21.7 Distributed Loss Compensation for Low-latency On-chip Interconnects

Anup P. Jose and Kenneth L. Shepard

Columbia University, New York, NY

The delay per unit length of on-chip wires, as determined by a diffusive RC-limited response and as measured relative to gate delays, approximately doubles every technology generation as wire resistances per unit length increase and gate delays decrease with scaling. Furthermore, these wire delays (D) grow quadratically with wire length, $D \propto R_{wire}C_{wire}L^2$. Wire bandwidths, which are inversely proportional to D, degrade.

This problem is traditionally managed by the addition of repeaters, which linearize the relationship between wire length and delay. Optimally repeated Cu wires of minimum width and spacing deliver a relatively constant delay per unit length, increasing from 55ps/mm for 0.18µm technology to approximately 80ps/mm in 35nm technology [1]. However, when measured proportional to gate delay, this delay per mm increases dramatically from 1 FO4 in 0.18µm to 7 FO4 in 35nm. As shown in Fig. 21.7.1 for the 14-mm-long Al wires in the 0.18µm technology used in this work, widening wires can improve the latency of optimally-repeated interconnect by reducing the RC delays associated with sidewall wire capacitance at the cost of degraded energy efficiency.

For off-chip communication, latency, bandwidth, and energy efficiency have been improved by increasing reliance on low-swing current-mode transmission-line-dominated interconnect. These techniques can be applied on-chip if the use of wider top-level wires are combined with signaling modes that emphasize transmission-line behavior [2, 3]. These approaches are ultimately limited in length by wire loss, which grows exponentially with interconnect length. In this paper, distributed negative transconductors (as also employed in distributed amplifiers and oscillators) are applied to compensate for signal attenuation allowing for low-swing transmission-line operation to arbitrary line lengths. Figure 21.7.1 shows the simulated delay and bit energy of such compensated lines (70mV far-end differential swing) assuming a constant attenuation of 0.2dB/mm at 500MHz. Latencies for 8µm-wide wires are ≈ 10ps/mm compared with 17ps/mm for traditional repeated lines. Bit energy, which improves from 6.2pJ/b to 2pJ/b, can be further improved by running the compensated link at higher data rates since most of its power is statically dissipated. Higher bandwidth copper wires will allow the advantages of distributed loss compensation to be achieved with narrower wires.

The basic negative transconductor element, as applied to a differential transmission line (also known as a negative impedance converter, NIC, when first employed in long-distance telephony), is shown in Fig. 21.7.2 and consists of two cross-coupled transistors with resistive and capacitive degeneration. Figure 21.7.2 also shows the expression for the differential admittance of the NIC frequencies $f << f_T$ when the gate-to-drain overlap capacitance is neglected. For g_m= 4mS, R=1kΩ (>> $1/g_m$) and C=600 fF, the zero at 1/2RC (~132MHz) gives a negative admittance that increases with increasing frequency (negative capacitance) until the pole is reached at approximately $g_m/2C$ (~660MHz), delivering loss compensation matching α for the uncompensated line that increases with increasing frequency. Unconditional stability of the compensated interconnect also requires a choice of R and C such that the attenuation constant α be greater than zero for all frequencies [4].

A prototype DDR 3Gb/s 14mm link (on the fifth metal level of a six-level metal process), as designed and fabricated in a 0.18µm CMOS technology, is shown in the die micrograph of Fig. 21.7.3 and schematically in Fig. 21.7.4. The interconnect has a coplanar waveguide topology with a line-width and spacing of 8µm with seven NICs evenly spaced on the line. Differential operation assures controlled inductance, high common-mode noise rejection, and reduced shielding requirements. The common-mode voltage at the driver sets the operating point for the NICs. Because of resistive losses in the line, bias currents decrease for NICs toward the far-end of the line, the devices of which are sized larger to provide uniform g_m. There is an identical link running below the main link (on third-level metal) to characterize the effect of coupled noise. The driver and receiver components must operate at the same frequency (mesochronous) although arbitrary skews are accommodated with an automated calibration at start-up. The link is doubly terminated with n-type diffusion resistors of value $2Z_0$.

A current-mode driver with pseudo-NMOS predriver, shown in Fig. 21.7.5, with input multiplexing to achieve DDR operation is employed. DDR operation, with the modest energy overhead of skewing and deskewing latches, helps to compensate for the reduction in bandwidth per unit wiring area associated with wider wires. There are multiple copies of the predriver/driver with varying sizes to dynamically control the drive current (I_d) from 3.0 to 6.0mA in steps of 0.35mA. Larger driver currents boost signaling levels as well as increase the g_m of the devices in the NICs, improving interconnect bandwidth.

