Post-Silicon Clock Deskew Employing Hot-Carrier Injection Trimming With On-Chip Skew Monitoring and Auto-Stressing Scheme for Sub/Near Threshold Digital Circuits

Yu Pu, *Member, IEEE*, Xin Zhang, *Member, IEEE*, Katsuyuki Ikeuchi, *Student Member, IEEE*, Atsushi Muramatsu, Atsushi Kawasumi, *Member, IEEE*, Makoto Takamiya, *Member, IEEE*, Masahiro Nomura, *Member, IEEE*, Hirofumi Shinohara, and Takayasu Sakurai, *Fellow, IEEE*

***Abstract*—Clock skew is a major cause of severe timing yield degradation for sub-/near-threshold digital circuits. We report for the first time on employing hot-carrier injection (HCI) for post-silicon clock-deskew trimming. An HCI trimmed clock buffer, which can be individually selected and stressed to adjust the clock edge, is proposed. In addition, it can be used in conjunction with on-chip skew monitoring circuits to achieve auto-stressing. Our approach is proven to be effective through a representative 1.1-mm $\times$ 0.8-mm clock tree in a 40-nm high-k complimentary metal–oxide–semiconductor process. On average, it reduces the clock skew by eight times at 0.4 V $V_{\rm dd}$. No significant recovery is noticed two weeks after trimming.**

***Index Terms*—Clock skew, hot-carrier injection (HCI), post-silicon tuning, sub-/near-threshold.**

I. Introduction

WHILE ultra-low-$V_{\rm dd}$ digital circuits are becoming increasingly popular, they suffer from severe timing yield degradation. A major cause of such a problem is clock skew. Although advanced electronic design automation (EDA) tools can synthesize clock trees with balanced RLC networks, achieving skew-matched clock trees in the presence of process variations is still difficult. To alleviate skew, prior art exploits clock buffers with post-silicon programmable size. As an example, the Intel's Itanium-family processors insert clock vernier devices (CVDs) at local clock buffers [1], as shown in Fig. 1(a). After comparing clock phases, the CVDs can add delay to any fast clock, hence accommodating the latest clock. However, gate sizing becomes neither effective nor efficient at an ultra-low-$V_{\rm dd}$, where the transistor's driving strength is influenced exponentially by its threshold voltage $V_{\rm th}$ variation. When migrating

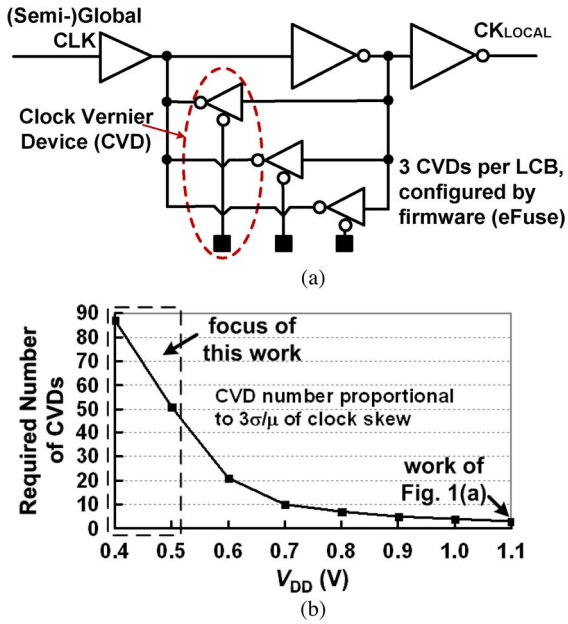

Fig. 1. (a) Clock buffer with CVDs [1]. (b) Required number of CVDs in a clock buffer quickly increases as $V_{\rm dd}$ scales.

Intel's approach to the ultra-low-$V_{\rm dd}$ regime, we see the following problems: 1) According to the simulated results, 3 σ/μ of clock skew becomes tens of times worse compared with the nominal $V_{\rm dd}$. Such a wide spread implies an impractically large number of CVDs embedded in a clock buffer, as shown in Fig. 1(b). 2) The CVDs require firmware (e.g., electrical fuse) to edit programming bits, which also exacerbates area overhead.

