REALITY

Reliable and Variability tolerant System-on-a-chip Design in More-Moore Technologies

Contract No 216537

Deliverable D3.4

Report: Techniques for enhancing interconnect variability tolerance

Version 1.1

Editor:	Andrea Acquaviva
Co-author / Acknowledgement:	Giacomo Paci, Luca Benini, Antonio Pullini
Status - Version:	V1.1
Date:	12/10/2009
Confidentiality Level:	Public
ID number:	IST-216537-WP3-D3.4-v1p1

The REALITY Consortium consists of:

Interuniversity Microelectronics Centre (IMEC vzw)	Prime Contractor	Belgium
STMicroelectronics S.R.L. (STM)	Contractor	Italy
Universita Di Bologna (UNIBO)	Contractor	Italy
Katholieke Universiteit Leuven (KUL)	Contractor	Belgium
ARM Limited (ARM)	Contractor	United Kingdom
University Of Glasgow (UoG)	Contractor	United Kingdom

Disclaimer

Acknowledgements

The editor Andrea Acquaviva, Giacomo Paci and Luca Benini acknowledges contributions by Antonio Pullini.

Preface

The scope and objectives of the REALITY project are :

- Development of design techniques, methodologies and methods for real-time guaranteed, energy-efficient, robust and adaptive SoCs, including both digital and analogue macro-blocks“

The Technical Challenges are :

- To deal with increased static variability and static fault rates of devices and interconnects.
- To overcome increased time-dependent dynamic variability and dynamic fault rates.
- To build reliable systems out of unreliable technology while maintaining design productivity.
- To deploy design techniques that allow technology scalable energy efficient SoC systems while guaranteeing real-time performance constraints.

Focus Areas of this project are :

- “Analysis techniques” for exploring the design space, and analysis of the system in terms of performance, power and reliability of manufactured instances across a wide spectrum of operating conditions.

- “Solution techniques” which are design time and/or runtime techniques to mitigate impact of reliability issues of integrated circuits, at component, circuit, architecture and system (application software) design.

The REALITY project has started its activities in January 2008 and is planned to be completed after 30 months. It is led by Mr. Bart Dierickx and Mr. Miguel Miranda of IMEC. The Project Coordinator is Mr Tom Tassignon. Five contractors (STM, ARM, KUL, UoG, UNIBO) participate in the project. The total budget is 2.899 k€.

Abstract

In this report we first analyse the impact of variability on chip interconnect and then we describe the design and implementation architectural solutions for compensating variability effects. Both systematic and random variations have been considered. Experimental results show the impact of the implemented techniques in terms of link delays and power consumption. This deliverable summarized the work done by UNIBO and STM as part of Task 3.4. The solutions have been implemented using 65nm ST technology library. They will be integrated in the final flow within WP5 and validated within WP6.

List of Abbreviations

REALITY	Reliable and Variability tolerant System-on-a-chip Design in More-Moore Technologies
CAD	computer aided design
DLC	
DMT	discrete multi-tone
DSP	digital signal processing
FFT	fast Fourier transform
HW	Hardware
IC	integrated circuit
MPSoC	Multiprocessor System-On-Chip
QoS	quality of service
SoC	system on chip
SOHO	small office/home environment
SW	Software

List of Figures

Table of contents

1 Introduction

In this deliverable we report the main achievements of the work conducted for Task 3.4: “Development of architectural solutions for variability management in on-chip interconnect fabrics”.

In this report we discuss the analysis of the impact of variability on on-chip interconnects and the innovative solutions we developed within the REALITY project to compensate its effect on power and performance. These solutions are based on AVS and FBB techniques, that have been discussed in D3.3. In this report we show the implementation of variability compensation techniques to on-chip interconnection links. We considered both full-swing and low-swing channels as they circuit characteristics also determine an increased or decreased sensitivity of their performance to the different compensation mechanisms.

To accomplish our work, we collaborated mainly with STM, that provided technology information and industrial feedback on the applicability of the proposed solutions on real systems.

The solutions described in this report have been tested on a 65nm technology. FBB and AVS techniques will be ported on target technology as part of WP5 and validated in WP6

2 Overview of Techniques for Variability Management in Interconnects

Post-silicon tuning allows to combat the impact of variations on performance and power consumption through the adjustment of device characteristics after a die has been manufactured to compensate for the specific deviations that occurred on that particular die [24], [25]. Two main techniques have been developed in the past and have been discussed in D3.3. One of the methods utilizes the transistor body effect to change transistor threshold voltage by applying an adaptive body bias (ABB) to chip devices to modulate performance and power [33], [24]. The other method of performing post-silicon tuning is to adjust the supply voltage (ASV) to trade performance with power, thus achieving a similar effect to ABB in spite of the different physical mechanism, implementation overhead and trade-off curves.

The effectiveness of ABB and ASV in reducing variability has been assessed and compared mainly on combinational logic circuits [27], key elements of microprocessor critical paths[25] and ring oscillators[23], sometimes achieving counterintuitive and even conflicting conclusions [27] [27] [25]. The reason for this is that the effectiveness of ASV and ABB cannot be generically assessed, but it has to be referred to the variance of a specific manufacturing process and to the performance and power tuning requirements of the design at hand.
In D3.3 we reported the implementation of two innovative post-silicon variability compensation techniques based on both AVS and ABB for microprocessor data-path.

However, with the advent of multi-core integrated systems, the assessment of post-silicon variability compensation techniques cannot be limited to the traditional test benches of past research any more, such as combinational logic circuits or even microprocessor circuit sub-blocks. In fact, the new architecture trend requires long (global) interconnects for the connection of system-level blocks with each other. Unfortunately, physical properties of these on-chip interconnects are not scaling well with feature sizes, and they are becoming a key limiting factor for performance, reliability and timing closure of the whole system. A common practice is to overcome the effects of interconnect reverse scaling by means of circuit-level techniques, so that on-chip interconnects cannot be viewed as simple on-chip wires any more, but rather as communication channels
including complex drivers and receivers [6], [22]. Analyzing the impact of process parameter variations on the performance and reliability of these communication channels and exploring effective means for their compensation is a key design issue.

The relative effectiveness of ABB and ASV in this domain may greatly depend on the specific circuit implementation of the communication channels. A traditional design technique for long links consists

of inserting equally spaced CMOS repeaters to deal with resistive loss along the wire. However, with the increase in number and density of the wires with each new technology, interconnect area and power are severely impacted [1]. The most effective technique for global interconnects to achieve significant power savings and energy-delay efficiency is to reduce the voltage swing of the signal on the wire [18] and, possibly, to avoid the use of repeater stages, like in [11]. On the other hand, low-swing signaling reduces noise immunity and poses non-trivial circuit design challenges.

Many previous works in the open literature, like [30], compare power, area and delay of full-swing vs low-swing
communication links. The novel contribution of this work is to compare the two signaling schemes from the viewpoint of their robustness to process variations. We distinguish between an inherent robustness, associated with the characteristics of the specific circuits building up the communication channels, and the robustness achieved as an effect of variability compensation.

The different circuit properties of full-swing and low-swing channels also determine an increased or decreased sensitivity of their performance to the different compensation mechanisms. Therefore, knowledge of the delay tuning range of ABB and ASV does not suffice to discriminate between them, since other effects need to be taken into account. First, for a given process variation scenario, the amount of induced delay variability is circuit-dependent, therefore making even the weakest (and typically most power saving) tuning mechanisms attractive for the most robust channels. Second, the sensitivity of channel performance to that of specific critical sub-blocks may be exploited to amplify the tuning capability of a variability compensation technique.

The work developed in this project considers the compensation efficiency - cost tradeoff by evaluating local circuit-level costs incurred by the compensation mechanisms. In practice, the power overhead of the compensated communication channels is considered, caused by the modified supply or body voltages. Other system level costs, associated with the availability of multiple biasing voltages or their distribution across the chip, are not considered here and are left for future work.

