International Journal of Electronics Communication and Computer Engineering
Volume 4, Issue 2, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

Enhancement Scheme of 1T DRAM for Low Voltage Operation in Video Processing

Dr. Mamta Sood, Prof. Ankur Beohar, Kasif khan

Abstract **- BJT-based 1T-DRAM cell has widely used mainly because of its full compatibility with the standard SOI(silicon insulator) processing and its potentially excellent performance Such as high sensing margin and retention characteristics [2] This Paper is mainly based on improving the 1T DRAM for Low voltage operation in video processing by reduction of DRAM Data access energy consumption using either selective epitaxial process or by an ion implementation technique. In this paper we presents a strategy for mapping pixels into the memory for video applications such as MPEG processing, thereby minimizing the transfer overhead between memory and the processing .DRAM architecture is better in video processing for unbalanced images. In this paper we have define techniques to reduce 1TDRAM energy consumption up to 96%.**

Keywords **- SRAM, Image Data Access, 1TDRAM Architecture, Energy Consumption in Video Processing.**

I. INTRODUCTION

The architecture of the present video processing units in consumer systems is usually based on various forms of processor hardware, communicating with an off-chip SDRAM memory [1]. Examples of these systems are currently available MPEG encoders and decoders, and high-end television systems.AS video processing becomes increasingly indispensable in mobile devices, its energy-efficient implementation is of great practical interest [1]. For typical video processing applications, the image frames are stored in a stand-alone DRAM, which are accessed and processed by a processing logic unit in a SoC[1].For double-data-rate SDRAM (DDR SDRAM), the proposed mapping strategy reduces the bandwidth in the system by even 50%. This substantial performance improvement can readily be used for extending the quality or the functionality of the system. In Tablet or Mobile devices or any other devices those supported Android application where energy consumption ratio is more. For typical video processing applications, the image frames are stored in a stand-alone DRAM, which are accessed and processed by a processing logic unit in an SoC. Although the continuous technology scaling helps to reduce energy consumption, DRAM image data access may not benefit as much as the video processing logic operation. This is because, as Tablet devices need to support increasingly diverse and more sophisticated functions, the storage capacity of DRAM has to accordingly increase. As a result, DRAM die size may not reduce or even increase in spite of technology scaling down. As we will elaborate later, DRAM access energy consumption is largely dominated by the routing interconnect energy consumption that is proportional to DRAM die size. Therefore, it is reasonable to expect that DRAM access energy consumption will play a more and more important role in determining the overall video processing energy consumption. Taking video decoding as an example, recent work has shown that, an H.264 decoder at 90 nm node only consumes 0.36 pJ per pixel, while the corresponding DRAM access energy consumption is 1.11 pJ per pixel. Nevertheless, most prior work on video decoding mainly focused on reducing the decoder logic power consumption.

This work is interested in reducing DRAM image data access energy consumption in video processing. Although reducing DRAM energy consumption has been widely most prior work focused on using DRAM in general purpose computing systems. Very intuitively, once we confine ourselves in certain specific application domain, it is possible to exploit its own unique characteristics to develop domain-specific techniques for reducing DRAM energy consumption. They can complement those general-purpose low-power design techniques to further push DRAM energy efficiency envelope for the specific application domain. To evaluate the effectiveness of the proposed data manipulation techniques and the heterogeneous DRAM architecture, we carried out extensive DRAM modeling and power consumption estimations using representative video sequences in the context of video decoding. The popular CACTI tool is used for DRAM modeling and estimating energy consumption induced by DRAM on-chip routing including both wire self-transition and coupling activities [1]. Simulation results show that the three data manipulation techniques together can reduce DRAM access power consumption by up to 40%. When incorporating the video decoding logic power consumption, we estimate that these data manipulation techniques can reduce the overall video decoding system power consumption by up to 30.4%.We also evaluated the effectiveness of the manipulation techniques when image frame recompression is being used, and the results show as high as 80% of power reduction. Regarding the heterogeneous 1TDRAM architecture, we considered different heterogeneity configuration scenarios for a 2 GB DRAM, and simulation results show that DRAM energy consumption can be reduced by up to 96% when using the proposed data manipulation techniques in a heterogeneous 1TDRAM architecture.

