Exploring the Use of Volatile STT-RAM for Energy Efficient Video Processing

Hengyu Zhao[1], Hongbin Sun[1], Qiang Yang[2], Tai Min[1], Nanning Zheng[1]
[1]Xi'an Jiaotong University, Xi'an, Shaanxi, P.R. China 710049
[2]Changhong Electric Co., Ltd, Mianyang, Sichuan, P.R. China 621000

Abstract—This paper explores the efficient use of STT-RAM in video processing system by choosing display processing as a case study. Through architectural analysis, we demonstrate that volatile STT-RAM rather than non-volatile STT-RAM is more appropriate for video processing system, as video data is mostly processed in streaming-style and relaxing the retention time can significantly reduce the write latency and energy of STT-RAM. Moreover, the error resilience of video display processing system is also evaluated to address the potential bit error rate increase when we relax the retention time of STT-RAM cells. Simulation results demonstrate that, the use of volatile STT-RAM and extra design techniques can significantly reduce the overall memory energy consumption to 38.24% with respect to that of baseline SRAM architecture, and video display processing system can tolerate up to 1×10^{-5} bit error rate without any visible image quality degradation.

Keywords— Volatile STT-RAM, selective buffer write, redundant bit write removal, video display processing

I. Introduction

Due to its great scalability, fast read access, low leakage power and non-volatility, spin-transfer torque RAM (STT-RAM) is recognized as a promising memory technology for on-chip embedded memory and its cache memory architecture design has been extensively explored [1–3]. Despite of its attractive features, STT-RAM also suffers from its slow and high energy write operation. Therefore, previous STT-RAM based cache designs mainly explore the SRAM&STT-RAM hybrid architecture to efficiently compensate its high write overhead. In the meanwhile, several works also try to reduce high write energy and long write latency by relaxing non-volatility of STT-RAM cells, and the performance and energy efficiency improvement by applying volatile STT-RAM to cache memory is also evaluated [4–6]. In addition, since part of cache blocks may have a very long lifetime and low retention volatile STT-RAM may lose data, the DRAM-style refresh policy is necessary for its cache memory design. However, to achieve optimal refresh scheme tends to complicate the architecture design, and it may also need the assistance from compiler [7].

Although STT-RAM-based cache design has attracted great attention, the use of STT-RAM to memory intensive video processing is also very promising, as video processing circuits always consume large capacity on-chip memories. Two previous works [8, 9] have explored the energy efficiency of the use of STT-RAM to video coding system. In particular, they use non-volatile STT-RAM as the mainstream cache design does, and complicated hybrid memory architectures are exploited to compensate the slow and energy consuming write. However, we claim that, volatile STT-RAM rather than non-volatile STT-RAM is more appropriate to be exploited in video processing system, as video data is mostly processed in streaming-style and does not need to be stored in a long period. More important, as lowering retention time is able to substantially reduce the write latency and energy of STT-RAM, the associated architecture design complexity can be alleviated and energy efficiency can be significantly improved.

This paper is interested in the efficient use of volatile STT-RAM to video processing system. In particular, we choose video display processing system as a case study to demonstrate its feasibility and efficiency. We should note that, this paper does not aim to propose any new architecture or technique. Instead, we try to demonstrate that, by analyzing a typical memory architecture requirement of video display processing module, the use of volatile STT-RAM can not only easily fit into the conventional memory architecture without any design effort, but also significantly outperform non-volatile STT-RAM in terms of energy efficiency. In addition, we present two simple yet effective techniques, i.e. selective buffer write and redundant bit write removal, to further reduce the write intensity and thus energy of volatile STT-RAM memory architecture. Simulation results demonstrate that the use of volatile STT-RAM and extra design techniques can significantly reduce the overall memory energy consumption to 38.24% with respect to that of baseline SRAM architecture. As lowering the retention time of STT-RAM cells tends to increase the bit error rate in STT-RAM memory, we also evaluate the error resilience of video display processing system. Both the objective and subjective testing results show that video display processing system can tolerate up to 1×10^{-5} bit error rate without any visible image quality degradation. Simulation results clearly demonstrate that the use of volatile STT-RAM is a feasible and energy efficient solution to memory design in video processing system.

II. Background and Motivation

A. Video Display Processing

The video display processor, together with multi-format video decoder and graphics processing unit (GPU), are three key modules for video/graphic processing in modern TV or mobile SoCs [10, 11]. In general, video display processing is dedicated to correcting artifact, modifying original video format and enhancing the image quality. Hence, it has a direct impact on the perceived image quality. Video display processing is not a single algorithm, which usually integrates a group of video processing modules. Fig. 1 illustrates a general video display processing chain, which usually integrates display processing modules from three main categories, i.e., corrective processing, format conversion and enhancement [12]. Corrective processing takes care of reducing noises and artifacts; format conversion includes

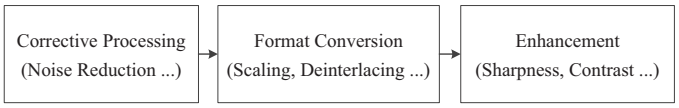

Fig. 1. Video display processing chain.

scaling, deinterlacing and scan rate conversion; enhancement aims at color, contrast, sharpness, and resolution enhancement.

