International Journal of Research in Computer Engineering and Electronics. 1 ISSN 2319-376X
VOl : 3 ISSUE :3(May-June'2014)

Latency and coherency reduction With Efficient Energy Utilisation using a Novel cache Architecture for NOC circuits

Ashwini Kulkarni[1] , S.P.Mahajan[2]

Abstract— For Network on Chip (NOC) Circuits where the fast computation and communication is required, memory architecture plays a vital role in deciding the efficiency of computation. Cache is the fastest memory amongst all available memories. Architecture and type of this cache memory has a crucial role to play in NOC circuits. Performance of cache memory plays an important role in deciding the overall performance of NOC circuit. This cache memory suffers with two major problem cache latency and cache coherency.As the number of cores are increasing power consumption of the circuit also increases and becomes a major constrain in the design phase. This paper takes a review of these problems associated with the existing cache memory of NOC chips and suggests a novel architecture of cache model to minimize cache latency and solve the coherency problem.

Index Terms— Network on Chip (NOC) , Cache Memory, Coherency, Latency, fast computation, energy saving, cache architecture

INTRODUCTION

Network on chip is an emerging technology today. It is used almost in all fields where high computational power is required. Memory hierarchy matters a lot in the process of data storage and retrieval. Cache memory is the closest memory of any processor. Sharing of data is often required especially in multi threaded applications when one task is divided into many threads, for joining data is always transferred and shared .

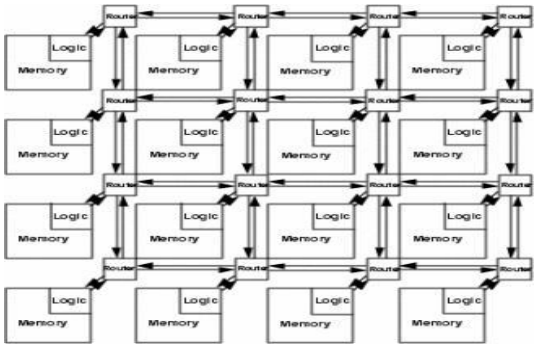


Fig 1. NOC Architecture

Overall speed of execution depends on the latency in fetching the data from memory. Cache memories is a well known choice for reducing the time lag and in turn minimizing latency of data access.

Type of cache and structure of cache plays a pivotal role in network on chip(NOC) circuits. If the latency is observed in fetching the cache data, then the actual purpose of cache design is not served. Another problem that is observed with the cache memory is coherency.

When two or more copies of the same data exist in different processors' memories, it leads to different values of the same variable in different processors. If there is inconsistency between the cached copy and the shared data copy or between the cache copies themselves because of multiple cache copies of data exists then it is a serious problem which challenges the accuracy of the entire system. In this paper we tried to resolve both the issues of latency as well as coherency of cache memories by introducing a new cache architecture for NOC networks.

In this architecture we are also integrating the power saving mechanism which is very useful for all the NOC circuits

Need of Latency Reduction:-

For any specific NOC application based on the requirement , cache architecture should also be precisely chosen. Though the choice of cache depends on the actual specific application, its structure surely influences the overall performance of the chip.

In network on chip circuits where multiple

- *[1]Ashwini Kulkarni is currently working as Assistant Professor at K.J. College of Engineering and Management Research,Pune university, India,*
- *[2]Dr. S.P. Mahajan is currently working as Associate Professor at College of Engineering , Pune,India.*

computational processors are used , different types of cache structures are possible. When fast access and speedy computation is required, private cache is the obvious choice. Still common global cache structure has it's own advantage of maximum utilization of memory. Static and dynamic types of cache memories are also available depending on the load of computation on the specific cores.

Not only choice of cache memory but it's updation is also equally important as it may cause the problem of latency. Number of cache misses are directly proportional to the off chip memory access and leads to cause latency so it is a crucial task to select the appropriate cache and it's refreshing mechanism.

There are various reasons for Cache latency. Cache architecture, path delay and connectivity amongst different components, design layout , cache Miss , cache access mechanism are various reasons for introduction of latency.

Different Approaches used to reduce latency

Previous work is done by many researchers to solve this latency problem.

