Microelectronics Journal 54 (2016) 138–149

Contents lists available at ScienceDirect

Microelectronics Journal

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

Off-chip bus power minimization using serialization with cache-based encoding

Khader Mohammad [a,*], Ahsan Kabeer [b], Tarek M. Taha [c], Muhsen Owaida [a], Mahdi Washha [a]

[a] Birzeit University, Palestine
[b] Clemson University, Clemson, SC, USA
[c] University of Dayton, USA

ARTICLE INFO

Article history:
Received 9 November 2015
Received in revised form
5 June 2016
Accepted 7 June 2016
Available online 14 June 2016

Keywords:
Memory data bus power
Frequent value encoding
Serialization
On and off-chip data bus minimization

ABSTRACT

The data bus is a major component of high power consumption in small process high-performance systems and in systems-on-chip (SoC) design. This paper presents an analysis of different state-of-the-art techniques for reducing the power of off-chip memory bus interface, with proposing an approach overcoming some limitations existing in the state-of-art methods. More precisely, the paper introduces a serialization (S) method combined with cache-based encoding scheme, aiming at saving the optimal possible power for off-chip memory bus. Bus serialization reduces the number of transmission wires, while a Serialization-Widening (SW) approach reduces the bus capacitance and the number of transmission wires. Experimental results show that, for off-chip data bus, the serialization approach with cache-based encoding achieves 31% and 52% power reduction for single-core and multi-core applications, respectively, when using fixed voltage and frequency with 128 bits data bus.

1. Introduction

In the last decades, performance, area, cost and reliability were the major system design concerns of the computer architects and very large scale integration (VLSI) designers, without giving significant attentions to the power reduction issue. However, in the recent years, with the appearance of small integrated circuits which are used to power hand-held devices and complex System-on-Chip (SOC) components used in server farms [1], the power issue has increasingly become a major concern. Also, the appearance of the computing devices (e.g. high-performance desktops, multi-processors, audio and video-based multimedia products) and the portable devices (e.g. laptops, cell-phones, space-based systems, personal digital assistants), which demands high-speed computations and complex functionalities with low power, forms an additional motivation. Moreover, there is a need for high-performance, high-end products to reduce their power consumption.

Power dissipation of Complementary Metal Oxide Semiconductor (CMOS)-based designs can be categorized into three types: (i) leakage power; (ii) short-circuit power; and (iii) switching power. Leakage power, so-called static power, is the result of the parasitic effect. Short-circuit power is the internal power produced because of the short-circuit connection between V_{DD} and the ground. Switching power is the result of charging and discharging of the load capacitance. Indeed, Static power is negligible for an off-chip interconnection bus compared to the other two power components (short-circuit power and switching power) [2]. The general Eq. (1) that represents the dynamic power dissipation in an off-chip interconnection bus is defined as

$$P = \alpha f C V_{DD}^2 \quad (1)$$

where P is the total dynamic power dissipation, α is the signal transition switching activity, f is the operating frequency of the bus, C is the load capacitance of the wire line, and V_{DD} is the swing voltage. Reducing the bus power consumption is usually achieved by using a low swing voltage and operating frequency as well as reducing the capacitance and the switching activity. As each system is configured with a fixed voltage and frequency, wire line reconfiguration can be used to reduce total capacitance of the bus. Load capacitance dominates the total capacitance of the off-chip interconnection bus [3]. Therefore, changing the structure of the wire (i.e., changing the wire width or spacing between the wire lines of a specified material) has little impact on total capacitance reduction. Table 1 [3] shows the capacitance values of different parts of an off-chip data bus. Capacitance reduction through changing wire reconfiguration is possible in only (3 pF+1 pF=4 pF) of the total 20 pF capacitance.

With using fixed voltage and frequency, the solution space of power saving is limited to reduce the amount of switching activity

* Corresponding author.
E-mail addresses: khamadawwad@birzeit.edu (K. Mohammad), ahsan.kabeer@gmail.com (A. Kabeer), ttaha@ieee.org (T.M. Taha), mowaidah@gmail.com (M. Owaida), mahdi.washaha@gmail.com (M. Washha).

http://dx.doi.org/10.1016/j.mejo.2016.06.002
0026-2692/

Table 1
Capacitance values of different parts of the off-chip bus.

Name	Capacitance (pF)
Memory I/O pin	6
Off-chip interconnect	3
ASIC I/O pin	5
On-chip interconnect	1
Bus driver	5
Total	20

for each data transition only. Many studies have been conducted in the field of reducing the amount of switching activity to reduce off-chip memory bus power and overall system power. In doing so, three basic techniques are used listed as: (i) bus encoding; (ii) bus serialization; and (iii) compression, functioning through changing the data values transmitted using the data bus to reduce the amount of switching activity, contributing thereby in reduction of overall power. Also, they can assist in reducing the number of wire lines, eliminating overhead expenses of area. The following subsections present detailed views of different off-chip bus power minimization techniques.