The receiver consists of two differential sense amplifier latches (Fig. 21.7.5) with a digitally trimmed capacitive load for input-offset calibration and with a typical aperture time of around 15ps. 4.5µm transistors at the data inputs give a low Z_0C_L time constant (2ps).

Other circuits used in the design to facilitate testing and characterization include a PRBS generator (consisting of a 17b LFSR) and a 512×16 bit SRAM to generate arbitrary input patterns. Picoprobe pads are included at both ends of the link for probing the waveforms and for direct network analysis.

The measured eye diagram at the far end of the line for both I_d = 3mA and I_d = 6mA is shown in Fig. 21.7.6 with a measured BER of $< 10^{-14}$ at 3Gb/s and I_d = 3mA. The effect of introducing the NIC elements on the measured transmission line parameters of the link is shown in Fig. 21.7.7 for both the real and imaginary parts of the propagation constant, $\gamma=\alpha+j\beta$. These values determine a wire delay of 10.7ps/mm with the NICs present but unbiased and 12.1ps/mm with the NICs biased on. This compares with a latency of more than 18.6ps/mm for optimally repeated RC signaling (Fig. 21.7.1) on the same wire (from S-parameter simulation). The link consumes 2pJ/b for I_d = 3mA, which is significantly lower than the 7pJ/b (from S-parameter simulation) consumed by an optimally buffered link (Fig. 21.7.1).

References

[1] R. Ho, K. Mai and M. A. Horowitz, "The Future of Wires," *Proceedings of the IEEE*, pp 490-504, Apr., 2001.

[2] A. P. Jose, G. Patounakis and K. L. Shepard, "Near Speed-of-light On-chip Interconnects using Pulsed Current-mode Signaling," *Symp. VLSI Circuits*, June, 2005.

[3] R. T. Chang, et al, "Near speed-of-light signaling over on-chip electrical interconnects," *IEEE J. Solid-State Circuits*, vol. 38, no. 5, May, 2003.

[4] M. Bussmann and U. Langmann, "Active Compensation of Interconnect Losses for Multi-GHz Clock Distribution Networks," *IEEE J. Solid-State Circuits*, vol. 39, no. 11, Nov., 1992.

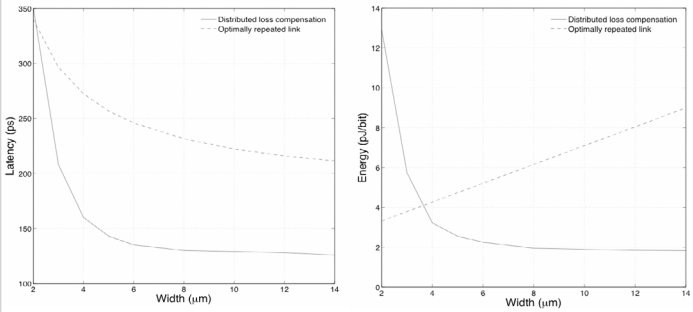

Figure 21.7.1: Wire latency and bit energy (pJ/bit) as a function of of wire width for two 14mm links operating at 3Gb/s -one with repeaters and one with distributed loss compensation.

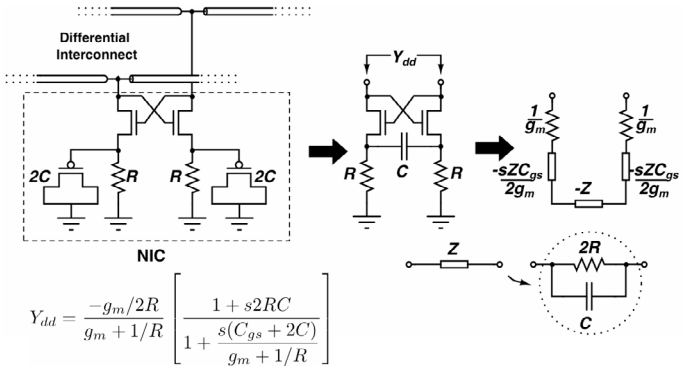

Figure 21.7.2: Circuit implementation of the NIC.

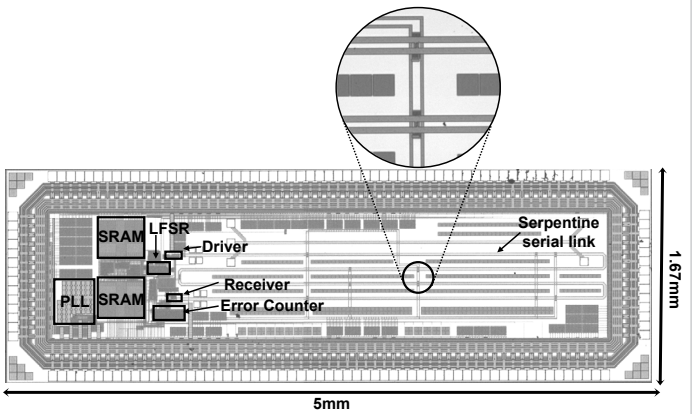

Figure 21.7.3: Die micrograph of the prototype chip in 0.18μm technology with two of the NICs shown in inset.

