International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 11 | Nov 2018 www.irjet.net p-ISSN: 2395-0072

Proposing a RTD-based block for on-chip GPU caches to reduce static power leaks

Richa Sharma

Netaji Subhas University of Technology, New Delhi, 110078

Abstract - *GPUs dominate the industry in the parallel processing. The architecture of GPUs supports parallel algorithms and process images at much higher rates than CPUs. Each generation of GPUs increases the use of on-chip caches and number of processor cores embedded on-chip. Also, with each generation CMOS technology gets significantly worse in its leakage energy consumption. This paper addresses the issues such energy leaks can cause in the efficiency of GPU computing. It further delves into Cache Power Management for GPUs. In this paper, the integration of Tunneling diode technology inside the embedded architecture of GPUs is proposed with a RTD switching block. Furthermore, the benefits of proposed integrations are discussed.*

Key Words: GPU Caches, Cache Power management, Static Leaks, Resonant Tunneling Diodes, Negative Differential Resistance, RTD switching block

1. INTRODUCTION

Graphical Processing Units are used to process images since their inherent parallel structure boosts the performance of parallel processing algorithms. In recent years, GPUs are being utilized for many general purpose computing applications of cloud computing, machine learning, and other cost-effective High-Performance Computing(HPC) clusters.[1] In this era of green computing, there has been a shift of focus in producing more high performance- energy efficient results than just achieving highest peak performance. Achieving energy efficiency has become important in the design of all range of processors, such as battery-driven portable devices, desktop or server processors to supercomputers.[2] The efforts to achieve this dual goal of performance and energy efficiency, researchers have suggested various architectural modifications across different components of the processor, such as processor core, caches, DRAM (dynamic random access memory) etc. Techniques like power grating, DTCMOS have been applied to reduce static power leaks and cache-decay, drowsy cache has been employed to reduce dynamic power leaks.[3] However, researchers face the challenge of scaling the devices with CMOS technology since the leakage currents increase significantly with the decrease in the thickness of the gate oxide. The increase of subthreshold currents is explained by thermodynamics, more specifically by Boltzmann distribution.[4] The problem of power consumption with CMOS technology gets worse with each generation of processors. The estimates of International Technology Roadmap for Semiconductors (ITRS) indicates that leakage power consumption could become a major threat for the survival of CMOS technology.[5] The competition in the development of GPUs and harnessing its computing power has driven companies to increase the number of processing cores on the processors and their number will continue to rise. Furthermore, to bridge the gap between the speed of the processor and main memory, the caches of larger sizes are being embedded on the chip.[6] Tables 1 and 2 list the cache memory requirements of current designs for different processors and fraction of energy consumed in cache power out of total power consumption. In this paper, I focus on static energy leaks of GPUs that is contributed by existing CMOS technology and propose to replace it by Tunneling Diode technology. Furthermore, I try to elaborate on the benefits of this upgrade in cache power management of GPUs. GPUs are yet to realize their full computing potential. Since energy management is one of the biggest challenges to overcome in the processor industry, our solution provides a new approach to this problem of excessive energy consumption in processor chips.

Table-1: On-chip cache memory size in the latest generations Processors.

Processor Type	Cache Memory Used
Desktop (CPU, GPU)	8 MB
Server Cores	24-32 MB
Mobile Processors	1 MB

Table-2: Power consumption of on-chip caches.

Processor Name	Percentage of Total Power Consumption
Alpha 21264	16%
StrongARM	30%
Niagra,Niagra-2 L-2 Caches	24%

2. CACHE POWER MANAGEMENT IN GPUs

The power consumption of CMOS circuits is mainly classified in two parts, namely dynamic power (also called active power) and leakage power (also called static power).

The mathematical modelling equations for both dynamic and leakage Power are given as the following

$$P_{leakage} = V_{DD} \times I_{leakage}$$
$$P_{dynamic} = \alpha \times C_{eff} \times F \times V_{DD}^2$$

Here, dynamic and leakage power can also be written as active and static power respectively. Dynamic Power is dependent on activity factor(number of cache accesses or number of bits accessed per cache access), effective capacitance, frequency of the operation and voltage supplied. Static Power is dependent upon supply voltage and leakage currents, thus its consumption can be reduced by decreasing either of these factors. Modern processors incorporate multi-level cache hierarchy with L1, L2, L3 caches. L1 is classified as First Level Caches (FLCs) and L2, L3 etc. are Lower Level Caches(LLCs).