In this report we identify the most promising compensation technique for each kind of communication channel and variability scenario, so to justify an effort for its system-level realization later on. In our study, the effectiveness of variability compensation techniques when applied to on-chip links is assessed in two steps. At first, the inherent effectiveness of the compensation mechanisms for the channel at hand is investigated. Later, it will be analyzed how such an effectiveness is impacted by layout effects in real life designs, especially crosstalk. Our objective here is to investigate the interaction between crosstalk effects and the behavior of the compensation mechanisms.

All our tests were conducted on an STMicroelectronics 65nm technology and our findings apply to generic on-chip communication channels. In is planned as part of WP5 the demonstration of these techniques on the target technology of REALITY project. Finally, without lack of generality, given the emerging role of networks-on-chip (NoCs) as reference interconnect fabrics for MPSoC platforms [29], we selected the links used for switch-to-switch connectivity in NoCs as our experimental case study.

3 Literature Review

Most research on low-swing interconnects is focused on designing circuit structures with minimal impact on delay, area and power, so the inherent advantages of low-swing signaling are not swamped by transmitter and receiver overhead. An overview of drivers and receivers is illustrated in [18], [4]. [4] makes a comparison with traditional CMOS circuits and is one of the few papers dealing with repeater stages in low-swing interconnects. The use of repeaters is avoided in [20] by means of a swing limiter and an interconnect accelerator at the receiver. Carefully engineered voltage level converters are proposed in [3], [21], while an optimized level restoration scheme based on bootstrapping can be found in [9]. Sense amplifiers are commonly used to detect a small voltage swing in reduced-swing buses [18], [5]. The minimum interconnect swing should be set by the need to overcome noise at the receiver. An adaptive sensing scheme is proposed in [17] to reduce the threshold voltage offset between a driver and a receiver and ensure low-swing reliable operation. An adaptive voltage swing is set at circuit initialization in [16] to drive interconnects based on their delay, thus coping with the increasing interconnect delay spread. To the limit, a self-calibrating interconnect can be designed [15],

[7]. Differential current-mode signaling schemes have a distinctive advantage over the single-ended ones in terms of noise immunity and signal integrity [13]. Neighbor-to-neighbor crosstalk can be reduced with twists in the differential interconnect pairs [10]. Differential low-swing interconnects come at the cost of a significant area and power overhead, therefore are not considered in this work. Current variation models tend to ignore variations in wires [32], however the spread of technology parameters may jeopardize functionality of transmitting and receiving circuits, causing communication performance degradation or even failure. The traditional techniques for post-silicon compensation of variability are adaptive body biasing (ABB) [24], [28] and adaptive supply voltage (ASV)[26]. Comparative studies of ABB vs ASV when put at work for variability compensation in microprocessor sub-circuits or generic combinational logic circuits have not reached a unique conclusion, proving that the choice is tightly design- and technology-dependent. In [25], [27], [23] there is consensus on the fact that ASV has a larger tuning range of circuit properties and the combined use of ASV and ABB further extends this range. However, the measured yield improvements are different depending on the technology and the design at hand, so it is not unambiguous whether hybrid approaches are worth the cost or not. In many cases, ABB seems to suffice for the required range of post-silicon compensation. Only for core-to-core variations ASV seems the best option [8]. [19] points out the dependence of ABB and ASV efficiency on the device type and operating temperature, while [27] emphasizes the role of biasing resolution as well.
The work conducted in REALITY project aims at extending the analyses performed so far to the link architectures for on-chip communication. First, the intrinsic robustness of full-swing vs low-swing signaling schemes to process variations will be explored. Second, ABB and ASV will be applied to find out to which extent they can restore the nominal performance of sample communication channels affected by process variations and what is the power cost incurred for this compensation. Third, it will be demonstrated how the above results are impacted when crosstalk is considered. For this purpose, a full 32 bit link was placed and routed on the target 65nm technology and a standard industrial tool for parasitic extraction was used. The ultimate objective of our analysis is to characterize the effectiveness of the traditional variability compensation techniques when applied to on-chip communication channels under real-life layout effects.

4 Communication channel design

We at first present the design of the communication channels that will be assessed later on in terms of robustness to process variations and suitability for traditional post-silicon compensation techniques. Without lack of generality, we restrict our analysis to an intermediate layer wire with a length of 2mm, which is already the typical length of a switch-to-switch link in a regular network-on-chip architecture [29]. Inserting repeaters to minimize delay of a wire is effective only when the wire is at least twice as long as the critical length of the technology and of the specific routing layer. In our target 65nm technology, a 2mm wire falls below this threshold and the choice is therefore for an unrepeated interconnect. Even for longer links, solid network-on-chip implementation works like [11] suggest the use of unrepeated wires for the point-to-point communication links between switches, unlike other scenarios where high fan-out nets are required. To the limit, link pipelining can be used to break long timing paths.

Following these indications, this work assumes the use of unrepeated wires for network-on-chip communication. We model the on-chip wire in HSPICE with a _3 distributed RC model. Interconnect parameters are taken from the predictive technology models for a 65nm node [14], while the transistor models to design link drivers and receivers are taken from the ST-Micro technology library. At first, we assume that cross-coupling capacitance is tackled by means of physical-level techniques such as shielding or proper wire spacing, therefore no crosstalk effect is modeled at this time. The interaction between crosstalk and variability compensation will be studied in section VI. The reference link architecture uses a 1V *full-swing signaling* (Fig.1.left). The driver consists of a (minimum sized) library flip-flop and a chain of buffers sized based on the exponential horn methodology for minimum delay. The receiver is yet another library flip-flop.

The alternative communication scheme is the *low-swing pseudo-differential (PDIFF) interconnect* architecture reported in Fig.1.right. The voltage swing is chosen to be 200mV. The basic circuit is taken from [18]. The driver is an NMOS-only push-pull driver which allows the use of very low power supplies and a quadratic energy reduction as a function of the voltage reference/swing Vref . The

receiver is still clocked but requires the voltage reference as an additional input. The original receiver circuit proposed in [18] is the clocked sense amplifier followed by a static latch illustrated in Fig.2.left. This pseudo-differential scheme uses single wire per bit while still retaining most advantages of differential amplifiers such as low input offset and good sensitivity. The major reliability degradation may come from the local device mismatch between the double input transistor pairs and from the variation between distant references of the driver and the receiver. In contrast, receiver operation is largely insensitive to Vdd supply noise, as opposed to other schemes. This was the basic motivation for selecting this scheme from [18].

However, we apply some improvements to this receiver, ending up with the circuit in Fig.2.right. First, PMOS transistor P6 in Fig.2.left has the task of equalizing the connected nodes, however it remains active even after the initialization, thus slowing down node transients. Moreover, it is not very conductive when the connected nodes reach an initialization value approaching its voltage threshold. In Fig.2.rigth it has been replaced by an NMOS transistor driven by the clock, thus achieving a better equalization and a faster node transition. Second, although the NOR static latch in Fig.2.left appears to be symmetric, it features uneven 0-to-1 and 1-to-0 switching times. Balancing rise and fall times makes the circuit actually asymmetric. The solution in Fig.2.right allows an easier balancing of these times while keeping the cross-coupled inverter pair fully symmetric: the outputs of the pseudo-differential receiver in fact directly drive the transistors (dis-)charging the flip-flop output capacitance, while the cross-coupled inverter pair keeps the sampled values. Output capacitance for the differential signal was tuned to be the same for POUT and POUTN signals. As a side effect, the flip-flop in Fig.2.right turns out to scale better from a performance viewpoint and enables higher operating frequencies for a comparable area than that of Fig.1b.left.

Transistor sizing for the low-swing communication channel is done to keep the same (maximum) performance of the full-swing interconnect (1.68 GHz): *driver sizing is used to achieve the same link propagation delay, while receiver and static latch sizing is used to enforce the same clock propagation*, so that the next logic stage fed by the communication channel is impacted in the same way. In particular, the technology library constraints for such propagation time have been enforced.

