Performance and Energy Efficient Cache System Design : Simultaneous Execution of Multiple Applications on Heterogeneous Cores

***Abstract*—Simultaneous execution of multiple applications across heterogeneous cores without space time sharing improves the overall resource utilization and hence performance. This has been well established. However, with regard to hierarchical cache system design for such heterogeneous many cores, capable of simultaneous execution of multiple applications, no investigation has been undertaken with regard to replacement policies and also to show how cache miss gets reduced in such execution environment. This paper presents replacement policies taking into effect, multiple applications being run without space time sharing.**

***Keywords*-Multiple Application Execution, Cache System Design, Replacement Policy**

I. Introduction

The research focus on future generation supercomputing clusters is towards designing cost effective, power and performance efficient heterogeneous core architectures. In this regard, intense research is being carried out [1] [2] on the trails of compiler and functional unit design considering simultaneous multiple application execution non-space time sharing.

1) Compilation Accelerator on Silicon (CAS): When the number of cores in a heterogeneous multi-core environment is increased in thousands, parallel issue of hundreds of instructions and their efficient scheduling becomes a bottleneck. This is overcome by using Compilation Accelerator on Silicon, a hierarchical hardware based compiler cum scheduler [1].

2) Algorithm Level Functional Unit (ALFU): Usage of judicious combination of ALFUs and scalars give energy and performance efficient core architectures [2].

3) Algorithm Level Instruction Set Architecture (ALISA): This is a superset of CISC and VLIW instruction. A single ALISA instruction is equivalent to several ALU, vector and VLIW instructions [2].

4) Simultaneous Execution of Multiple Applications (SMAPP): It is well established that simultaneous execution of multiple applications non space time sharing improves resource utilization [2]. The architecture of ALFUs are evolved in such a manner to support simultaneous multiple application execution [2].

While research has been carried out in the above directions with regard to cache system design, there has not been enough stress on replacement policies and the associated performance-energy relationship in the context of simultaneous execution of multiple application non space time sharing. The Least Recently Used (LRU) cache replacement policy [6] allocates resources based on current need but does not perform replacement based on cache utility. This leads to substantial reduction in performance in heterogeneous multi-core environment. Adaptive per-Thread Least Attained-Service [4] is a memory scheduling technique that prioritizes threads that have attained least service from the memory controller over other threads for scheduling job. [7] proposes a set of heuristics in which pages are allocated to applications based on their needs and these requirements fall under three main categories: latency, bandwidth and power.

The focus of this paper is on investigating the replacement policies and energy models to suit simultaneous execution of multiple application non space time sharing. Section 2 proposes the cache organization and its associated heuristics while section 3 describes cache controller and its associated architecture. Section 4 deals with the associated energy model which is followed by the results and analysis section

II. Cache organization and replacement strategies

All data which is being stored in the dynamic memory are in the form of data packets, and these data packets are clubbed together to be called a block of size equal to a single cache line and the group of lines are termed cache set. For mapping the data packets from DRAM to the cache employing a combination of mapping strategies based on the cache size at the respective levels as in the conventional system cache design. For mapping the data packets from the L3 cache to the L2/L1 cache, the "weighted data packet strategy" is being proposed.

It is noticed from the Figure 1 that the data packets associated with multiple applications co-exists together on a single cache line. Thus the application statistics is given a serious consideration besides the instruction/data, to improve the cache performance in simultaneous execution environment. Further, it is the replacement policies based on the collective statistics that will influence cache performance which is the focus of the paper. On the other hand, the conventional mapping strategies are readily used even in this case, as these mapping strategies are mainly based on cache sizes at respective levels and hence a combination of conventional mapping strategies are used based on the respective cache size.