First Level caches have higher dynamic power consumption rates and Lower Level Caches have higher leakage power consumption rates. Since the design of FLCs and LLCs incorporates latency of caches differently in each of them. Techniques developed to improve energy efficiency with Cache includes developing a high-impedance path, so that it can turn "off" the cache line when they hold data not likely to be reused. The period of non-usage of caches is referred to as "dead time". The technique used to predict "dead time" is cache decay which incorporates a counter to keep the count of pre-set cycles since its last usage. [7]

3. RESONANT TUNNELING DIODE AND INTEGRATION OF RTD SWTICHING BLOCK IN GPU

3.1 Resonant Tunneling Diode(RTD)

Resonant Tunneling Diode(RTD), a variant of Tunneling Diode, invented by Esaki in 1957, is one of the most promising quantum effect devices that are operational at room temperature. These diodes produce enhanced quantum tunneling effect, which in turn produces very high-speed currents, which can be fine-tuned. Resonant Tunneling Diodes similar to CMOS transistor turns "on", conducts a current under a specific external bias voltage range.

The inherent difference between these two devices is that current in RTDs tunnels through quasi-bound states within a double barrier structure, instead of going through a channel between drain and source. In other Tunneling Diodes like Esaki Tunneling Diodes current tunnels through depletion region. Shown in the Fig. 1 is the IV characteristic curve of RTDs. The figure represents the dynamic voltage vs current behavior.

The peak current occurs at the resonance at a specific voltage, then the device exhibits Negative Differential Resistance (NDR) as the current continues to decrease with the increasing voltage and as it reaches a minimum "valley" current at a specific voltage it starts to rise again. This minimum "valley" current can be classified as leakage current is significantly less than in TDs. The bistability of RTDs gives them an advantage over TDs, since TD's produce very high leakage currents when the reverse bias voltage is applied.

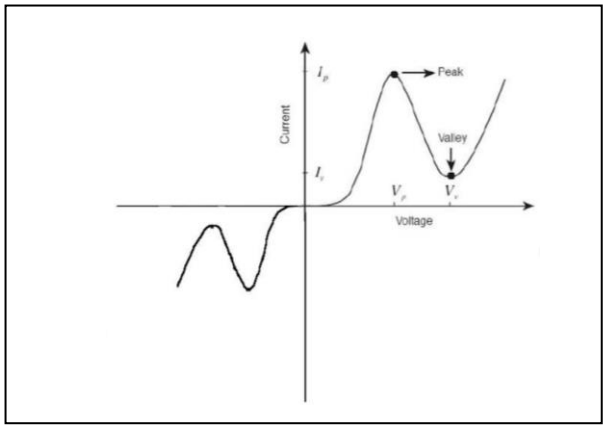

Figure-1: I-V characteristic of RTDs. Leakage currents are kept to a minimum due to the resonance and bistability of the device, therefore providing advantage over TDs.

3.2 Applications of Negative Differential Resistance(NDR) in High-Switching Blocks

A huge advantage of RTDs is the ease of their integration with other technologies like complementary metal-oxide-semiconductor (CMOS) and high electron mobility transistors (HEMTs). We can utilize the Negative Differential Resistance of the RTDs and create a switching block that will switch off and on at the application of specific voltages. This switching block gives two operational points to switch the device on and off as shown in Figure. 2

These RTDs, resistor and CMOS blocks can provide high speed switching method for cache lines. [8] This provides smooth integration that can work in tandem with cache decay technique. The counter will indicate when the cache line should be deactivated, sending the high voltage signal to RTD, RTD will enter NDR region and turn "off" Cache line, hence implementing cache decay. But due to this method, we could easily reactivate and switch back to low-impedance path way for cache line.

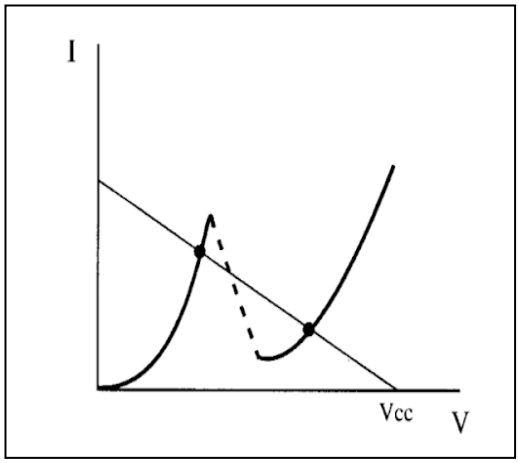

Figure-2: NDR property of RTD allows two stable operation points for switching.

