1. Introduction

The restrictions of the 2-D IC issues together with the memory-wall problem using (TSVs) vertical connections passing through silicon substrate is the (3-D) three-dimensional integration. 3-D integration provides become shorter on the chip overall wires due to extra spacious output/inputs as well as vertical routing paths due to intense TSVs as compared with (2-D) two-dimensional integration. Also 3-D integration uses a last-level on-chip cache to permit stacking multiple small blocks of IC which is unrelated memories against a multiprocessor die. As many as cache memory layers are stacked to improve the performance is estimated since to diminish the off-chip and cache miss memory access, the cache capacity is increased.

On the other hand, one needs to makes use of the high cell density and low leakage power characteristics of the rising NVMs is data is stored even power is off such a memory are PCRAM (Phase-Change RAM) and MRAM (Magnetic RAM). As a result, it is essential to improve energy efficiency and power consumption against a Chip-Multi Processor (CMP) we need to think about the stacking SRAM/NVM hybrid cache surrounded by important performance overhead, the energy efficiency is recovered by NVMs while much to perform a critical write access can be tolerated by means of the lower-latency SRAM and also Pseudo open drain IO standards based energy efficient solar charge sensor design on 20nm FPGA and Modified Corner List (MCL) algorithm is proposed and properties of MCL algorithm is derived on the other hand low power decimal adder based on prediction technique for decreasing the energy/power consumption of parallel decimal multiplication is discussed.

Abstract

Objective: Technology trend are one of the most important to use the bulky on-chip memories such as level-2 (L2) and level-3 (L3) cache. In this paper, Non-Uniform Cache access Architectures (NUCA) has been used in order to decrease the effects of the high access latencies. For the systems which operate at high frequency, the latencies of such caches are conquered by wire delays. Methods: In this architecture, several banks are partitioned and which is connected to the controller through switches and links. This larger cache memory increases overall wire delays across the chip due to very bulky latencies. The majority of the access time will not absorb them self and that occupy in the routing to and from the several banks. The predictable lower-level caches have the best ever access time among the entire the sub arrays, although such uniform cache access don’t make the use of differences in latencies of the entire sub arrays. Findings: Hence there is a necessitate for non-uniform cache architectures where the cache memories is broken down into several banks that can be analyzed at dissimilar latencies and resulting in diminish of energy and power consumption. Applications/Improvements: An experimental result shows the comparison of NUCA and UCA using SRAM/MRAM in terms of power and energy.

Keywords: Chip Multiprocessors, Non Uniform Cache Access, Uniform Access
As regards power methods for CMPs among the mapped cache only, a power reduction method can eliminate needless inconstant cool down operations so 3-D stacked DRAM cache is projected. To reduce the thermal variations and peak temperature, temperature awareness mapping the DRAM banks to cache accesses are proposed. The mapped caches are limited to consistent cache equipment. There are a number of studies which uses the hardware-based techniques which reduces the power density of hybrid cache stacked resting on a CMP by a data maintenance scheme is used to reduce power utilization is deploy to shrink the power leakage is not including the degrading performance.

A smaller block of the integrated circuit uses a high threshold voltage, when dies are placed far away from the heat sink to protect sub-threshold leakage power consumption. Run-time optimum solution to pick up the implementation mode of a piece core from the available frequency and voltage settings by means of predictions a failure based model to exploit the best throughput while maintaining the temperature and power constraints are discussed.

A run-time is to discover voltage/frequency and an energy-minimal cache configuration surrounded by some predetermined configurations which uses a greedy algorithm. A run-time solution is used for single core with a dynamism adopts to the core’s frequency/voltage and the cache configuration. Voltage control has a fixed cache capacity have been used to both MRAM-based L2 cache and processor cores chain of command is assigned to cores.

2. Background

There are a number of integrated circuits manufactured by stacking approach which is individual implement and explore.

2.1 Die-to-Die

Die is the small block of integrated circuits Electronic apparatus which is built on the multiple of small blocks in IC are bonded and aligned. Thinning or removal of some parts and TSV (through silicon vias) creation can be ended subsequent to or previous to bonding. One advantage of small block of IC is that a piece element die can be experienced initial; hence that one bad small block of IC does not demolish a whole stack. And also, a piece of small block of IC used in the 3D can be enclosed space in advance, so to facilitate they may be blended and mixed to diminish the power consumption and performance.

