A Bit of Analysis on Self-Timed Single-Bit On-Chip Links

Jonathan Tse, Benjamin Hill, and Rajit Manohar
Computer Systems Laboratory
Cornell University
Ithaca, NY, U.S.A.
{jon,ben,rajit}@csl.cornell.edu

***Abstract*—We present a study of five different self-timed single-bit on-chip links implemented in 90 nm, 65 nm, and 45 nm process technologies. These include representative examples of Quasi Delay-Insensitive, single-track, ternary, and voltage-scaled links, as well as a link of our own design intended to minimize wire usage. We characterize the tradeoffs between throughput, energy, and area for planar wiring as well as 3D through-silicon vias. We also describe our multi-objective optimization framework for exploring this parameter space.**

***Index Terms*—asynchronous; on-chip links; QDI; WCHB; relaxed QDI; ATLS; STFB; ternary; single-track; TSV; 3D integration**

I. Introduction

As technologies scale and power envelopes tighten, it is time to revisit the design of the humble single-bit on-chip link. Of course, designers will still use the highest bandwidth links within their energy and area budgets, but other considerations have grown in importance. Synchronous designers have long felt the additional constraints associated with clock distribution [1], and asynchronous designers have seen a host of other issues arise as well, such as the need to pipeline long planar links [2] and variability-related problems [3]. Even the link wires themselves present design challenges.

Increasing system complexity has begun to put serious pressure on planar wiring resources [4]. At first glance, new process nodes and better back-end-of-line (BEOL) manufacturing have kept the problem mostly at bay. Unfortunately, while designers might have enough wires to meet connectivity requirements in all but the most wire-starved designs, the RC characteristics of the wires have not scaled with transistors. In order to keep shrinking BEOL features without dramatically increasing wire resistance, chip foundries have increased the cross-sectional height of wires. The resistance of long wires can no longer be ignored—the lumped capacitor model is no longer valid in deep-submicron technologies [5]. Furthermore, taller, more closely spaced wires have resulted in large coupling capacitance values, increased crosstalk, and decreased performance. Some designers of high-frequency systems have resorted to increasing planar wire spacing to decrease wiring capacitance and crosstalk, thereby preserving performance. Over-reliance on this technique can artificially increase pressure on wiring resources, especially for wide buses.

Regardless of bus width, wire spacing, or signaling protocol, the energy of intra-chip communication represents a non-trivial portion of total chip energy consumption [6]. Some projections show wide, cross-chip links consuming a hundredfold more energy in wire transitions alone than in computation [7]. One way to alleviate this problem is to move to 3D integration, for both energy [8] and performance [9], as transmitting data inter-die through a through-silicon-via (TSV) is lower in energy and delay than transmitting data through planar wires across a die. 3D integration has its own problems, such as variability [10] and thermal management [11]. However, for the purposes of this study we focus on the fact that TSV resources are quite limited in comparison with planar wire resources. TSV pitch is at least 1 µm and is often much larger, on the order of 25 µm [12], well over tenfold the pitch of modern planar wiring.

Self-timed single-bit links are uniquely situated in this complex design space. While they are robust to delay variations, the encodings used incur additional overheads in transition counts and wiring resources—especially important in the TSV case. In comparison, synchronous links make efficient use of wiring resources but suffer from clock distribution and recovery problems. As such, the benefits they provide in comparison with self-timed links are largely dependent on usage case [13].

In light of the pressures on planar and TSV wiring resources by today's asynchronous designers, we present an analysis of self-timed single-bit links. We evaluate representatives from the various classes of self-timed links on the metrics of throughput, energy per bit (token) transmitted, and circuit area. We also present our Single-Track Asynchronous Ternary Signaling (STATS) single-bit link design, which is a single-wire link intended for use in wire/TSV limited environments. However, evaluating each link type at a single point in the throughput/energy/area space is unfair, as factors such as transistor sizing, V_{DD}, and circuit topology can easily change that point. As part of this work we present an optimization framework to obtain throughput/energy/area Pareto efficiency fronts. While this work focuses solely on self-timed single-bit links, multi-bit or even synchronous protocols are on our future-work road map.

II. Single-Bit Signaling Protocols

Table I shows the self-timed single-bit signaling protocols we chose to study, representative of the various classes of competing schemes. Figure 1 shows the wire transitions required for each to send the same token pattern. We provide a brief description of each protocol and justification for our choices below.

Other self-timed techniques, such as bundled data [14] and GaSP [15], leverage traditional clock-based datapath elements like flip-flops and latches for pipelining. They generate "clock signals" for each pipeline stage locally, and amortize the cost of this control circuitry over the many bits of a wide datapath. We plan to revisit these techniques in a multi-bit study, which we believe is a more appropriate comparison space. We also omit link protocols which do not include any handshaking flow control, such as [16].

TABLE I
SELF-TIMED SINGLE-BIT SIGNALING PROTOCOLS

Name[1]	Handshake	Timing	Voltage	Wires
ATLS	4-Phase	QDI	Ternary	2
RQDI	2-Phase NRTN	RQDI	Full-Swing	3
STATS	2-Phase RTN	Single-Track	Ternary	1
STFB	2-Phase RTN	Single-Track	Full-Swing	2
WCHB	4-Phase	QDI	Full-Swing	3

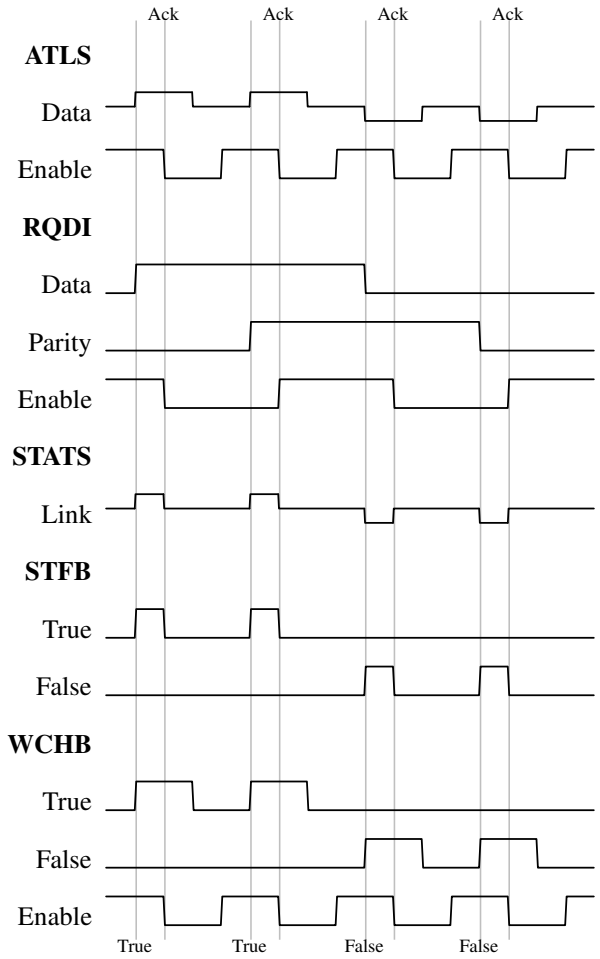

Fig. 1. Signaling Protocols. Transitions are aligned in time for readability; in general the different buffer types will *not* have the same latencies.

