ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 10, October 2016**

# Clustered Look Ahead Prefetching Mechanism for DRAM NUCA Structure

R.Meera[1], T.Perarasi[2]

Student (M.E -VLSI Design), Dept. of ECE, KVCET, Anna University, Tamil Nadu, India[1]

Associate Professor, Dept. of ECE, KVCET, Anna University, Tamil Nadu, India[2]

**ABSTRACT:** DRAM is one of the main memory for energy consumption. Clustered DRAM accesses a popular first-ready, first-come, first-serve is memory request scheme for energy consumption. We propose a 3-D Integration Technology to combine SRAMs and MRAMs together onto chip multiprocessors. SRAMs have a more complicated structure with active connection between power and ground. The active connection between power and ground provide a much more robust signal. SRAMs is much easier to read resulting in a smaller latency. In this 3-D Integration technology combines resource allocation, power gating, data migration and frequency scaling of processors. NUCA architecture is proposed to reduce cache access latency. By running multiple programs, a multiple threads to exploit many resources. In a static NUCA architecture is proposed where the cache is broken into banks which can be accessed at different banks may proceed in parallel. A runtime cache management scheme improves the system performance and energy efficiency. This method yields on average of 61% performance improvement in terms of IPS. The proposed architecture can be used in real time system, mobile application and memory application.

**KEYWORDS:** DRAM, NUCA, energy efficiency, cache management scheme.

## I.INTRODUCTION

**E**nergy efficiency is more critical in battery operated mobile systems. This become more apparent in high performance mobile system such as smart phone and tablet PCs. DRAM with a capacity of several gigabytes [1], [2]. DRAM is most popularly used as main memory because of high density and low cost. Multi core processors are popular in computer system owing to poor scalability of single core-processors. Programs are getting bigger and trending to larger DRAM accommodation for larger programs in main memory. However larger DRAM capacity is accommodate with higher amounts power and energy, thus increasing cooling cost and reducing the lifetime of the battery. By increasing power and energy consumption of DRAM it is focused on over fetching problem and static power consumption. We propose a 3-D integration technology to combine SRAMs and MRAMs together onto chip multiprocessors. A memory clustering traffic is proposed which focuses on the energy conservation of RAM, called Clustered Look Ahead Prefechting (CLAP) to reduce the activate/ precharge and idle energy consumption of the system. The system memory is predicted using look-ahead prefechting (LAP). In CLAP, prefechting accesses are postponed until normal memory accesses are generated at data path. In this way we can increase the probability of row buffer hits and idle periods with first-ready, first-come, first serve (FR-FCFS). By using this memory scheduling technique we can reduce the number of row activation and idle power consumption.

High Latency of off-chip memory accesses has been critical in thread performance. Inter thread memory contention, If not properly managed can have individual thread performance as well as overall system throughput which leads to system underutilization and thread starvation [11]. Previously proposed memory scheduling algorithms are biased in system performance. By using this approaches cannot provide the high fairness and system throughput at same time. Cache performance is discussed for system performance and energy consumption. Cache is a hardware or software components that stores data so that future request for that data can be served faster, the data stored in a cache might be the result of an earlier computation. DRAM has some problem is that the switch is not a perfect valve, so electrons often “leak” away which can cause the device to lose information. In this paper we propose 3-D integration technique the SRAMs provide a active connection between power and ground. The active connection between power and ground provide much more robust signal. SRAM is much easier to read resulting in a smaller latency. MRAM

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 10, October 2016**

stores information in nano scale magnets. MRAM makes a lot of sense for on-chip cache. Cell area is also comparable to DRAM. This technique combines resource allocation, power gating, data migration and frequency scaling of processors. NUCA architecture is proposed to reduce cache access latency. By running multiple programs, a multiple threads to exploit many resources. NUCA architecture which allows nearer cache banks to lower access latencies than further banks. NUCA architecture was initially proposed for uniprocessor systems. They consider L2 cache that has a single cache controller feeding one processor core.

## II.SURVEY

1. Energy efficient hardware data prefetching mechanism is introduced in recent year for energy and power efficiency in embedded system. The prefetching instructions are supported by most contemporary microprocessors. The most commonly used Hardware prefetching techniques use additional circuitry for prefetching data on access patterns. Hardware prefetching yields better performance than software prefetching. In this hardware prefetching mechanism a net leakage energy reduction is due to performance improvement.