2.2 Die-to-Wafer

Die to wafer means small blocks of IC is transferred to thin slice of semiconductor. Electronic apparatus which is usually built on the two thin slice semiconductor materials. One wafer becomes the smaller blocks; the process of removing the thin slice die are bonded and aligned against the die site of the other wafer. Since in the wafer-on-wafer method, Removal of some parts and TSV (through silicon vias) creation can be ended subsequent to or previous to bonding. Further die can be added to the stacks previous to the dicing.

2.3 Wafer-to-Wafer

Wafer is the thin slice of semiconductor materials. Electronic apparatus is usually constructed on two or more semiconductor materials that they are be diced, arranged and strong connection keen on 3D ICs. Each of semiconductor materials may be removal of some parts previous to or subsequent to bonding. Perpendicular connections are moreover fashioned in the stack later than bonding or else built into the semiconductor materials previous to bonding.

2.4 Dynamic Voltage and Frequency Scaling (DVFS)

Dynamic Voltage and Frequency Scaling (DVFS) which explains the utilize of the two power reduction techniques they are change frequency scaling and change voltage scaling which is second-hand to save the power in the surrounded system. This category of power reduction is dissimilar from the mostly what we commonly imagine regarding like hideaway power states or else standby. We can diminish the power which is consumed by your surrounded system and by lowering the voltage along with/before frequency of the CPU and attached to the outer edge of particular area.

One more advantage of diminishing the power utilization, which produces a smaller amount of heat after generated by your device; this has reimbursement near the automatic devise. It can compose the dissimilarity between needing an inactive or proactive cooling system. If you can able to stay away from a fan, you should improve a amount of objects together with Mean Time
D. Priya, J. Arunarasi and A. Arul Mary

To Failure (MTTF) and cost per device. The good object is the hardware vendor’s covers by adding up these power reduction capabilities to devices so some part of this work is previously done for us, we want to be aware of our devices necessities. Changing the frequency and/or voltage does have any negative aspect; this is a diminution to perform a device. Reduce frequency or voltage and the CPU processes fewer information in a particular time. But nowadays there is a CPU but also attached to the outer part devices. So the entire of these play a component role in the entire surrounded method.

Dynamic voltage and frequency scaling consists of a device that does not do anything or some fundamental communication, functionality and multimedia processing. For a piece of these let us imagine that we maintain the electrical energy of the system regular but we know that CPU frequency can be altered depending on what is the phenomenon happening in the classification:

- fundamental mode with frequency 63 MHz
- Communication mode with frequency 297 MHz
- Multimedia mode with frequency 454 MHz

The reason for selecting those 63MHz, 297MHz and 454MHz frequencies is supported by the i.MX 28 that they are definite operating points. To reduce the power of my method this uses a lowest one to locate as the in use frequency. It can be shown in Figure 1 that, the operating frequency #1 at 63 MHz requires less than partially of the running voltage at operating frequency # 3 at 454 MHZ than current. And operating frequency # 2 at 297MHz requires less than operating frequency #3.

**Figure 1.** Operating Point Vs Soc Current Consumption..

The following application runs on the devices, the drivers and the OS:

- Op#0(200MHz)–fundamental Mode or Communication Mode
- Op#1(230 MHz) – fundamental Mode Only
- Op#2(370MHz)– Communication Mode Only
- Op#3(470MHz) – Multimedia Mode Only

### 3. Proposed Method

#### 3.1 Introduction

The designs of large caches memories changes when the wire delays increase substantive. Conventional Uniform cache access architecture accepts at each stage in the cache hierarchy have a particular and identical time to hit the bulk of on-chip caches memories that are gathered together of all the line’s substantial which is located surrounded by the cache will make increasing the on-chip processors communication delays. And also, cache access times will develop into a particular isolated (discrete) latency rather than a constant (continuous) of latencies. This non-uniformity occupy closer to the processor that provides a more rapidly to access the cache lines which is located in the portions of the cache memory. In this paper, substantial designs for these (NUCAs) Non-Uniform Cache Architectures are projected. Therefore, substantial designs surrounded by the identical level of the cache were enlarged with the logical policies which permit the most essential data to move around toward the processor.