A. WCHB

The Weak-Conditioned Half Buffer (WCHB) [17] is a handshake reshuffling of the 4-phase dual-rail Quasi Delay-Insensitive [18] protocol, which we refer to as *e1of2* (*e* for "*enable*", an inverted-sense *acknowledge* signal). While there are other possible reshufflings such as the PCHB and PCFB [17] used for logic, we chose the WCHB variant because the buffer implementation is small, simple, and fast. Of all the schemes we study in this paper, the e1of2 protocol is the most conservative. The other link types relax timing assumptions or use more aggressive signaling techniques (e.g. low swing, single track). Evaluating the WCHB allows us to compare the effects of those decisions on throughput, energy, or area.

[1]For brevity, we use the same initialism to refer to both the signaling protocol and the buffer that implements it.

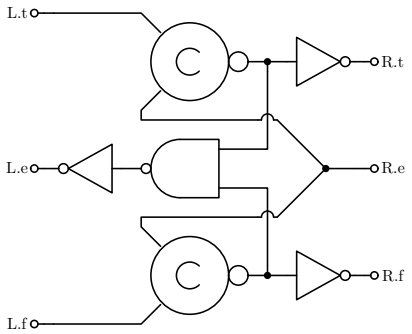

Fig. 2. WCHB Buffer

B. RQDI

The Relaxed Quasi Delay-Insensitive (RQDI) buffer design [19] implements a 2-phase, non-return-to-null (NRTN) protocol. It leverages a timing assumption already present in QDI circuits to reduce circuit complexity. We have implemented the LEDR [20] 2-phase encoding, although RQDI supports other 2-phase encodings. Our future multi-bit work will examine LETS [21] as well. We use RQDI to represent the state of the art in 2-phase, single-bit QDI links.

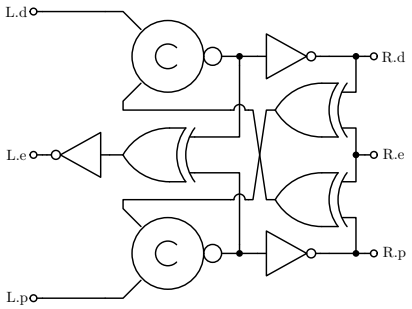

Fig. 3. RQDI Buffer

C. ATLS

Asynchronous Ternary Logic Signaling (ATLS) [22,23] is a 4-phase, QDI signaling protocol with a ternary delay-insensitive data encoding. This encoding compacts the dual-rail data wires into a single wire. V_{DD} encodes a *true* token, *GND* a *false* token, and $\frac{1}{2}V_{DD}$ represents the *null* state of the dual-rail encoding. The half-swing encoding reduces the energy cost of data rail transitions, which is attractive as a power saving measure but lowers static noise margins. The enable rail is still full-swing. ATLS simultaneously attacks the

problem of limited wiring resources and power consumption, hence its inclusion in our study.

Our implementation of ATLS differs from the proposed circuits in [22] and [23] as the proposed ternary decoding structures have not scaled well into deep submicron technologies. As in the original proposed circuits, we assume a $\frac{1}{2}V_{DD}$ power supply is available and account for it in our power measurements. We use the circuits described in Section II-E to encode/decode the ternary data rail. Since ATLS as proposed does not include any pipelining, we use an additional WCHB buffer when necessary as a pipelining element.

D. STFB

The Single-Track Full Buffer (STFB) [24] is designed for throughput. It uses a 2-phase, return-to-null (RTN) protocol with no control wires. It is, however, dual-rail, using a total of two wires to transmit a single bit. An upgoing transition on the *true* (*false*) rail encodes a *true* (*false*) token, and a downgoing transition on the rail signals an RTN. The sending process is responsible for raising a rail and the receiving process is responsible for lowering it. The single-track timing assumption requires that the sender and receiver are not simultaneously driving the rail, to avoid shorting the chip power supplies across a link.

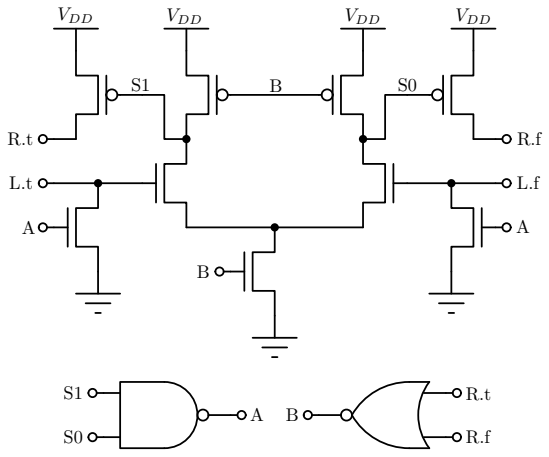

Fig. 4. STFB Buffer

E. STATS

Single-Track Asynchronous Ternary Signaling (STATS) is a single-track buffer template of our own design. The design goal was to reduce the total wiring resource requirements to a single wire. It combines the ternary encoding of ATLS with the 2-phase RTN, single-track handshake of STFB. To send a *true* (*false*) token, the sending process sets the wire to V_{DD} (*GND*). The receiving process returns the state to *null* by driving the wire to $\frac{1}{2}V_{DD}$. As with STFB, the single-track timing assumption requires that the sending and receiving process do not simultaneously drive the link.

To decode the ternary link, we use the pair of level shifter structures shown in Figure 5. The cross-coupling ensures full rail-to-rail swing, minimizing static power dissipation.

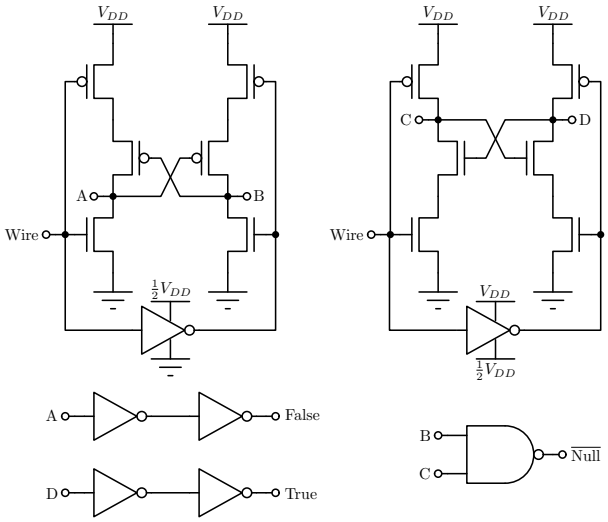

Fig. 5. Ternary Voltage Decoder

While the level shifters are fragile to pathological imbalances in pullup/pulldown network sizings, weakening the pullup/pulldown cross-coupled stacks with respect to their pulldown/pullup counterparts to a ratio of 1:2 is sufficient. Further increasing the drive strength disparity by changing transistor thresholds is recommended. The inverters and NAND gate should be sized to equalize load capacitances on nodes A, B, C, and D.

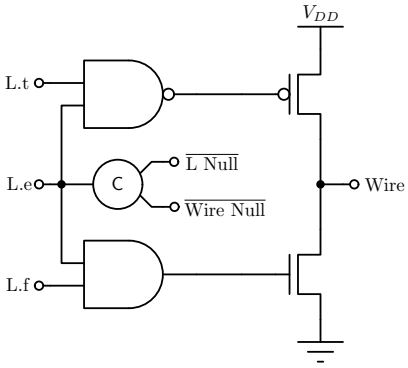

Fig. 6. STATS Transmit Stage: The null calculation for the Wire and L are obtained from the NAND from Figure 5 and the traditional NOR dualrail calculation (not pictured), respectively.