Ideally, we want to find a low-cost post-silicon tuning method that can adjust the $V_{\rm th}$ of the fabricated transistors in each clock buffer individually. Although body biasing can tune $V_{\rm th}$, applying it to individual buffers is impossible due to an incredibly large overhead. In this brief, hot-carrier injection (HCI) deskew trimming is proposed. To the best of our knowledge, we are the first to apply HCI trimming to logics, whereas previous work [2] had applied it to static random-access memory cells. The remainder of this brief is organized as follows: Section II introduces the concept of the proposed HCI trimmed clock buffer (HTCB). Section III presents the experimental clock tree diagram and circuit implementation. Section IV presents the

Manuscript received October 18, 2010; revised January 17, 2011; accepted February 26, 2011. Date of publication May 23, 2011; date of current version June 8, 2011. This work was supported by the New Energy and Industrial Technology Development Organization (NEDO), Japan, under the Extremely Low Power (ELP) Project. This paper was recommended by Associate Editor K. Chakrabarty.

Y. Pu, X. Zhang, K. Ikeuchi, M. Takamiya, and T. Sakurai are with the University of Tokyo, Tokyo 153-8505, Japan (e-mail: ypu@iis.u-tokyo.ac.jp).

A. Muramatsu, A. Kawasumi, M. Nomura, and H. Shinohara are with the Semiconductor Technology Academic Research Center, Yokohama 222-0033, Japan.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSII.2011.2149050

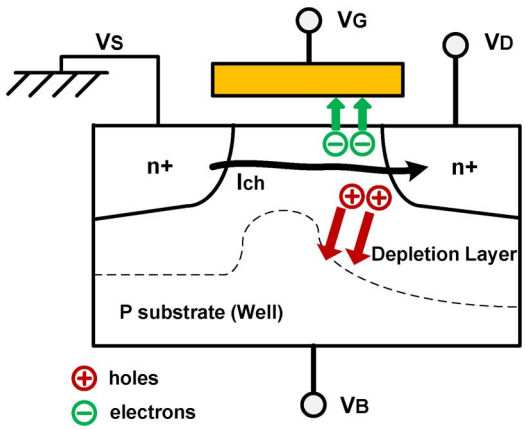

Fig. 2. HCI mechanism on nMOS transistor.

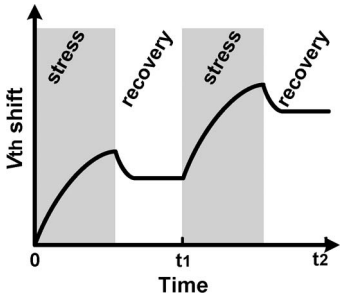

Fig. 3. $V_{\rm th}$ shift on nMOS transistor.

deskew trimming procedures. Measurement results in a 40-nm high-k CMOS process are shown in Section V. Finally, Section VI draws the conclusions of this brief.

II. Concept of HTCB

The HCI mechanism is illustrated in Fig. 2. Typically, the HCI effect is much stronger on nMOS devices than on pMOS devices. The necessary and sufficient conditions of HCI stressing are the following: 1) A large current flows through the channel. Some electrons are energized by the horizontal channel electric field so they may become hot-carriers and are injected into the gate oxide. 2) A high $V_{\rm ds}$ exists, so the channel carriers are accelerated by a high electric field of the drain. Ionization collision and avalanche multiplication can occur and cause electron–hole pairs. Some of these pairs are also injected into the gate oxide. The net result of HCI stressing is an increase in $V_{\rm th}$. This $V_{\rm th}$ increase can be memorized because a significant amount of carriers can be trapped for the device's entire lifetime [3], as shown in Fig. 3.

The proposed HTCB is shown in Fig. 4. In the HTCB, only the nMOS transistor M_1 in the first inverter needs to be stressed because the timing elements in our circuits are triggered by positive clock edges and don't care clock duty changes. The second inverter can reshape the delayed clock signal to improve its slew rate. During stressing, the switch is on, and a high $V_{\rm dd}$ is applied. Meanwhile, an AC signal (with 0-$V_{\rm dd}$ amplitude) or a DC signal (at $V_{\rm dd}$ level) is applied to the gate terminal. In this brief, we choose an AC signal as it can be more effective compared with a DC signal [4]. After stressing, $V_{\rm dd}$ goes back to normal, and the switch is off. Since

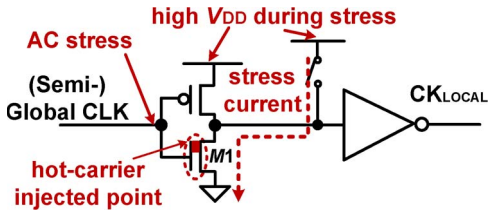

Fig. 4. Proposed HTCB.

