CHAPTER 1

Introduction

Interconnect is a set of wires in the circuit which is responsible for communication, clock distribution, and power distribution. Interconnect is a major source of design problems [7]. As predicted by Moores law [23], processor performance is doubling every eighteen months. Increased speed of processor implies that the communication rate of data and instructions must be increased as well. This is not trivial, especially in the global interconnect which is responsible for carrying information between distant locations on the processor. The properties of the wires, such as width, thickness, height, do not scale with technology in a favorable way [10]. Even in the current technologies, signals require several clock cycles to traverse across advanced processor chips.

The speed of light becomes a limiting factor as the clock frequency and chip sizes are increasing. Clock distribution is becoming an issue for the same reasons. The interconnect problem is so severe and so general that Semiconductor Industry Association [2] has devoted a lot of resources to address it.

This thesis concentrates on the communication aspect of interconnect. On-chip communication is realized by circuit devices called buses. A bus consists of parallel, aligned, similar wires that in most cases are on the same metal layer. Drivers and receivers are connected at the two ends of the wires, and in some cases there are intermediate repeaters as well. In modern deep sub-micron technologies, the communication problem has become severe and complex. A reason for this is the increased
coupling between the bus lines [10, 34]. This is because of the smaller distances be-
tween the lines as well as the higher aspect ratio (height/width) necessary to maintain 
a linear resistance of reasonable size. Another problem is the distributed nature of 
the thin and long lines. Both these issues are expected to become even more se-
vere in the future technologies [15, 34]. Transmitting data and instructions on these 
long and wide buses requires a lot of energy and many clock cycles [7, 5, 34]. In 
some widely used circuits [22], the communication power may exceed half of the total 
power consumption.

As on-chip buses in deep sub-micron technology designs have large propagation 
delays and consume significant amounts of energy, minimizing propagation delay and 
energy consumption are one of the most important design objectives.

On-chip interconnect delay and energy consumption depend on the transition 
patterns of data that is transmitted over interconnects as well as the interconnect 
features. Data transition patterns determine both self and coupling transition activ-
ity. Self transitions are due to switching activity on each individual interconnects, 
while coupling transitions change based on the relative switching activity between 
adjacent wires [40]. Coupling transition activity is high when adjacent interconnects 
are transitioning in opposite direction as compared to transitioning in the same di-
rection. Coupling transition activity mainly determines the propagation delay while 
both the self and coupling transition activity determine the energy consumption. As 
CMOS process technology scales down to nano-meter region, the interconnect resis-
tance increases with shrinking interconnect width and the coupling capacitance ($C_I$) 
increases with reducing spacing between interconnects. In deep-submicron technolo-
gies, the coupling capacitance ($C_I$) between long parallel interconnects is of magnitude 
several times higher than the wire-to-substrate capacitance ($C_L$) [40]. The coupling 
capacitance depends on the structural factors such as wire spacing [14], wire width,
wire length [14], wire material, coupling length, driver strength [16], signal transition time. The dominant coupling capacitance has two significant effects: large propagation delay due to opposite transitions on adjacent wires [9, 48, 50] and energy dissipation associated with driving on-chip buses [48].

As propagation delay limits the performance of the system and high energy consumption may lead to reliability and packaging problems, it is very much important to devise techniques for both delay and energy minimization. One way to reduce the propagation delay and energy consumption is to minimize the relative switching activity between adjacent interconnects. Several bus encoding techniques have been proposed in the literature to minimize the relative switching activity [11, 12, 18, 19, 21, 20, 29, 33, 40, 41, 42, 44, 49, 53]. All these techniques use *spatial redundancy*, i.e., encode n-bit data using m bits such that \( m > n \), to minimize power or delay and hence have large area overhead. For example, *crosstalk avoidance codes* [49] eliminate opposite transitions on adjacent wires completely but require 53 wires to encode 32-bit data and a coding technique proposed in [44] is both area and energy-efficient and eliminates opposite transitions on adjacent wires completely but it requires 48 wires for encoding 32-bit data. Most of these techniques consider uniformly distributed random data to study their delay efficiency. But in general, the propagation delay is data dependent and data is application dependent. Different applications may have different data behavior. So, the existing techniques which consider uniformly distributed random data may not exploit the exact data behavior of an application and hence can give inferior performance results. By exploiting data transition behavior of SPEC2000 CPU benchmarks, techniques are proposed in [20, 29], which use variable data transmission delay for propagation delay minimization. None of these techniques actually exploit the data behavior of an application.

3
1.1 Contribution of the Thesis

By exploiting similarity in the data transmitted on on-chip interconnects, we propose three techniques, i.e., data packing \cite{35}, data permutation \cite{35}, and data replication with shielding and two-phase transmission \cite{36}, to minimize on-chip data transmission delay. By focusing on the L1 cache address/data buses, we evaluate the effectiveness of our techniques using SPEC2000 CPU benchmark suit in terms of minimizing delay, energy, and energy-delay-product (EDP). We show that our techniques outperform several existing bus encoding techniques. By combining our techniques with the wire spacing \cite{6} technique, we achieve significant delay and EDP savings over the existing techniques.

Our next contribution is proposing a heuristic solution \cite{38} to the data sequencing problem. Experimental validation reveals that for complete sequences, our technique achieves the minimum switching activity. For incomplete sequences, our technique significantly reduces the switching activity as compared to the random case, where datawords of a given sequence are transmitted in a random order. We show that the running time of our algorithm is $O(n\log_2 n)$, where $n$ is the number of datawords in a sequence.

1.2 Overview of the Thesis

Chapter 2 provides analytical models for propagation delay and energy consumption.

Chapter 3 presents related work in literature.

Chapter 4 discusses our first contribution \cite{35, 36} of the thesis. In this chapter, we describe the motivation to exploit data behavior for delay and energy-delay-product minimization. We then propose three different techniques which exploit data behavior for delay and energy-delay-product minimization. We validate our techniques
using SPEC 2000 CPU benchmarks suite. We also study the impact of wire spacing technique on our techniques.

Chapter 5 discusses our second contribution [38] of the thesis. In this chapter, we solve the data sequencing problem by proposing a heuristic solution.

Chapter 6 discusses future work and concludes the thesis.