Although video decoding usually attracts much more attention, video display processing is also a critical and memory intensive design module in multimedia SoC. Video display processing algorithm is usually designed as window-based 2D local filtering, and SRAM line buffers are necessary to construct the window. As video display chain usually integrates a variety of unique display processing modules for high image quality and each processing module has to consume several line buffers, the overall on-chip memory capacity can be considerably large. For full HD display processor, on-chip SRAM memory can take several hundred KBs [13]. As video resolution standard has been steadily increased to 4K*2K, the capacity of SRAM will continue to increase and leakage power of the SRAM line buffers will become a critical design bottleneck. Moreover, all of these SRAM line buffers have to constantly work in active mode as long as display panel is turned on. Although idle-mode low leakage techniques have been substantially improved recently [14–16], leakage power in active mode remains very challenging for SRAM design [17].

STT-RAM has the great potential to address the high leakage power issue of large capacity SRAM. However, non-volatile STT-RAM suffers from its slow and energy consuming write operation. Therefore, a variety of SRAM&STT-RAM hybrid architecture designs [2,3,8,9] have been explored to address write overhead. The hybrid architecture may be suitable for cache memory, but is actually inefficient for video processing. Since video memory is usually designed as small memory blocks distributed into a variety of processing modules, hybrid architecture tends to complicate the circuit design and increase implementation cost. Moreover, video data is usually processed in streaming-style and its temporal and spatial locality are relatively very low, hence hybrid architecture may be not capable to address high write energy issue especially when write intensity is high. In the meanwhile, the streaming-style of video processing can significantly relax the retention time requirement, resulting in fast and low energy write of STT-RAM. These observations motivate this work to explore the use of volatile STT-RAM in video display processing for energy efficiency improvement.

B. Volatile STT-RAM Basics

Fig. 2(a) shows the typical 1T1MTJ STT-RAM cell structure. Each cell contains one MTJ as the storage element and one nMOS transistor as the access control device. As the basic storage element in STT-RAM, each MTJ has two ferromagnetic layers separated by one oxide barrier layer. The resistance of each MTJ depends on the relative magnetization directions of the two ferromagnetic layers, i.e., when the magnetization is parallel (or anti-parallel), MTJ is in a low (or high) resistance state, as illustrated in Fig. 2(b). In STT-RAM, parallel and anti-

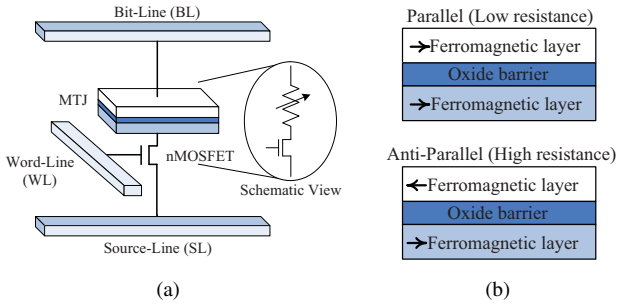

Fig. 2. (a) The structure of a 1T1MTJ STT-RAM cell and (b) the resistance state of MTJ.

parallel magnetization are realized by steering a write current directly through MTJs along opposite directions. For non-volatile STT-RAM, write to STT-RAM cells is much longer than SRAM and also consumes much more energy dissipation, which consequently limit the use of STT-RAM as the embedded memory in microprocessors or memory intensive signal processing system to some extent.

Although STT-RAM is usually exploited as non-volatile memory, it can also be designed as volatile memory by shrinking the planar area of magnetic tunnel junction (MTJ). Moreover, lowering the retention time of MTJ can consequently reduce the write latency and energy. This interesting feature has been exploited by several previous works [4–6] in cache memory design. The retention time of a MTJ is determined by the thermal stability factor Δ, which can be calculated from Equation 1 using the effective activation volume (V), the in-plane anisotropy field (H_k), the saturation magnetization (M_s) and the absolute temperature in kelvin (T). The retention time also can be calculated from Equation 1, where f_0 is chosen to be 1 GHz according to [18]. In this paper, we choose to reduce the effective activation volume (V) and the associated retention time by shrinking the planar area of MTJ as discussed in [4]. Although the retention time can also be controlled by saturation magnetization (M_s), this tends to complicate the fabrication process.

$$T_{retention} = \frac{1}{f_0} \times e^{\Delta}, \ \ where \ \Delta \propto \frac{V \cdot H_k \cdot M_s}{T} \tag{1}$$

At the determined retention time node, there is a relationship between write current and switching time that the required switching current rises exponentially as the MTJ switching time is reduced. The required switching current density J_C of a MTJ operating in different three working regions, i.e. thermal activation J_C^{The}, the dynamic reversal J_C^{Dyn}, and the precessional switching J_C^{Pre}, can be approximated as Equation 2 [19].