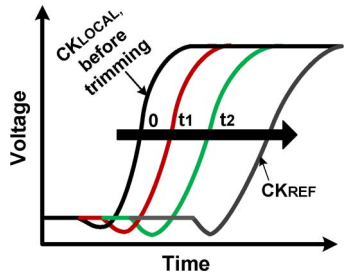

Fig. 5. Increased delay of $CK_{\rm LOCAL}$ after HCI trimming.

the $V_{\rm th}$ of M_1 is increased, consequently the HTCB delay is also increased, which is especially prominent at the sub-/near-threshold region. Therefore, by controlling the switch, any fast local clock $CK_{\rm LOCAL}$ can be delayed to match the latest reference clock $CK_{\rm REF}$, as illustrated in Fig. 5.

III. Experimental Clock Tree Diagram and Circuit Implementation

A grid-based H-type clock tree with balanced loadings is implemented, as shown in Fig. 6. Conventional clock deskew requires iterative pair-by-pair comparison and individual clock edge adjustment. After which, all the clocks can be synchronized with the latest clock. To reduce deskew effort and time, we create the latest clock $CK_{\rm REF}$ artificially. This is accomplished by using delay elements to generate the $CK_{\rm REF}$s, which are predetermined to be slightly later than any $CK_{\rm LOCAL}$ in the worst device mismatch case during design time. When deskew, the phase information of a $CK_{\rm REF}$ is forwarded in its region, and the HTCBs are stressed to lock their $CK_{\rm LOCAL}$s with the $CK_{\rm REF}$.

Please be aware that, although the deskew trimming scheme is only applied to each region in our experimental clock tree, just to prove the effectiveness of our concept, the proposed scheme can be easily extended to other clock tree levels. Inter-region clock deskew is also possible by using a hierarchical trimming method, that is, first deskew the $CK_{\rm REF}$s across regions and then deskew $CK_{\rm LOCAL}$s according to their $CK_{\rm REF}$s inside each region. This strategy helps reduce (semi-) global metal wire utilization and save power consumption in operational mode.

The key circuit components in our scheme include phase comparators for skew monitoring, HTCBs, and a scan-chain. They are shown in Fig. 7. Many kinds of phase comparators for skew monitoring exist, such as the fully customized phase comparator for ultra-high-speed clocks described in [1]. To ease the design, the adopted phase comparator consists of only standard cells and thus is synthesizable in EDA flow.