The leading module of latency for large caches memories will maintain to grow the wire delay. By exploit the difference in the right to use the crossways widely-spaced sub arrays, NUCA allows slower way to access the far sub arrays whereas recovering faster way to access the slow sub arrays.

#### 3.2 Level 1 and Level 2

The expectations development of Level 2 and Level 3 caches is to obtain superior because of the reasons are:

- Bandwidth difficulty have been enlarged on this package
- Bigger Program sizes.
- Minor technologies sanction more bits per mm³.
- Spatial locality requires multi-threading for large caches.

L1 is "level-1" cache memory is usually built inside the microprocessor chip or memory chip for example, the Intel microprocessor comes through a 32000 bytes of L1(level-1).

L2 is “level-2” cache memory is built on a separate chip that can be accessed more rapidly than the major memory.
These larger cache memories increases overall wire delays across the chip because will have very bulky latencies. Most of the access time will not absorb them self and that occupy in the routing to and from the several banks. The predictable lower-level caches have the best ever access time among the entire the sub arrays, although such uniform cache access don't make the use of differences in latencies of the entire sub arrays.

### 3.3 Static Non-Uniform Cache Access Architecture

Most of the performance is vanished via using bad cases of uniform cache access memories into the wire-delay subjugated cache. If a piece of bank in the cache memories be able to access by dissimilar speeds, comparative to the distance between the bank from the cache memory organizer, performance would be enhanced. Indeed, on the figure below, the worst-case 41 cycles delay, compared to the closest banks can be accessed in 17 clock cycles.

Data are stacked into several banks, the bank is determined with the lower-order bit of the index. These (S-NUCA) static-non-uniform cache access architectures have two recompense when compared with the UCA organization. (1) Cache accesses to dissimilar banks results to reduce contention which may progress in parallel; (2) Cache access to the several banks results to lower latency which is closer connected to cache. Therefore, cache memories is known as Static-NUCA, because the data are stacked into several banks are stationary, and banks will not have identical access times.

![Figure 2. S-NUCA-1 cache design.](image1)

As shown in the Figure 2, a piece of address bank in the S-NUCA-1 cache design consists of data channels and private address. Since each address banks have several private addresses and a piece bank can be access at its upper limit speed individually. While smaller address bank which may give a greater reliability as well as much further concurrency of the cache design, the repeated per-bank channels insert more region to overcome the number of several banks array that constrains.

### 4. Results and Discussion

The experimental results show the Static- Non Uniform cache access architecture using SRAM/MRAM in given below. The Figure 3 (a) and (b) shows the results in terms of energy consumption using static NUCA architecture compared with UCA architecture and in Figure 4 shows the results in terms of power consumption using static NUCA architecture compared with UCA architecture. Table 1 shows the difference between two architecture and parameters used in UCA and NUCA.

![Figure 3. (a) Output of Static NUCA.](image2)

![Figure 3. (b) Output of Static NUCA.](image3)
As shown in Table 1, S-NUCA yields up to 32 numbers of banks when compared with UCA. UCA architecture uses lower bank size, bank rows and bank columns compared with S-NUCA.

**Table 1. Configuration Parameters**

| PARAMETERS         | UCA   | S-NUCA |
|--------------------|-------|--------|
| Size               | 8MB   | 8MB    |
| Line size          | 64MB  | 64MB   |
| No. of bank        | 1     | 32     |
| No. of sub bank    | 2     | 1      |
| No. of bank row    | 1     | 8      |
| No. of bank column | 1     | 4      |
| Bank size          | 8MB   | 256MB  |
| Bank associatively | 4-way | 4-way  |
| Bank latency       | 18    | 5      |
| Link latency       | -     | 2      |
| Link size          | -     | 2*128  |
| Energy bank access | 1928.pJ | 287.8pJ |
| Power bank static  | 32205.9mW | 1007.0mW |

Figure 4 shows the comparison of NUCA and UCA in terms of line size, size, Number of bank, Number of sub-bank, Number of row and Number of column.

Figure 5 (a) shows in SRAM cell memory address, data, chip select, write enable and output enable is taken as input and 8-bit data out is taken to be an output. When write enable is higher than chip select, a read operation occurs. In Figure 5(b) same as read operation, but when write enable is lower than chip select, a write operation occurs.