$$\begin{cases} J_C^{The}(T_{sw}) = J_{C0}(1 - \frac{1}{\Delta} ln(\frac{T_{sw}}{\tau_0})) & (T_{sw} > 10ns) \\ J_C^{Dyn}(T_{sw}) = \frac{J_C^{The}(T_{sw}) + J_C^{Pre}(T_{sw}) e^{(-k(T_{sw} - T_{PIV}))}}{1 + e^{(-k(T_{sw} - T_{PIV}))}} & (3ns < T_{sw} < 10ns) \\ J_C^{Pre}(T_{sw}) = J_{C0} + \frac{C ln(\frac{\pi}{2\theta})}{T_{sw}} & (T_{sw} < 3ns) \end{cases} \tag{2}$$

Where *k*, *C*, and T_{PIV} are the fitting parameters, T_{sw} is the switching time of MTJ resistance. The critical current density J_{c0} is calculated as following, according to [18].

$$J_{C0} = \frac{2e\alpha M_s t_F (H_k \pm H_{ext} + 2\pi M_s)}{\hbar \eta} \tag{3}$$

Where e is the electron charge, α is the damping constant, τ_0 is the relaxation time, t_F is the free layer thickness, $\hbar$ is the reduced Planck's constant, H_{ext} is the external field and η is the spin transfer efficiency.

Although volatile STT-RAM can significantly reduce write latency and energy consumption, the decrease of thermal stability factor Δ tends to increase read disturb error rate of STT-RAM [20, 21] in the meanwhile. The influence of reading operation and thermal stability factor Δ on the read disturb rate can be estimated by the following equation.

$$P_{fail} = 1 - exp\{-\frac{\tau_r}{\tau_0} exp[-\Delta(1 - \frac{I_r}{I_{C0}})]\} \quad (4)$$

Where P_{fail} is the reading disturb rate, t_r is the read pulse width, I_r is the read current, τ_0 is the attempt period which is equal to $\frac{1}{f_0}$.

III. Energy Efficiency Exploration

A. Architectural Requirement Analysis

Image processing algorithms in video display processing chain are usually design as 2D local filters, where the output is certain function of the pixel values within a local neighbouring window of input image. Moreover, as video display processing module is directly connected with display panel, it should meet certain timing standard, e.g. VESA (Video Electronics Standards Association) standard. Therefore, the VLSI architecture of video display processing module has to process the image line-by-line in a raster-scanned manner from top left to bottom right and be fully synchronized to each pixel clock. The memory hierarchy design is critical to real-time implementation of this fully synchronized circuit.

Fig. 3 illustrates the conventional memory hierarchy design in video display processing modules as discussed in [22]. To support the window-based image processing, the memory hierarchy is designed as two-level buffering, including a line buffer and a register window. The line buffer is a set of RAMs which is able to contain several lines of the input video, and the window buffer is a set of shift registers (represented as 'R') which contain the pixels in the processing window. Since the window buffer consists of registers, it is able to guarantee instant access to all its elements. The line buffer is generally designed with dual-port RAMs to support the simultaneous read and write. The number of required DPRAM is $m - 1$, and the length of each DPRAM equals to the horizontal resolution of input video frame (maximum video format). The working frequency of DPRAM is synchronized to pixel clock that the line buffer has be written in and read out one pixel simultaneously within each pixel period.

The memory architecture clearly indicates that all the line buffers have to work in active mode as long as the display panel is turned on, hence leakage power can be considerably large and the use of low-leakage STT-RAM to replace SRAM becomes a natural choice. However, for high resolution video formats, such as 1080P@60Hz or 4K*2K@60Hz, the write latency of these line buffers should be sufficiently low. The use of hybrid memory architecture may reduce the working frequency requirement of STT-RAM line buffers, this nevertheless tends to incur the multi-clock domain issue and complicate the architecture design. More important, as all the video data should finally be written into line buffers and the write intensity is relatively high, the write power consumption of non-volatile STT-RAM may be extraordinary high and overtake the leakage energy save. As reducing the retention time of MTJ can significantly reduce the write latency and energy, volatile STT-RAM has the great potential to well fit into this memory architecture without any design effort and achieve reduced energy consumption in the meanwhile. Nevertheless, the retention time of MTJ is not the smaller the better. Video data stored in MTJ should sustain for at least one image line as illustrated in Fig. 3. Therefore, how to determine the appropriate retention time of MTJ is critical to the use of volatile STT-RAM in video display processing system.

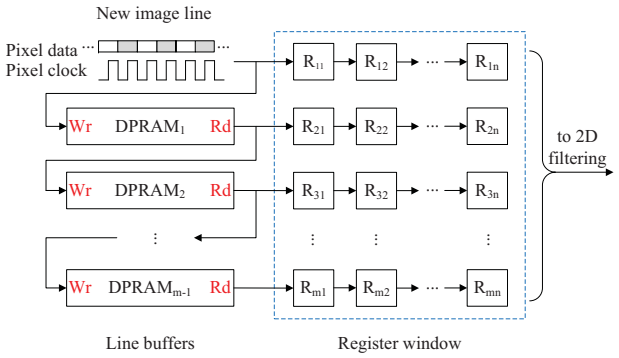

Fig. 3. Memory hierarchy design in video display processing modules.

TABLE I
Video format specifications.

