Extended Abstracts of the 2004 International Conference on Solid State Devices and Materials, Tokyo, 2004, pp. 134-135

I-1-1 (Invited)

CMOS Scaling on I/O Design

Chih-Kong Ken Yang, Mayank Garg, Jason C.S. Woo,
University of California, Los Angeles
56-147A Engr IV, 420 Westwood Plaza, Los Angeles, CA, 90095-1594, USA
Tel: 310-206-3665, E-mail: yang@ee.ucla.edu

1. Introduction

CMOS transistor scaling has provided increased transistor density and digital switching speeds. However, the performance of large digital systems-on-chip (SoC) depends greatly on key mixed-signal blocks that are considerably analog in nature such as clock generation blocks, voltage references and regulators, and high-speed I/O.

The focus of this paper is on the I/O subsystem. Because of the data-bandwidth bottleneck, the I/O subsystem is critical to the system performance. In order for system performance to improve, I/O data rates must continue to increase with device scaling. Section I reviews the key components of an I/O subsystem and describes the historical scaling trend of the off-chip data rate based on published results. The data rates have been strongly related to the speed of CMOS transistors.

Digital scaling focuses on improving switching characteristics such as I_{off}/I_{on} and $R_{S/D}$. For mixed-signal performance, this paper extends existing analysis of device scaling to circuit characteristics such as g_m, r_o, and intrinsic gain. Section II describes the scaling trends down to the 45-nm channel lengths. Section III draws implications of the scaling trends on the circuit components. Whether or not I/O data rates continue to scale as predicted by the ITRS roadmap depends on how well some key challenges are addressed.

2. I/O Subsystem

Figure 1 illustrates a simplified block diagram of an I/O subsystem. The two primary elements are the transmitter and receiver. The transmitter comprises of a clock-generation block that produces a high-frequency timing signal, and a signal-conditioning block that converts the digital data into an analog output waveform. Typical signal conditioning includes a driver that produces the proper output levels and impedance, a predriver that shapes the waveform, and a serializer that sends the signal with the appropriate timing. The receiver comprises of a sampling-and-amplifying block that converts the weak received signal into quantized digital values and a closely-related clock-and-data recovery block that produces the correct sampling clock edge(s).

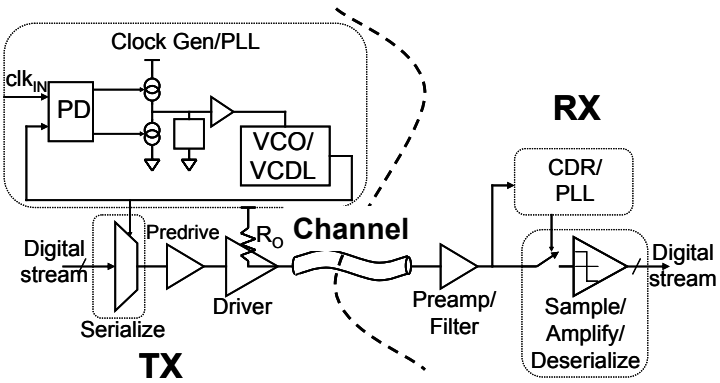

Figure 1 I/O subsystem

Both clock generation and recovery blocks typically use a phase-locked loop (PLL or DLL). The circuit uses a variable oscillator (VCO) or delay-line (VCDL) that produces an arbitrary clock phase. To control the VCO/VCDL, the phase difference between the input data/clock is compared with the internally-generated clock. The phase error is filtered and typically integrated into a control voltage.

Each component has a wide variety of circuit designs and architectures. Within the past 10 years, data rates in CMOS exceeded a gigabit per second (Gbps); the subsequent increase in data rate led to additional complexity in order to cope with bandlimited channels such as signal modulation and equalization. Most of the designs share the basic requirements: 1. high bandwidth to support the data rate, 2. fine voltage resolution at the receiver, 3. small sampling aperture, 4. small offsets and low noise for both voltage and timing, 5. voltage and power gain for signal amplification and drive strength, 6. accurate integration and capacitive switching for filtering, 7. reasonable linearity, and 8. low supply and substrate sensitivity.

The scaling of I/O performance relies ultimately on the scalability of analog device characteristics. Historically, I/O performance has scaled relatively well. Figure 2 illustrates a scatter plot of published link-related performances.