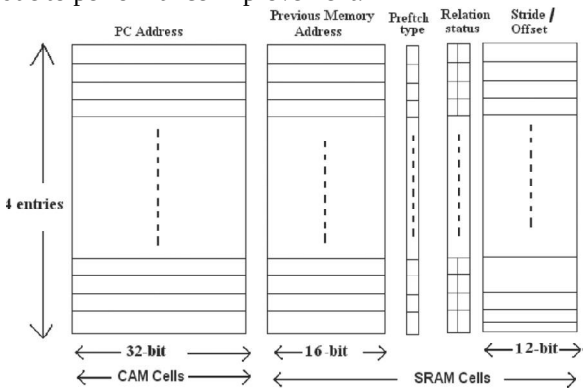
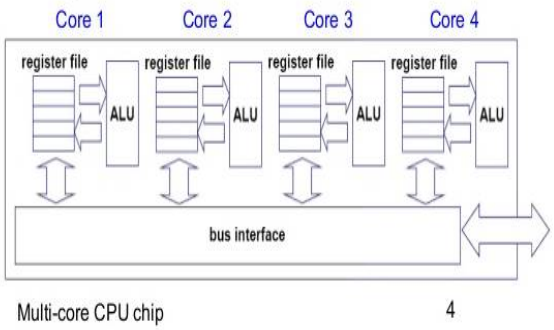



Fig.1 Baseline design for hardware prefetch.

2. DRAMs is one of the most popular used to implement the main memory due their high densities and low prices. Multicore processors are popular in embedded system and high performance systems due to poor scalability of single-core processors. They have try to reduce energy consumption by using conventional DRAMs. Larger DRAMs are demanded due increase in program sizes. This trend increases power and energy consumption of DRAMs. Here th proposed Skinflint DRAM system ha lower performance, area, delay and energy because it is implemented by conventional cache and modified conventional DRAM system.




3. DRAM memory plays an important role in overall power of latest-generation servers with multicore processors. There is a need to fully evaluate the memory power of contemporary DRAM memories. DDR2 and FB-DIMM are also included which shows that DRAM system configuration, including page policy, power mode, device configuration, burst length, channel organizations, selection of DRAM technology affects the memory power consumption.
A contemporary high-performance memory system can perform multiple independent channels. Each channel connect several DIMMs that can include multiple ranks and each rank provides a group of DRAM chips to support memory

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 10, October 2016**

data bus width. A bank is a two dimensional array whose data is based on row buffer status, DRAM access may require different numbers of operations and different access time to consumes different amount of power. A row buffer access requires precharge access if the bank has not been precharged. The performance of DRAM memory systems is sensitive to system organization.

We thoroughly evaluate and compare representative contemporary DRAM architectures performance under multicore processor systems. By increasing memory power consumption is becoming a severe concern for modern high-performance in computing systems. Simple low power mode management policies that used as an idle rank into power-down modes.

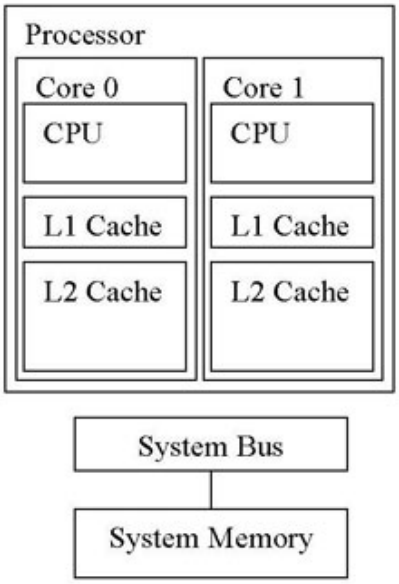



Fig.2 Multicore processor separate L2 cache

## III. EXISTING

In this paper clustered DRAM accesses exploit a popular first-ready first-come first-serve memory request scheduling which is more effective and increase the system performance. DRAM is one of the most main sources of energy consumption in computer systems. DRAM also reduces the energy consumption which prolong the lifetime of the battery operated. A new prefetching scheme is introduced to increase row buffer hits and idle periods of DRAM to improve the system performance and utilizing them for energy conservation. This scheme predicts the energy uses in system and clusters future DRAM access. In this prefetching, the memory traffic clustering scheme is to reduce the power and energy consumption of DRAM.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 10, October 2016**