Video format	Resolution	Pixel frequency	Horizontal frequency
VGA@60Hz	640×480	25.2 MHz	31.5 KHz
1080P@60Hz	1920×1080	148.5 MHz	67.5 KHz
4K×2K@60Hz	3840×2160	594 MHz	135 KHz

B. Volatile vs. Non-Volatile STT-RAM

This subsection explores the design space of volatile STT-RAM in video display processing application. As video display processing systems are usually designed to support a variety of video formats, the retention time of volatile STT-RAM certainly should locate between the following two margins.

- upper margin: the retention time should be below this margin to make the write latency of volatile STT-RAM short enough that it can be seamlessly synchronized to pixel clock of the maximum video format V_{max}.
- lower margin: the retention time should be above this margin to guarantee that the pixel data stored in line buffer can sustain long enough until it is useless, i.e. the period of one image line of the minimum video format V_{min}.

These two margins can also be described as the following equation, where the F_h and F_p represent the horizontal and pixel frequencies, respectively.

$$1/F_h(V_{min}) \leq T_{retention} \leq Func(F_p(V_{max})) \quad (5)$$

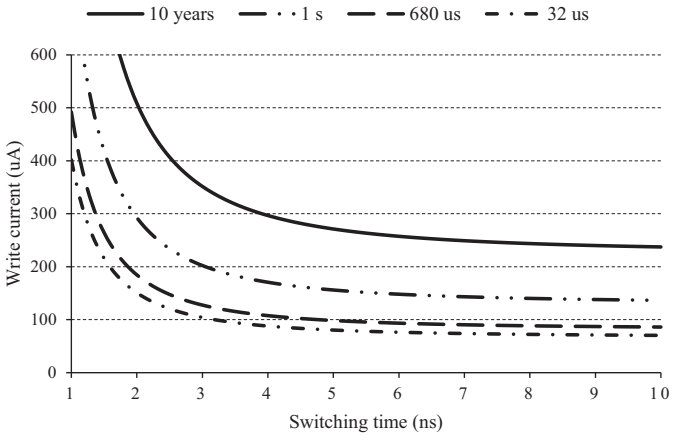

Fig. 4. Write current vs. switching time at different retention time nodes.

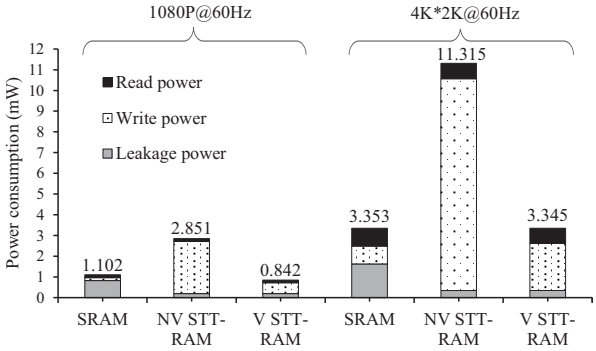

Fig. 5. Power consumption comparison among SRAM, non-volatile and volatile STT-RAM technologies.

In this paper, we assume the maximum and minimum video formats supported by the targeting video display processing system are 4K*2K@60Hz and VGA@60Hz, respectively. Their specifications are listed in Table I. We can conclude that the upper and lower margins of retention time of MTJ are $680\mu s$ and $32\mu s$, respectively. Accordingly, we simulate four retention time nodes, i.e. 10 years for non-volatile STT-RAM and $1s$, $680\mu s$ and $32\mu s$ for volatile STT-RAM, and further make curve fitting using the MTJ device equations discussed in subsection II.B. The curve-fitted results for four different MTJ retention time nodes are shown in Fig. 4.

By implementing a STT-RAM cell model into the CACTI memory modeling tool [23], we simulate a series of line buffers with SRAM, non-volatile STT-RAM and volatile STT-RAM technologies. For each technology, we simulate two memory configurations, i.e. 2K×8b for 1080P@60Hz and 4K×8b for 4K*2K@60Hz. In addition, we simulate three working points for volatile STT-RAM in terms of retention time, i.e. A for $680\mu s$, B for $130\mu s$ and C for $32\mu s$. The detail simulation results are listed in Table II. We can see that write latency of non-volatile STT-RAM is more than 10ns, and hence it can not be used in memory hierarchy of high-end video display processing system without complicated hybrid memory architecture design. Nevertheless, volatile STT-RAM has the write latency of less than 1.4ns and can be easily integrated into the memory hierarchy without any architecture effort. This is very important for replacing SRAM with STT-RAM in video processing system. In addition, we can notice that the write energy of non-volatile STT-RAM is almost 12 times larger than its SRAM counterpart. Therefore, the use of low-leakage STT-RAM does not necessarily mean low power. For write-intensive memory system, such as the line buffers in video display processing, the energy efficiency may be even degraded considerably.