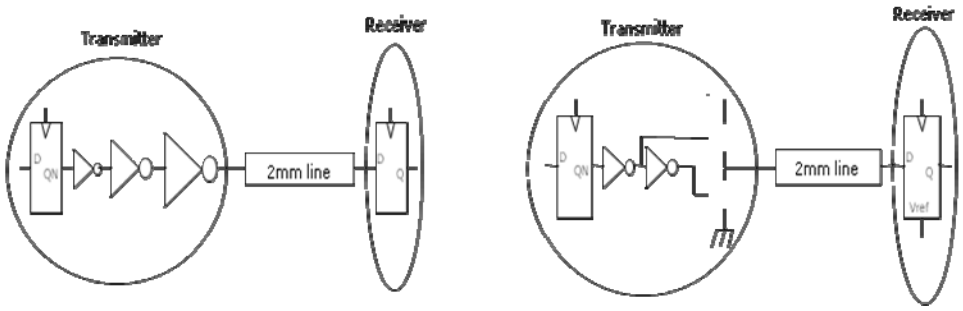

Figure 1: Link Model. On the left: the state of the art solution with high noise margin and high power consumption. On the right: low-power consumption solution with low noise margin.

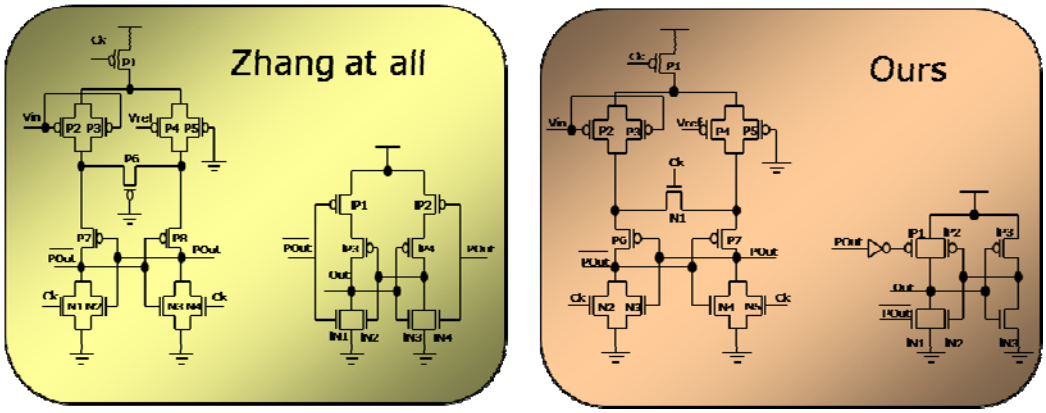

Figure 2: Pseudo-differential flip flop. On the left: PDIFF low-swing receiver from [18]. On the right: Optimized PDIFF low-swing receiver.

4.1 Link Characterization

This section characterizes power and area of the full-swing vs PDIFF low-swing signaling schemes in order to provide the same target performance (1.68 GHz). Power results with 100% input switching activity are reported in Fig.3. Our low-swing channel consumes almost 5x less power than the full-swing one, confirming its power efficiency. Most of the power savings obviously come from the driver and from its reduced. reference voltage. The input flip-flop is the same, and so is the power. Moreover, the PDIFF receiver almost equals the power of the library flip-flop in the full-swing scheme, which was chosen with the minimum driving strength. By measuring idle power (0% input switching activity and clock on), the low-swing channel turns out to have higher
power consumption due to the dynamic and differential nature of the PDIFF scheme: at each clock cycle, one of the two branches of the receiver has to switch. By progressively increasing the switching activity of the link, we found out that the superior power efficiency of the low-swing link over the full-swing one is materialized when the switching activity is above 5%.

Low-swing signaling also achieves 28.5% lower leakage power. Most of the savings come again from the driver, but also the PDIFF receiver has a lower leakage than the library flip-flop, due to the power gating PMOS transistor in precharge mode and to the minimum area NMOS transistors that are switched off in evaluation mode. As regards area, the low-swing channel has a negligible 1% increase in area. The low-swing receiver has a slightly larger area than the library flip-flop, which is counterbalanced by the lower area footprint of the low-swing driver. Please observe that the PDIFF receiver consumes the same total power of the library flip-flop with more area, and this is due to the fact that some of its internal nodes switch with a lower swing.

Finally, by modeling and simulating wire lengths larger than 2mm, we got almost the same quadratic delay increase for the full-swing and the low-swing interconnects as shown in Fig.4, since the charging/discharging time constant stays the same. Given a target frequency for a network-on-chip design, the NoC must ensure a maximum link length, eventually enforced by applying link pipelining techniques

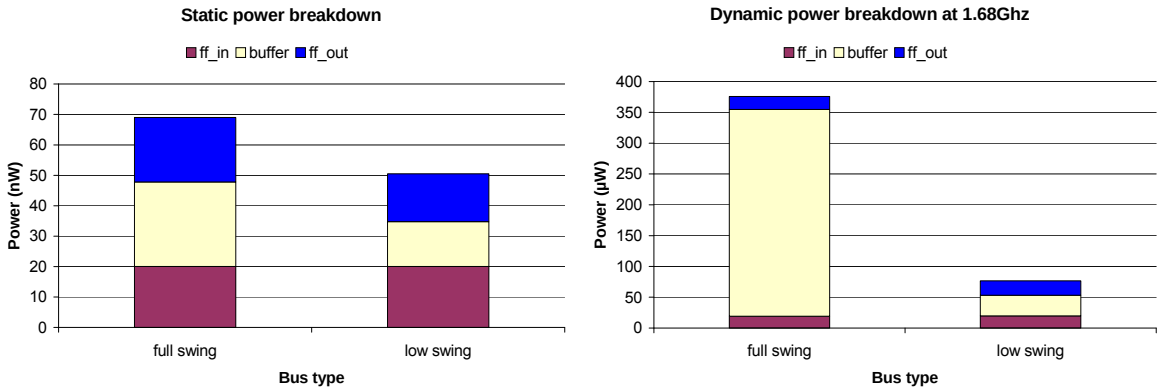

Figure 3: Static (left) and dynamic (right) power breakdown with 100% switching activity at 1.68 GHz, i.e. the maximum performance achievable by full-swing signaling

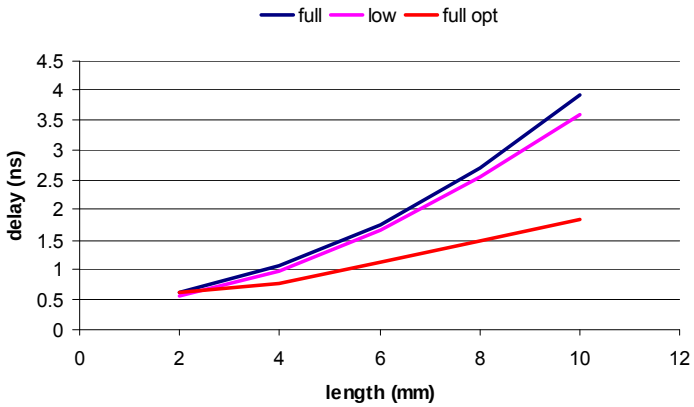

Figure 4: Channel delay vs line length

4.2 Inherent robustness to process variations

The first objective of this work is to compare the inherent robustness of full-swing and PDIFF low-swing signaling schemes to process variations, while compensation techniques will be addressed in Section V and VI.

Our focus is on within-die variations, which happen at the length scale of a die, and that can be further divided into two contributors: systematic and random. Systematic variations can be predicted prior to fabrication and exhibit space locality. In contrast, random variations are due to the inherent unpredictability of the semiconductor technology itself. In our tests, we inject effective gate length variations, which have implications on the threshold voltage as well, as computed by the SPICE device models of our target library. HSPICE is used as our simulation engine. We ignore variations in wires, in agreement with current variation models (e.g., [32], [26]). Fig.5 shows the sensitivity of the signaling schemes to an increasing amount of *systematic variations*. The sensitivity is measured as the variation-induced deviation of the clock propagation time of the receiver from the nominal value. The propagation time goes from the clock sampling edge to the 50% voltage swing of the receiver output, and its nominal value is the same for both full- and low-swing channels, since they were designed to impact the next stage of the design in the same way. Systematic variations have been applied selectively to the transmitter, to the receiver and to the whole channel, so the bars in Fig.5 should be read pairwise. It can be clearly observed that low-swing signaling proves a far more robust scheme to systematic variations. By restricting the analysis to the full-swing channel, its transmitter

turns out to be the weak point of this scheme. The reason lies in the high sensitivity of the library flip-flop (i.e., the receiver) to the settling time of its input signal. This latter significantly deviates from nominal conditions when systematic variations affect the transmitter, and this explains the large degradation of the whole full-swing channel performance. In contrast, the receiver seems much more robust, and variations affecting the whole channel introduce only an incremental degradation with respect to the one caused by the transmitter.