Latency can be caused by multiple hops in the path. So the location of cache memory on chip is also equally important. It is important especially in globally declared cache memory being utilized by multiple cores.[11][7]

Various attempts are continuously being made to reduce this latency. Some of the approaches are reducing execution time[1] , reducing miss rate [2] thereby increasing hit ratio by proper bank replacement policies, cache bank replacement like spilling [3] where the exact data which may not be required by the core is spilled and speculation[4] in which the data likely to be required by the processor core is initially speculated and accordingly the data replacement is done causing reduction in latency or another approach is to dynamically allocate shared and private caches near processor core[5] , circuit pinning is also suggested [6] to reduce latency and another approach is to compress data [7] so that data storage capacity is increased and hit ration increases. Co-operative partitioning is also used wherein the nodes cooperate among themselves and they decide how much memory to be used by which node depending on their load [8]

All the above mentioned techniques have their specific application domain. There is no single technique exist which perfectly reduces latency as the causes of latency are numerous. We hereby suggest a general architectural module of cache memories of NOC circuits which would reduce the latency and totally eliminate the chances of cache miss giving 100% hit ratio.

Effect of coherency

In multiprocessor application shared data must be the updated one. But the speed at which it gets updated must be synchronized with the cores that are using this data. This generally doesnot happen and it leads to problem of coherency.

Cache coherency is a commonly observed problem especially in chip multiprocessors it is observed to be occurring very frequently.

Coherency problem occurs when the data of one core is accessed by other core communicating with it. Generally when the private cache memories which support the high speed communication and which are the preferred choices in cache latency reduction problem , there , the data to be processed is transferred into the desired cache memory of the target core but the updation in the record is not keenly done. So in this situation, old data is processed somewhere and the new data which is required to be actually processed remains in the previous cache.

This leads to the situation of having two values of the same variables at two different locations. It is the most unsafe situations for computing the correct results.

Different approaches used to avoid /reduce Coherency

Previously cache coherency problem is addressed and tried to be solved by two main approaches software approach and hardware approach. In hardware approach , tags [9] or dedicated control network [1] are used or hardware cache coherence protocol [1] is suggested. The sharing tracker approach [9] is a directory based approach which deals with this problem. Software approach basically considers techniques like snooping[10] , directory based approach[11], virtual tree directory[12][13]etc they are protocol based programs.

In the proposed design we suggest a mechanism which does not let this problem arise. The architecture , interconnection and communication mechanism of the suggested module enable NOC cache to totally get rid of this problem.

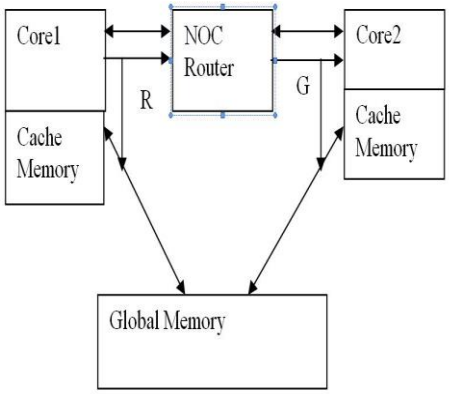
Energy Saving Mechanism

Power saving is crucial in NOC circuits. As the number of cores are increasing day by day, power consumption of NOC increases. Various attempts are made previously to tackle this condition. Removing global clock[14] , introducing low voltage swing[15], keeping the unused path off using power gating[16], , using speculation[12] and thereby reducing cache miss to saving power, using clock signal with reduced swing [15] approaches which are previously used to save the power of NOC circuit.

Global clock removal need separate controller, for adjusting voltage swings separate design consideration

is required whereas tracing of unused paths needs continuous monitoring. Though speculation to some extent is helpful still it is not a perfect solution of power saving.
In our architecture we are integrating the clock gating mechanism where power is utilized only when the core and cache is active otherwise if both of them are unused, power is not consumed and hence saved. So the circuit accepts energy as and when required.

Block Diagram



This new cache module consists of a global memory which is shared and accessed by cache memories of the processor cores. Whenever processor1 wants to communicate with processor2, through NOC router request signal (R) is sent. Based on the priority mechanism of router, request is either granted or it has to wait. But whenever request signal is sent out at the same time data of cache memory is written into global memory and it's address is sent on data bus of the core. The moment request is granted, (G) signal is enabled and cache data which is replicated in global memory is accessible to cache memory of the requested processor2.

This technique actually saves the time in accessing the data of the requesting processor. As cache memory of processor core1 is updated, immediately global memory and then the cache of the requested processor is refreshed and updated. Upon the successful transmission of the requested data, data from global memory is erased and the freed memory can be utilized for other applications. The additional features which we are introducing with this new architecture a power saving facility associated with the computing functionality, the priority computation supportive architecture and reconfigurable memory utilization technique.