Fig. 5 illustrates the power consumption comparison of display processing line buffers among SRAM, non-volatile and volatile STT-RAM technologies. The volatile STT-RAM is chosen as working point B and the targeting video formats include both 1080P@60Hz and 4K*2K@60Hz. The power consumption is calculated with the simulation results in Table II. As aforementioned, we can see that leakage power occupies the majority of overall energy consumption for SRAM and the use of STT-RAM can significantly reduce the leakage power. However, as the write intensity to line buffer in display processing is relatively high and non-volatile STT-RAM write is really energy consuming, the overall energy consumption of non-volatile STT-RAM is even much higher than its SRAM counterpart. Hence, the use of volatile STT-RAM is actually necessary for write-intensive memory hierarchy of video display processing. We can see that, by employing volatile STT-RAM, the overall energy consumption can be significantly reduced to below that of SRAM. Nevertheless, the energy efficiency is still not satisfactory that, compared with its SRAM counterpart, the overall energy consumption of volatile STT-RAM for 1080P@60Hz and 4K*2K@60Hz are 76.4% and 99.8%, respectively. This is mainly because that the write energy of volatile STT-RAM is still 3 times larger than its read energy. Therefore, extra technique explorations are necessary to further reduce the write intensity and energy in addition to the direct use of volatile STT-RAM.

C. Techniques to Further Reduce Write Energy

Although write to volatile STT-RAM consumes much less energy than non-volatile STT-RAM, its energy consumption is still much higher than SRAM. Therefore, when the write intensity is very high, the energy reduction of the use of volatile STT-RAM tends to be marginal. Actually, by employing small architecture and circuit modification, the write intensity to volatile STT-RAM line buffers can be largely reduced, and hence the energy efficiency can be significantly improved. In this paper, we presents two simple yet effective techniques, i.e. selective buffer write and redundant bit write removal.

The memory hierarchy design presented in Fig. 3 is the most common architecture [22] for display processing and can be easily implemented. Nevertheless, its major drawback is write intensity to line buffers. For STT-RAM line buffers, we can significantly reduce write intensity by making small modification to memory hierarchy design, as shown in Fig. 6. For a new image line, pixel data can be only written to one of line buffers and data stored in the rest of line buffers remain unchanged. This will certainly make image lines in buffer out of order. Nevertheless, we can reorder image lines after pixel data are read out

TABLE II

Characteristics of dual-port 2K×8b and 4K×8b memory designs with SRAM and STT-RAM technologies.

	Configuration	Retention time	Cell size (F^2)	Area (μm^2)	Read latency (ns)	Write latency (ns)	Read energy (pJ)	Write energy (pJ)	Leakage power (mW)
SRAM	Dual-port 4K×8b 32nm	N/A	146	8481.13	0.257	0.257	1.451	1.451	1.629
NV STT-RAM		10 years	28	5122.98	0.167	10.157	1.263	17.211	0.342
V STT-RAM A		680μs	9.4	4373.74	0.192	1.399	1.243	4.369	0.340
V STT-RAM B		130μs	8.2	4283.35	0.187	1.397	1.235	3.832	0.341
V STT-RAM C		32μs	7.2	4199.12	0.186	1.394	1.231	3.373	0.341
SRAM	Dual-port 2K×8b 32nm	N/A	146	4270.20	0.202	0.202	0.938	0.938	0.824
NV STT-RAM		10 years	28	2827.17	0.177	10.148	0.924	16.943	0.198
V STT-RAM A		680μs	9.4	2430.50	0.168	3.251	0.910	3.966	0.197
V STT-RAM B		130μs	8.2	2374.80	0.163	3.238	0.896	3.479	0.197
V STT-RAM C		32μs	7.2	2319.83	0.158	3.230	0.882	3.063	0.196

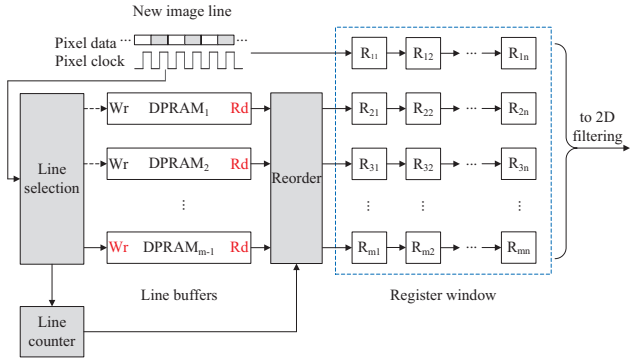

Fig. 6. Memory hierarchy design employing selective buffer write technique.

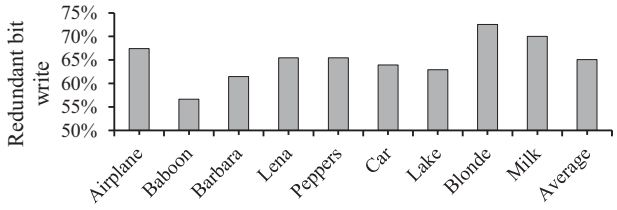

Fig. 7. Redundant bit write statistics in nine different testing images.

and make them in order to register window. One simple way to select line buffer for current image line is in circular order, then we can just use a 'line counter' to record the order and control the reorder module, as illustrated in Fig. 6. Moreover, as the reordering logic tends to significantly complicate the circuit design in some cases, the architecture presented in Fig. 6 is less popular than that in Fig. 3 as discussed in [22].