The link is driven to V_{DD} or *GND* by a single appropriately-sized pMOS or nMOS transistor, respectively, as shown in Figure 6. A parallel combination of one or more of the circuits in Figure 7 returns the link to the *null* state at $\frac{1}{2}V_{DD}$. We allow our analysis framework, described in Section III-B, to permute the combination and sizing of the RTN circuits to fully explore the tradeoff space.

- **Passgate** (Figure 7a): This circuit drives the link to $\frac{1}{2}V_{DD}$ using the least energy, by connecting to the $\frac{1}{2}V_{DD}$ supply. It is the most conservative of the three, but also the slowest.
- **Self-Invalidating Driver** (Figure 7b): The self-invalidating driver is the most aggressive of the three

designs, as it is essentially a full rail-to-rail transition interrupted halfway. While it offers the best slew-rate (a single RC time constant is more than a $\frac{1}{2}V_{DD}$ swing), it relies on the level-shifter structures in Figure 5 to resolve the state of the wire quickly and switch the True/False signals depicted in Figures 5 and 7b. A slow transition on either of those two signals will result in an overshoot of $\frac{1}{2}V_{DD}$ and potentially a spurious token on the link.

- **Shorted Inverter** (Figure 7c): The shorted inverter makes use of the CMOS inverter voltage transfer curve behavior to drive the wire very quickly to $\frac{1}{2}V_{DD}$. It is faster than the Passgate technique, but very energy inefficient as it essentially shorts V_{DD} to *GND* while enabled.

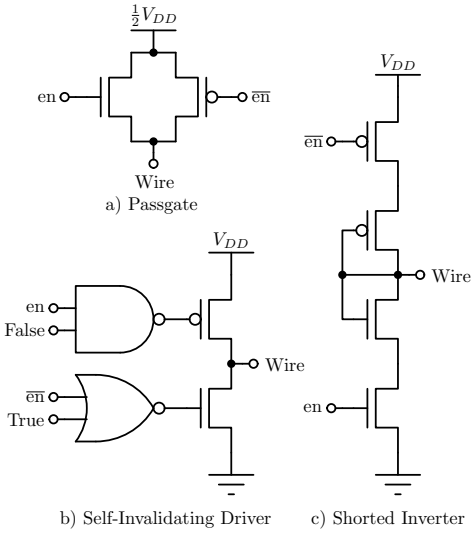

Fig. 7. Ternary Return to Null Schemes. en high starts the RTN process, and True and False are signals from the decoder shown in Figure 5.

III. Methodology

We constructed a framework to evaluate the various link types described in Section II across a wide range of operating points. Links were studied in two contexts: on-chip planar communication, and 3D signaling through TSVs.

A. Link Simulation

We used SPICE simulation to determine the throughput and energy for each link type. Figure 8 shows the basic Device Under Test (DUT) for these simulations. The link DUT is a FIFO pipeline, implemented at the transistor level. It is driven by an environment that generates pseudorandom tokens as fast as the link can accept them.

The link DUT also includes a distributed RC interconnect model (planar wire or TSV). Planar wires of a given length may be broken up into several shorter wires by adding extra buffers as pictured. TSV links cannot be so divided, as there is no way to insert a buffer in the middle of a TSV. We discuss interconnect models in more detail in Section IV.

In our simulations, the environment communicates using e1of2, and the cost of conversion to the protocol used by the DUT is included as part of the energy and area costs of the link. This simulates a fully-asynchronous system where

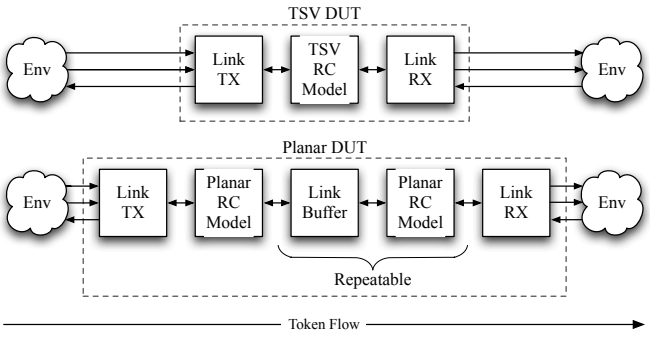

Fig. 8. Link DUT, for both planar and TSV contexts. Double-headed arrows represent link channels (e.g. STFB); 3-wire channels from the environment are e1of2. "Link TX" and "Link RX" convert to and from the link protocol, respectively, while "Link Buffer" is a native buffer for the protocol.

computation is done with islands of 4-phase QDI logic and the links are used to shuttle data across planar links or TSVs [25]. This assumption penalizes 2-phase protocols, but it is generally accepted that 2-phase computation is unwieldy in comparison to 4-phase [26] (with the possible exception of STFB [24]).

Figure 9 shows the complete simulation harness. Since the environment source and sink are implemented in Verilog, we use two WCHBs to decouple the DUT from any digital boundary effects. The harness is designed to operate faster than the DUT, so that link throughput is governed primarily by the DUT itself and the RC characteristics of the interconnect.

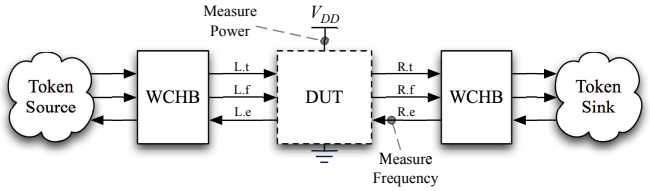

Fig. 9. SPICE simulation harness for the DUT. Average frequency is measured using the right-side enable signal, and power dissipation is measured for the DUT alone using a dedicated power supply.

ATLS and STATS buffers require an additional $\frac{1}{2}V_{DD}$ supply. In order to be fair, we allow RQDI, STFB, and WCHB links to run at a voltage lower than the harness V_{DD}. To support this, we implemented pipelined level shifters based on the WCHB template, shown in Figure 10. These are considered part of the environment, so they are not counted against the link energy and area. The usual protocol converters are still required in addition to these level shifters for non-e1of2 links.

B. Optimization Framework

The goals for our links are to maximize throughput, minimize energy dissipation, and minimize buffer silicon area. Because these measures are not independent, the solution to this multi-objective problem is a Pareto front of different buffer configurations situated in a three-dimensional tradeoff space between throughput, energy, and area.

To explore this space, we can apply multi-objective heuristic optimization algorithms. While heuristic optimization algorithms are not guaranteed to find the true Pareto front of

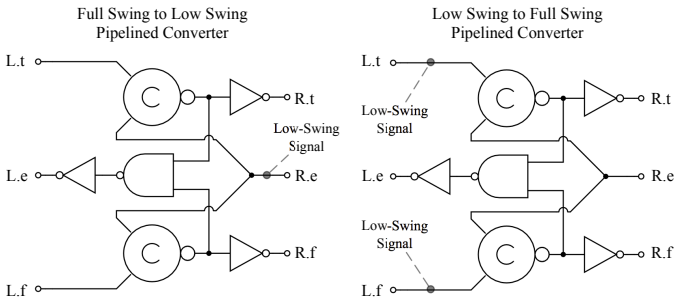

Fig. 10. WCHB Level Shifters. The C-elements have one low-swing input and one full-swing input. In the C-elements, the nMOS transistors connected to the low-swing input are LVT and sized double-width, and the pMOS transistors are HVT. All other transistors are standard VT.

a given space, i.e. the globally optimal front, in practice a reasonable approximation can be obtained.