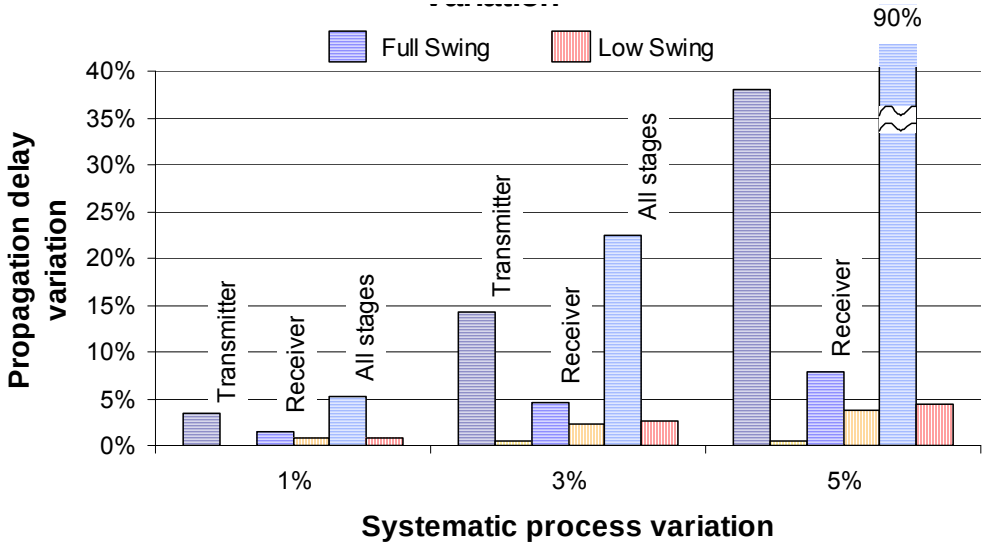

Figure 5: Sensitivity to systematic variations

The only exception occurs for channel-wide 5% systematic variations, where nominal delay is degraded by 90% (height of the last bar for full-swing is truncated to preserve the scale). This is much more than one could expect by looking at the transmitter-degraded case, but this is due to the fact that we are working close to the point where full-swing channel operation fails: in this region, delay is highly sensitive to process parameter variations.

The opposite holds for the low-swing channel. The PDIFF receiver does a good job in providing a noise margin to the perturbations of its input signal induced by systematic variations in the transmitter. However, when variations affect directly the receiver, the PDIFF scheme suffers from increased switching delay. Clock propagation delay variations are much smaller for low-swing channels with respect to the full-swing ones anyway, and these latter might more easily induce the following stage in the design to fail, since it may be impossible to leave a 90% performance degradation margin for 5% systematic variations. We detected a failure of the full-swing channel when the transmitter is affected by 6/7% variations (tolerating a maximum propagation delay degradation by 90%), while the low-swing channel can keep working also under 70% systematic variations affecting both transmitter and receiver, after that the channel fails. At that time, however, propagation delay is degraded by 40%. The sensitivity of the channels under test to random variations (3_/μ=15%) is illustrated in Fig.6. Delay variability is similar in the two cases, with a slightly more tightened distribution for the low-swing channel. Again, we found the transmitter to be the most critical part of the full-swing channel, while the receiver is obviously the weak point of the low-swing channel. In fact, its pseudo-differential behavior makes it very sensitive to random process variations, although we found only a negligible amount of malfunctioning channels with 3_/μ lower than 20%. This indicates that under such variations, the unbalancing of the differential branches remains within the noise margin of the receiver and correct 1/0 sampling takes place in due time. Delay variations pointed out in Fig.5 and Fig.6 indicate that compensation is apparently more challenging in full-swing channels, though the effectiveness of compensation depends not only on the delay spread, but also on the sensitivity of such delay to the compensation mechanism and to the interaction between the sub-blocks of the communication channel, as illustrated hereafter.

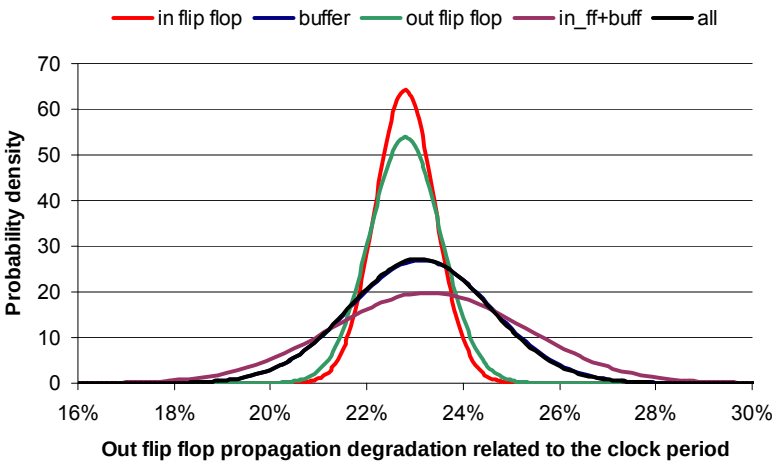

Figure 6: Sensitivity to random variations

5 Post-silicon compensation

Next, we explore the effectiveness of ABB (and forward body bias, FBB, in particular) vs ASV in bringing channel instances slowed down by process variations back within nominal performance. Compensation is applied to both the driver and the receiver for channel-wide tuning, but also selectively to individual sub-circuits to capture sensitivity of channel performance to that of these sub-circuits and eventually come up with a better trade-off between compensation efficiency and cost.

5.1 Experimental framework

Since our target 65nm manufacturing process does not provide a triple well, we apply forward body biasing only to PMOS transistors. Our analysis aims to capture whether this lower cost solution suffices for compensation purposes in onchip communication channels. In addition, it is not possible to selectively apply ASV only to the receiver of a full-swing channel, since this would require a voltage level shifter which is not there. In contrast, such level shifter comes for free in a low-swing channel, which therefore allows PDIFF receiver selective compensation with ASV. FBB does not have any kind of constraints in any signaling scheme. Our experiments encompass the compensation of a representative subset of variation scenarios. Similarly to [8], [26], worst-case systematic variations of +5% of parameter nominal value are assumed and superimposed to random variations. For these latter, the 3_/μ of channel length distribution is varied from 10, 15 to 20%, thus giving rise to three scenarios featuring the same amount of worst-case systematic variations and an increasing parameter spread associated with random variations.

Systematic variations were applied to the *whole channel but also selectively* to the receiver and to the transmitter to account for place&route effects. In fact, transmitter and receiver might be far apart from each other, thus suffering from systematic variations to a different extent, or they might be placed close to each other. In this latter case, physical parameters of the whole communication channel would be skewed by the same amount. We hereafter report only this latter case and the differences (if any) with the other variation scenarios are discussed in the text. We also recall that *random variations were always applied* to the circuits of the *whole channel*. Recently, advanced modeling frameworks have been proposed to propagate variation information from the transistor compact model up to the system level, offering a correlated view on yield, timing, dynamic and static energy [31]. They also improve the traditional Monte Carlo statistical static timing analysis techniques by accounting for rare events in variability distributions. Since this work focuses on a relatively small yet critical amount of logic, we developed an ad-hoc and simplified methodology based on Monte Carlo analysis to study the impact of systematic and random variability and how effectively it can be compensated. For each signaling scheme, variation scenario and compensation technique, we perform Monte Carlo simulations with a statistically significant sample set. Each Monte Carlo run (i.e., a channel instance with different random variability injections) goes through the compensation methodology illustrated in

Fig.9. At first, we check for nominal performance requirements. If met, a new instance is analyzed. If not, a compensation step is applied. In practice, if FBB is under test, an incremental reduction step of the body bias is applied so to improve performance. Similarly, the supply voltage is increased when ASV is assessed. Decrements/Increments are applied with steps of 100 mV both for ASV and FBB. This choice stems from the conclusion of previous works [24] and from considering realistic resolutions of low-cost voltage regulators.