Considering the block of data to be comprising of data packets, the statistics of the data are analyzed and correspond-

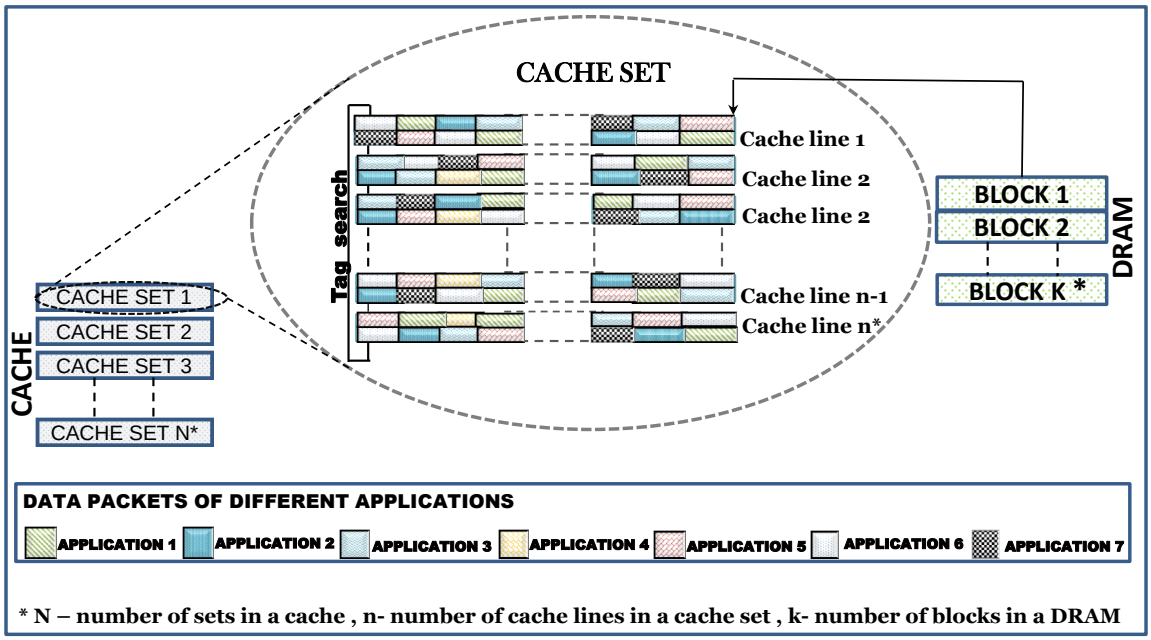

Fig. 1. Proposed cache organisation for multiple applications

ingly a weight is assigned to them namely critical, moderately critical or non-critical. When these data packets are being mapped, one parameter which is of prime importance is the hit to miss ratio.

Under each of these mapping strategies data packets from DRAM to cache are mapped in a predictive manner by analyzing the current instructions in the queue. This data packet mapping is performed by setting threshold in terms of in terms of number of time stamps (of the WIMAC simulator) by which the concerned instructions will get executed. A suitable value for this threshold is fixed to avoid the replacement of data packets mapped very recently in a predictive manner. However during the cold start, the threshold could be kept higher to avoid compulsory misses.

Application Based Replacement Parameters ⟷ Data Based Replacement Parameters

App Utilization Statistics	App Frequency Statistics	Stalling Statistics	L1 Occupancy	L2 Occupancy	L3 Occupancy	Completion Of Application Execution (%)	APP I.D	Data Packet I.D	Data Utilization Statistics	Data Frequency Statistics

Fig. 2. Application and Data Statistics Table

There is an immediate need for application aware replacement strategy. In conventional replacement schemes are tracking only the data statistics but not application statistics. But in case of simultaneous multiple application execution both the application and data statistics must be taken into consideration and hence replacement strategies should be based on collective decision of data and application statuses.