References

1. Lodde, M. ; Roca, T. ; Flich, J.,"Built-in fast gather control network for efficient support of coherence protocols" , IET,Computers & Digital TechniquesVolume: 7 , Issue: 2, pp 69 - 80 ,May 2013
2. Chaturvedi, N. ; Gurunarayanan, S.,"An Adaptive Block Pinning Cache for Reducing Network Traffic in Multi-core Architectures" , 5th International Conference on Computational Intelligence and Communication Networks (CICN), 2013 pp. 446 - 450 ,Sept 2013
3. Qureshi, M.K. , "Adaptive Spill-Receive for robust high-performance caching in CMPs", IEEE 15th International Symposium on High Performance Computer Architecture, 2009. HPCA 2009. pp 45 - 54 ,Feb 2009
4. Kim, Hyungjun ; Grot, Boris ; Gratz, Paul V. ; Jimenez, Daniel A. , "Spatial Locality Speculation to Reduce Energy in Chip-Multiprocessor Networks-on-Chip" , IEEE Transactions on ComputersVolume: 63 , Issue: 3 , pp-543 – 556, March2014
5. Hyunjin Lee ; Sangyeun Cho ; Childers, B.R. ; "SimulusCache: Boosting performance of chip multiprocessors with excess cache ", IEEE 16th International Symposium on High Performance Computer Architecture (HPCA), 2010 ,pp-1, January 2010
6. Abousamra, A. ; Melhem, R. ; Jones, A., "Winning with Pinning in NoC", 17th IEEE Symposium on High Performance Interconnects, 2009. HOTI 2009. pp 13 – 21,Aug 2009
7. Enright Jerger, N.D. ; Li-Shiuan Peh ; Lipasti, M.H., "Virtual tree coherence: Leveraging regions and in-network multicast trees for scalable cache coherence" , 41st IEEE/ACM International Symposium on Microarchitecture , MICRO-41. pp35-46 ,Nov 2008
8. Qadri, M.Y. ; McDonald-Maier, K.D. ; "A Fuzzy Logic Reconfiguration Engine for Symmetric Chip Multiprocessors " , International Conference on Complex, Intelligent and Software Intensive Systems (CISIS), 2010 , pp 973 , February 2010.
9. Tarjan, D. ; Skadron, K. , " The Sharing Tracker: Using Ideas from Cache Coherence Hardware to Reduce Off-Chip Memory Traffic with Non-Coherent Caches" .International High Performance Computing, Networking, Storage and Analysis (SC), Conference for , pp1 - 10 ,2010
10. Daehoon Kim, Hwanju Kim, and Jaehyuk Huh, "Virtual Snooping: Filtering Snoops in Virtualized Multi-cores" , 43rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), pp 459-470 , Dec 2010
11. Jinglei Wang ;Yibo Xue ; Haixia Wang ; Dongsheng Wang, " Network Victim Cache: Leveraging Network-on-Chip for Managing Shared Caches in Chip Multiprocessors",4th International Conference on Embedded and Multimedia Computing, 2009. EM-Com 2009, pp 1-5, Dec 2009
12. Lenjani, M. ; Hashemi, M.R. , "Tree-based scheme for reducing shared cache miss rate leveraging regional, statistical and temporal similarities , Computers & Digital Techniques" , IET Volume: 8 , Issue: 1, pp 30 - 48 , Jan 2014
13. Roy, A. ; Jeevan, S. ; Jingye Xu ; Chowdhury, M.H., "Impact of cache power reduction techniques in multi-core processor using network on-chip paradigm" ICM 2008. International Conference on Microelectronics, 2008. pp 163 - 166 , Dec2008
14. Gebhardt, D.,Junbok You ; Stevens, K.S. ,"Comparing Energy and Latency of Asynchronous and Synchronous NoCs for Embedded SoCs",Fourth ACM/IEEE International Symposium on Networks-on-Chip (NOCS), 2010 ,pp-115-122, May 2010

15. Yi Liu ; Gang Liu ; Yintang Yang ; Zijin Li , "A novel low-swing transceiver for interconnection between NoC routers ", 7th International Conference on Digital Content, Multimedia Technology and its Applications (IDCTA), pp 39 - 44 ,Aug 2011
16. Sundararajan, K.T. ; Porpodas, V. ; Jones, T.M. ; Topham, N.P. ; Franke, B. . , "Cooperative partitioning: Energy-efficient cache partitioning for high-performance CMPs" IEEE 18th International Symposium on High Performance Computer Architecture (HPCA), pp 1-12, Feb 2012