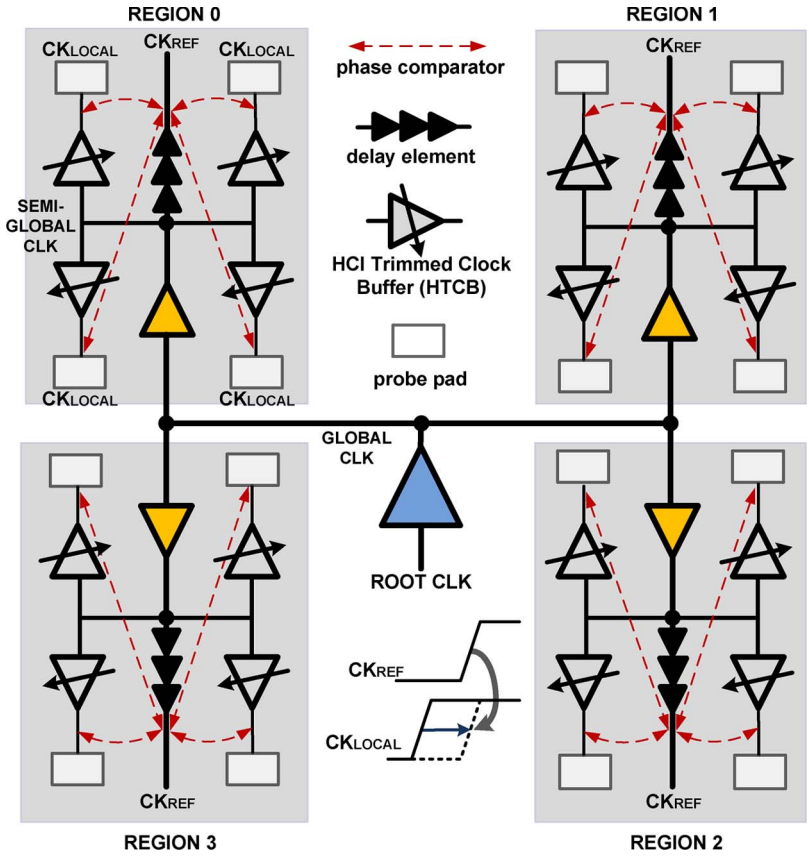

Fig. 6. Diagram of the experimental H-type clock tree, with region-based, on-chip skew monitoring, and auto-stressing scheme. The scheme can be extended to other clock tree levels and inter-region deskew.

The CONFIG bits in the phase comparators are connected by the scan-chain. During initialization, all the CONFIG bits are initialized to "1" by scan operation. During evaluation, the cross-sampling D-flip-flops I_1 and I_2 compare the phases of $CK_{\rm LOCAL}$ and $CK_{\rm REF}$ and reset the CONFIG bit to "0" if any phase difference is detected. Note that the setup times of I_1 and I_2 provide a guardband to avoid meta-stability, so the resolution of added delay to $CK_{\rm LOCAL}$ should be confined to this guardband. Once pulled down, the CONFIG bit cannot restore back to "1" until the next initialization. The low-active CONFIG bit indicates whether the connected HTCB should be stressed. During stressing, the global signal STRESS is asserted to "0." $V_{\rm dd}$ is raised to be significantly higher than the nominal $V_{\rm dd}$ level. A "0" CONFIG bit remains at ground level, and a "1" CONFIG bit follows the $V_{\rm dd}$ level accordingly, thanks to the latch (essentially two cross-coupled inverters) in I_3. Therefore, a level-shifter is not needed. An AC signal with 0-$V_{\rm dd}$ amplitude is applied to ROOT CLK and SEMI-GLOBAL CLK. If CONFIG is "0," then M_1 and M_2 in the HTCB are turned on simultaneously, so a large current goes into the nMOS transistor M_4, and the HCI stressing conditions are fulfilled.

It is worth mentioning that the phase comparators, HTCBs, and other logic gates share the same power grids in order to minimize the area overhead. Comparing our circuits with the previous works, the overhead of phase comparators remains the same. However, the clock buffer size is greatly reduced. A huge number of programming bits are also avoided.

IV. Proposed Trimming Procedures

Fig. 8 proposes the automated trimming procedures. Our programming method takes "trial and error"-based iterations, so a maximum iteration number $N_{\rm max}$ is set beforehand to avoid an exceedingly long trimming time. In addition, we are aware of a large current during stressing (refer to Fig. 10 in Section V). The stressing current per HTCB exceeds 1 mA. For large size circuits with hundreds or even more HTCBs, the total stressing current may exceed the maximum current tolerated by power grids. To solve this issue, Steps III–V scan out the CONFIG bits of phase comparators, mask part of "0" (stressing enable) bits to "1" (stressing disable) bits, and scan in them again, hence allowing stressing a chip sequentially. All the scan operations, i.e., Steps I, III and V, are operated with the nominal $V_{\rm dd}$ (e.g., 1.1 V) to guarantee correct scan functioning. Step II is performed at the ultra-low-$V_{\rm dd}$ (e.g., 0.4 V) in operational mode to capture correct skew information.

As indicated in Fig. 8, our method also contains a manual trimming mode for selective stressing, which can be used in design rectification, such as hold-time debugging. In such mode, Steps V and VI are used standalone to scan-in predetermined bits, hence selecting targeted HTCBs for stressing.

V. Measurement Results

The clock tree regions 0, 1, and 2 in Fig. 6 are fabricated in a 40-nm high-k CMOS process. The tree is distributed on a 1.1-mm $\times$ 0.8-mm die. Fig. 9 shows the die photo.