In this paper, this technique is referred to as selective buffer write (SBW). By employing the selective buffer write technique, the write intensity to line buffers can be significantly reduced to $1/(m-1)$, where m is vertical size of register window. Hence, the write energy of volatile STT-RAM line buffers can be significantly reduced. In addition, additional control modules incurred by SBW technique are really simple, and hence the implementation overhead is actually negligible compared with large capacity memory comsuption. Another issue is that, SBW technique requires a $m-1$ times larger retention time, and this tends to increase write energy simultaneously. Nevertheless, as we illustrated in Fig. 4, this influence is marginal especially when compared with write intensity reduction of SBW technique.

We can also reduce the write energy by employing circuit techniques. One simple yet effective write energy reduction technique at circuit level is to avoid redundant bit write and referred to as redundant bit write removal (RBWR) in this paper. This technique is based on the observation that many bits are written with the same values that are already stored in the STT-RAM memory. Those writes are therefore unnecessary, and should be removed to save energy. This technique has been previously explored in STT-RAM based cache memory design, and several bit-wise early write-completion circuits have been proposed [24, 25]. This paper does not aim to propose any new early write-completion circuit, but instead, to evaluate the effectiveness when applying the RBWR technique to volatile STT-RAM line buffers in video display processing system. For the memory architecture illustrated in Fig. 6, the new image line k is written to the line buffer that stores image line $k-m$. As spatial redundancy is an inherent feature of video frame and has been extensively explored for intra-frame video compression, the probability of redundant bit write is expected to be considerably high.

Assuming that vertical size (m) of register window is 5, we collect redundant bit write statistics at the vertical interval of 5 in nine well-known testing images and illustrate the results in Fig. 7. Although the percentage of redundant bit write in each different testing image varies, the lowest value is still above 56% and the average value is as high as 66%. The results clearly demonstrate the effectiveness of redundant bit write removal technique in video display processing system. According to the modeling results in [24], the write energy consumption of unchanged cells is only around 5.3% compared with that of changed cells. Therefore, according to the calculation using the following equation, the write energy consumption of volatile STT-RAM line buffers can be further reduced to around 37.5% by employing RBWR technique.

$$E_{write} = E_{changed} \times P_{changed} + E_{unchanged} \times P_{unchanged} \quad (6)$$

Fig. 8 illustrates the normalized power consumption reduction of 4K×2K@60Hz video format when employing SBW and RBWR techniques. We continue to choose working point B for volatile STT-RAM and assume the evaluated memory hierarchy includes 4 volatile STT-RAM line buffers and a 5×5 register window. For the display processing of 4K×2K@60Hz, the overall memory energy consumption is significantly reduced to 38.33% and 38.24% with respect to that of baseline volatile

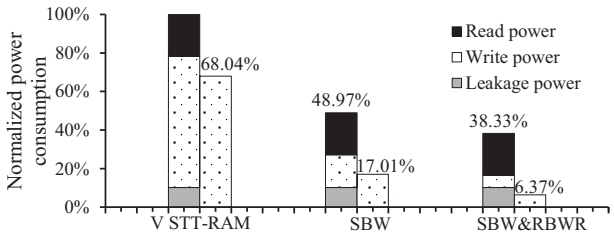

Fig. 8. Normalized power consumption reduction when employing SBW and RBWR techniques.

TABLE III

PSNR and SSIM results at four bit error rates.

Images /Videos	1×10^{-6}		1×10^{-5}		5×10^{-5}		1×10^{-4}	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Airplane	59.40	1.000	54.91	1.000	43.31	0.997	40.27	0.992
Baboon	59.17	1.000	54.94	1.000	43.24	0.997	40.37	0.996
Barbara	59.27	1.000	55.34	0.999	43.44	0.999	40.43	0.996
Lena	59.37	1.000	54.82	0.999	43.60	0.997	40.52	0.994
Foreman	59.53	1.000	55.76	1.000	44.83	0.997	41.17	0.993
Waterfall	59.35	1.000	55.58	1.000	44.88	0.998	41.08	0.994
Akiyo	59.65	1.000	55.48	0.998	44.97	0.996	41.21	0.991
Bus	59.29	1.000	55.70	0.999	45.08	0.998	41.59	0.995
Flower	60.03	1.000	55.95	0.999	44.52	0.997	40.69	0.993
Average	59.45	1.000	55.27	0.999	43.99	0.998	40.81	0.994

STT-RAM and SRAM architectures. In particular, by employing these two effective techniques, the write energy becomes much less than read energy and is no longer a critical issue for memory hierarchy in video display processing system any more. We should note that, the SBW and RBWR techniques can also be employed in SRAM to reduce write intensity. However, as the write energy only takes a small percentage of the overall energy consumption in SRAM, the energy reduction by employing the above two techniques to SRAM is actually limited.

IV. Error Resilience Evaluation

Video processing system is inherently error-tolerate, and its error resilience capability has been evaluated in several video decoding systems [27]. In this section, we evaluate the error resilience of video display processing system to address the potential read disturb error increase due to the decrease of thermal stability factor Δ. According to Equation 4, when the retention time of MTJ is reduced to 130μs, the read disturb error rate may reach up to 10^{-6}.