II. DRAM ENERGY CONSUMPTION

The computing landscape is undergoing a major change, primarily enabled by ubiquitous wireless networks and the rapid increase in the usage of mobile devices which access the web-based information infrastructure. It is expected that most CPU-intensive computing may either happen in servers housed in large datacenters, e.g., cloud computing

and other web services, or in many-core high-performance computing (HPC) platforms in scientific labs. In both situations, it is expected that the memory system will be problematic in terms of performance, reliability, and power consumption. The memory wall is not new: long DRAM memory latencies have always been a problem. Given that little can be done about the latency problem, DRAM vendors have chosen to optimize their designs for improved bandwidth, increased density, and minimum cost-per-bit. With these objectives in mind, a few DRAM architectures, standards, and interfaces were instituted in the 1990s and have persisted since then. However, the objectives in datacenter servers and HPC platforms of the future will be very different than those that are reasonable for personal computers, such as desktop machines. As a result, traditional DRAM architectures are highly inefficient from a future system perspective, and are in need of a major revamp. Consider the following technological trends that place very different demands on future DRAM architectures:

Energy: While energy was never a first-order design constraint in prior DRAM systems, it has certainly emergedas the primary constraint today, especially in datacenters. Energy efficiency in datacenters has already been highlighted as a national priority. Many studies attribute 25-40% of total datacenter power to the DRAM system. Modern DRAM architectures are ill-suited for energy-efficient operation because they are designed to fetch much more data than required. This over fetch wastes dynamic energy. Today's DRAMs employ coarse-grained power-down tactics to reduce area and cost, but finer grained approaches can further reduce background energy.

Reduced locality: Single-core workloads typically exhibit high locality. Consequently, current DRAMs fetch many kilobytes of data on every access and keep them in open row buffers so that subsequent requests to neighboring data elements can be serviced quickly. The high degree of multi-threading in future multi-cores] implies that memory requests from multiple access streams get multiplexed at the memory controller, thus destroying a large fraction of the available locality. The severity of this problem will increase with increased core and memory controller counts that are expected for future microprocessor chips. This trend is exacerbated by the increased use of aggregated memory pools ("memory blades" that are comprised of many commodity DIMMs) that serve several CPU sockets in an effort to increase

Queuing Delays: For several years, queuing delays at the memory controller were relatively small because a single core typically had relatively few pending memory operations and DRAM systems were able to steeply increase peak memory bandwidth every year [20]. In the future, the number of pins per chip is expected to grow very slowly. The 2007 ITRS Road-map expects a 1.47x increase in the number of pins over an 8-year time-frame – over the same period, Moore's Law dictates at least a 16x increase in the number of cores. This implies that requests from many cores will be competing to utilize the limited pin bandwidth. Several studies have already highlighted the emergence of queuing delay as a major bottleneck. A DRAM architecture that is geared towards higher parallelism will likely is able to de-queue requests faster and better utilize the available limited data bandwidth.

Efficient Reliability: Recent studies have highlighted the need for DRAM architectures that are resilient to single faults or even failure within an entire DRAM chip, especially in datacenter platforms. Because these fault tolerant solutions are built upon commodity DRAM chips, they incur very high overheads in terms of energy and cost. New DRAM architectures can provide much more efficient reliability if fault-tolerant features are integrated into the DRAM chip micro architecture at design time.

Lower relevance of DRAM chip area: DRAM vendors have long optimized the cost-per-bit metric. However, given that datacenters consume several billion kilowatt hours of energy every year it has been shown that the 3-year operating energy costs of today's datacenters equal the capital acquisition costs. Therefore, it may now be acceptable to incur a slightly higher cost-per-bit when purchasing DRAM as long as it leads to significantly lower energy footprints during operation.The design of DRAM devices specifically addressing these trends has, to the best of our knowledge, not been previously studied and is now more compelling than ever. We attempt to fundamentally rethink DRAM micro architecture and organization to achieve highly reliable, high performance operation with extremely low energy footprints, all within acceptable area bounds. In this work, we propose two independent designs, both attempting to activate the minimum circuitry required to read a single cache line. We make the following three significant contributions:

We introduce and evaluate Posted RAS in combination with a Selective Bit line Activation (SBA) scheme. This entails a relatively simple change to DRAM micro architecture, with only a minor change to the DRAM interface, to provide significant dynamic energy savings.

We propose and evaluate a reorganization of DRAM chips and their interface, so that cache lines can be read via a Single Sub-array Access (SSA) in a single DRAM chip. This approach trades off higher data transfer times for greater (dynamic and background) energy savings.

In order to provide chip kill-level reliability even though we are reading a cache line out of a single DRAM device, we propose adding a checksum to each cache line in the SSA DRAM to provide error detection. We then evaluate the use of RAID techniques to reconstruct cache lines in the event of a chip failure.

