Abstract—This paper addresses cache organization in Chip Multiprocessor (CMPs). We introduce Nahalal, a novel non-uniform cache (NUCA) topology that enables fast access to shared data for all processors, while preserving the vicinity of private data to each processor. Our characterization of memory accesses patterns in typical parallel programs shows that such a topology is appropriate for common multi-processor applications. Detailed simulations in Simics demonstrate that Nahalal decreases the shared cache access latency by up to 54% compared to traditional CMP designs, yielding performance gains of up to 16.3% in run time.

Index Terms—cache memories, chip-multiprocessors, shared memory systems

I. INTRODUCTION

Emerging and future computer architectures are designed as Chip Multi-Processors (CMPs), which leverage the parallelism of multi-threaded applications to achieve higher performance within a given power envelope. Data access is typically a bottleneck in such systems, as multiple threads compete for limited on-die memory resources. Hence, the organization and management of on-chip cache memory become critical to system performance.

Two major factors impact the latency of on-chip memory access: wire delays and contention over shared memory. Global wire delays are becoming a dominant factor in VLSI design [1][2][3], and on-chip cache access time increasingly depends on the distance between the processor and the data. Concurrent access by multiple processors to a shared cache further increases the access time, as additional delays are incurred for resolving contention on the cache. Some communication fabric such as a Network-on-Chip [4] is used for interconnecting the