INTEGRATION, the VLSI journal 51 (2015) 10–20

Contents lists available at ScienceDirect

INTEGRATION, the VLSI journal

journal homepage: www.elsevier.com/locate/vlsi

NBTI alleviation on FinFET-made GPUs by utilizing device heterogeneity

Ying Zhang*, Sui Chen, Lu Peng, Shaoming Chen

Division of Electrical & Computer Engineering, School of Electrical Engineering and Computer Science, Louisiana State University, Baton Rouge, LA 70803, United States

ARTICLE INFO

Article history:
Received 10 December 2014
Received in revised form
26 March 2015
Accepted 15 April 2015
Available online 25 April 2015

Keywords:
NBTI
FinFET
Reliability
Heterogeneity

ABSTRACT

Recent experimental studies reveal that FinFET devices commercialized in recent years tend to suffer from more severe NBTI degradation compared to planar transistors, necessitating effective techniques on processors built with FinFET for endurable operations. We propose to address this problem by exploiting the device heterogeneity and leveraging the slower NBTI aging rate manifested on the planar devices. We focus on modern graphics processing units in this study due to their wide usage in the current community. We validate the effectiveness of the technique by applying it to the warp scheduler and L2 cache, and demonstrate that NBTI degradation is considerably alleviated with slight performance overhead.

1. Introduction

As we shift into the deep submicron era, innovative materials and device architectures are becoming ever demanding to continue the trend toward smaller and faster transistors. Among all candidates in investigation, the fin field-effect-transistor (FinFET) stands as one of the most promising substitutes for traditional devices at the ensuing technology nodes, since it presents several key advantages over its planar counterpart [1–4]. By wrapping the conducting channel with a thin vertical "fin" which forms the body of the device, the gate is coupled tighter with the channel, increasing the surface area of the gate-channel interface and allowing much stronger control over the conducting channel [1]. this effectively relieve the so-called short channel effects (SCE) that are observed on planar transistors manufactured with sub-32 nm technology, which in turn implies that FinFET device can provide superior scalability in the deep submicron regime [1].

Another cornerstone motivating the realization of FinFET is the potential performance gain. FinFET transistors can be designed with lower threshold voltage (V_t) and operate with higher drive current, leading to faster switching speed compared to conventional planar devices [1]. Released documents from industry demonstrate that the FinFET transistor persistently demonstrates shorter delay than the planar one while the support voltage is varying, enabling the design and manufacturing of faster processors. Public documents from leading manufacturers also show that the FinFET structure is capable of largely decreasing leakage when the transistor is off [1]. Recently, the Ivy Bridge [5] and Haswell central processing units [6] released by Intel have commercialized this structure (i.e., referred to as "Tri-gate transistor" by Intel), which is also expected to be adopted by other semiconductor manufacturers on their upcoming products [7].

Nonetheless, FinFET is not an impeccable replacement of traditional devices as it raises many challenges to the current industry. One of the most daunting conundrums is the increasing aging rate caused by negative bias temperature instability (NBTI). Recent experimental studies demonstrate that FinFET transistors are more vulnerable to NBTI, leading to a shorter lifetime than a planar device [8,9]. The NBTI aging rate is evaluated by the increase of delay on the critical path after a certain amount of service time. A chip is considered as failed when the delay increment exceeds a pre-defined value after which the timing logic of the processor cannot function correctly. Under the same operation condition, the FinFET device is observed to degrade much faster than the planar counterpart, implying a significantly reduced service lifespan of the target processor. This clearly spurs the development of new techniques to circumvent this problem and prolong the lifetime of FinFET-made processors.

Fortunately, a brief comparison between the main features of FinFET and planar devices sheds some light on alleviating the NBTI effect on future processors. By effectively exploiting the device heterogeneity and leveraging the higher NBTI immunity of planar transistors, the aging of the FinFET structures can be largely suppressed. In this paper, we propose a technique built on top of this principle to improve the durability of FinFET processors.

* Corresponding author.
E-mail addresses: ying.esz.zhang@gmail.com (Y. Zhang), csui1@lsu.edu (S. Chen), lpeng@lsu.edu (L. Peng), schen26@lsu.edu (S. Chen).

http://dx.doi.org/10.1016/j.vlsi.2015.04.003
0167-9260/

In general, our technique is implemented by replacing an existing structure with a planar-device equivalent. Along with minor modifications at the architectural level, our proposed technique is essentially transferring the "aging stress" from the vulnerable FinFET components to the more NBTI-tolerable planar structures, which in turn lower down the temperature on the structure in study, and thus considerably mitigate the NBTI degradation. Note that the proposed scheme is practically feasible because of the good compatibility between the FinFET and planar process technology [10–12].