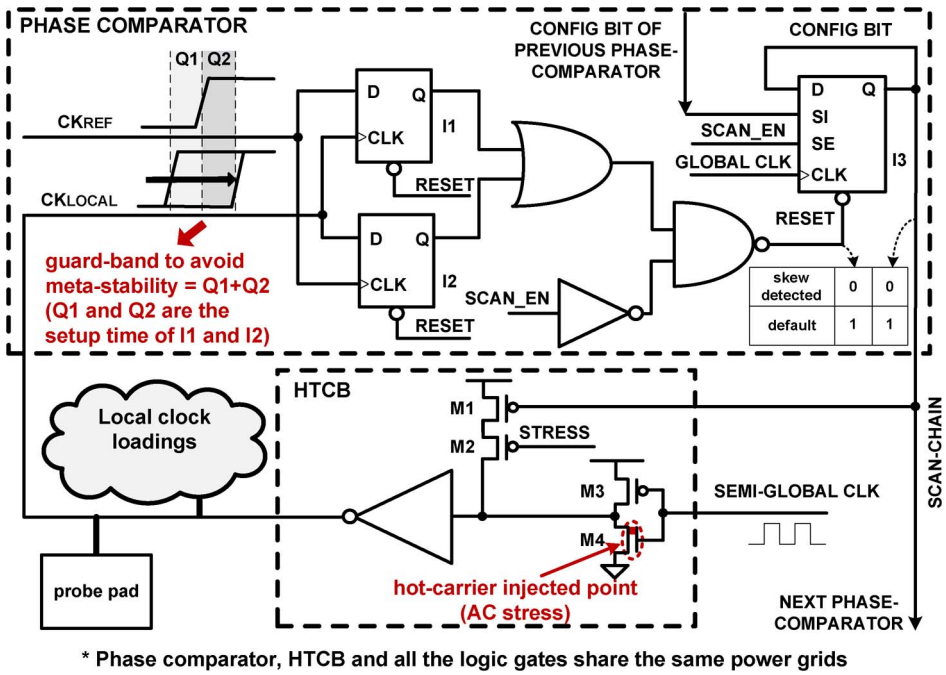

Fig. 7. Phase comparator for on-chip skew monitoring, HTCB implementation, and scan-chain.

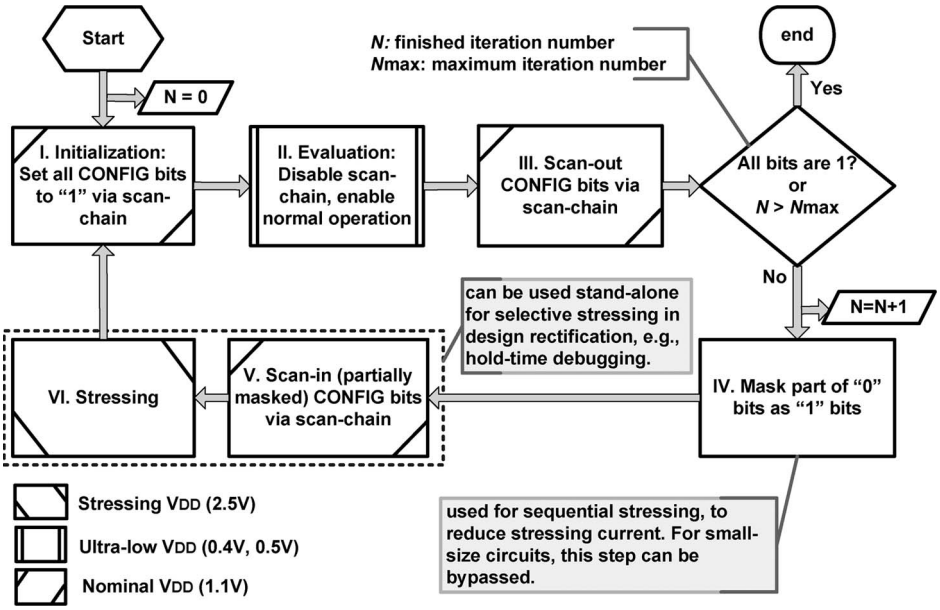

Fig. 8. Automated trimming procedures. Steps V and VI can also be used standalone for manual and selective stressing.