The advantage of using NDR to turn on Cache lines lies in its low static power consumption since the leakage currents in RTDs are minimal and the switching speeds achieved are in picoseconds.

3.2 Execution of Cache Decay by utilizing Negative Differential Resistance(NDR)

Cache line decay technique as proposed by Kaxiras Et al. [7] utilizes a pre-set counter to count a number of inactive cycles of cache line and cuts-off that particular line thus dramatically saving leakage power. In this implementation cut-off signal will be sent to switching block a cycle earlier than proposed in Kraxis model.[7] Kraxis model utilizes gated Vdd CMOS transistor as a switch as proposed by Powell Et al.,2000.[8]

Switching block of RTD will provide a stable point to cut-off power supply to cache line. Implementation of the block as demonstrated in Figure 3 will provide a switch that will turn off at a certain input voltage and will resume operation, according to output Voltage.

Negative Differential Resistance property of RTD will execute in a manner of capacitance, and switching speed of block is dependent upon Load voltage, shown in Figure 3 as Vout. Thus, it provides us with a probe to control switching speed of RTD.

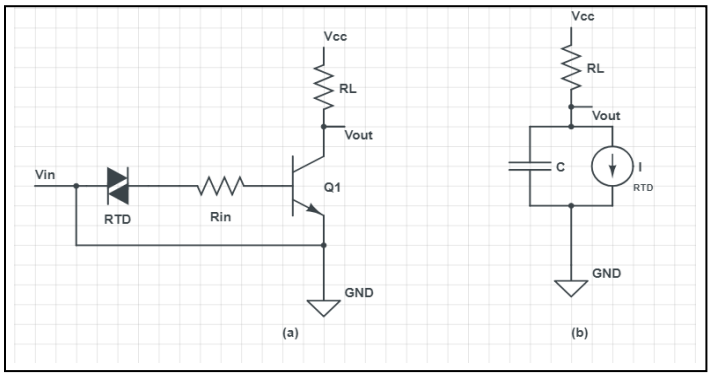

Figure 3: (a) RTD- switching Block with CMOS transistor and Vout as a probe to control switching speed. **(b)** Demonstration of RTD-block as capacitance and current source

4. CONCLUSION

The integration of RTD based structure in SRAMs will be beneficial on many accounts. The proposed utilization of NDR for cache decay technique to reduce power will speed up the operation and therefore not compromise the performance of GPUs. The leakage current can be maintained at very low levels and hence reducing leakage power substantially. Hence, I have proposed the integration of RTDs with CMOS to improve the power efficiency of caches in GPUs and discussed the benefits of utilizing Negative Differential Resistance(NDR) in reducing total power consumption of GPUs.

REFERENCES

[1] OLCF, http://www.olcf.ornl.gov/titan/, 2012
[2] S. Murugesan, “Harnessing green IT: Principles and practices,” IT professional, vol. 10, no. 1, pp. 24–33, 2008
[3] D. J. Frank, "The Limits of CMOS Scaling from a Power-Constrained Technology Optimization Perspective," Nanohub, 4 Oct 06.
[4] “International technology roadmap for semiconductors (ITRS),”http://www.itrs.net/Links/2011ITRS/2011Chapters/2011ExecSum.pdf, 2011.
[5] Sparsh Mittal, “A Survey of Architectural Techniques For Improving Cache Power Efficiency”.
[6] Stefanos Kaxiras, ”Cache-line Decay: A Mechanism to Reduce Cache Leakage Power”
[7] J. L. Huber, J. Chen,” An RTD/Transistor Switching Block and Its Possible Application in Binary and Ternary Adders”
[8] Powell, M. D. et al. 2000. Gated-Vdd: A Circuit Technique to Reduce Leakage in Deep Submicron Cache Memories. In Proc. of the Int'l Symp. On Low Power Electronics and Design'00 (2000).

BIOGRAPHY

Richa Sharma received the Bachelors of Engineering from Netaji Subhas University of Technology, New Delhi formerly Netaji Subhas Institute of Technology. Richa is working on technology intensive areas of physics research and creating innovation in educational methods of these subjects. Currently she works with Society of Applied MicrowaveElectronics Engineering & Research.