Considering that the general-purpose graphics processing unit is becoming an increasingly important component in a wide spectrum of computing platforms, we choose a modern GPU as the target architecture to evaluate the effectiveness of our proposed strategy. In this paper, we mainly concentrate on optimizing the reliability of the warp scheduler because of its importance. However, the technique described in this paper can be simply applied to CPU for NBTI mitigation as well. In general, the main contributions of this work are as follows:

- We propose a hybrid-device warp scheduler for reliable operation. By decoupling the warp scheduling into two steps of operations and conducting the prerequisites evaluation in a planar-device structure, we eliminate a large amount of read accesses to the FinFET scheduler hardware and considerably alleviate the NBTI effect.
- We develop a hybrid-device sequential-access cache architecture. All memory requests to this cache hierarchy are handled in a serialized fashion that the tag-array made of planar transistors is probed first and the matching block in the FinFET data array is only accessed on a cache hit. This significantly reduce the activity on the cache data array and improve its reliability.

2. Background

2.1. NBTI degradation mechanism

Negative bias temperature instability is becoming one of dominant reliability concerns for nanoscale P-MOSFETs. It is caused by the interaction of silicon–hydrogen (Si–H) and the inversion charge at the Si/oxide interface [13,14]. When a negative voltage is applied at the gate of PMOS transistors, the Si–H bonds are progressively dissociated and H atoms diffuse into the gate oxide. This process eventually breaks the interface between the gate oxide and the conducting channel, leaving positive traps behind. As a consequence, the threshold voltage of the PMOS transistor is increased, which in turn elongates the switching delay of the device through the alpha power law [15]:

$$T_s \propto \frac{V_{dd}L_{eff}}{\mu(V_{dd}-V_t)^{\alpha}} \tag{1}$$

where μ is the mobility of carriers, α is the velocity saturation index and approximates to 1.3. L_{eff} denotes the channel length. The process described above is termed the "stress" phase where the threshold voltage is persistently increasing with the service time, modeled by the following equation [9].

$$\Delta V_{tstress} = \left(\frac{qT_{ox}}{E_{ox}}\right)^{1.5} \cdot K \cdot \sqrt{C_{ox}(V_{gs}-V_t)} \cdot e^{\frac{-E_a}{4kT}+\frac{2(V_{gs}-V_t)}{T_{ox}E_{01}}} \cdot T_0^{-0.25} \cdot T_{stress} \tag{2}$$

However, when the stress voltage is removed from the gate, H atoms in the traps can diffuse back to the interface and repair the broken bond. This results in a decrease in the threshold voltage, thus termed the "recovery" stage. This iterative stress-recovery processes lead to a saw-tooth variation of the threshold voltage throughout the device's lifespan. The final V_t increase taking both stress and recovery into account can be computed as:

$$\Delta V_t = \Delta V_{tstress} \cdot \left(1 - \frac{2\xi_1 T_{ox} + \sqrt{\xi_2 e^{\frac{-E_a}{kT}} T_0 T_{stress}}}{(1+\delta)T_{ox} + \sqrt{e^{\frac{-E_a}{kT}}(T_{stress}+T_{recovery})}}\right) \tag{3}$$

Note that in Eqs. (2) and (3), T_{stress} and $T_{recovery}$ respectively denote the time under stress and recovery. Other parameters are either constants or material-dependent variables and are listed in Section 4.

That FinFET devices are more vulnerable to NBTI is generally attributed to its unique non-planar architecture, which is visualized by Fig. 1. As can be seen, compared to a traditional planar transistor, the FinFET structure is designed with additional fin sidewall surface with higher availability of Si–H bonds [8,9], implying larger chances of forming interface trap and consequently expediting the device degradation.

The NBTI aging rate depends on multiple factors including both circuit parameters and workload execution patterns. In general, it is acknowledged that voltage, temperature, and the stress/recovery time have strong impact on the aging rate [16,17]. In this work, our proposed techniques significantly reduce the accesses to the target structures, thus lowering down the localized activity and temperature, which is beneficial in enhancing the structure durability.

2.2. Target GPU architecture

The prevalence of unified programming language (e.g., CUDA and OpenCL) has made the general-purpose graphics processing unit a core component in a large variety of systems ranging from personal computers to high-performance computing clusters. Therefore, it is highly important to alleviate the NBTI degradation on this ever increasingly important platform.

Fig. 2 shows the architectural organization of a representative GPU. Note that we follow the Nvidia terminology to depict the processor architecture. As can be seen, the major component of a modern GPU is an array of Streaming Multiprocessors (SMs), each of which contains an amount of CUDA cores (SPs), load/store units and special function units (SFUs). A CUDA core is responsible for performing integer ALU and floating point operations while the

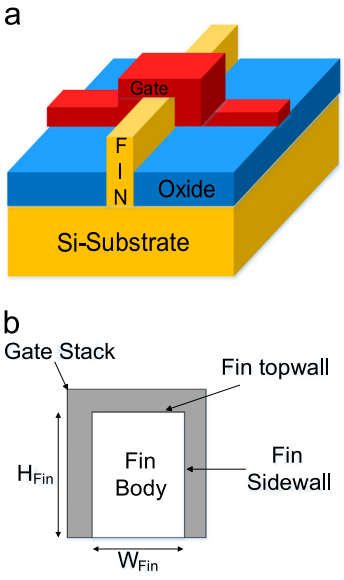

Fig. 1. FinFET transistor structure: (a) overview (b) side view.