Performance Analysis of On-Chip Communication Structures under Device Variability

Faiz-ul-Hassan, University of Glasgow, UK
Wim Vanderbauwhede, University of Glasgow, UK
Fernando Rodriguez-Salazar, University of Glasgow, UK

ABSTRACT

On-chip communication is becoming an important bottleneck in the design and operation of high performance systems where it has to face additional challenges due to device variability. Communication structures such as tapered buffer drivers, interconnects, repeaters, and data storage elements are vulnerable to variability, which can limit the performance of the on-chip communication networks. In this regard, it becomes important to have a complete understanding of the impact that variability will have on the performance of these circuit elements in order to design high yield and reliable systems. In this paper, the authors have characterized the performance of the communication structures under the impact of random dopant fluctuation (RDF) for the future technology generations of 25, 18, and 13 nm. For accurate characterization of their performance, a Monte Carlo simulation method has been used along with predictive device models for the given technologies. Analytical models have been developed for the link failure probability of a repeater inserted interconnect which uses characterization data of all communication structures to give an accurate prediction of the link failure probability. The model has also been extended to calculate the link failure probability of a wider communication link.

Keywords: Communication Structures, Dopant Fluctuation, Link Failure Probability, Network on Chip, Repeaters, Tapered Buffers, Variability

INTRODUCTION

As transistor gate lengths continue to shrink according to Moore’s law (Moore, 1965), designers are now able to incorporate complete systems with complex functionalities in a single System-on-Chip (SoC). As a result, Embedded System design is migrating from board level integration to on-chip level integration, where components are often-times developed by third parties and integrated by the system designer. The increase in complexity of these systems has
introduced a number of challenging problems in the design, project management, simulation and verification of these devices. For instance, according to the ITRS (ITRS, 2003), as the technology scales, the devices are increasingly becoming faster relative to the interconnect. Therefore, modern SoC and Embedded designers are faced with the difficult task of orchestrating the computation of a large number of fast local third party hardware blocks, across the whole chip, by using (relatively) progressively slower interconnects. Consequently, on-chip communication is rendering major constraints to the performance of the system.

Network-on-Chip (NoC) has been proposed as a suitable design methodology (Benini & Micheli, 2002; Dally & Towles, 2001) for modern SoCs which offers an interconnect-centric system architecture to manage the problems of on-chip communication. It works as a small on-chip communication network consisting of typical communication layers such as physical, data link, network, transport and application, each one geared for particular functions. The design of the physical layer is of vital importance as it provides a physical communication media from router-to-router and from router-to-functional units (FUs) for the exchange of data between them. The global interconnect between the routers and semi-global interconnect between the routers and functional units put many challenges for designing them for high performance, especially in deep sub-micron region (Lee & Yoo, 2004). These communication links can either be (a) multi-bit parallel, (b) partially parallel where an \( n \)-bit packet or flit is divided in to a small number of \( m \)-bits or (c) source-synchronous serial (Kim & Sobelman, 2006).

The increase in clock speed and chip size means that synchronization can no longer be obtained by the use of a global clock, and that other models of synchronization (such as the use of multiple clock domains or by the Globally Asynchronous Locally Synchronous (GALS) approach) are required. Effectively the maximum synchronous area is determined by the clock skew. The individual synchronous blocks can communicate with each other synchronously, asynchronously or in a self timed manner. In this paper, we have chosen to focus on cross clock domain communication structures, since due to their high performance and low design effort, it is likely that they will be used in future designs, in which an overlap of different synchronous blocks on the chip communicate with each other synchronously. However, these structures are not completely immune to skew which can develop in the global clock and wide data communication links.

On-chip communication in these systems generally consists of clock distribution network and data links between the functional units. Their performance depends on the performance of the constituting elements like, line drivers, repeaters, interconnect and data storage elements (Flip-flops or latches). Owing to the technology scaling, the variability in the devices and interconnect has emerged as one of the major challenges faced by the silicon industry (Asenov, 2007; Vishvanathan, Ravikumar, & Menezes, 2005; Borkar, 2005; Lin, Spanos, Milor, & Lin, 1998) and the impact of variability on the performance of the circuits has become exceptionally critical (Nassif, 2000; Keutzer & Orshansky, 2002). The variability at deep sub-micron (DSM) region has resulted in the transition from a deterministic design paradigm to probabilistic methods (ITRS, 2005; Narasimhan & Sridhar, 2007) and so there could be large discrepancies between the design and the products being manufactured (Agarwal, Sylvester, & Blaauw, 2006). The impact of process variations is increasing with every new technology generation because process tolerances are not scaling exactly in the same proportion with the design dimensions (Bowman, Duvall, & Meindl, 2002). The variability in the devices and interconnect produces delay uncertainty in the communication structures which can introduce timing uncertainty in the clock and data signals. The technology scaling has pushed the clock frequencies in the gigahertz region in the continuous quest for higher circuit performances and as a result, clock period has reduced to below nanoseconds. Therefore, uncertainty in
Emerging Trends in Digital Libraries: Mobile Technology and Mobile Learning
www.igi-global.com/chapter/emerging-trends-in-digital-libraries/105681?camid=4v1a

A Buffered Dual-Access-Mode Scheme Designed for Low-Power Highly-Associative Caches
www.igi-global.com/article/a-buffered-dual-access-mode-scheme-designed-for-low-power-highly-associative-caches/89261?camid=4v1a