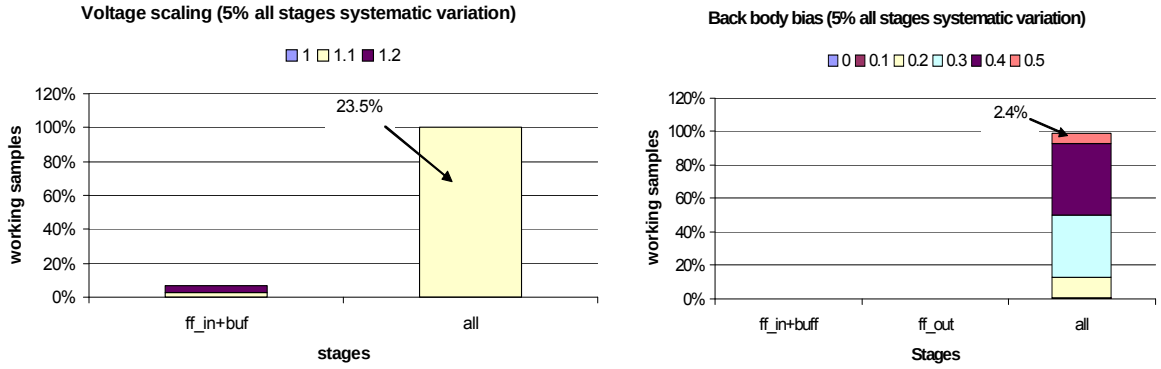

Figure 7: Working samples after compensation of *full-swing channels*. x-axis indicates the channel circuits to which compensation was applied

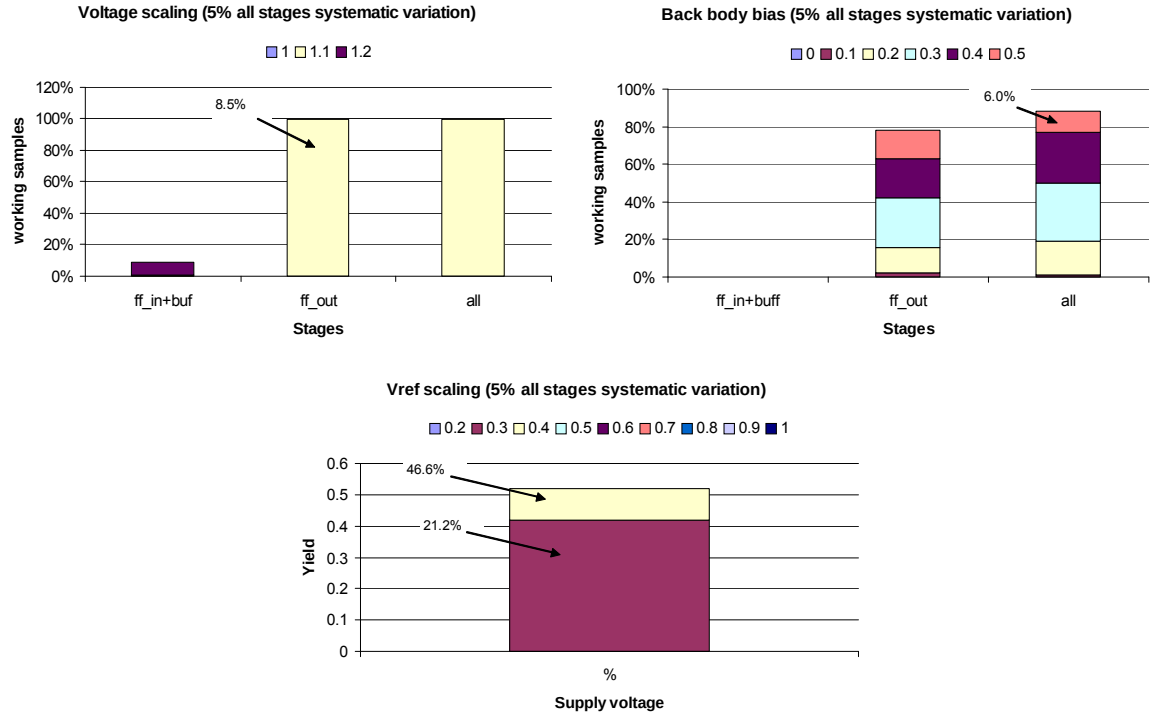

Figure 8: Working samples after compensation of *PDIFF low-swing channels*. x-axis indicates the channel circuits to which compensation was applied

After the compensation step, performance is re-evaluated and eventually an additional compensation step is applied. The process completes when nominal performance is finally met OR when the voltage range limit is reached: 500 mV for forward body bias (to avoid turning on the source pn junction of transistors) and 200mV for ASV (for reliability and technology library constraints). Effectiveness of a technique is expressed as the percentage of the sample set that can be brought back within nominal performance by the compensation technique under test. We denote those successfully compensated samples as *working samples*. Nominal performance means correct sampling at 1.68GHz, with clock propagation time constraints met at the output of the receivers. Moreover, the average power overhead for compensating channel instances with the highest power supply value (lowest PMOS body bias value) is measured, denoting *power efficiency* of the compensation techniques. For low-

swing signaling, we also explore adaptive voltage swing as an additional and built-in compensation technique by raising the voltage swing in increments of 100mV. In the first set of experiments 3_/μ is assumed to be 15%. See subsection V-D for different values. When systematic and random variations are injected into the entire channel, we find almost no channel instances in the sample set working without compensation, both for full-swing and low-swing channels. So, in the experiments that follow, the entire sample set needs to be compensated.

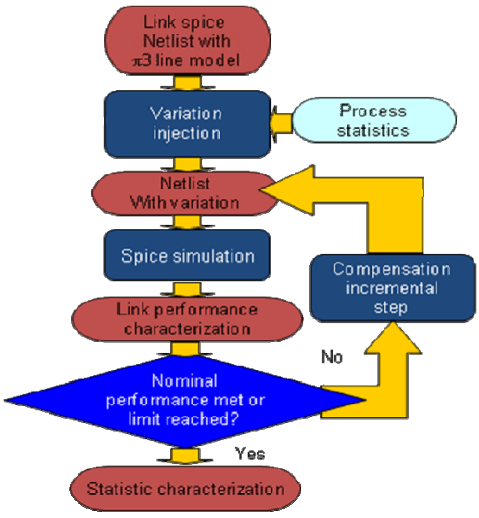

Figure 9: Framework for assessing the effectiveness of variability compensation techniques.

5.2 Compensation efficiency in full-swing links

As can be observed from Fig.7, neither ASV nor FBB are able to restore functionality of all working samples by
only acting upon the transmitter or (for FBB) the receiver. The compensation in this case would be totally ineffective. Variability can only be compensated by tuning all the circuits of the channel. In fact, performance of full-swing channels is highly sensitive to the interaction between the signal provided by the transmitter and the requirements imposed by the receiver on the timing and shape of this signal [12]. Moreover, systematic variations (recall Fig.5) significantly impact both the transmitter and the receiver. As a consequence, an effective compensation can only be carried out by acting upon both modules at the same time. However, while ASV requires a single voltage step to reach 100% working samples, FBB needs its entire voltage range to achieve the same objective. Anyway, the large variations taking place in full-swing channels can be successfully compensated by FBB in spite of its inherently weaker performance tuning capability. In practice, the sensitivity of channel performance to transmitter-receiver interaction was found to be an amplifying effect of FBB tuning capability. The main difference between the compensation techniques lies however in their power efficiency. When ASV raises the supply voltage to 1.1V, the communication channel instances on average exhibit a 23% power overhead with respect to the variation-free scenario. In contrast, a 500mV forward body bias incurs only an average power overhead of 2.4%, almost negligible.