Figure 6. (a) MRAM cell memory using read operation.
In Figure 6(a) shows read cache enable, read address, read data, write cache enable, write address, write data is taken as input and output as read output and write output. If condition is true read output (rq) is stored in memory and whereas in Figure 6(b) if condition is false write output (wq) is stored in memory. The memory capacity runs in “i” variable.

Figure 6. (b) MRAM cell memory using write operation.

5. Conclusion

Implementing Non-volatile memory such as SRAM and MRAM in the Static Non-uniform cache access architecture rather than in uniform cache access, runs a multiple programs and a piece of which consists of multiple threads to lightly make use of the various compute assets and which will diminish the cache access latency and. Simulation results using MODELSIM software shows that Static NUCA architecture achieves energy consumption per access, power consumption per access and reduces the number of sub-banks when compared with existing SRAM/MRAM memory cell.

6. Acknowledgement

The authors gratefully acknowledge the contribution of Govt. of India for Financial Assistance, DST-FIST F.NO:SR/FST/College-189/2013

7. References

1. Loh GH. 3D-stacked memory architectures for multi-core processors. IEEE Proceedings 35th Annual Conference on Computer Architecture. 2008. p. 453–64.
2. Park H, Yoo S, Lee S. Power management of hybrid DRAM/PRAM-based main memory. Design Automation Conference (DAC), IEEE. 2011. p. 59–64.
3. Xu W, Sun H, Wang X, Chen Y, Zhang T. Design of last level on-chip cache using spin-torque transfer RAM (STT RAM). IEEE Transactions on Very Large Scale Integration (VLSI) Systems. 2011; 19(3):483–93.
4. Sun G, Dong X, Xie Y, Li J, Chen Y. A novel architecture of the 3D stacked MRAM L2 cache for CMPs. 15th IEEE International Symposium on High Performance Computer Architecture, HPCA, Raleigh, NC. 2009; 239–49.
5. Kalia K, Pandey B, Nanda K, Malhotra S, Kaur A, Hussain DMA. Pseudo open drain IO standards based energy efficient solar charge sensor design on 20nm FPGA. 11th IEEE International Conference on Power Electronics and Drive Systems (PEDS), Sydney, NSW. 2015. p. 367–72.
6. Sivaranjani P, Senthilkumar A. 3D VLSI Non-Slicing Floor planning using Modified Corner List Representation. Indian Journal of Science and Technology. 2014; 7(3):276–81.
7. Ghoshand H, Lee HS. An enhanced memory controller design for reducing energy in conventional and 3D die-stacked DRAMs. 40th Annual IEEE/ACM International Symposium on Microarchitecture. 2007; 134–45.
8. Mengand J, Coskun AK. Analysis and runtime management of 3D systems with stacked DRAM for boosting energy efficiency. Design, Automation and Test in Europe Conference and Exhibition (DATE). 2012. p. 611–16.
9. Wang Y, Zhang C, Yu H, Zhang W. Design of low power 3D hybrid memory by non-volatile CBRAM-crossbar with block-level data retention. Proceedings of the 2012 ACM/IEEE International symposium on Low Power Electronics and Design, 2012; 197–202.
10. Chenetal Y T. An energy-efficient last level cache design. Proceedings of the Design, Automation and Test in Europe Conference and Exhibition (DATE). 2012. p. 45–50.
11. Chakraborty K, Roy S. Stack aware threshold voltage assignment in 3-D multi core designs. IEEE Transactions on Very Large Scale Integration (VLSI) Systems. 2011; 20(3):512–22.
12. Ac, Givargis T. Dynamic voltage and cache reconfiguration for low power. Proceedings of the Design, Automation and Test in Europe Conference and Exhibition (DATE). 2004. p. 1376–77.
13. Wang W, Mishra P. Leakage aware energy minimization using dynamic voltage scaling and cache reconfiguration in real-time systems. IEEE Transactions on Very Large Scale Integration (VLSI) Systems. 2010; 357–62.
14. Zhao J, Dong X, Xie Y. An energy efficient 3D CMP design with fine grained voltage scaling. Design, Automation and Test in Europe Conference and Exhibition (DATE), Grenoble. 2011. p. 1–4.