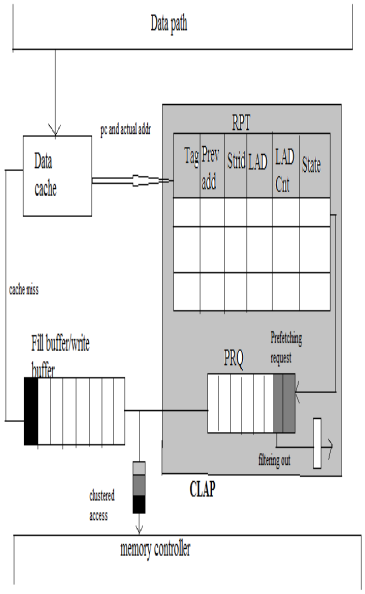



Fig.3 CLAP overall architecture.

In the existing system prefetching scheme is developed to boosts the system performance by bring the data to lower level memory before they are required. Prefetching technique basically predicts the future data and determines the previous data. By this future work it implies that prefetching can be used for system performance as well as memory traffic clustering. While previously proposed memory traffic clustering scheme degrades the system performance due to demanded memory accesses are delayed. Power consumption is reduced, the energy may be counter balanced if the overall system execution time is increased. Clustering will reduce the power consumption in future memory accesses.

### *A. PREFETCHING MECHANISM*

There are three types of prefetching schemes which plays there own role for various system performance and energy consumption. First it is based on sequential prefetching scheme. This scheme is simplest method and prefetch the data sequentially i.e initiates a prefetch for block s+1 when block s is accessed. In sequential prefetching it is divided into prefetch on miss and tagged prefetch. Prefetch on miss scheme initiates a prefetch for block s+1 when block s results in cache miss, in tagged prefetch scheme fetches blocks s+1 when block s is referenced first time.

### *B. CLAP MECHANISM*

Look Ahead Prefetching technique is proposed to track the future instructions which are known as program counter and prefetching technique is called as Look Ahead program counter. To generate a prefetching addressed from a previous memory reference address with a stride limits the prefetching to a single loop forward iteration. While in LAPC prefetching the prefetch address is created using previous memory reference and time valve. The prefetchimg data are useless when branch is incorrectly predicted which causes unnecessary memory traffic and cache pollution.

In this existing work larger and faster DRAM systems were demanded due to increase in program sizes and popular thread level parallelism. This trend however increases the power and energy consumption of DRAM. In this paper it is based on data prefetching scheme for memory traffic clustering which can achieve a large improvement in the row buffers and power down mode utilization for system performance.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 10, October 2016**

## IV. PROPOSED

DRAM in the main memory access which is mainly used as a memory clustered traffic system to reduce power, energy and system performance. In this we proposed a new prefetching scheme called stride prefetching which monitors the memory access patterns to detect constant gaps in the addresses of memory accesses. Stride prefetching is implemented normally by comparing addresses used by memory instruction. The stride prefetching requires the address of previous memory access to be recorded with the last detected stride called reference prediction table (RPT). This is used to maintain the information regarding to the recently used load instructions. Prefetching request queue (PRQ) is responsible for clustering prefetching requests generated by RPT. A prefetching request is filtered out to avoid duplicate prefetching for same data.

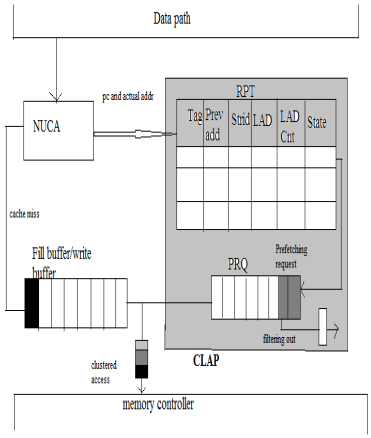



Fig..4 CLAP overall structure for NUCA technique.

The CMPs cache structure is designed to have uniform cache access time regardless of the block being accessed. Such Uniform Cache Access (UCA) architectures certainly simplify the cache access policies. Today cache become larger and also partitioned into multiple banks, maintaining uniform accesses time for the entire cache is not good choice. The banks nearer to a core can actually be accessed much faster than the furthest bank. Wire delay plays an increasing significant role in cache design. While increase in wire delay it is difficult to provide uniform access latencies to all L2 cache banks. NUCA architecture was initially proposed for uniform systems. They consider a large L2 cache that has a single cache controller feeding one processor core.