When we applied systematic variations only to the transmitter (flip-flop and driver), we observed that a selective tuning of the transmitter circuits only partially solved the problem. ASV could restore about 80% of the samples, while FBB about 60% by remaining in the voltage range limits. This indicates the impact of random variations, which require a tuning of the receiver as well to restore 100% working samples. The situation is even worse when only the receiver is affected by systematic variations: while no selective tuning of the flip-flop is feasible with ASV due to a lack of a voltage level shifter, only 20% of working samples were achieved by selective FBB. Again, the only option was to tune the entire channel, finding again the same power efficiency gap between FBB and ASV.

5.3 Compensation efficiency in low-swing links

Quite different considerations hold for variability compensation in low-swing channels. This time, ASV can be selectively applied to the receiver since the level shifter is built-in in the signaling scheme. Fig.8 clearly shows that a selective tuning of the receiver with both ASV and FBB reaches a high percentage of working samples. With just one voltage increment step applied to the output flip-flop, ASV can restore performance of the entire sample set. More interestingly, the average power overhead is limited to 8.5%, much lower than in a full-swing channel. In low-swing channels, the transmitter is marginally impacted by systematic variations (recall Fig.5). At the same time, receiver performance is much less sensitive to the perturbations of its input signal than in full-swing channels. Therefore, acting upon the receiver proves an effective compensation method.
Unfortunately, FBB cannot reach 100% working samples with a selective compensation at the receiver, and neither a channel-wide compensation can (90% is the best result achieved with a 500mV FBB). This is essentially due to the weak performance tuning knob represented by FBB, which is not boosted by any circuit level property in this case (for instance, no high sensitivity of channel performance to transmitter-receiver interaction). The average power overhead incurred for the worst-case FBB compensation is around 6%, comparable with that of ASV. Considering the cases where systematic process variations affect only the transmitter or the receiver, we found that FBB is not able again to reach 100% of working samples (best
coverage is 90%). ASV instead works effectively. However, in all cases and for both ASV and FBB, selective compensation at the receiver turns out to be as effective as full channel compensation. Power overhead for ASV is around 7 and 8%, while for FBB is around 3%. Fig.8 also shows the efficiency of an intuitive compensation
technique which stems from the possibility to tune the voltage swing in the low-swing channel. Although intuitive, this technique proves highly ineffective to restore channel performance. By increasing the voltage swing from 200mV to 400mV, only 50% of the non-working samples can be saved. Interestingly, further increasing the swing proves useless, and no further improvements can be achieved, thus spending power uselessly. This is due to the fact that compensating process variations is not just an issue of speeding up signal propagation across the link, but to restore functionality at the transmitter, at the receiver and their correct interaction. Only when the transmitter is impacted by systematic variations while the receiver is not, then speeding up the link with a swing of 400mV achieves 82% working samples. Compensating receiver variability proves more difficult (about 60% working samples). Another argument against reference voltage scaling is power. The measured average power overhead for the worst case compensations (those at 400mV) amounts to a significant 46%. This confirms the results of the work in [7], showing that using the voltage swing to speed up a low-swing link is highly power inefficient.

5.4 Role of random variations

When we repeated the experiments with a 3_/μ = 10% and below, the minor role played by random variations translated into a better compensation efficiency of FBB in low-swing channels, since working samples were always close to 100%. The lower delay spread makes the worst-case compensation scenario affordable also for the tuning capability of FBB, so that this latter can be considered also for low-swing signaling as the impact of random variations decreases. Finally, 3_/μ was set to 20%. In this case, even for full-swing channels FBB could not bring all samples within nominal performance bounds, although still achieving around 95% working samples. Interestingly, in low-swing channels the effectiveness of FBB was as low as 70% working samples.

6 Variability compensation with cross-talk

Robustness to delay variability and its compensation have been evaluated for link models ignoring crosstalk effects so far. However, as technology scales down to the nanoscale regime, coupling capacitance plays a dominant role in determining signal integrity. Moreover, this work also points out

the implications of crosstalk effects on the effectiveness of variability compensation with ASV and FBB.

6.1 Link parasitic extraction

In order to capture realistic layout effects of on-chip interconnects, we synthesized a 32-bit unrepeated link with Synopsys Physical Compiler on the target 65nm STMicroelectronics library. Placement and routing were performed by Cadence SoCEncounter. Transmitter and receiver were placed in two fences 2mm far apart, since the driver and the receiver had been sized in Section 4 based on this wirelength. This work targets a network-on-chip application, therefore STALL and VALID flow control wires were routed together with the 32 bit flit. They are used to implement a stall/go flow control policy in NoCs [2]. The importance of control wires for
communication reliability is such that they might be operated at full swing even though the flit is inferred with low-swing links. In this case, capacitive coupling between full- and lowswing wires within the same channel might be a serious concern. We analyzed both cases: a fully low-swing link and an hybrid one. Finally, the clock tree was synthesized. Again, its coupling with the flit lines needs to be carefully monitored.

We then extracted the parasitic resistance and capacitance with the STAR RCXT tool, enabling the extraction of coupling capacitance as well. The result was the generation of an HSPICE link netlist (modeling parasitics), which was connected with both the full-swing and low-swing transmitters and receivers designed in Section III. The two new link models are equivalent to the one analyzed so far, except for the inclusion of coupling capacitance and the use of STMicroelectronics technology instead of the predictive one to quantify interconnect resistance and capacitance, as extracted from the synthesized link.

6.2 Signal integrity

Fig. 10 reports the communication channel routed by the SoCEncounter tool. Clearly, the routing pattern does not
consist of fully parallel wires as often assumed in abstract link analyses, but encompasses some wire crossings and metal layer switchings. This implies a non-trivial crosstalk interaction among the wires and a hardly controllable signal integrity. In fact, if we look at the capacitance breakdown of the wire named flit 28 (28th wire of the 34 bit link), we can clearly observe that the cross-coupling capacitance with the clock signal is 30.5% of the whole flit 28 line capacitance (Fig.10).

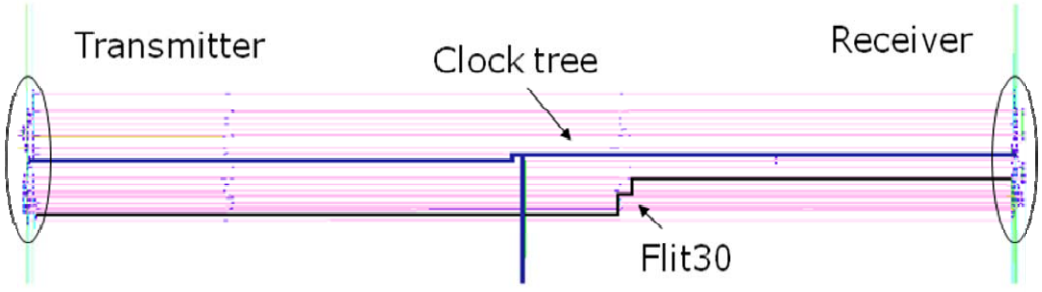

Figure 10: 32bit communication channel layout

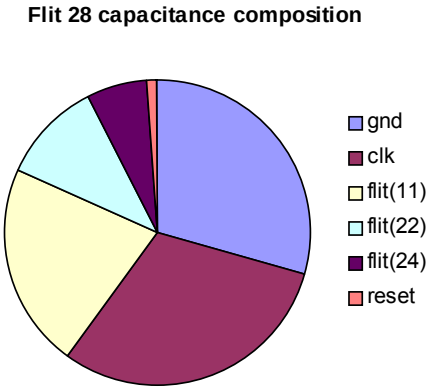

Figure 11: Breakdown of flit_28 capacitance

In these conditions, the low-swing receiver even fails to correctly sample flit 28 where the full-swing one instead succeeds, as Fig.12 illustrates. The clock signal samples input data (second row), which is then output by the driver (third row - full swing). Monitoring the corresponding input of the receiver (fourth row - full swing, third row low-swing) clearly indicates a relevant cross-coupling effect with the clock signal, resulting in the sampling failure (last row - all schemes) of the low-swing receiver. Although the full-swing channel still works, its maximum speed when comparing the clock enabled with the disabled case (the clock is in this case given with a simulation trick, not via the synthesized line) degrades by 16.5%. More in general, whenever in the same communication channel full-swing wires interact with low-swing ones, the signal integrity concern for these latter arises. In order to further prove this, we designed another low-swing link where the two flow control wires (STALL/GO and VALID) were operated at full-swing. The results showed a loss in maximum performance of 31.7% with respect to a fully low-swing link. The critical path was located across the wire denoted as flit 7, whose capacitance breakdown is illustrated in Fig.13. Clearly, the cross coupling capacitance with the STALL control wire accounts for 43.5% of total line capacitance, thus leading to a significant performance degradation of the communication channel. The key take-away here is that in order to materialize the power efficiency of low-swing signaling, reliability concerns caused by coupling with the clock signal and/or with other full-swing control wires need to be tackled by enforcing new routing constraints (e.g., wire extra spacing or shielding). This consideration is of the utmost importance for source synchronous communication schemes, where the clock signal has to be transmitted together with data signals while experiencing the same routing delay. This calls for an advancement of routing scripts and/or techniques in commercial place&route tools that future work has to address.

