Adaptive Robustness Tuning for High Performance Domino Logic

Bharan Giridhar[1], David Fick[1], Matthew Fojtik[1], Sudhir Satpathy[1], David Bull[2], Dennis Sylvester[1], David Blaauw[1]
[1]University of Michigan, Ann Arbor, MI, [2]ARM, Cambridge, United Kingdom

Abstract

A new domino design style is proposed that provides performance gains of up to 71% over conventional domino, and is demonstrated in a 32b multiplier in 65nm CMOS. The design dynamically tunes domino gates to trade surplus noise margins at nominal conditions for performance by detecting stability errors during runtime while guaranteeing correct operation.

Introduction

While many chips are constrained by power, speed critical circuit portions in a design continue to benefit from targeted use of a high performance logic design style [1]. Domino logic [2, 3] has been the mainstay for this purpose. However, increasing process variation has made conventional domino design more complex and less beneficial, forcing designers to revert back to static CMOS design [4]. Fig. 1 shows that margining the keeper for robustness under worst-case PVT conditions can result in a 32% delay increase. Increasing PVT sensitivities with process scaling and more frequent use of low voltage operation are expected to further increase these design margins. However, these margins can be reduced and traded for performance gains under typical PVT conditions. This motivates a new design style called Adaptive Robustness Tuning (ART) that shrinks robustness margins with minimal design overhead and enables performance gains of up to 34% using robustness speculation. Similar to recently proposed adaptive approaches [5, 6], the robustness margins are reduced until functionality errors are detected. Failures are used to guide robustness tuning and are corrected to guarantee forward progress in computation. In addition to robustness speculation, ART also removes timing margins, increasing the total speed improvement to 71% over conventional domino in a 32b multiplier in 65nm CMOS.

Proposed Approach

ART Domino performs two evaluations of the domino gate (Fig. 2): a fast, speculative evaluation followed by a slower, safe evaluation with sufficient margins to guarantee correct operation under worst-case conditions. The safe evaluation is performed in the background and does not impact latency of the computation. To allow for the larger delay of the safe evaluation, we introduce a technique to split each pipeline stage at its middle point during the safe evaluation phase, effectively doubling the time available for safe evaluation. The results of the two operations are compared and in case of errors, the errant computation is flushed from the pipeline and the result of the safe evaluation is propagated, guaranteeing forward progress.

The ART Domino gate operates in four phases: (a) Speculative Precharge (SP), where the gate is precharged with margins removed; (b) Speculative Evaluate (SE), where the gate performs a fast, speculative evaluation; (c) Checker Precharge (CP), where the gate is precharged with restored margins; (d) Checker Evaluate (CE), where the gate performs a slower “always correct” evaluation. During the Speculate (SPEC) phase, precharge voltage V_X is lowered to TVDD and voltage V_Y on the output inverter is raised to TVSS speeding critical transitions at both nodes by reducing voltage swings. Raising V_Y also speeds the following gate by trading its noise margin for speed. During the Check (CHECK) phase, robustness margins are restored and a safe evaluation checks for errors. The values of TVDD and TVSS are tuned to operate the design at the edge of failure, thereby maximizing performance gains and automatically tracking PVT conditions.

Fig. 3 shows an ART Domino pipe stage. The headers/footers are shared across gates in a pipeline stage to minimize design overhead relative to conventional domino circuits. An overlapping clock generator provides overlapping clocks which eliminate latches between pipe stages and provide skew tolerance [7]. To allow the slower, safe evaluation to complete, each pipe stage is split during CE by inserting a fully margined domino latch DOMBUF in the middle of the stage. During SE, this latch is bypassed and the delay overhead is limited to only a single transmission gate. The output of the gate preceding the latch is copied onto DOMBUF, and during CE this value is propagated forward, cutting the stage depth by half.

Both halves of each pipe stage perform safe evaluations simultaneously. The second half passes its result via phase overlap to the next stage, which then in turn performs simultaneous safe evaluations on its two halves. Fully-margined gates are used in the error logic for “always correct” operation. The speculated and checked results of the segment are copied to domino latches and the two values are compared in the following SPEC cycle.

As in all design styles incorporating timing speculation, metastability can occur in ART Domino design on the error signal and cause error detection failures. We propose solutions to suppress the two sources of metastability in the latch DOMBUF: 1) Metastability due to genuine timing violations during SE is minimized by providing an additional half cycle of slack for the latch to evaluate. 2) Unintentional leakage in the preceding gate (Gate marked X in Fig. 3) during SE can also cause DOMBUF to become metastable. To address this, DOMBUF is given the full CP to resolve and is further latched through BUF1 during CE prior to using this value in the error and data paths, thereby reducing the probability of metastability to acceptable levels ($\sim 2.5 \times 10^{-21}$ or once every 12,700 years).