We chose the DEAP [27] toolkit and its implementations of the $(\mu + \lambda)$ genetic algorithm[2] (GA) [28] and the widely-used NSGA-II [29] population selection algorithm. NSGA-II-based genetic algorithms are designed to provide a well-distributed family of points along a Pareto front, allowing us to capture the engineering tradeoffs in the design space. Some commercial tools [30] converge to a near-globally-optimal Pareto front faster than NSGA-II-based algorithms, but the same result can be obtained by NSGA-II given enough design space samples—typically, a few thousand is sufficient, and we sample at least 2850 points in the space for each link.

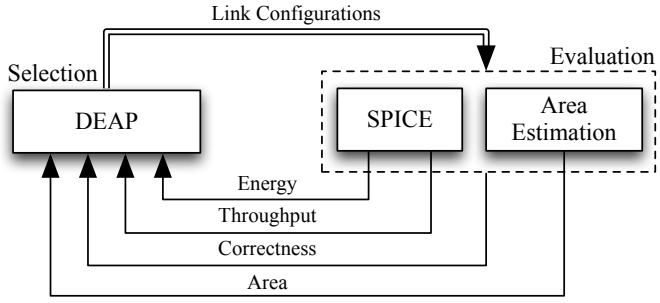

Fig. 11. Heuristic optimization framework for evaluating link circuit designs

To find the Pareto front for each link design, we built the optimization tool shown in Figure 11. Candidate link configurations are selected by DEAP, simulated, and evaluated based on relevant fitness criteria. Throughput and energy are measured using the SPICE harness described in Section III-A. Area is estimated as the total transistor area, i.e. the sum of $W \times L$ for all MOSFETs in the DUT. While this does not account for routing, etc., it provides a lower bound to make reasonable direct comparisons. Some link configurations may deadlock, send spurious tokens, violate dual-rail encodings, or present other failure modes. The environment checks for this and removes any failing configurations. All these evaluation results are fed back to DEAP and used to direct the selection of further candidate configurations.

[2]Two-Point Crossover ($c_p = 0.7$), Gaussian Mutation ($m_p = 0.2$), $\mu = 20$, $\lambda = 60$, $n_{gen} = 60$

It is worth noting that this optimization approach is not limited to on-chip links. The same general framework can be applied to any system with a parameterized configuration (genome) and a set of performance metrics (fitness).

TABLE II
LINK CONFIGURATIONS EXPLORED BY FRAMEWORK

Link Type	Transistor Sizing	Circuit Topology	Link Voltage
ATLS	✓	✓	
RQDI	✓		✓
STATS	✓	✓	
STFB	✓		✓
WCHB	✓		✓

The specific configuration parameters selected by our tool depend on the link type being optimized, and are summarized in Table II. For all link types, the framework selects transistor sizes for the circuits. Transistor sizing is usually handled by convex optimization algorithms, but since we want to explore the multi-objective space we allow DEAP to choose sizing, from minimum transistor width up to 100 times minimum.

For the ternary link types (ATLS, STATS), the framework can alter the circuit topology by choosing which combination of RTN schemes to use (Figure 7). For the other types, it can voltage-scale the link as described in Section III-A. Finally, in the planar context our tool can vary the number of buffer stages used (Figure 8). We account for the area and energy consumed by multiple planar buffers, but not the additional pipeline slack they provide (which may or may not be desirable depending on the specific system).

IV. EVALUATION AND DISCUSSION

We evaluate each link on the metrics of throughput, energy per token, and planar buffer area. In order to definitively conclude that one link protocol is "better" than another on these metrics, the Pareto front of the better link must completely dominate the other front—the three dimensional surfaces of the fronts must not intersect. Intersecting surfaces imply that the links being compared are situationally better than one another.

The rest of this section is devoted to examining the throughput/energy/area tradeoff space. To present the data in the most readable format, we have chosen to show two-dimensional projections of the three-dimensional Pareto front, omitting the dominated points on each plot. The bottom right quadrant of each plot represents the most attractive link configurations, as we are trying to maximize throughput and minimize energy/area.

In this study we look at three different technology nodes: a low-power $90\,\mathrm{nm}$ bulk process, a low-power $65\,\mathrm{nm}$ bulk process, and a high-performance $45\,\mathrm{nm}$ Silicon-on-Insulator (SOI) process. While we did not fabricate test structures in all three technologies, we were able to build WCHB and STATS planar links in our $90\,\mathrm{nm}$ process. We did not obtain isolated power measurements, but the SPICE-predicted frequency numbers for our test structures were within 7% of the actual silicon measurements. Since we have performed the same technology characterization steps for all three technologies when building our SPICE environments, we are reasonably confident in the simulation results presented in this section.

Our methodology does not directly account for robustness to noise or process, voltage, and temperature (PVT) variation. We provide a qualitative analysis of these factors here, but a complete characterization is pending in our future work.

A. Planar Links

In our planar link simulations, we model wires using a 100-segment π-model with RC parameters obtained from extracted layout. It is vital to use distributed wire models when studying long links, in order to avoid unrealistically optimistic results [5]. As a concrete example: STATS transceivers sense the state of the wire locally to determine when to stop driving. A lumped wire model would yield misleading results, since sender and receiver observe the same voltage. In reality, charge relaxation across the wire means that the voltages may differ and the sender may stop driving too soon—this places a restriction on the maximum slew rate possible for a given wire length.

Figure 12 shows the energy/throughput Pareto front for each buffer type in a $90\,\mathrm{nm}$ process. Each point in the front represents a different buffer configuration (transistor sizing, V_{DD}, number of buffer stages, etc.). The relative merit of each link type is similar across process technology generations, so we omit the $65\,\mathrm{nm}$ and $45\,\mathrm{nm}$ plots for brevity. Our evaluation framework allows us to examine the configuration of each individual point on a Pareto front and uncover Pareto-front-wide trends.

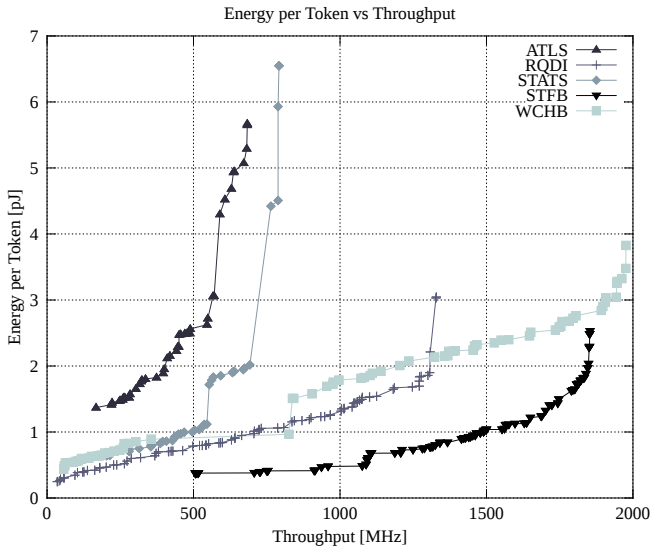

Fig. 12. Energy vs Throughput in $90\,\mathrm{nm}$ for a $1000\,\mu\mathrm{m}$ planar link

From Figure 12, we can see that RQDI and STFB are more energy efficient than WCHB with only a few exceptions. This is unsurprising, as 2-phase protocols like RQDI and STFB expend less energy by halving the number of transitions on the RC link. STFB goes one step further by removing the acknowledge wire and the associated drive circuitry, offering additional energy savings. STATS and ATLS are almost completely dominated by the full-swing protocols (RQDI, STFB, and WCHB). The obvious conclusion for the designer is that ternary signaling is a poor choice for planar wiring, which essentially behaves like an RC lowpass network and limits the throughput of low-swing signals. This is borne out by the fact that the high-throughput Pareto-optimal points for RQDI, STFB, and WCHB all run at full V_{DD} for all technologies, in spite of the capability to aggressively reduce V_{DD}.