To evaluate the error resilience of video display processing system, we establish a experimental video display processing chain with three classical algorithms, i.e. Gaussian filtering for noise reduction, bi-cubic image upscaling for format conversion and high-pass filtering for sharpness enhancement. The Gaussian filtering and high-pass filtering are calculated in a 5×5 window and hence each uses 4 volatile STT-RAM line buffers, while the bi-cubic upscaling uses 3 line buffers to build a 4×4 interpolation window. The bit errors are injected into STT-RAM line buffers of all the three processing stages, as illustrated in Fig. 9. In addition, to measure the image quality degradation when bit errors occur in the STT-RAM line buffers, we choose

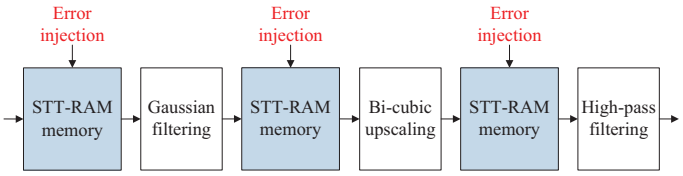

Fig. 9. Error injection in the experimental video display processing chain.

four 512×512 well known test images, such as 'Lena' and 'Barbara', and five 352×288 well known test video sequences, such as 'Foreman' and 'Akiyo'. The selected image files are RGB color images with a depth of 8 bits per channel. These color images are firstly converted to YCbCr color space, and employed the three-stage error injection and image processing algorithms. Then the YCbCr images are re-converted to RGB color space for the following evaluation. The selected video sequences are evaluated directly in YUV color space. Both objective and subjective tests of these color images and video sequences are performed in order to quantitatively compare the quality of the images generated at different bit error rates.

The objective test compares error-free and error-injected images after the experimental video display processing chain, and the comparison is performed on all the three RGB channels. The differences are evaluated by the metrics of peak signal noise ratio (PSNR) and structural similarity (SSIM) [26]. Table III lists the PSNR and SSIM results at the bit error rates of 1×10^{-6}, 1×10^{-5}, 5×10^{-5} and 1×10^{-4}, respectively. We can see that video display processing system shows strong error resilience capability that the PSNR and SSIM results remain very high (55.27dB and 0.999 on average) even at the bit error rate of 10^{-5}. Although the PSNR and SSIM results decrease slowly when the bit error rate is above 10^{-5}, the error-injected images after display processing can still maintain high objective quality. At the bit error rate of 10^{-4}, the average PSNR and SSIM results are as high as 40.81dB and 0.994, respectively.

It should be noted that images with high PSNR or SSIM are not necessarily corresponding to high visually perceived quality. Therefore, the visual quality of images at different bit error rates is further evaluated. Fig. 10 shows image quality comparison using 'Lena' image as an example among different bit error rates. At bit error rates of 1×10^{-6} and 1×10^{-5}, we can hardly observe any artifact in 'Lena' images, as bit error rates are relatively low and only errors on the significant bits are visible. At bit error rates of 5×10^{-5} and 1×10^{-4}, we begin to see visible artifacts in 'Lena' images by careful observation. We also make error tags on images to demonstrate the artifacts more clear. According to the image visual quality evaluation, we can conclude that video display processing system can tolerate errors in STT-RAM without visible image quality degradation when the bit error rate is below 1×10^{-5}. The error resilience evaluation results further confirm that, it is feasible and promising to explore the use of volatile STT-RAM in video display processing system for significant energy efficiency improvement.

V. Conclusion

This paper proposes to use volatile STT-RAM rather non-volatile STT-RAM for video processing system, as video da-

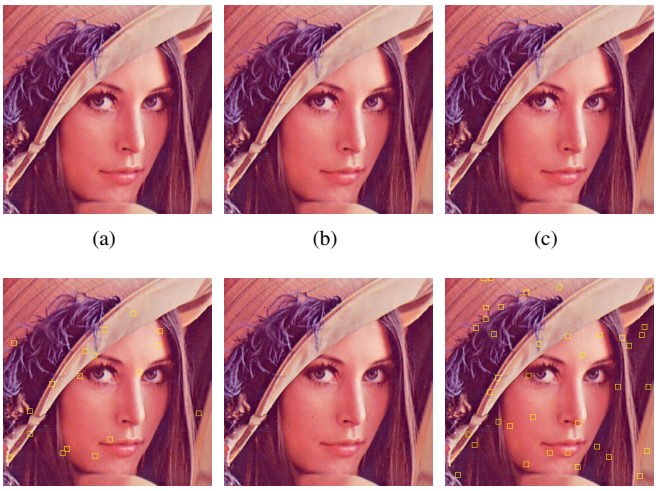

Fig. 10. Image quality comparison using 'Lena' image among different bit error rates: (a) 1×10^{-6}, (b) 1×10^{-5}, (c) 5×10^{-5}, (d) 5×10^{-5} with error tag, (e) 1×10^{-4}, and (f) 1×10^{-4} with error tag.

ta is processed in streaming-style and do not need to be stored in a long period. Targeting for the video display processing of 4K*2K@60Hz format, we demonstrate that, the direct use of non-volatile STT-RAM can dramatically increase the memory energy consumption due to its high write energy. Simulation results demonstrate that, the use of volatile STT-RAM and extra design techniques can significantly reduce the overall memory energy consumption to 38.24% with respect to that of baseline SRAM architecture.