While this study focuses on DRAM as an evaluation vehicle, the proposed architectures will likely apply just as well to other emerging storage technologies, such as phase change memory (PCM) and spin torque transfer RAM (STT-RAM)

International Journal of Electronics Communication and Computer Engineering
Volume 4, Issue 2, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

III. DRAM BASICS AND BASELINE ORGANIZATION

We first describe the typical modern DRAM architecture for most of the paper, our discussion will focus on the dominant DRAM architecture today: JEDEC-style DDRx SDRAM, an example is shown in Figure 1.

Modern processors often integrate memory controllers on the processor die. Each memory controller is connected to one or two dedicated off-chip memory channels. For JEDEC standard DRAM, the channel typically has a 64-bit data bus, a 17-bit row/column address bus,and an 8-bit command bus [38]. Multiple dual in-line memory modules (DIMMs) can be accessed via a single memory channel and memory controller. Each DIMM typically comprises multiple ranks, each rank consisting of a set of DRAM chips. We will call this a rank-set. Exactly one rank-set is activated on every memory operation and this is the smallest number of chips that need to be activated to complete a read or write operation. Delays on the order of a few cycles are introduced when the memory controller switches between ranks to support electrical bus termination requirements. The proposed DRAM architecture is entirely focused on the DRAM chips, and has neither a positive or negative effect on rank issues. Figure 1 shows an example DIMM with 16 total DRAM chips forming two rank-sets .Server Power Consumption:

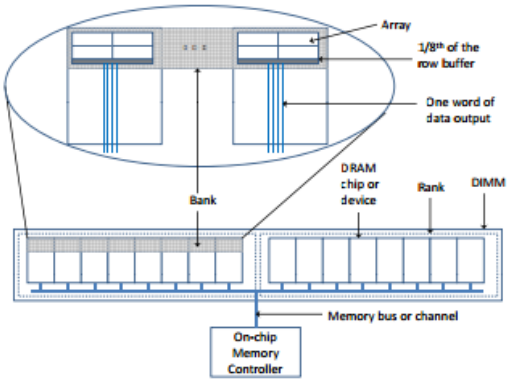

Fig.1. An example DDRx SDRAM architecture with 1 DIMM, 2 ranks, and 8 x4 DRAM chips per rank.

DRAM energy consumption contributes significantly to the total power usage of computing systems

Understanding DRAM energy consumption will help design more efficient systems

DRAM Power Modeling

- Levels of DRAM power modeling
- Data sheet power values
- DRAM must be available on market
- Typically based on hardware measurements
- Full chip schematics and simulator (Spice, Nanosim etc.)
- Only at DRAM vendor
- Requires completed circuit design
- Model with DRAM architecture and technology as part of the program code
- Not flexible enough for quick thought experiments
- A flexible DRAM model which can cover a large variety of architectures and technologies is needed to experiment with power saving ideas for DRAMs

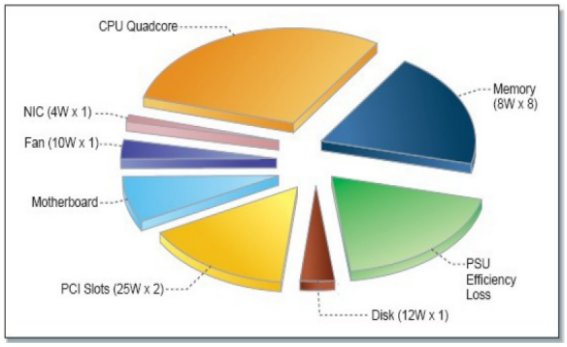

Fig.2 DRAM Energy Consumption System

IV. SELECTIVE BITLINE ACTIVATION (SBA)

In an effort to mitigate the over fetch problem with minimal disruption to existing designs and standards, we propose the following two simple modifications: (i) we activate a much smaller segment of the word line and (ii) we activate only those bit lines corresponding to the requested cache line. Note that we will still need a wire spanning the array to identify the exact segment of wordline that needs to be activated but this is very lightly loaded and therefore has low delay and energy. Thus, we are not changing the way data gets laid out across DRAM chip arrays, but every access only brings down the relevant cache line into the row buffer. As a result, the notion of an open-page policy is now meaningless. After every access, the cache line is immediately written back. Most of the performance difference from this innovation is because of the shift to a close-page policy for workloads with little locality, this can actually result in performance improvements as the page precharge after write-back is taken off the critical path of the subsequent row buffer miss. Next, we discuss the micro architectural modifications in more detail.