The rest of the paper is organized as follows. Section 2 discusses a set of compression techniques used for off-chip power minimization. Section 3 describes the framework for possible bus power minimization approaches (existing approaches and a proposed new approach) and their efficacy. Section 4 presents our experimental methodology. Section 5 details experimental results and their analysis. Finally, Section 6 concludes the paper with providing future directions in this area.

2. State-of-the-art compression techniques for off-chip bus power minimization

The encoding techniques used for on-chip bus power minimization are also applicable for off-chip memory bus power reduction, with minor modification in switching activity computation. Since the off-chip memory bus power Eq. (1) does not consider coupling capacitance, the computation of coupling transitions can be ignored using these encoding techniques. Compression techniques reduce the number of bus lines, which allows to use widening bus lines and to increase the distance between bus lines. As a result, coupling capacitance effect is reduced significantly, making switching activity the dominant power consumption leak.

For an off-chip memory bus, it is possible to use bus serialization (Section 3) by increasing the operating frequency. Jacob et al. [4] explored DRAM and memory bus organization, finding that it is possible to use different memory bus widths with higher frequencies to maintain the overall bus bandwidth. Rambus [5] is a second implementation of a narrow bus using a higher frequency. Since the off-chip bus does not consider coupling capacitance, bus serialization reduces the number of wire lines, and hence minimizing the area overhead. The off-chip data bus requires calculating the signal transitions to compute the power consumption. Hence, bus serialization can assist in reducing the signal transition switching activity. This reduction in the amount of switching activity primarily depends on the data transition patterns of the application programs and how those transitions are affected by serialization. Serialization can also assist in reducing the number of signal transitions if the serialized data can be used with cache-based encoding, providing the advantage of more data matching probability with serialized data than with full length conventional data [6]. The limitation of bus serialization in increasing the switching activity for a particular data sequence (Fig. 1b) is applicable for an off-chip data bus as for an on-chip data bus. Such increase in switching activity degrades the power budget.

Fig. 1a gives an example of switching activity reduction using serialization for an off-chip data bus. As the Figure shows, a 16-bit data stream passing through a conventional 8-bit wide data bus requires 4 transitions, whereas a serialized 4-bit wide data bus requires only 2. Fig. 1b shows the increase in the switching activity for a 16-bit data stream transmitting through a conventional 8-bit wide data bus and a serialized 4-bit wide data bus. Using the data stream shown, the conventional bus requires 4 transitions, whereas the serialized bus requires 8 for transmitting the same data sequence.

2.1. Compression techniques

There are at least two methods used to reduce off-chip data bus power, summarized as: (i) reduce the number of active data bus lines during each data transmission; and (ii) reduce the number of signal transitions on the active data bus lines. Bus serialization falls into the former method whereas encoding techniques fall into the latter. More power savings can be achieved by using a combination of both. In the research presented here, compression techniques are proposed and implemented to utilize both methods to achieve more power savings. These techniques are applied to the data bus to maximize the bus bandwidth by reducing the number of wires, thus reducing both power and area overhead expenses. Among the several compression techniques widely

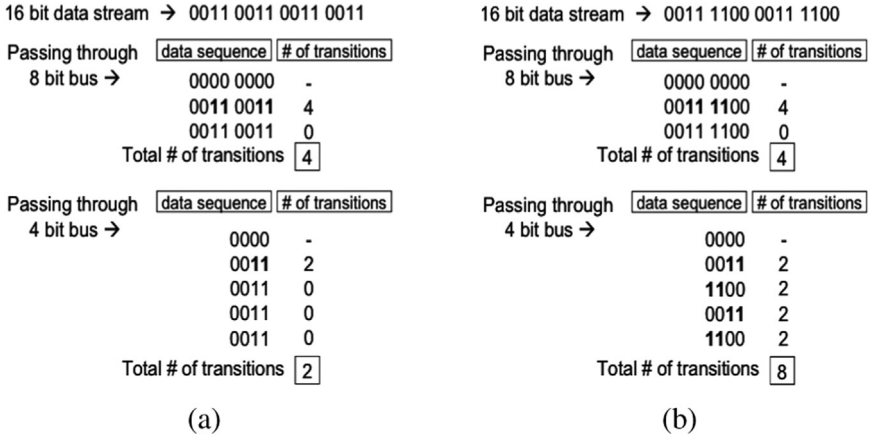

Fig. 1. An example of (a) switching activity decrease and (b) switching activity increase using a serialized off-chip data bus.