CCIT Report #674 December 2007

Parallel vs. Serial On-Chip Communication

Rostislav (Reuven) Dobkin, Arkadiy Morgenshtein, Avinoam Kolodny, Ran Ginosar
VLSI Systems Research Center, Electrical Engineering Department
Technion – Israel Institute of Technology
Haifa, Israel
[rostikd@tx.technion.ac.il]

***Abstract*—Synchronous parallel links are widely used in modern VLSI designs for on-chip inter-module communication. Long range parallel links occupy large area and incur high capacitive load, high leakage power and cross-coupling noise. The problems exacerbate for applications having low utilization of the links or suffer from congestion of the interconnect. While standard synchronous serial links are unattractive due to limited bit-rate, novel high performance serial links may change the balance. In this paper we show that novel serial links provide better performance than parallel links for long range communications, beyond several millimeters. We analyze the technology dependence of link performance. An example for 65 nm technology is presented, and compare wave-pipelined and register-pipelined parallel links to a high performance serial link in terms of bit-rate, power, area and latency.**

Index Terms—Serial Link, Parallel Link, Asynchronous Circuits, Dual-Rail, Long-Range Interconnect

I. Introduction

Transistor size scaling drastically improves on-chip clock rates, practically doubling the performance every five years [1]. While local interconnect follows transistor scaling, global lines do not, challenging long range on-chip data communications in terms of latency, throughput and power [1]. In addition, as Systems-on-Chip (SoC) integrate an ever growing number of modules, on-chip inter-modular communications become congested and the modules must turn to serial interfaces, similar to the trend from parallel to serial inter-chip interconnects.

Long-range bit-parallel data links provide high data rates at the cost of large chip area, routing difficulty, noise and power. In addition, such links are often utilized only a small portion of the time, but dissipate leakage power at all times. Leakage is incurred at the line drivers and also at the repeaters, which are often necessary for long interconnects [2][3]. Parallel link performance is bounded by available clock rate and by clock skew, delay uncertainty due to process variations, cross-talk noise, and layout geometries.

Bit-serial communications offer an alternative to bit-parallel interconnects, mitigating the issues of area, routability, and leakage power, since there are fewer wires, fewer line drivers, and fewer repeaters. However, to provide the same throughput as an N-bit parallel interconnect, the serial link must operate N times faster. Simple synchronous serial links that employ the system clock are incapable of providing the required throughput. Recently proposed novel wide-bandwidth serial link circuits [4]—[14], which operate faster than the system clock, may deliver the required bandwidth.

Synchronous serial links are typically employed for off-chip communications, where pin-out limitations call for a minimal number of wires per link. Source-synchronous protocols are often used for these applications [15]—[20]. A common timing mechanism for serial interconnects injects a clock into the data stream at the transmitting side and recovers the clock at the receiver. Such clock-data recovery (CDR) circuits often require a power-hungry PLL, which may also take a long while to converge on the proper clock frequency and phase at the beginning of each transmission. If the receiver and transmitter operate in different clock domains, the transaction must also be synchronized at both ends, incurring additional delay and power. Alternatively, an asynchronous data link employs handshake instead of clocks. Traditional asynchronous protocols are relatively slow due to the need to acknowledge transitions [14][21]. In [22] asynchronous protocols share data lines, but their performance depends on wire delays.

High-speed serial schemes, having data cycle of a few gate delays (down to single gate-delay cycle), have been recently proposed [4]—[14]. These fast schemes exploit wave-pipelining, low-swing differential signaling, fast clock generators and asynchronous protocols. In addition, these schemes require channel optimization to support wide-bandwidth data transmission over the link wires. A wave-front train serialization scheme was presented in [11]. The serializer is based on a chain of MUXes (similar to [23]). The link is single-ended and employs wave-pipelining. The link data cycle is approximately $7 \cdot d_4$ (3Gbps@180nm), where d_4 is an inverter FO4 delay. Wave-pipelined multiplexed (WPM) routing