The appropriate stressing $V_{\rm dd}$ depends on individual process technology. We use a 2.5-V $V_{\rm dd}$, whereas our device can tolerate over 3.0-V $V_{\rm dd}$ according to measurement. A shorter stressing period per iteration can provide a higher resolution of added delay and prevent over-stressing HTCBs. We choose 5 s as the stressing period per iteration. Raising the temperature can accelerate the HCI, but stressing at a very high temperature for a long time (e.g., 10^3 s at 100 °C) is not suggested. Otherwise, bias temperature instability (BTI) [5] may occur and degrade other logic gates that share the same power grids since we do not distribute separate power grids for the HTCBs. The main difference between HCI and BTI is that HCI needs large current injection, whereas BTI does not need current. Typically, the HCI effect happens much faster and more intensive than the BTI effect. However, BTI can become an important concern when 1) a high bias voltage is imposed on a transistor for a long time, e.g., many hours/days, and 2) a high temperature exists. To mitigate BTI on other gates, short-time AC stressing at 25 °C room temperature is applied to HTCBs in the experiment. Fig. 10 shows the measured stressing current per HTCB at different stressing $V_{\rm dd}$s.

The deskew results of 10 chips for 0.4-V operational $V_{\rm dd}$ are provided in Fig. 11. On average, HCI trimming consumes 135 s of stressing time per die [Fig. 11(a)] and renders an over eight times skew reduction [Fig. 11(b)]. Note that we use a high-k process. For processes without high-k dielectric, the stressing

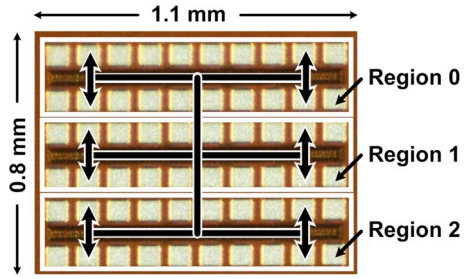

Fig. 9. Die micrograph of the experimental clock tree in a 40-nm high-k CMOS.

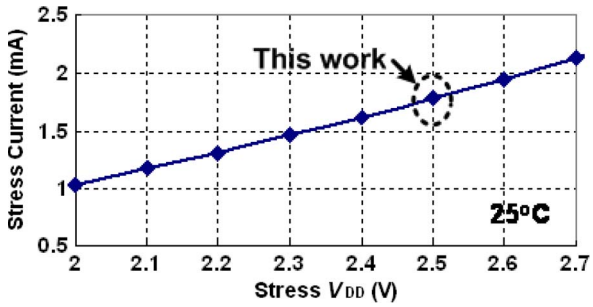

Fig. 10. Stressing current at different stressing $V_{\rm dd}$s.

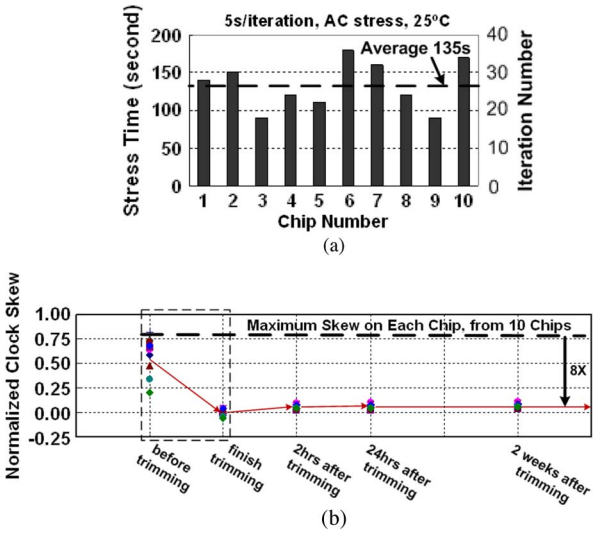

Fig. 11. (a) Stressing time of 10 chips. (b) Normalized skew at 0.4-V $V_{\rm dd}$ before and after HCI trimming.

time is expected to be significantly shorter. A very important concern is the retention property of the injected charge. In Fig. 11(b), two weeks after trimming, no significant recovery has been observed. By extrapolating the line in Fig. 11(b), the injected carriers could be trapped for many years, which is long enough to cover the lifetime of many consumer electronic devices.