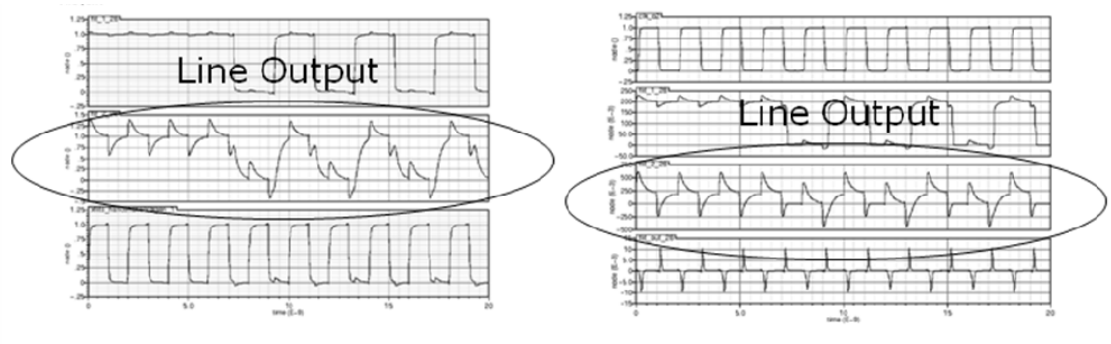

Figure 12: Sampling failure due to wire coupling with clock for full-swing (left) and low-swing (right).

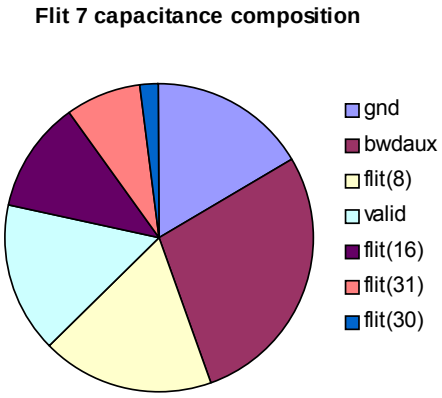

Figure 13: Breakdown of flit_7 capacitance

6.3 Compensating cross-talk affected links

We repeated the variability compensation tests of Section V with the crosstalk-augmented wire models, so to assess how crosstalk interferes with the compensation tasks of ASV and FBB. Based on the results of the previous analysis, these mechanisms are applied to the whole channel in full-swing links, while in low-swing links ASV is selectively applied to the receiver and FBB again to the whole channel. Assuming our routing requirements derived in previous section are met, we consider links clocked by a simulation clock (to avoid
considering the destructive crosstalk induced by the routed clock signal) and, in the low-swing channel, the low-swing operation even for the flow control wires (corresponding, in real-life layouts, to an increased spacing for these lines or to a shield between them and low-swing ones). The same variability injection is operated like in Section 5. The only difference is that this time the entire 34-bit link is compensated, not just a single wire, since interaction between wires is of interest to this experiment. Link speed is characterized as in 1.

Even in the presence of crosstalk, the capability of ASV to restore 100% of the sample set with only one voltage increment for the full-swing link remains unchanged, as suggested by Fig.13. The average power overhead for compensation is 22.4%, similar to the overhead required without crosstalk. Whereas the low-swing link requires two ASV increment steps to restore nominal performance in all cases, while requiring 19.2% power overhead (it was around 8% without crosstalk). These results clearly indicate that crosstalk effects make variability compensation in low-swing channels more expensive for ASV. However, the power overhead for ASV to compensate a low-swing channel is

lower than a full-swing one, indicating that whenever ASV is the only available compensation mechanism, low-swing signaling is more amenable to it.

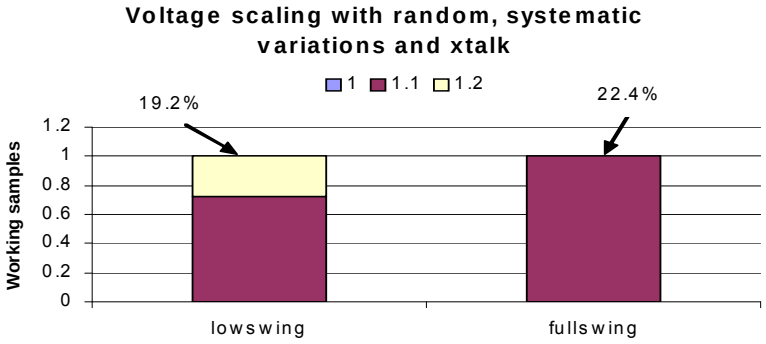

Figure 14: Working samples after ASV compensation with random, systematic variations and crosstalk

Thereafter we ran the tests with the ABB performance tuning technique. In this case the presence of crosstalk affected the compensation of both full- and low-swing links, in fact only a statistically irrelevant percentage of the sample set was brought back to the nominal performance in both cases. The limitation posed by crosstalk on the tuning capability of ABB is extremely severe. We found the restricted applicability of ABB to p-MOS transistors only (due to the single well technology) a very limiting factor for this scenario. In fact, by artificially extending compensation to n-MOS transistors as well, the results of Fig.15 were obtained. For the full-swing channel, the entire sample set can be successfully restored at minimum power overhead, while for the low-swing channel ABB proves an even more ineffective alternative than in the absence of crosstalk.

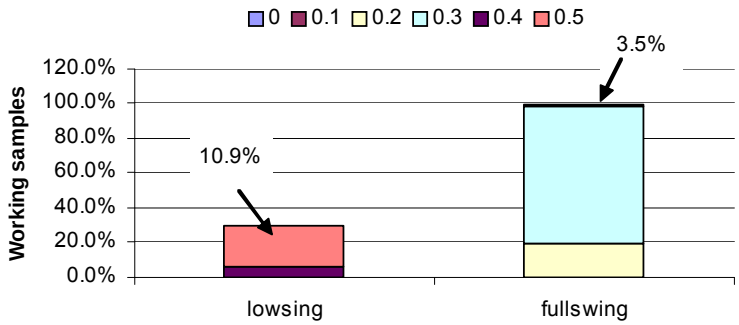

Figure 15: Working samples after p-mos n-mos ABB compensation with random, systematic variations and crosstalk

Conclusion

The work performed in Task 3.4 explored the effectiveness of ASV and FBB as post-silicon variability compensation techniques for on-chip communication channels. Our work shows that FBB is effective for tuning performance of full-swing channels with minimum power overhead. In contrast, when applied to low-swing channels, FBB proves not capable of compensating all variation patterns, since its limited performance tuning capability is not amplified by any circuit property. On the other hand, ASV can exploit the built-in voltage level shifter in low-swing channels and achieve an effective and low cost selective compensation. Crosstalk effects do not change the best compensation technique for each scenario, but make compensation more expensive. For full-swing channels, ABB remains the technique of choice for its minimum power impact, but coupling capacitance makes the tuning of both p-MOS and n-MOS transistors necessary. For low-swing links, ASV selectively applied at the receiver seems the best trade-off between compensation power overhead and yield both with and without crosstalk.