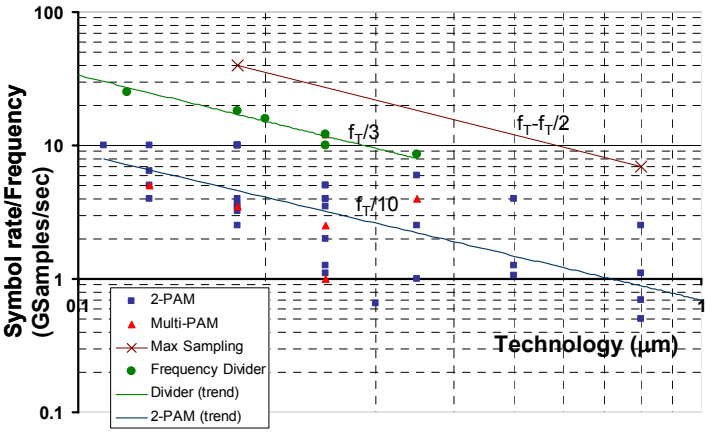

Figure 2 Scaling of published speed of links and frequency dividers with technology

The trend has been favorable because the previously listed design requirements for up to a few Gbps have not been severe. For instance, resolution of 100mV is adequate for non-equalized I/O channels. With symbol time (or bit duration) of 1ns, aperture and 6σ jitter of a few hundred-ps still leaves sufficient timing margins. The amount of excess bandwidth needed by the electronics is not large because little equalization and modulation is needed. Hence, I/O performance has been dominated by how quickly transistors can switch which is then strongly coupled to f_T.

To the first order, we can use f_T scaling to estimate an upper bound on the scaling of I/O data rates. Based on device simulations, Figure 3 shows the predicted f_T for CMOS devices down to 50-nm channel lengths. Note that f_T scales inversely proportional to L_{eff} (as opposed as to L^2). As will be shown, the slower scaling is due to the sub-linear g_m scaling and also the increasing impact of overlap capacitance (C_{GOV})

on the total gate capacitance. f_T only leads us to an upper bound for data rate because design requirements have tightened considerably as data rates approach 10s of Gbps.

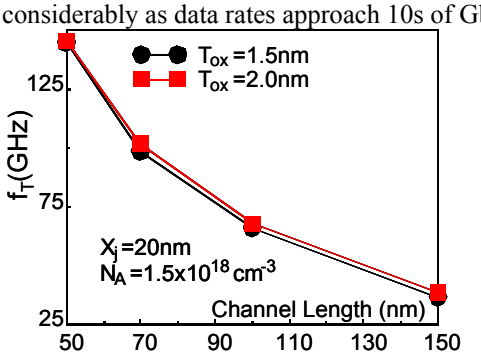

Figure 3 Impact of scaling on f_T

3. Scaling of Device Characteristics

As I/O data rates increase, link-design requirements have similarly become more stringent. Especially when large amount of equalization is necessary and multiple voltage levels are needed, the timing requirement, σ_{jitter}, can be sub-picosecond, and voltage resolution is only a few millivolts. The analog characteristics of the scaled devices must be included in our analysis. This paper extracts the analog characteristics from device simulations. The device simulation data are fitted to BSIM-like equations. The analysis includes a range of values for many device parameters such as oxide thickness (T_{ox}), substrate doping concentration (N_A), and junction depth (X_j). These parameters allow proper modeling of short-channel effects (SCE) and threshold voltage (V_{th}).