The statistics related to the data and application are recorded and listed in table shown in figure 2 are the inputs for replacement strategy. Thus, the algorithm calculates the weights of the associated data packets and decides whether the packet could be dropped or not based on its associated weights.

```
For each block within a set
    For each packet within a cache line

        Assigning weights to data packets with respect to following parameters
        {
        Read the number of clock cycles after which the data is required for execution;
        Read the Spatial and temporal locality statistics (LRU, MFU);
        Read the data dependencies field i.e.  based on the number of data dependent on the particular data;
        }
        For each application within a cache line
        {
        Read the Application priority of the particular data packet;
        Read the Remaining amount of the data that is pending to get executed;
        Read the stalling statistics of the related data packets;
        Maintain the ratio of data packets of a application in L1:L2:L3 of a particular application and ensure the fair mapp
        of the data packets all the applications;
        Calculate the associated data packet's utilization statistics (LRU. MFU);
        }

    Calculating  weights of the data packet with a collaborative decision from application and data statistics,
    Based on the weights of the particular data, If a data value goes above/below a
            threshold value then the data packet is termed as High-weight/low-weight data;
    Assigning weights to the blocks=(Number of high weight blocks)-(Number of low weight blocks)
    Notifying the cache controller regarding the details about the low weight blocks to be replaced;

    End
End
```

Fig. 3. Cache replacement heuristics algorithm

III. Cache controller :Implementing replacement policies

Resorting to simultaneous multiple application execution across heterogeneous multi cores, replacement strategies play a vital role to achieve overall cache hit and energy efficiency. These replacement policies are discussed in detail bringing into effect of both application characteristics and associated

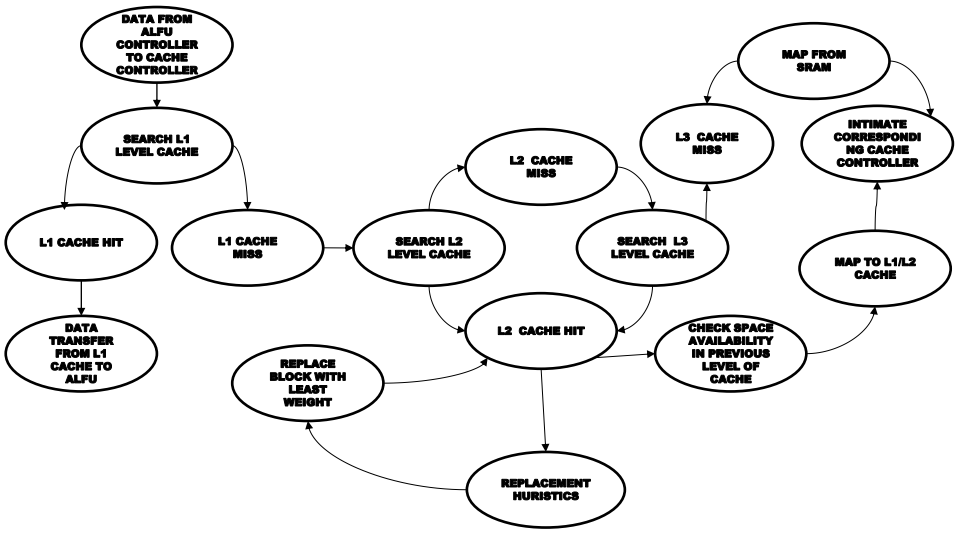

Fig. 4. Proposed Cache Controller: Finite State Machine

Fig. 5. Sample output shows least recently used data packet shown in waveform no.2

instruction/data sets. On the other hand, with regard to cache organization and mapping, conventional techniques are used. Cache mapping is primarily decided by cache size and application characteristics will not play a vital role. In the same way, cache mapping strategies are dependent on cache size and the levels. Either a single strategy or a combination of direct, associative or set associative strategies is used.

A. Cache Controller: Finite State Machine

The cache controllers are present at each levels of cache. The cache controller synchronizes the operations in a particular cache level. The working of the cache controller is similar in all the 3 levels of cache. When there is a miss at a lower level of cache, controller at the lower level sends a control signal to the controller at next level in order to trigger a search operation at that level. The cache controller involves different execution paths for different replacement strategies. Hence a Finite State Machine needs to be designed such that various state transitions depend on the replacement policies. The different replacement policies involve different statistics and almost similar operations. Hence, the different policies can be modeled as a single state. The design of the cache controller revolves around designing a finite state machine the hardware implementation of different replacement policies. From the above section it is evident that most of the operations involve the replacement policies across all cache levels are almost similar leading to a simpler cache controller design. However, for parallel access of different cache levels, either independent cache controller or a single controller can be shared across all levels.