An added downside to ternary signaling is that the energy cost of voltage level conversion in ATLS and STATS is quite high, especially when replicated many times in a multi-hop link. The sharp increases in energy per token in Figure 12 represent the addition of more buffers on a planar link. Examining the trends across links, STFB and WCHB gradually increase the number of buffers on the link as throughput increases—more buffers driving shorter links allows for higher frequencies. Conversely, STATS and ATLS increase the number of buffers only if aggressive transistor sizing is unable to achieve additional throughput. Figure 12 demonstrates the much greater energy cost of adding buffers to a STATS or ATLS link compared to a similar addition for STFB or WCHB. RQDI also mainly uses transistor sizing to achieve higher throughput. The vertical jump in energy and area at the very highest throughputs represents the addition of more buffers when aggressive sizing is not enough.

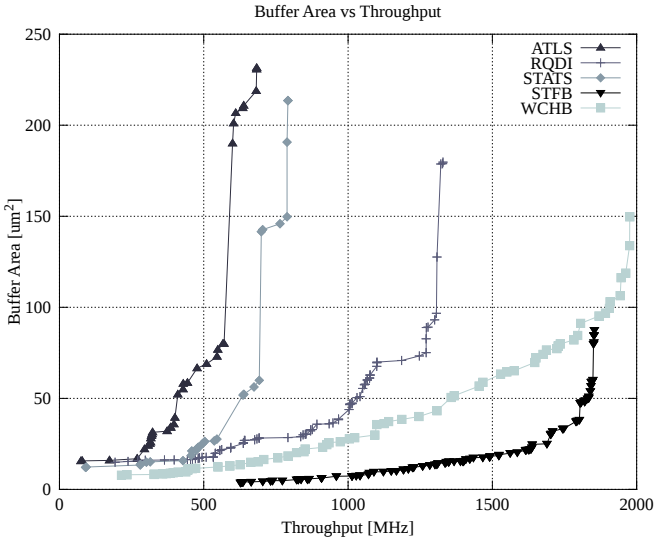

Fig. 13. Area vs Throughput in $90\,\mathrm{nm}$ for a $1000\,\mu\mathrm{m}$ planar link

Figure 13 shows the area/throughput Pareto front for each buffer type in our $90\,\mathrm{nm}$ process. The aggressive sizing of STATS and ATLS buffers can be seen here—the almost $100\,\mu\mathrm{m}^2$ increase in area around $400\,\mathrm{MHz}$ and $550\,\mathrm{MHz}$ represents the addition of a single ATLS or STATS buffer stage, respectively. STFB is best in area, as it has the lowest transistor count per buffer of any link.

As discussed in Section II-E, the Passgate RTN scheme is used for low-throughput, energy-efficient STATS and ATLS configurations, while the faster, more aggressive Self-Invalidating Driver is used in high-throughput links. For average throughput, a mix of these two schemes is used. The Shorted Inverter RTN scheme is only used for the highest throughput link configurations, where energy costs are already high.

As a general observation (that also holds for TSV links as seen in Section IV-B), ATLS is never optimal and almost always dominated by every other buffer. The link, as proposed by [22], is more of a data encoding than an actual link design. While we improve some of the circuit designs as described in Section II-C, we still use WCHBs as a pipelining element. Including WCHBs in series adds extra transitions/cycle and power, adversely affecting throughput and energy. In an attempt to maximize the frequency, DEAP selected large transistor sizes, which makes ATLS look unattractive in area as well. A redesign of ATLS that combines the pipelining element with the encode/decode structures could improve its Pareto efficiency performance.

Figures 14 and 15 show composite Pareto fronts across all technology nodes, for energy/throughput and area/throughput respectively. In other words, the curves on these plots represent the best buffers in that technology at each given operating point. In order to compare results across technology nodes, we scale link length by the technology feature size. Results presented below are for a link length of 20,000λ, equivalent to $1000\,\mu\mathrm{m}$ in a $90\,\mathrm{nm}$ technology.

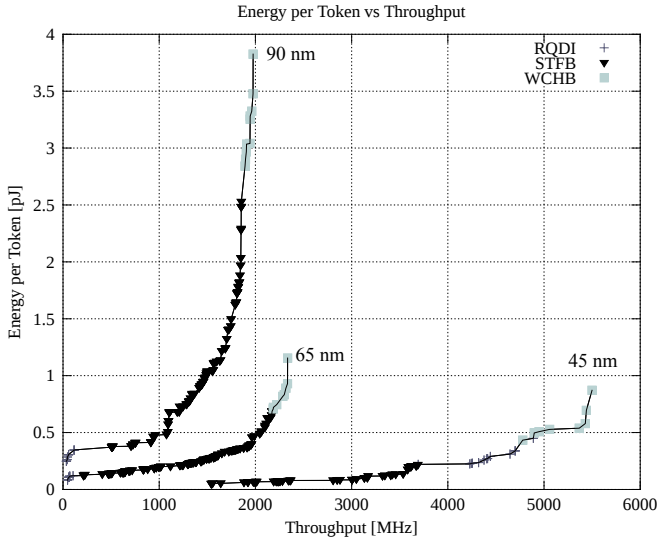

Fig. 14. Energy-throughput Pareto-dominant points across planar technologies

The results are consistent across technologies: STFB buffers are the most energy- and area-efficient for planar signaling across most of the range. At the very highest throughputs WCHB (and $45\,\mathrm{nm}$ RQDI) buffers continue to operate after STFB fails, but at a greatly increased cost in energy per token. This high energy is due to aggressive transistor sizings, reflected in extravagant area usage as shown in Figure 15. From these results alone, STFB is the clear winner in the planar context for all but the most aggressive throughput targets. However, the single-track timing assumption makes STFB less robust than QDI buffers, as we discuss in Section IV-C. This presents a tradeoff to the designer between energy/area usage and ease of design. The additional cost of "robustness" is not prohibitive, as can be seen by comparing STFB against the QDI buffers (WCHB for high throughput, RQDI for lower) in Figures 12 and 13.