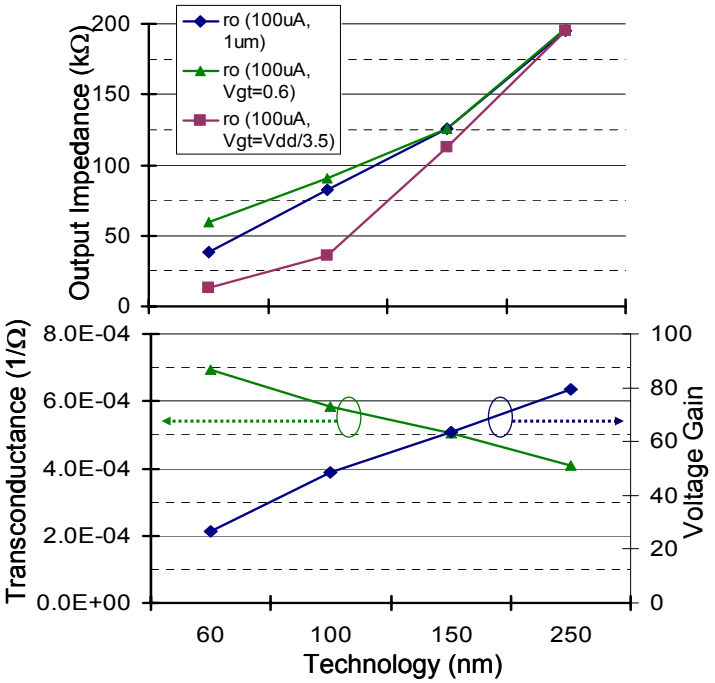

Figure 4 Scaling of intrinsic r_o (top) and g_m and voltage gain (bottom). r_o scaling includes several conditions.

Figure 4 illustrates the impact of scaling on two fundamental circuit parameters, g_m and r_o.[1] The figure shows the severe impact of SCE. g_m no longer scales proportionally with length scaling. The more dramatic trend is the reduced r_o.

4. Design Implications

Because the f_T of devices still scales, we can expect the performance of many of link building blocks to scale. Ring-oscillator frequency relates to the unity-gain bandwidth; the regeneration gain of clocked comparators and frequency dividers depends on g_m/C; samplers have sampling bandwidth and intrinsic aperture related to $1/R_{on}C$. However, as one would expect, the degrading gain and r_o can be expected to impact the scaling of I/O data rates.

A well-known problem is the increased power dissipation due to noise. In order to sustain the smaller dynamic range (from V_{DD} scaling) and higher bandwidth, bias currents (and hence power) must increase to reduce the input-referred voltage noise, $\propto\sqrt{\gamma/g_m}$, proportional to the signal voltage. Our analysis of scaled device characteristics indicates that the situation is even worse since g_m scales sub-linearly. A similar equation can be written for jitter, i.e. $\propto 1/i_{DSAT}\sqrt{(g_m\gamma)T_{ref}}$ for inverter oscillators where T_{ref} is the period of the reference clock. Assuming that the reference frequency increases, power must still be increased in order to scale jitter proportionally. While the sub-linear g_m scaling helps jitter scaling, because of limited crystal frequencies, T_{ref} is not likely to scale and power must increase substantially.

The dramatic decrease in r_o has several important implications on I/O design. The most apparent is that intrinsic gain is <20. The low intrinsic gain limits the application of simple differential pairs. Many current designs of pre-amplifiers and equalizers must be reconsidered to provide sufficient high-frequency gain. Additionally, many mixed-signal I/O blocks rely on the r_o of current sources for supply-noise rejection. With a g_m/g_{ds} <20 for a simple different amplifiers, the PSRR is <26dB. Lastly, the low r_o incurs large error currents in integrating elements (i.e. integrating receivers or charge-pumps in the PLL). Circuit solutions exist but they have additional costs and requirements. Multi-stage amplifiers can be used but not for feedback structures. Longer channel lengths, cascoding, and gain-boosting can increase output impedance. However, they require more voltage headroom and significantly more power. As shown in Figure 4, if voltage margins scales, r_o decreases even more quickly.

5. Conclusion

The paper describes how links have scaled historically and how such scaling is not likely to continue with current scaling trends of digital devices. An analytical model has been developed to understand the analog circuit characteristics of deeply-scaled devices. Our result shows f_T scaling by 1/L which enables the performance scaling of many link components. However, from our model of the device r_o and g_m, serious challenges face I/O designers. In an increasingly power-constrained environment, the current approach to noise and offset problems is not viable. Voltages can not continue scaling leading to device issues. Future solutions will involve the joint effort of devices (both passive and active elements), circuits (designs with lower inherent noise), and architectures (for power and signal integrity).

[1] Note that while device characteristics can be highly susceptible to the device parameters, we show only a very small subset of the data that properly illustrate the trends (i.e i_{bias}=100μA or V_{gt}=0.6V, X_j=10nm and N_A=1.5x10^{18}/cm^3). In fact, our choice of N_A implies scaling that is favorable to analog characteristics in maintaining a higher r_o at the expense of V_{th}.