The replacement strategies for simultaneous multiple application execution are simulated using Verilog.The corresponding Verilog output waveforms are shown in figure 5.There are 3 data packets represented by their IDs in waveforms no.3,4,5.The bottom 3 waveforms no.6,7,8 represent the utilization of these data packets.The utilization statistics of data packets and applications are obtained from their statistics table shown in figure 2 present in section 2.Using these statistics as input,the cache controller shown in figure 4 decides either data packets or applications are LRU, MFU etc.In the above verilog example,waveform no.1 shows the least recently used datapacket. In the above example,the utilization of data packets changes frequently whereas the application utilization waveform no.2,remains constant over a period of time.

IV. Multiple Application Execution based power model for the cache system

A unified energy model that captures the execution dynamics of the cache system is essential for the design of the cache system. Conventional cache system design methodologies make use of cache system simulation tools that are integrated with standard energy models. This method of energy

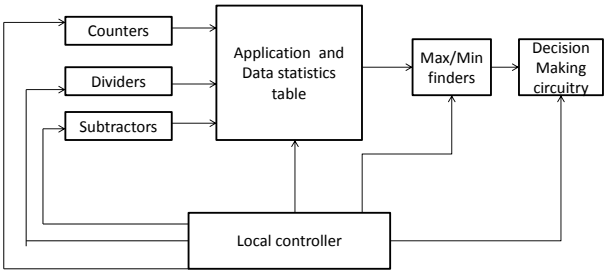

Fig. 6. Hardware implementation of cache replacement strategies

estimation holds good for simple mapping and replacement policies. For the replacement policies that have been presented in the previous section for multiple application execution, there is need for a more rigorous method for estimation of energy by taking into consideration the energy consumed by the controllers that implement the working of the replacement policies. This section presents an energy model based on replacement policies that become essential when simultaneous execution of multiple applications without space-time sharing is in perspective.

The model can be viewed at two levels: the replacement policy execution level and the read/write level. This level encompasses all the operations that are associated with the working of the heuristics based replacement policy. The energy paths taken by the various states in the replacement policies are first listed as shown in figure 7. The operations involved such as search, counting, comparator operations, subtraction etc. are individually traced during execution and based on the number of times the respective units are utilized, the energy associated with the input transitions are computed. Conventional replacement policies do not involve many operations as the proposed approach and hence an execution based power model is often not given as much importance.

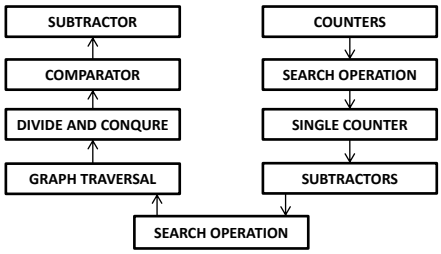

Fig. 7. Operations involved in cache replacement heuristics

The read/write level of the model makes use of the conventional cache models to estimate the energy consumption of the cache. Based on technology and size, the energy associated with wordline, read/write and leakage power of every cache level is integrated with the execution level model and the energy associated with the entire cache system in the replacement scenario is calculated using this model.

V. Results and Analysis

To illustrate the efficiency of our proposed cache paradigm we use an in-house built simulator, Warft India MAny Core (WIMAC) [15] to capture and compare the cache dynamics of various classes of applications. The WIMAC simulator is made cycle accurate to mimic the behaviour of the underlying architecture with great precision. An integrated optimized engine helps to prune the architectural design space to select the most suitable architectural configuration to meet the power and performance requirements of the application. The architectures of Algorithm Level Functional Unit (ALFU) [2] and Compilation Acceleration on Silicon (CAS) [1] are part of this simulation framework.