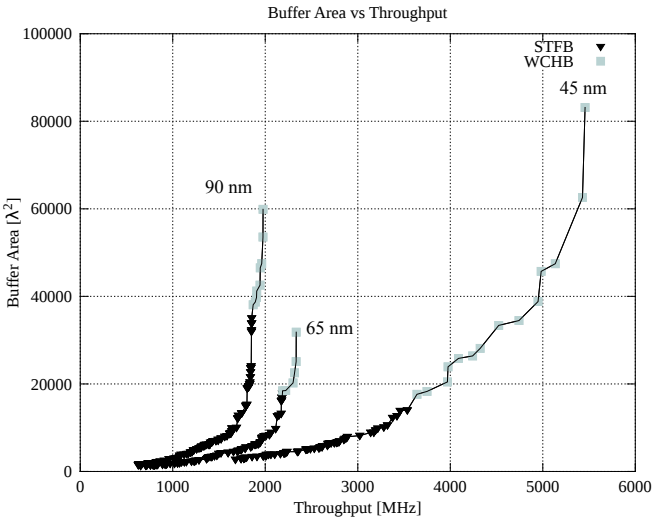

Fig. 15. Area-throughput Pareto-dominant points across planar technologies

In the planar context, designers also have control over interconnect wire spacing, which has a direct effect on coupling capacitance. We found that a change from minimum to sparse spacing (twice minimum) chiefly impacted energy per token. For brevity, we report the average energy improvements at a given frequency for each link in Table III, as opposed to including additional Pareto fronts. Note that this table does not capture the additional benefits of reduced crosstalk due to the increased spacing.

TABLE III
PERCENTAGE IMPROVEMENT IN SPARSE WIRING ENERGY

Link	90 nm	65 nm	45 nm
ATLS	47.36	16.93	-24.67
RQDI	33.71	7.22	13.98
STATS	27.42	-92.28	-112.87
STFB	39.04	18.11	12.26
WCHB	49.66	28.43	20.99

For most link/technology parings, the results shown in Table III are as expected. To first order, wire resistance remains constant with increased wire spacing while coupling capacitance decreases. This decrease in capacitance leads to lower $\frac{1}{2}CV^2$ energy dissipation.

Strangely, the sparse wiring energy *increases* for STATS and ATLS. The root cause of this energy increase is that for high-throughput link configurations, DEAP selects more highly pipelined links. As an example, in $45\,\mathrm{nm}$, the fastest running ATLS and STATS configurations divided the planar wiring into 10 sections for the sparse wire spacing case, and only 5-6 for the minimum spacing case.

In order to implement single-track timing (sender and receiver must not drive a wire simultaneously), STATS inspects the local voltage to determine when to cut off the driving transistors. If the interconnect resistance is high relative to its capacitance (as in sparse wiring), the buffer may see the local voltage change and turn off before moving enough charge to resolve the state transition at the remote end of the wire. This tends to favor shorter wires with more buffers, leading to greater energy consumption. High-throughput ATLS

configurations use the fast Self-Invalidating Driver, which has the same property.

B. TSV Links

To simulate 3D links between stacked dies, we use the TSV model from [12], modified to have distributed rather than lumped RLC components. It represents a $20\,\mu\mathrm{m}$ diameter copper TSV with $25\,\mu\mathrm{m}$ pitch in a digital process. We also model coupling capacitance to Manhattan neighbor TSVs. TSV fabrication is usually a separate step from the rest of the CMOS process and scales at a different rate, so we use the same TSV model for all process technologies in this study.

Because TSV pitch is much larger than the standard via pitch, we assume that TSVs are a scarce resource. As a result, we report throughput per-TSV below (by scaling using wire counts from Table I). This penalizes buffers that require more wires to send a single bit of data.

Figures 16 and 17 show the energy/throughput and area/throughput Pareto fronts in a $90\,\mathrm{nm}$ process for each buffer type communicating vertically through a TSV link. Since buffers in each technology must drive the same TSV structure, reported buffer area is not scaled as it was for the planar results.

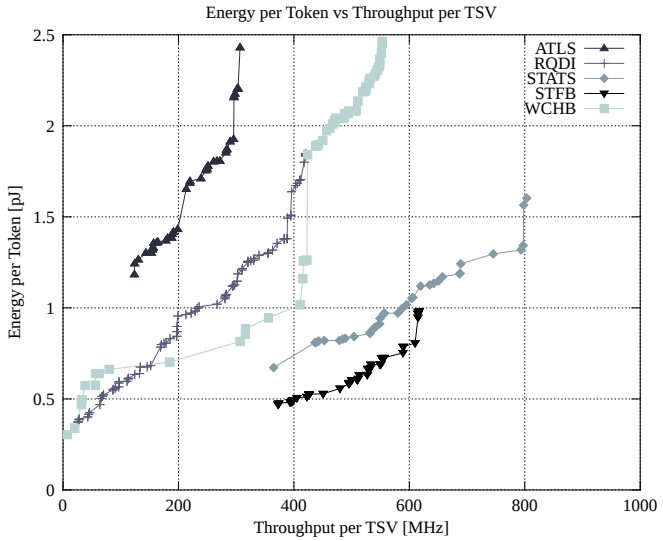

Fig. 16. Energy vs Throughput in $90\,\mathrm{nm}$ for a $25\,\mu\mathrm{m}$ pitch TSV model

In the TSV context, STATS is a strong contender due to its efficient use of TSV resources. Furthermore, TSVs have high capacitance but low resistance compared to planar wires, due to the sheer amount of conductive material. This environment approaches the ideal lumped capacitance case where a low-swing link such as STATS excels—theoretically, a half-swing protocol would expect to see 4x savings in $\frac{1}{2}CV^2$ switching energy. In practice, the ternary conversion energy cost cuts into this savings, but STATS is more attractive in energy/throughput-per-TSV than all other links save STFB.

An interesting phenomenon is the sharp energy increase for WCHB buffers in Figure 16 around $400\,\mathrm{MHz}$. Examining the link configurations that straddle this increase revealed essentially identical configurations save for one gate: the

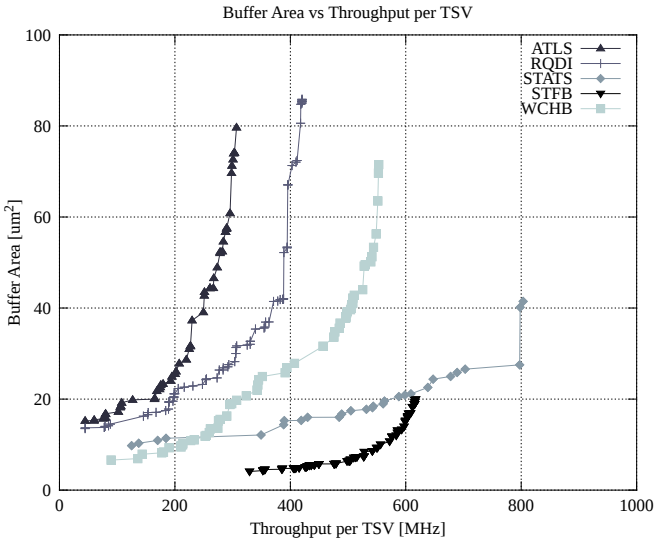

Fig. 17. Area vs Throughput in $90\,\mathrm{nm}$ for a $25\,\mu\mathrm{m}$ pitch TSV model

inverter driving the returning L.e acknowledge signal (shown in Figure 2) for the Link RX unit (shown in the TSV DUT section of Figure 8). The higher-energy configuration was a *maximal* sizing of this inverter, whereas the lower-energy point was a *minimal* sizing of the same inverter. This phenomenon is present in all three technologies. It occurs in the planar case as well: the slight discontinuity in WCHB energy per token seen in Figure 12 around $800\,\mathrm{MHz}$ displays the same jump in driver sizing. There are other effects at work in the planar case (e.g. number of planar buffers), but the trend is still noticeable.