VI. ACKNOWLEDGMENT

This research was partially supported by the National Science and Technology Major Project of China (No. 2013ZX01033001-001) and National Natural Science Foundation of China (No. 61103048).

REFERENCES

[1] W. Xu *et al.*, "Design of last-level on-chip cache using spin-torque transfer RAM (STT RAM)," *IEEE TVLSI*, no. 3, vol. 19, pp. 483–493, 2011.
[2] H. Sun *et al.*, "Using magnetic RAM to build low-power and soft error-resilient L1 cache" *IEEE TVLSI*, no. 1, vol. 20, pp. 19–28, 2012.
[3] G. Sun *et al.*, "A novel architecture of the 3D stacked MRAM L2 cache for CMPs" *HPCA*, pp. 239–249, 2009.
[4] C. Smullen *et al.*, "Relaxing non-volatility for fast and energy-efficient STT-RAM caches," *HPCA*, pp. 50–61, 2011.
[5] A. Jog *et al.*, "Cache revive: architecting volatile STT-RAM caches for enhanced performance in CMPs," *DAC*, pp. 243–252, 2012.
[6] Z. Sun *et al.*, "Multi retention level STT-RAM cache designs with a dynamic refresh scheme," *Micro*, pp. 329–338, 2011.
[7] Q. Li *et al.*, "Compiler-assisted refresh minimization for volatile STT-RAM cache" *IEEE TC*, pp. 1–14, 2015.
[8] F. Sampaio *et al.*, "Energy-efficient architecture for advanced video memory," *ICCAD*, pp. 132–139, 2014.
[9] M. U. K. Khan *et al.*, "AMBER: Adaptive energy management for on-chip hybrid video memories," *ICCAD*, pp. 405–412, 2013.
[10] A. Wang *et al.*, "Heterogeneous multi-processing quad-core CPU and dual-GPU design for optimal performance, power, and thermal tradeoffs in a 28nm mobile application processor" *ISSCC*, pp. 180–181, 2014.
[11] J. E. Caviedes, "The evolution of video processing technology and its main drivers," *Proc. of the IEEE*, no. 4, vol. 100, pp. 872–877, 2012.
[12] J. E. Caviedes, "Intelligent sharpness enhancement for video post-processing," *EUSIPCO*, 2006.
[13] H. Sun *et al.*, "Design and implementation of a video display processing SoC for full HD LCD TV," *ISOCC*, pp. 297–300, 2012.
[14] T. Fukuda *et al.*, "A 7ns-access-time 25uW/MHz 128kb SRAM for low-power fast wake-up MCU in 65nm CMOS with 27fA/b retention current," *ISSCC*, pp.236–238, 2014.
[15] Y. Wang *et al.*, "A 1.1GHz 12uA/Mb-leakage SRAM design in 65nm ultra-low-power CMOS with integrated leakage reduction for mobile applications,"*ISSCC*, pp. 324–326, 2007.
[16] Y. Wang *et al.*, "A 4.0 GHz 291Mb voltage-scalable SRAM design in 32nm high-K metal-gate CMOS with integrated power management," *ISSCC*, pp. 456–458, 2007.
[17] N. Verma, "Analysis towards minimization of total SRAM energy over active and idle operating modes," *IEEE TVLSI*, no. 9, vol. 19, pp. 1695–1703, 2011.
[18] Z. Diao *et al.*, "Spin-transfer torque switching in magnetic tunnel junctions and spin-transfer torque random access memory," *JPCM*, no. 16, vol. 19, pp.165209, 2007.
[19] A. Raychowdhury *et al.*, "Design space and scalability exploration of 1T-1STT MTJ memory arrays in the presence of variability and disturbances" *IEDM*, pp. 1–4, 2009.
[20] W. Zhao *et al.*, "Failure and reliability analysis of STT-MRAM" *Microelectronics Reliability*, no. 9, vol. 52, pp. 1848–1852, 2012.
[21] Y. Zhang *et al.*, "STT-RAM cell design optimization for persistent and non-persistent error rate reduction: a statistical design view" *ICCAD*, pp. 471–477, 2011.
[22] D. G. Bailey, "Design for embedded image processing on FPGAs" *John Wiley&Sons*, 2011.
[23] N. Muralimanohar *et al.*, "CACTI 6.0: A tool to model large caches," *HP Lab.*, 2009.
[24] P. Zhou *et al.*, "Energy reduction for STT-RAM using early write termination" *ICCAD*, pp. 264–268, 2011.
[25] T. Zheng *et al.*, "Variable-energy write STT-RAM architecture with bit-wise write-completion monitoring" *ISLPED*, pp. 229–234, 2013.
[26] Z. Wang *et al.*, "Image quality assessment: from error visibility to structural similarity" *IEEE TIP*, no. 4, vol. 13, pp. 600–612, 2004.
[27] J. Kwon *et.al*, "Hetergeneous SRAM cell Sizing for low-power H.264 applications" *IEEE TCSI*, no. 10, vol. 59, pp. 2275–2284, 2012.