System Implementation

ART Domino was incorporated in a 32×32-bit multiplier (Fig. 4) in 65nm CMOS. The design is split into two pipe stages and four tunable voltage domains. With ART disabled, the multiplier runs at 890MHz (at 1.2V, 27ºC and consuming 184mW). Measured frequency contours as a function of the tunable voltages (Fig. 5) show that performance with ART improves to 1.192GHz (34% increase) by eliminating robustness margins at nominal PVT conditions. Fig. 6 top plots measured minimum ART power/energy overheads with achieved performance. The power initially reduces due to reduced voltage swing and rises at higher frequencies due to increased leakage. Fig. 6 bottom shows the tuning voltage to achieve the measured power-frequency points. Measured error rate due to robustness failures (Fig. 7) indicate higher sensitivity to TVSS tuning than TVDD. Temperature dependence of gains due to robustness speculation show higher gains for lower temperatures. Measurement gains due to timing speculation (Fig. 8, left) at nominal temperature (27ºC) and voltage (1.2V) range from 20% to 33% compared to performance of the slowest die at 85ºC with 10% supply droop. Tuning robustness margins provides further gains (24% to 34%) resulting in measured total gains of 49% to 71% over conventionally margined designs. ART Domino provides an overall performance improvement of 3.2× over static CMOS putting it in the category of fastest reported logic families such as Output Prediction Logic (with a speedup of 3.03× over CMOS [8]), although OPL does not rely on PVT margins to achieve this gain.

References

[1] S. Wijeratne et al, ISSCC, 2006
[2] G. Yee et al, ICCD, 1996
[3] N. Weste et al, CMOS VLSI Design : A circuits and systems perspective
[4] R. Kumar et al, ISSCC, 2009
[5] J. Tschanz et al, ISSCC, 2010
[6] S. Das et al, JSSC, April 2006
[7] D. Harris et al, JSSC, Nov 1997
[8] L. McMurchie et al, ICCD, 2000

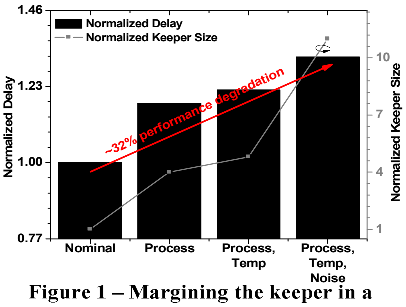


Figure 1 – Margining the keeper in a Domino gate for robustness under worst-case PVT conditions