While further investigation is warranted, this suggests two Pareto efficient operating regimes for the WCHB link. In the first mode, throughput is unaffected by a slow transition on the enable signal—the link is not token-hole-limited. At some throughput threshold, however, the system becomes token-hole-limited and a fast acknowledge transition (with associated energy cost) is required to see further improvement. Examination of the dominated points in the DEAP runset revealed that DEAP had tried many similar configurations, more or less holding all other parameters constant and varying the sizing of the L.e inverter across the range of allowable sizings—in other words, this phenomenon is quite unlikely to be an artifact of the heuristic optimization algorithm. Intuitively, a small increase in the inverter sizing would offer negligible throughput gains with an energy penalty. Conversely, a downsizing of a maximal inverter would penalize throughput without much benefit to energy. Furthermore, it is likely that an algorithm that sizes transistors based on their electrical environment alone would not have discovered these two operating regimes. Such an algorithm would have sized the L.e inverter to drive the large TSV capacitance and missed out on the low-energy WCHB configurations.

A cross-technology examination of TSV links, plotted in Figures 18 and 19, is slightly more complicated than the planar scenario. We use the same TSV structure across all technologies, so the electrical characteristics of the physical link remain constant while the transistors shrink. This leads to

STFB failure (described in Section IV-C) and its disappearance from the Pareto front after 90 nm.

Measured on a throughput/TSV basis, STATS (which uses only one TSV) dominates. The QDI buffers (RQDI, WCHB) are penalized due their 3-wire interface, but also appear on the Pareto fronts at low throughput/TSV.

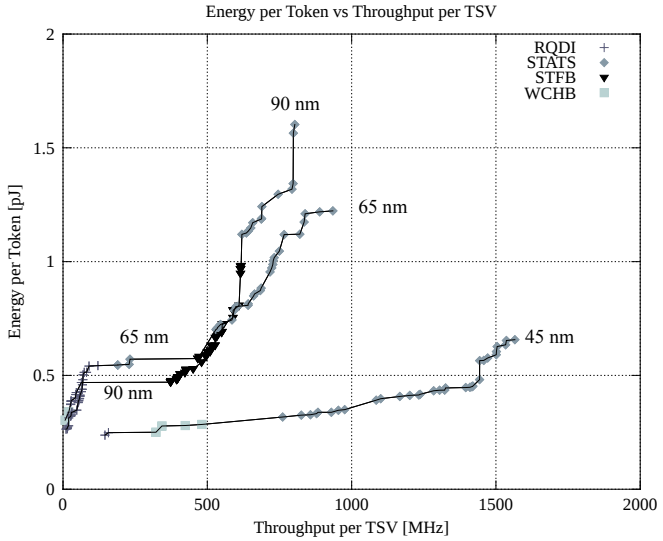

Fig. 18. Energy-throughput Pareto-dominant points for 25 µm pitch TSV

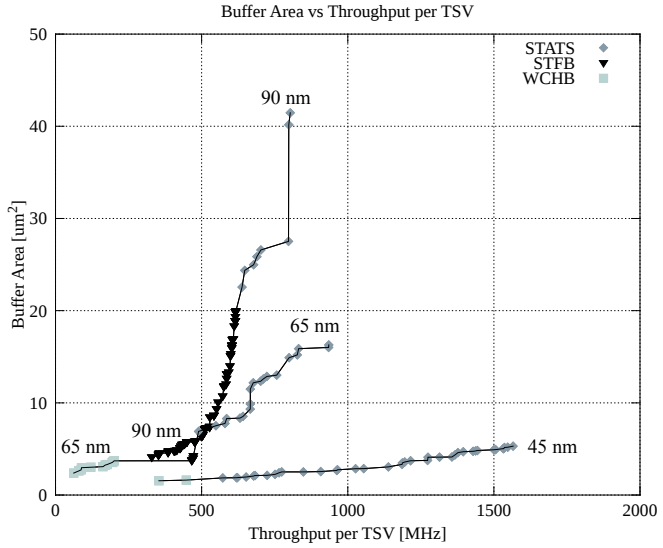

Fig. 19. Area-throughput Pareto-dominant points for 25 µm pitch TSV

Examining silicon area (Figure 19) instead of energy per token, WCHB (not RQDI) is the smallest for low per-TSV throughput, for similar reasons to the planar case. The other rankings are the same as for energy.

C. Link Failures and Reliability

In addition to throughput, energy, and area, we record the failure rate of individual configurations selected by DEAP. These statistics are reported in Table IV for planar and TSV links across our three process technology nodes. As discussed in Section III, we verify that links send tokens correctly and do not deadlock. Failures are typically due to poorly-sized transistors driving large RC loads, since DEAP can choose sizes at random. Roughly speaking, these failure rates provide information about how easy a link is to design and how robust it is to sizing variation. While a significant amount of additional work is required to quantify link robustness, we believe the failure rate is of use in building an intuitive understanding of a link's timing assumptions and relative design difficulty.

TABLE IV
LINK FAILURE RATES

Link	% Planar Failure			% TSV Failure		
	90 nm	65 nm	45 nm	90 nm	65 nm	45 nm
ATLS	23.94	16.34	19.23	17.72	20.83	15.54
RQDI	25.60	23.93	17.80	19.72	21.52	24.68
STATS	42.40	36.26	45.45	33.26	33.96	33.31
STFB	28.18	21.99	33.63	29.19	99.33	100.00
WCHB	10.67	8.49	12.43	12.79	12.80	25.32

Note: $2856 \leq n \leq 11158$

STATS has the highest failure rates of all buffer types in planar and the second highest in the TSV context. This is not surprising, as STATS combines both ternary encoding and the single-track timing assumption to achieve its single-wire goal, and each of these techniques reduce reliability compared to a delay-insensitive link.

Ternary decoders (Figure 5) are particularly sensitive to sizing variations, and their failure prevents the buffer from sensing the link state correctly. Even an accurate but slow decoder may cause link failure, for example by causing a Self-Invalidating RTN Driver (Figure 7b) to overshoot $\frac{1}{2}V_{DD}$. This impacts the failure rates of both STATS and ATLS.

As discussed in Section IV-A, single-track timing can cause STATS to fail if the interconnect resistance is too high (planar wiring). This is less of an issue in the high-capacitance, low-resistance TSV context, so we see correspondingly lower failure rates in Table IV. ATLS, RQDI, and WCHB do not suffer from this problem, due to the QDI timing of their handshake. Reasonably slow transitions are acceptable, as they will be not be acknowledged until the receiving end can resolve the wire state.

STFB uses an even more aggressive single-track timing assumption. The STATS level shifter structures offer a better inspection of the wire state due to hysteresis, whereas the link wire directly drives the STFB handshaking logic as seen in Figure 4. As a result, STFB has simpler circuits and better throughput, but at the expense of robustness. The traces shown in Figure 20 were selected from the fastest five STFB and STATS TSV link configurations in 90 nm. The STFB *true* and *false* rails do not complete full-swing transitions. Examining the figure, it takes at least two tokens traversing a link (and driving the link pulldown network) to return the wire state to *GND*. In contrast, STATS transitions cleanly between V_{DD}, $\frac{1}{2}V_{DD}$, and *GND* because it inspects the voltage before cutting off the transistors driving the wire.