Cores	9
L3 Size	2 MB
L1 Size	16 KB
Number Of L1 Lines	448
Number Of L2 Lines	448
Number Of L3 Lines	3800
L1 Associativity	4
L2 Associativity	4
Number Of Packets	4

Fig. 8. Sample architecture table

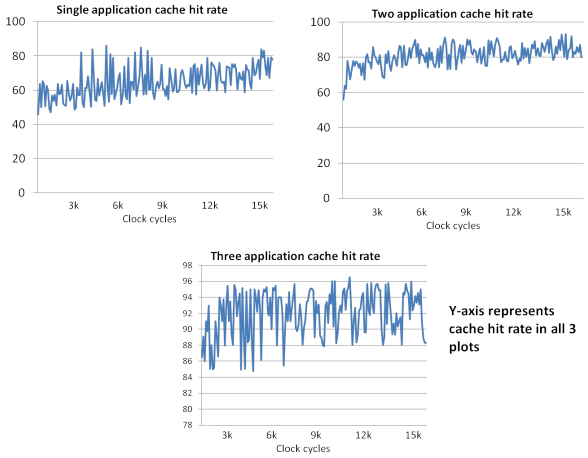

Fig. 9. Variation of Cache Hit Ratio across entire simulation

Figure 8 shows the architectural specifications we have taken for our simulations. We have used SPEC-INT equivalent benchmarks to analyse the effectiveness of the proposed heuristics by scaling up the number of applications in every simulation run. Executing multiple applications simultaneously without space time sharing increases the independent instruction count, thereby improving resource utilization which indirectly contributes to better cache hit rate. From the graphs,

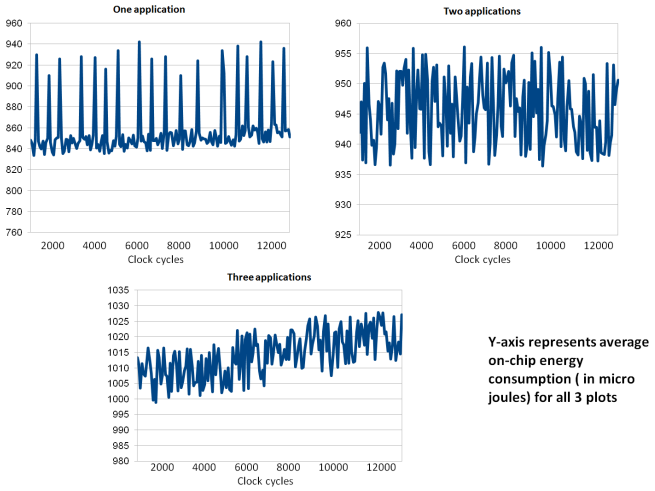

Fig. 10. Single and Multiple power model results for a specific sample of simulation after cold start

it is quite evident that our heuristics has aided in improving the cache hit rate by 8-11% for simultaneous multiple application execution.

For the purpose of energy estimation, an execution sample of about 200 clock cycles much after the cold start period is considered. As discussed earlier, the cache hit ratio for a single application is about 70%. The energy consumption is inherently lesser for this case and shoots up occasionally as can be seen in Figure 10. This shoot up in energy can be attributed to the employment of the replacement policies. In the cases where two and three applications are simultaneously executed together, the overall energy consumption increases due to better cache hit ratio which is a consequence of the proposed mapping and replacement policies. However the rise in energy consumption is not very sharp and is significantly kept constrained by the mapping and replacement policies which reduce the number of misses.

VI. Conclusion

The paper focuses on cache organization taking in to consideration simultaneous multiple application execution without space time sharing. The effectiveness of the replacement strategies related to application as well as data statistics is stressed upon. The cache controller architecture and its interaction is distinctively brought out to accurately predict the execution delay in incorporating the heuristics and the associated cache execution energy model is proposed to accurately capture the energy spent on cache execution paths. Thus, the paper comprehensively captures the cache system paradigm catering to the multiple application execution with accurate execution energy model.