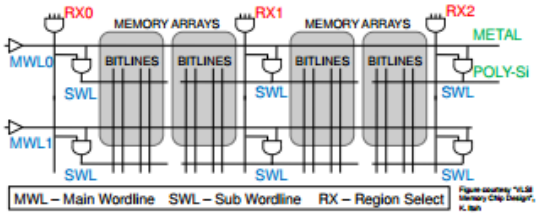

Fig.4. Hierarchical wordline with region select.

Memory systems have traditionally multiplexed RAS and CAS commands on the same I/O lines due to pin count limitations. This situation is unlikely to change due to technological limitations and is a hard constraint for DRAM optimization. In a traditional design, once the RAS arrives, enough information is available to activate the appropriate wordline within the array. The cells in that row place their data on the corresponding bit lines. Once the row's data is latched into the row buffer, the CAS signal is used to return some fraction of the many bits read

from that array. In our proposed design, instead of letting the RAS immediately activate the entire row and all the bit lines, we wait until the CAS has arrived to begin the array access. The CAS bits identify the subset of the row that needs to be activated and the wordline is only driven in that section. Correspondingly, only those bit lines place data in the row buffer, saving the activation energy of the remaining bits. Therefore, we need the RAS and the CAS before starting the array access. Since the RAS arrives early, it must be stored in a register until the CAS arrives. We refer to this process as Posted-RAS 1 because we are now waiting for the CAS to begin the array access, some additional cycles (on the order of 10 CPU cycles) are added to the DRAM latency. We expect this impact to be relatively minor because of the hundreds of cycles already incurred on every DRAM access. Note again that this change is compatible with existing JEDEC standards: the memory controller issues the same set of commands; we simply save the RAS in a register until the CAS arrives before beginning the array access.

V. CONCLUSION

We proposed a BJT-Based 1TDRAM to achieve low voltage operation. This paper concerns how to reduce DRAM image data access energy consumption in video processing applications. Its objective is two-fold: 1) video processing is one of the most energy consumption functions compact Tablets as well as mobile devices, and 2) DRAM image data access energy consumption increasingly outweighs video processing logic energy consumption. The objective of this paper is to develop domain-specific techniques to reduce 1TDRAM energy consumption up to 96% by appropriately exploiting image data access characteristics, in particular the image data spatial and temporal correlation and unbalanced image access in most video processing. We propose to use three simple yet effective data manipulation techniques to exploit the spatial/temporal correlation to reduce 1TDRAM data access energy consumption.

REFERENCES

[1] Yiran Li, and Tong Zhang, senior member *IEEE* "Reducing DRAM Image Access Energy Consumption in video processing," *IEEE Transaction on Multimedia,Vol.14, No. 2, April 2012*, pp. 303–313.

[2] Kyung-suk shim, In-Young June Park, Senior member *IEEE* "A BJT-Based Heterostructure 1TDRAM for Low-Voltage operation," in *IEEE Electron Device Letter*, Vol.33, No.1, Jan. 2012, pp. 14–16.

[3] Shuhei Tanakamaru,Student Member,*IEEE*, and Ken Takeuchi,Member "A 0.5 V Operation V_{TH} Low compensated DRAM Word-Line Booster Circuit for Ultra-Low power VLSI System ," *IEEE J. Solid-State Circuits*, vol.46, no. 10, pp. 2406–2415, oct. 2011.

[4] Dong-su Lee, Young-Hyun Jun, and Bai-Sun Kong "Simultaneous Reverse Body and Negative Word-Line Biasing Control Scheme For Leakage Reduction of DRAM," *IEEE J. Solid-State Circuits*, vol. 46, no. 10, pp.2396–2405, Oct. 2011.

[5] M. Ghosh and H. Lee, "Smart refresh: An enhanced memory controller design for reducing energy in conventional and 3D die-stacked DRAMs," in *Proc. 40th Annu. IEEE/ACM Int. Symp. Microarchitecture*, 2007, pp. 134–145.

[6] H. Zheng, J. Lin, Z. Zhang, E. Gorbatov, H.David, and Z. Zhu, "Minirank: Adaptive DRAM architecture for improving memory power efficiency," in *Proc. 41st Annu. IEEE/ACM Int. Symp. Microarchitecture*, 2008, pp. 210–221.

[7] A. Udipi, N. Muralimanohar, N. Chatterjee, R. Balasubramonian, A. Davis, and N. Jouppi, "Rethinking DRAM design and organization for energy-constrained multi-cores," in *Proc. 37th Annu. Int. Symp. Computer Architecture*, 2010, pp. 175–186.