Fig. 12(a) and (b) show the measured clock rise edges and skews at 0.5-V $V_{\rm dd}$ before and after trimming. Because the clock signals drive heavy probe pads, the absolute values of the rise time and skews are magnified. To examine whether other logic gates that share the same power grids are affected by imposing a high stressing $V_{\rm dd}$ in Fig. 12(c) and (d), Fig. 12(a) is superimposed on Fig. 12(b). After HCI trimming, the rise time

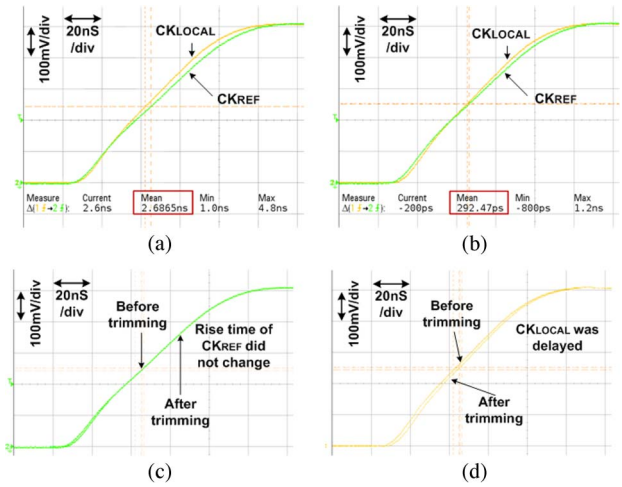

Fig. 12. Measured clock waveforms at 0.5-V $V_{\rm dd}$. (a) Before trimming. (b) After trimming. (c) By superimposing $CK_{\rm REF}$ of (a) on (b), the rise time of $CK_{\rm REF}$ is almost unaffected. (d) Delayed $CK_{\rm LOCAL}$.

and the shape of $CK_{\rm REF}$ are almost unaffected (degradation is only less than 1%), as shown in Fig. 12(c). Since ultra-low-$V_{\rm dd}$s typically apply to medium-/low-speed applications, this slight degradation can easily be covered by design margin. Meanwhile, $CK_{\rm LOCAL}$ is delayed, as shown in Fig. 12(d). Therefore, the effectiveness of selective stressing at HTCBs is confirmed.

VI. Conclusion

We have reported for the first time on employing HCI for post-silicon clock-deskew trimming. An HTCB, which can be individually selected and stressed to adjust the clock edge, is proposed. In addition, it can be used in conjunction with on-chip skew monitoring circuits to achieve auto-stressing. Our approach is proven to be effective through a representative clock tree in a 40-nm high-k CMOS process.

The future CMOS technologies that scale into very deep sub-micrometer regions increase the electric fields in MOSFETs. The chance for hot-carriers to get enough energy and be injected into gate oxide gets higher, yielding a more significant HCI effect. Currently, we are exploring the optimum stressing conditions to reduce the trimming time while maintaining device reliability. The reversibility of HCI mechanism is also worth further investigation. Looking to the future, this trimming method can possibly be incorporated in a wafer burn-in test.

References

[1] P. Mahoney, E. Fetzer, B. Doyle, and S. Naffziger, "Clock distribution on a dual-core, multi-threaded itanium-family processor," in *Proc. ISSCC Dig. Tech. Papers*, Feb. 2005, pp. 292–293.

[2] K. Miyaji, S. Tanakamaru, K. Honda, S. Miyano, and K. Takeuchi, "Margin enhancement by VTH mismatch self-repair in 6T-SRAM with asymmetric pass gate transistor by zero additional cost, post-process, local electron injection," in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2010, pp. 41–42.

[3] T. Ong, M. Levi, P. Ko, and C. Hu, "Recovery of threshold voltage after hot-carrier stressing," *IEEE Trans. Electron Devices*, vol. 35, no. 7, pp. 978–984, Jul. 1988.

[4] Y. Shiyanovskii, F. Wolff, C. Papachristou, D. Weyer, and W. Clay, "Hardware Trojan by hot carrier injection," ArXiv e-prints, Jun. 2009.

[5] J. Hicks, D. Bergstrom, M. Hattendorf, J. Jopling, J. Maiz, S. Pae, C. Prasad, and J. Wiedemer, "Transistor reliability," *Intel Technol. J.*, vol. 12, no. 2, pp. 131–144, Jun. 2008.