In short, STFB is releasing the pull-up network and pull-down networks too early. Large capacitances exacerbate this problem. Because the TSV RC characteristics in our model

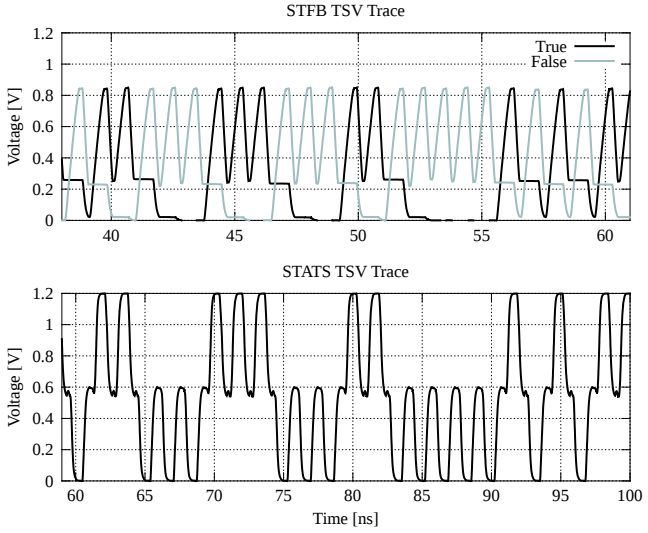

Fig. 20. Trace of STFB and STATS TSV links in 90 nm. V_{DD} is 1.2 V. Times shown are matched by sent tokens.

do not scale with technology—we assume they are separately fabricated—but transistor switching speed does, the timing margins become progressively worse for STFB with each smaller technology. This is reflected in the increased number of failing configurations, shown in Table IV, and the complete failure of STFB to drive the TSV link in 45 nm.

One side effect of this timing failure is that STFB becomes a de facto low-swing signaling protocol, which artificially improves its energy efficiency. While this is certainly not without merit, it comes at the cost of noise margins and robustness. While we do not model noise sources, the STFB traces shown in Figure 20 are more susceptible to noise than a full-swing signal would be. Note that ternary encodings (ATLS, STATS) also have reduced noise margins compared to a full-swing signal.

V. Conclusion

We studied five self-timed single-bit signaling protocols with widely varied properties (timing assumption, wire count, voltage swing), including our proposed STATS single-wire link design. We developed a multi-objective optimization tool to evaluate the performance (throughput, area, and energy per bit) of these protocols for both traditional planar wiring and 3D inter-die communication using TSVs.

Pareto front analysis is a powerful framework for evaluating competing objectives. From this study, we draw several conclusions useful for circuit designers. For planar links, STFB offers the best performance across the range of process technologies studied, though its tight timing requirements do not cope well with non-ideal wires. WCHB is also a good choice, trading some energy and area for increased robustness. STATS performs poorly with planar wiring but makes efficient use of scarce TSV resources, and is a good match for TSV electrical characteristics.

Future work will include a quantitative approach to characterizing signaling protocol robustness, especially for single-track and ternary links. We also plan to expand our analysis to include multi-bit links.

Acknowledgements

This research was supported in part by AFRL award FA8750-12-2-0035, by NSF award CCF-1065307, and by the TRUST STC center CCF-0424422.

References

[1] J. L. Neves and E. G. Friedman. "Optimal clock skew scheduling tolerant to process variations." "IEEE DAC," 1996.
[2] K. Stevens, *et al.* "Energy and Performance Models for Synchronous and Asynchronous Communication." *IEEE VLSI*, 19(3):369–382, 2011.
[3] A. Martin. "Asynchronous logic for high variability nano-CMOS." "IEEE ICECS," pp. 69–72. 2009.
[4] R. Ho, *et al.* "The future of wires." "Proc. IEEE," pp. 490–504. 2001.
[5] S. Gilla, *et al.* "Long-Range GasP with Charge Relaxation." *IEEE ASYNC*, pp. 185–195, 2010.
[6] D. Liu and C. Svensson. "Power consumption estimation in CMOS VLSI chips." *IEEE SSC*, 29(6), 1994.
[7] S. W. Keckler, *et al.* "GPUs and the Future of Parallel Computing." *IEEE Micro*, 31(5):7–17, 2011.
[8] S. Borkar. "3D integration for energy efficient system design." "IEEE DAC," pp. 214–219. 2011.
[9] K. Banerjee, *et al.* "3-D ICs: a novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration." "Proc. IEEE," 2001.
[10] F. Akopyan, *et al.* "Variability in 3-D integrated circuits." *IEEE CICC*, pp. 659–662, 2008.
[11] T. R. Harris, *et al.* "A Transient Electrothermal Analysis of Three-Dimensional Integrated Circuits." *IEEE CPMT*, 2(4), 2012.
[12] R. Weerasekera, *et al.* "On signalling over Through-Silicon Via (TSV) interconnects in 3-D Integrated Circuits." "IEEE DATE," pp. 1325–1328. 2010.
[13] K. Stevens. "Energy and performance models for clocked and asynchronous communication." "IEEE ASYNC," pp. 56–66. 2003.
[14] I. Sutherland. "Micropipelines." *CACM*, 32(6), 1989.
[15] I. Sutherland and S. Fairbanks. "GasP: a minimal FIFO control." *IEEE ASYNC*, pp. 46–53, 2001.
[16] C. H. Svensson and J.-R. Yuan. "A 3-level asynchronous protocol for a differential two-wire communication link." *IEEE JSSC*, 29(9), 1994.
[17] A. Lines. *Pipelined Asynchronous Circuits*. Master's thesis, California Institute of Technology, 1995.
[18] A. Martin. "The Limitations to Delay-Insensitivity in Asynchronous Circuits." "6th MIT Conference on Advanced Research in VLSI," Proceedings of the 6th MIT Conference on Advanced Research in VLSI, 1990.
[19] C. LaFrieda and R. Manohar. "Reducing Power Consumption with Relaxed Quasi Delay-Insensitive Circuits." *IEEE ASYNC*, pp. 217–226, 2009.
[20] M. E. Dean, *et al.* "Efficient self-timing with level-encoded 2-phase dual-rail (LEDR)." "Advanced Research IN VLSI," pp. 55–70. IEEE ICCD, 1991.
[21] P. B. McGee, *et al.* "A Level-Encoded Transition Signaling Protocol for High-Throughput Asynchronous Global Communication." *IEEE ASYNC*, pp. 116–127, 2008.
[22] T. Felicijan and S. Furber. "An asynchronous ternary logic signaling system." *IEEE VLSI*, 11(6):1114–1119, 2003.
[23] J.-M. Philippe, *et al.* "An energy-efficient ternary interconnection link for asynchronous systems." *IEEE ISCAS*, pp. 4 pp.–1014, 2006.
[24] M. Ferretti and P. Beerel. "Single-track asynchronous pipeline templates using 1-of-N encoding." "IEEE DATE," pp. 1008–1015. 2002.
[25] A. Martin and M. Nystrom. "Asynchronous Techniques for System-on-Chip Design." *Proc. IEEE*, 94(6):1089–1120, 2006.
[26] W. F. McLaughlin, *et al.* "Asynchronous Protocol Converters for Two-Phase Delay-Insensitive Global Communication." *IEEE VLSI*, 17(7):923–928, 2009.
[27] F.-A. Fortin, *et al.* "DEAP: Evolutionary Algorithms Made Easy." *Journal of Machine Learning Research*, 2012.
[28] T. Back, *et al.* "A survey of evolution strategies." *Proceedings of the Fourth International Conference on Genetic Algorithms*, 1991.
[29] K. Deb, *et al.* "A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II." *Kan-GAL report*, 200001, 2002.
[30] N. Chase, *et al.* "A Benchmark Study of Multi-Objective Optimization Methods." Technical Report BMK-3021, Red Cedar Technology.