will reduce the number of misses in the cache and hence can improve performance. The utilisation issue of LLC can be divided into two categories: (a) local utilisation and (b) global utilisation. The memory accesses within a bank are not distributed uniformly among the sets. Some sets are used heavily while some others remain idle. Such utilisation issue is termed as the local utilisation issue. It has also been observed that the banks of the LLC are not carrying equal loads during the execution. Some banks are loaded heavily while some other banks remain almost unused. Better load distribution among the banks may improve the utilisation factor of the cache. Such inter-bank utilisation issue is termed as the global utilisation issue.

In this work we propose architectures to increase both the local and global utilisations of LLC for TCMP. Our first three proposals are for improving the local utilisation. We do this by allowing the heavily used set to use the idle ways of lightly used sets. Hence the associativity of each bank is managed dynamically. The three architectures we propose have different performance benefits and hardware requirements. To improve global utilisation we propose two DNUCA based TCMP architectures capable of distributing loads among multiple banks. Experimental evaluation using full-system simulations has validated our claim of performance enhancement. The improvements in local utilisation give better performance in the range of 6.3-13.5% and those in global utilisation give 6.1-13% performance enhancement.