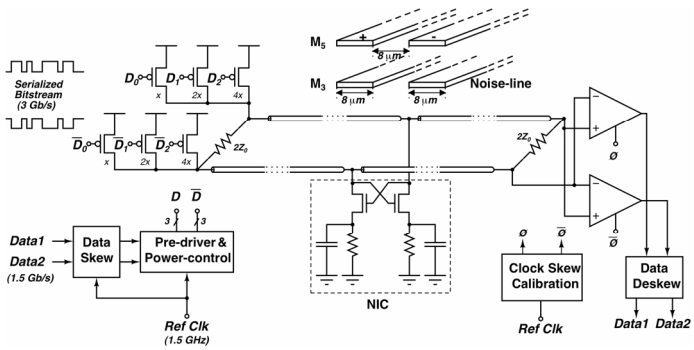

Figure 21.6.4: Overall system architecture of the prototype link.

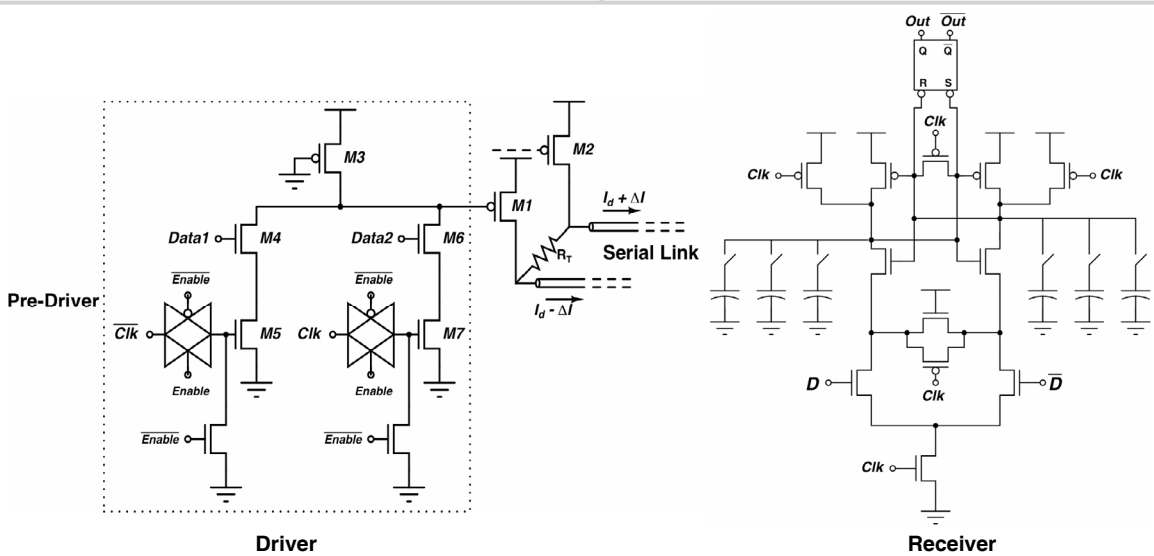

Figure 21.7.5: Driver and receiver design.

Continued on Page

21

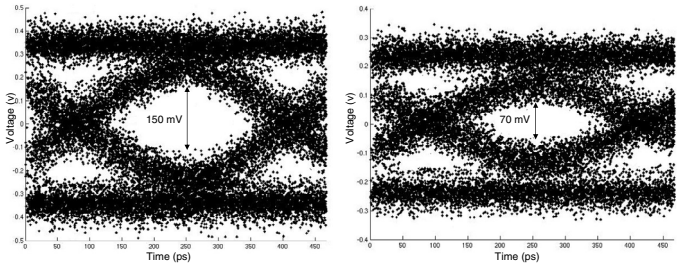

Figure 21.7.6: Measured far end eye-diagram at 3Gb/s for maximum and minimum power.

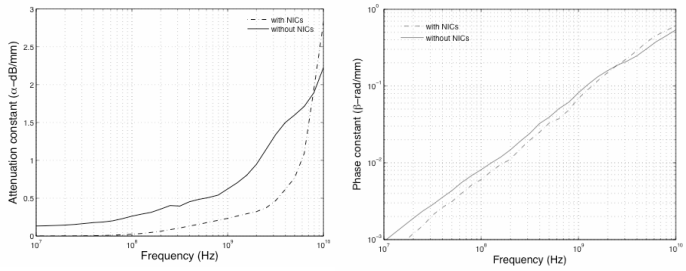

Figure 21.7.7: Measured attenuation (α) and phase constant (β).