The results of this work point out the superior robustness of low-swing channels to process variations. After considering a realistic range of systematic and random WID process variations, it is evident that low-swing channels (i) can better cope with systematic variations (lower delay deviations and functional correctness guaranteed over a wider range of variations), (ii) feature a lower delay spread under random variations. After exploring all the possible countermeasures based on FBB and ASV, it can be also derived that low-swing channels can be compensated with success against delay variability at a low power cost. These features add up to the reference characteristic of low-swing channels, which is their inherent low power consumption.

References

[1] D. Sylvester and K. Keutzer, ”Getting to the bottom of deep sub-micron II: A global paradigm”, Proc. IEEE Int. Symp. Physical Design, pp.193- 200, 1999.

[2] A. Pullini, F. Angiolini, D. Bertozzi, L. Benini, ”Fault Tolerance Overhead in Network-on-Chip Flow Control Schemes”, Proceedings of 18th Annual Symposium on Integrated Circuits and System Design (SBCCI) 2005, Florianpolis, Brazil, Sep 4-7, 2005, pp. 224-229.

[3] M.Karlsson, M.Vesterbacka, L.Wanhammar, ”Low-Swing Charge Recycle Bus Drivers”, ISCA ’98, pp.117-120, 1998.

[4] Rjoub, A.; Koufopavlou, O.; ”Efficient drivers, receivers and repeaters for low power CMOS bus architectures”, ICECS ’99, pp.789 - 794 vol.2, 1999.

[5] Byung-Do Yang; Lee-Sup Kim; ”High-speed and low-swing on-chip bus interface using threshold voltage swing driver and dual sense amplifier receiver”, ESSCIRC ’00, pp.105 - 108, 2000.

[6] R. Dobkin, A. Morgenshtein, A. Kolodny, R. Ginosar ”Parallel vs. serial on-chip communication”, SLIP 2008, pp.43-50.

[7] S.Medardoni, M.Lajolo, D.Bertozzi, ”Variation tolerant NoC design by means of self-calibrating links”, DATE’08, pp.1402-1407, 2008.
[8] E.Humenay, D.Tarjan, K.Skadron; ”Impact of process variations on multicore performance symmetry”, DATE ’07, pp.1653 - 1658, 2007.

[9] Garcia, J.C.; Montiel-Nelson, J.A.; Nooshabadi, S.; ”High performance bootstrapped CMOS low to high-swing level-converter for on-chip interconnects”, ECCTD 2007, pp.795 - 798, 2007.

[10] Mensink, E.; Schinkel, D.; Klumperink, E.A.M.; van Tuijl, E.; Nauta, B.; ”Optimal Positions of Twists in Global On-Chip Differential Interconnects”, IEEE Transactions on VLSI Systems, Volume 15, Issue 4, April 2007, Page(s):438 - 446.

[11] Kangmin Lee; Se-Joong Lee; Hoi-Jun Yoo; ”Low-power network-onchip for high-performance SoC design”, IEEE Transactions on VLSI Systems, Volume 14, Issue 2, pp.148 - 160, 2006.

[12] N. Terrassan, D. Bertozzi, and A. Bogliolo, ”Spice-Accurate SystemC Macromodels of Noisy on-Chip Communication Chnnels”, in Proceedings of SPI-07, 2007.

[13] A.Narasimhan, B.Srinivasaraghavan, R.Sridhar, ”A low-power asymmetric source driver level converter based current-mode signaling scheme for global interconnects”, Int.Conf.on VLSI Design, 4 pp., 2006.

[14] Predictive Technology Models, ”Interconnect”, http://www.eas.asu.edu/ ptm/

[15] F.Worm, P.Ienne, P.Thiran, G.De Micheli, ”A robust self-calibrating transmission scheme for on-chip networks”, IEEE Trans. on VLSI
Systems, pp.126 - 139, 2005.

[16] Jeong, W.; Paul, B.C.; Kaushik Roy; ”Adaptive supply voltage technique for low swing interconnects”, ASP-DAC 2004, pp.284 - 287, 2004.

[17] Chang-Ki Kwon; Kwang-Myoung Rho; Kwyro Lee; ”High speed and low swing interface circuits using dynamic over-driving and adaptive sensing scheme”, ICVC ’99, pp.388 - 391, 1999.

[18] H.Zhang, V.George, J.M.Rabaey, ”Low-swing on-chip signaling techniques: effectiveness and robustness”, IEEE Trans. on VLSI Systems, pp.264-272, Vol.8, no.3, June 2000.

[19] von Arnim, K.; Borinski, E.; Seegebrecht, P.; Fiedler, H.; Brederlow, R.; Thewes, R.; Berthold, J.; Pacha, C.; ”Efficiency of body biasing in 90- nm CMOS for low-power digital circuits”, IEEE Journal of Solid-State Circuits, Volume 40, Issue 7, July 2005, Page(s): 1549 - 1556.

[20] Venkatraman, V.; Anders, M.; Kaul, H.; Burleson, W.; Krishnamurthy, R.; ”A Low-swing Signaling Circuit Technique for 65nm On-chip Interconnects”, International SOC Conference, pp.289 - 292, 2006.

[21] Garcia, J.C.; Montiel-Nelson, J.A.; Nooshabadi, S.; ”High performance CMOS symmetric low swing to high swing converter for on-chip interconnects”, IEEE Northeast Workshop on Circuits and Systems, 2007, pp.566 - 569, 2007.

[22] Joonsung Bae; Joo-Young Kim; Hoi-Jun Yoo; ”0.6pJ/b 3Gb/s/ch transceiver in 0.18 um CMOS for 10mm on-chip interconnects”, ISCAS 2008, pp.2861 - 2864, 2008.

[23] Meijer, M.; Pessolano, F.; Pineda De Gyvez, J.; ”Technology exploration for adaptive power and frequency scaling in 90nm CMOS”, ISLPED ’04, pp.14 - 19, 2004.

[24] Tschanz, J.; Kao, J.; Narendra, S.; Nair, R.; Antoniadis, D.; Chandrakasan, A.; Vivek De; ”Adaptive body bias for reducing impacts of dieto- die and within-die parameter variations on microprocessor frequency and leakage”, IEEE Journal of SSCs, pp.1396 - 1402, vol.37, no.11, 2002.

[25] Tschanz, J.W.; Narendra, S.; Nair, R.; De, V.; ”Effectiveness of adaptive supply voltage and body bias for reducing impact of parameter variations in low power and high performance microprocessors”, IEEE Journal of SSCs, pp.826 - 829, vol.38, Issue 5, 2003.

[26] Bonesi S., Bertozzi D., Benini L., Macii E., ”Process variation tolerant pipeline design through a placement-aware multiple voltage island design style”, DATE 2008, pp.967 - 972, 2008.

[27] Chen, T.; Naffziger, S.; ”Comparison of adaptive body bias (ABB) and adaptive supply voltage (ASV) for improving delay and leakage under the presence of process variation”, IEEE Transactions on VLSI Systems, Volume 11, Issue 5, Oct. 2003, Page(s):888 - 899.

[28] Gregg, J.; Chen, T.W.; ”Post silicon power/performance optimization in the presence of process variations using individual well adaptive body biasing (IWABB)”, 5th International Symposium on Quality Electronic Design, Page(s):453 - 458, 2004.

[29] I.Hatirnaz, S.Badel, N.Pazos, Y.Leblebici, S.Murali, D.Atienza, G.DeMicheli; ”Early wire characterization for predictable network-on-chip global interconnects”, SLIP’07, Page(s):57-64, 2007.

[30] D.Bertozzi, L.Benini, B.Ricc, ”Parametric timing and power macromodels for high level simulation of low-swing interconnects”, ISLPED 2002: pp.307-312.

[31] A.Papanicolaou et al., ”At Tape-out: Can System Yield in Terms of Timing/Energy Specifications be Predicted?”, IEEE Custom Integrated Circuits Conference, pp.773-778, 2007.

[32] E.Humenay, D.Tarjan, K.Skadron; ”Impact of Parameter Variations on Multi-Core Chips”, Int.Workshop on Architectural Support for Gigascale Integration, 2006.

[33] H. C.Wan et al., ”Channel doping engineering of MOSFET with adaptable threshold voltage using body effect for low voltage and low power applications”, Int. Symp. VLSI Technology, Systems, and Applications, 1995, pp.159-163.