References

[1] N.Venkateswaran et al *Complation Accelerator on Silicon* published in the proceedings of IEEE International Symposium on VLSI 2012

[2] N.Venkateswaran et. al *SCOC IP cores for custom built supercomputing nodes* published in the proceedings of IEEE International Symposium on VLSI 2012

[3] Keshavan Varadarajan et. al *Molecular Caches: A caching structure for dynamic creation of application-specific Heterogeneous cache regions* , In Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO 39), 2006

[4] Kim, Yoongu, et al. *ATLAS: A scalable and high-performance scheduling algorithm for multiple memory controllers* High Performance Computer Architecture (HPCA), 2010 IEEE 16th International Symposium on. IEEE, 2010.

[5] Chang Joo Lee et al *Prefetch-Aware Memory Controllers* , IEEE Transactions on computers, VOL. 60, NO. 10.(October 2011)

[6] Aamer Jaleel et al *CRUISE: cache replacement and utility-aware scheduling* , In Proceedings of the seventeenth international conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS XVII).

[7] Phadke, Sujay, and Satish Narayanasamy *MLP aware heterogeneous memory system*, Design, Automation & Test in Europe Conference & Exhibition (DATE), 2011. IEEE, 2011.

[8] R. Kumar et al*Single-ISA Heterogeneous Multi-Core Architectures for Multithreaded Workload Performance* , SIGARCH Comput. Archit. News 32, 2 (March 2004)

[9] Lee, Chang Joo, et al. *Improving memory bank-level parallelism in the presence of prefetching*, Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture. ACM, 2009.

[10] Ipek, Engin, et al. *Self-optimizing memory controllers: A reinforcement learning approach*, Computer Architecture, 2008. ISCA'08. 35th International Symposium on. IEEE, 2008.

[11] Subramanian, Lavanya, et al. *MISE: Providing Performance Predictability and Improving Fairness in Shared Main Memory Systems*, Proceedings of the 19th International Symposium on High Performance Computer Architecture.

[12] Ausavarungnirun, Rachata, et al. *Staged memory scheduling: Achieving high performance and scalability in heterogeneous systems*, Proceedings of the 39th International Symposium on Computer Architecture. IEEE Press, 2012.

[13] Das, Reetuparna, et al. *Application-aware prioritization mechanisms for on-chip networks*, Microarchitecture, 2009. MICRO-42. 42nd Annual IEEE/ACM International Symposium on. IEEE, 2009.

[14] Zhuravlev, Sergey, Sergey Blagodurov, and Alexandra Fedorova.*Addressing shared resource contention in multicore processors via scheduling*, ACM SIGARCH Computer Architecture News. Vol. 38. No. 1. ACM, 2010.

[15] Warft India MAny Core (WIMAC) simulator *http://www.warftindia.org/joomla*

[16] Venkatraman Govindaraju et al *Dynamically Specialized Datapaths for Energy Efficient Computing*Proceedings of the 17th International Symposium on High-Performance Computer Architecture (HPCA), February 2011

[17] Sanjeev Kumar et al., *Carbon: Architectural Support for Fine-Grained Parallelism on Chip Multiprocessors*In the Proceedings of IEEE/ACM International Symposium on Computer Architecture (ISCA), San Diego, California, June 2007

[18] Xiao Zhang, Sandhya Dwarkadas, and Kai Shen *Towards practical page coloring-based multicore cache management*, In Proceedings of the 4th ACM European conference on Computer systems (EuroSys '09). ACM, New York, NY, USA, 89-102. DOI=10.1145/1519065.1519076

[19] Mohammad Abdullah Al Faruque et al., *ADAM: run-time agent-based distributed application mapping for on-chip communication*, In Proceedings of the 45th annual Design Automation Conference (DAC '08). ACM, New York, NY, USA, 760-765