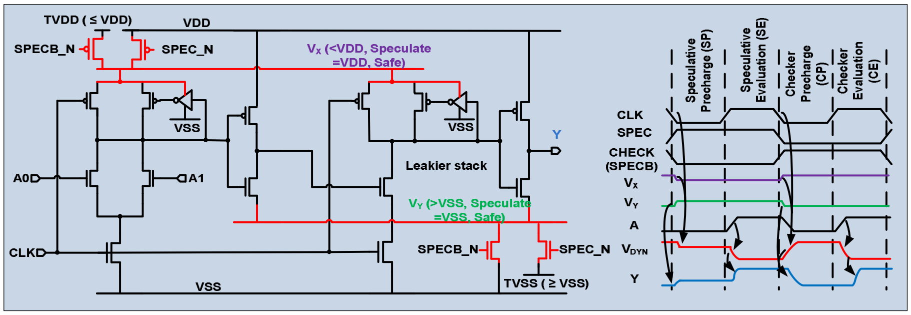


Figure 2 – ART Domino Gate with timing waveforms

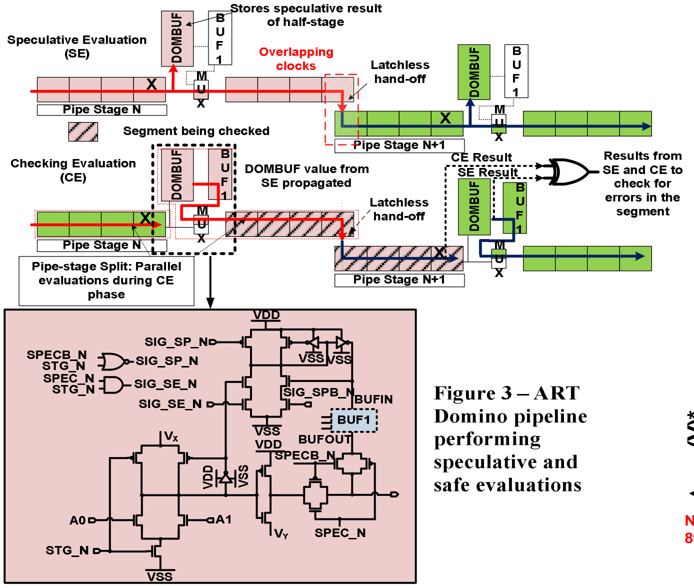


Figure 3 – ART Domino pipeline performing speculative and safe evaluations

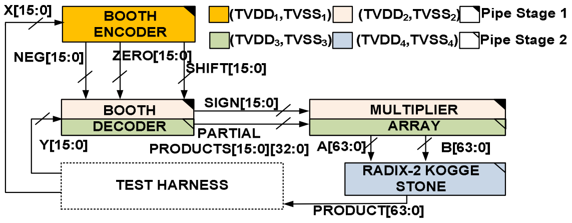


Figure 4 – Test prototype – 32b multiplier

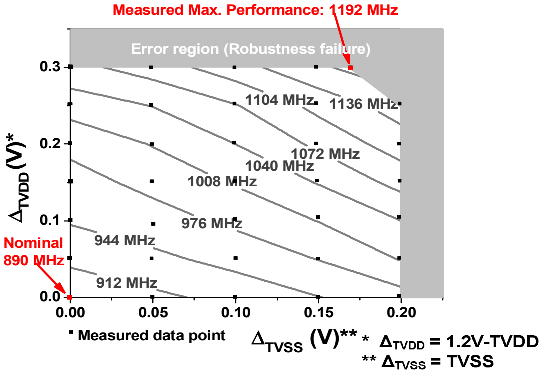


Figure 5 – ART Domino increased multiplier performance from 850MHz to 1.192 GHz

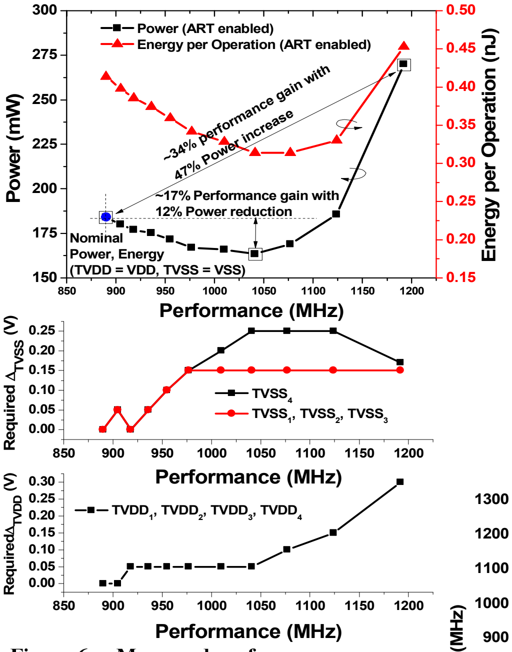


Figure 6 – Measured performance improvement and tunable voltage profiles.

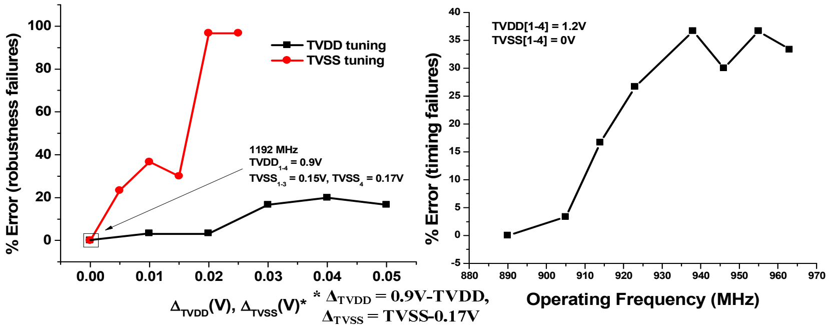
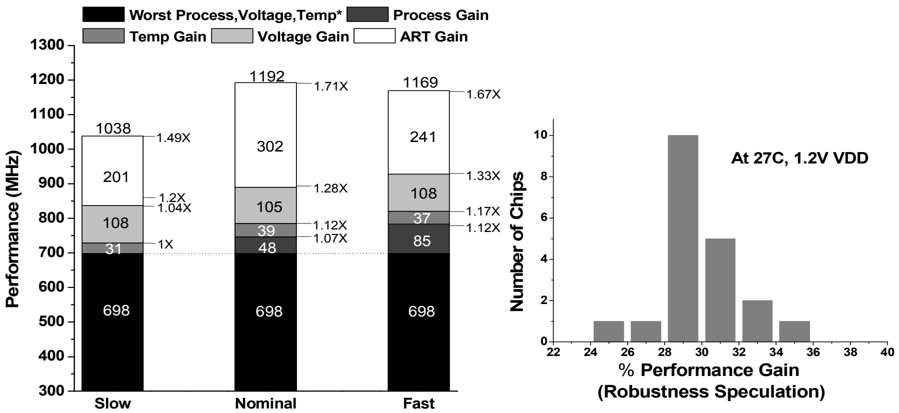


Figure 7 – Measured error rates due to robustness and timing failures.



* Worst process was set by the slowest die. Temperature was set to 85C and Supply was degraded by 10% to 1.08V

Figure 8 – Performance improvement due to robustness and timing speculation

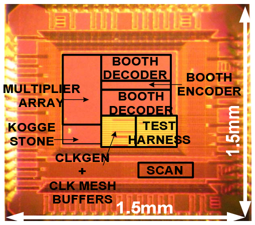


Figure 9 – Die Micrograph