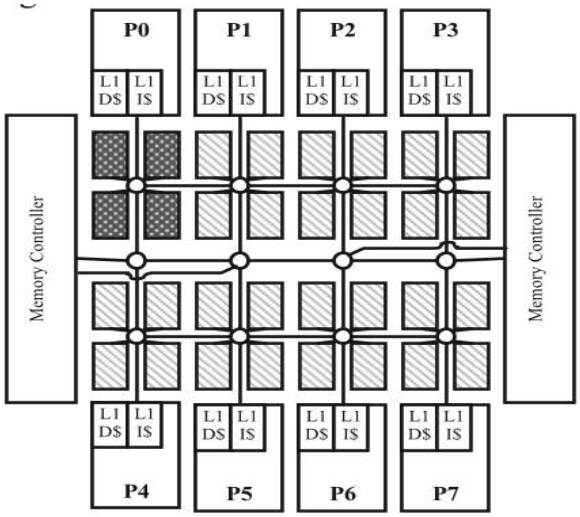
Increasing wire delay makes it difficult to provide uniform access latencies to all L2 cache banks. One alternative is NUCA designs, which allow nearer cache banks to have lower access latencies than further banks. NUCA architecture was initially proposed for uniprocessor systems. They consider a large L2 cache that has a single cache controller feeding one processor core. Larger L2 cache is divided into multiple banks and all the banks are connected between them and cache controller

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 10, October 2016**




.Fig. 5 NUCA architecture.

NUCA caches are the only candidate to be the most performing and energy saving systems. In design of NUCA caches which considers energy consumption, the reduction of static power consumption. For dynamic components, the switched network dissipation is the most critical one to be considered in NUCA designs.

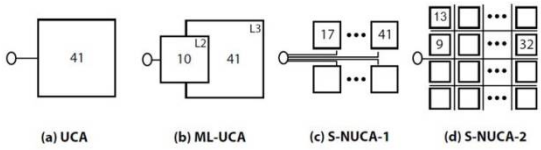
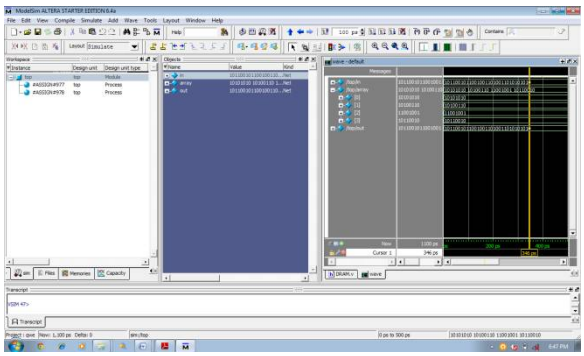



Fig. 6 Different cache banks with different latencies.

By considering the NUCA architecture and stride prefetching scheme in DRAM system the energy and power consumption is reduced.

## V. RESULT

The simulation result is shown in Modelsim 6.2c and synthesized using ISE design suite 14.5 and are shown below.



ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 10, October 2016**

The simulated result is discussed about the NUCA scheme implemented in memory system for energy consumption and power.

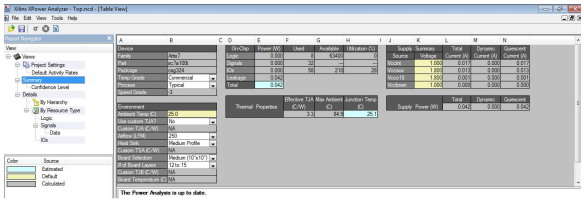


Fig.7 Power analysis of NUCA architecture.

The amount of power used in this architecture is 0.042W.

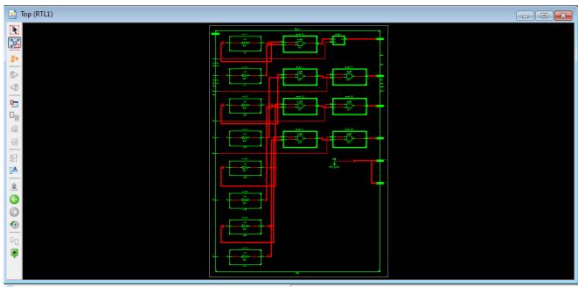


Fig.8 RTL schematic viewer.

Fig.9 Technology schematic viewer.

## VI. CONCLUSION

The proposed method has an advantage of using the stride prefetching scheme and NUCA architecture by reducing the efficiency of energy and power. The usage of memories are also controlled by reuse the memories in base of cache. The cache stores the data so future requests for that data can be served faster, the data stored in a cache might be the result of an earlier computation. The uses of cache allows for higher throughput. L2 cache is partitioned into multiple banks to enable parallel operations. This NUCA architecture can be used in real time systems, mobile application and memory application.

## REFERENCES

[1] M. T. Schmitz, B. M. Al-Hashimi, and P. Eles, *System-Level Design Techniques for Energy-Efficient Embedded Systems*. New York, NY,USA: Springer-Verlag, 2004.
[2] V. Konstantakos, A. Chatzigeorgiou,S.Nikolaidis, and T. Laopoulos, "Energy consumption estimation in embedded systems," *IEEE Trans. Instrum. Meas.*, vol. 57, no. 4, pp. 797–804, Apr. 2008.
[3] B. Jacob, S. W. Ng, and D. T. Wang, *Memory Systems: Cache, DRAM, Disk*. San Francisco, CA, USA: Morgan Kaufmann, 2010.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 10, October 2016**

[4] Micron Technology, Inc. (2009). *1Gb: x16, x32 LPDDR SDRAM Datasheet—MT46H64M16LF-5:B*. [Online].Available:http://www.micron.com/products/dram/mobile-lpdram
[5] K. Lim, P. Ranganathan, J. Chang, C. Patel, T. Mudge, and S. Reinhardt,"Understanding and designing new server architectures for emerging warehouse-computing environments," in *Proc. 35th Annu. Int. Symp. Comput. Archit.*, 2008, pp. 315–326.
[6] D. Meisner, B. T. Gold, and T. F. Wenisch, "PowerNap: Eliminating server idle power," in *Proc. 14th Int. Conf. Archit. Support Program. Lang. Oper. Syst.*, 2009, pp. 205–216.
[7] M. R. Guthaus, J. S. Ringenberg, D. Ernst, T. M. Austin, T. Mudge, and R. B. Brown, "MiBench: A free, commercially representative embedded benchmark suite," in *Proc. IEEE Int. Workshop Workload Characterization*, Dec. 2001, pp. 3–14.
[8] Á. Beszédes, R. Ferenc, T. Gergely, T. Gyimóthy, G. Lóki, and L. Vidács, "CSiBE benchmark: One year perspective and plans," in *Proc. GCC Developers' Summit*, 2004, pp. 7–15.
[9] M. C. Carlisle and A. Rogers, "Software caching and computation migration in Olden," *J. Parallel Distrib. Comput.*, vol. 38, no. 2, pp. 248–255, 1996.
[10] T.-F. Chen and J.-L. Baer, "Effective hardware-based data prefetching for high-performance processors," *IEEE Trans. Comput.*, vol. 44, no. 5, pp. 609–623, May 1995.
[11] Z. Herczeg, D. Schmidt, Á. Kiss, N. Wehn, and T. Gyimóthy, "Energy simulation of embedded XScale systems with XEEMU," *J. Embedded Comput.*, vol. 3, no. 3, pp. 209–219, Aug. 2009.
[12] P. Rosenfeld, E. Cooper-Balis, and B. Jacob, "DRAMSim2: A cycle accurate memory system simulator," *Comput. Archit. Lett.*, vol. 10, no. 1, pp. 16–19, Jan./Jun. 2011.
[13] C. J. Lee, O. Mutlu, V. Narasiman, and Y. N. Patt, "Prefetch-aware memory controllers," *IEEE Trans. Comput.*, vol. 60, no. 10, pp. 1406–1430, Oct. 2011.
[14] *Intel 80200 Processor Based on Intel XScale Microarchitecture: Developer's Manual*, Intel Corp., Mountain View, CA, USA, 2003.
[15] D. Tarjan, S. Thoziyoor, and N. P. Jouppi, "CACTI 4.0," HP Lab., Palo Alto, CA, USA, Tech. Rep. HPL-2006-86, 2006.

## BIOGRAPY

**R.Meera** is an M.E scholar in VLSI design in Electronics and Communication Department, Karpaga Vinayaga College of Engineering and Technology, Affiliated to Anna University. She received B.E degree in 2015 from Anna University, Chennai, Tamil Nadu, and India. Her research interests are Low Power VLSI and Verilog.

**T.Perarasi** is working as associate professor in Electronics & Communication Engineering Department, Karpaga Vinayaga College of Engineering and Technology, Affiliated to Anna University, Chennai, Tamil Nadu and India. Her research interests include cognitive radio networks and communication systems.