Research Summary

July 14, 2018

Prof. Steven Nowick
Department of Computer Science (and by courtesy, Electrical Engineering)
Columbia University, New York, NY

OVERVIEW

Introduction

My main research is on asynchronous and mixed-timing digital design. Asynchronous circuits have no centralized or global clock. Instead, they are distributed hardware systems where multiple components coordinate and synchronize at their own rate on communication channels. Effectively, they can be viewed as "object-oriented hardware," where components can be designed separately, and flexibly connected by channels using systematic protocols, in a "Lego-like" manner, without any notion of global timing.

As chips grow increasing larger and faster, power and design-time requirements become more aggressive, and timing variability becomes a critical factor, there are increasing challenges in assembling centrally controlled synchronous systems.

Asynchronous design has the potential to offer significant improvements in performance, energy, reliability and scalability, since it eliminates the rigidity and overhead of the fixed-rate clock, and allow flexible and distributed assembly and communication of components. In particular, it can provide *low power* (components activated only on-demand, without the need to instrument clock gating, and entirely eliminating the global clock); *high performance* (some asynchronous systems have significantly lower latency and increased average throughput, rather than be bound to a worst-case clock rate); *great robustness to timing variability and unpredictability*; and *modularity and composability*. There is also a recent surge of interest in industry in hybrid designs, which connect standard synchronous components (e.g. processors, accelerators, caches/memories) through flexible asynchronous interconnection networks, forming *globally-asynchronous locally-synchronous or GALS systems*, where the asynchronous network provides a scalable and reliable integration medium.

However, asynchronous design poses several unique challenges: (i) circuits must be hazard-free (i.e. avoiding the potential for glitches); (ii) developing correct and effective computer-aided design tools and optimization techniques has been difficult, as well as harnessing existing synchronous tools to be applied to asynchronous systems; and (iii) the analysis and verification of such concurrent systems poses particular obstacles**.** These have been impediments to industrial acceptance, but in the last few year there have been major breakthroughs in overcoming these challenges.

My key goal is to make asynchronous and GALS digital design a viable foundation for hardware system design.

Recent Advances: Asynchronous and GALS Design

There has been a remarkable resurgence of interest in asynchronous design since the late 1990's, and especially in the last decade.

Several application areas are flourishing. While asynchronous design is not yet fully mainstream, there are multiple points of use, both experimentally and industrially, which are motivated both by the challenges of organizing modern large-scale systems with rigid fixed-rate clocks, and the opportunities, cost benefits and flexibility provided by un-clocked digital systems.

Large-scale system integration:

The use of mixed synchronous components integrated by asynchronous networks, i.e. GALS, as well as fully asynchronous systems, has become widespread. The introduction of asynchronous on-chip networks (i.e. "networks-on-chip" [NoCs]) to integrate these systems is flourishing, with applications ranging from STMicroelectronics' "STHORM" parallel processor, to several recent "neuromorphic processors (see below), as well as research designs from Intel, experimental evaluation by AMD Research with our own NoC designs, and many academic efforts.

Our own research advances in designing asynchronous NoC's demonstrate significant cost savings are possible over leading synchronous NoC's (area, power, latency), along with greater ease of integration. We have also recently developed a complete automated computer-aided design (CAD) tool synthesis flow, using synchronous CAD tools. (See below.)

Neuromorphic processors:

Several recent visible efforts are exploring the development of non-von Neumann parallel computer architectures, using "brain-inspired" organization. In these systems, processing elements correspond to models of neurons, while the interconnection network corresponds to synapses. These processors target rapid and energy-efficient processing for human-oriented applications: vision, pattern recognition, etc. A number of processors in the last five years have been developed: (i) *Intel's Loihi* (IEEE Micro, 2018), (ii) *IBM's TrueNorth* (Science, cover story, 2014), (iii) *Stanford's Neurogrid* (Proc. of IEEE, 2014; Scientific American, cover story, 2005), and (iv) *Manchester's Spinnaker* (multiple publications). *All of these neuromorphic processors use asynchronous on-chip networks, and several have nearly entirely asynchronous processing elements.* Asynchrony has been demonstrated to be a critical enabler, both for the large-scale system integration (TrueNorth had > 5B transistors and 4K cores) as well as for extremely low power.

Our own highly-efficient codes for robust asynchronous channel communication, called "LETS" codes, have been incorporated into the Stanford "Neurogrid" processor, to enable its low-energy inter-synaptic communication. (See section below.)

Industrial advances:

Recent industrial uptake, beyond the above NoC's and neuromorphic chips, include: (i) a complete asynchronous commercial tool flow and design methodology at Philips Semiconductors (1990's-early 2000's), with over 700 million asynchronous chips sold, for low- to moderate-performance applications such as contactless smart cards, digital passports, automotive, pagers, cell phones, etc.; (ii) Intel acquired Fulcrum Microsystems in 2011, for the development of low-latency Ethernet switch chips; (iii) Achronix Semiconductor developed a series of Speedster commercial asynchronous FPGAs, which could operate at 1.5 GHz, and were claimed as the world's fastest.

Our own collaborations -- with AMD Research, NASA Goddard, and IBM -- have contributed significantly to demonstrating the benefits of asynchronous design, in industrial settings, with direct experimental comparisons to commercial synchronous designs. (See "Technology Transfer" below.)

Applications to emerging areas, architectures and technologies:
Asynchronous design has been showing great promise in the last decade or so, for applications to a variety of emerging areas. The resiliency of robust-style asynchronous circuits to noise, large temperature and voltage variations, and uncertain or extreme environments, without the requirement to meet fixed clock rates, has shown substantial cost and reliability benefits in multiple domains. These include: *ultra-low energy systems* (sub-/near-threshold), *energy harvesting* (with highly-variable power availability), and *handling of extreme temperature ranges* (applications for space missions). Significant benefits have also been shown for asynchronous circuits for alternative computing paradigms, which include large timing deviations and challenges in clock distribution: *flexible electronics, cellular nano-arrays,* and *nano-magnetics*.

Our own research in emerging areas includes the development of *"continuous-time digital signal processors"* (CT-DSPs). This work demonstrates, for the first time, a complete and full-functionality architecture that allows amplitude-based sampling, with no aliasing, through the use of asynchronous control and datapath, coupled with real-time delay lines. (See section below.)

Many of these asynchronous advances, and others, are summarized in our recent 2-part overview article:

S.M. Nowick and M. Singh, "Asynchronous Design-Part 1: Overview and Recent Advances," IEEE Design & Test of Computers, vol. 32(3), pp. 5-18, 2015.
S.M. Nowick and M. Singh, "Asynchronous Design-Part 2: Systems and Methodologies," IEEE Design & Test of Computers, vol. 32(3), pp. 19-28, 2015.

Summary
While the above applications are in a wide variety of areas, they exhibit a few commonly recurring themes, representing beneficial and unique opportunities for asynchrony:

(i) extreme fine-grain pipelining:
The ability to implement and exploit extremely fine-grain – even gate- and bit-level – pipelines, unconstrained by the need to distribute a high-speed fixed-rate clock;

(ii) exploiting data-dependent completion times:
To support and micro-architect systems which can exploit subtle and fine-grain differences in data-dependent completion, i.e. at a sub-cycle granularity;

(iii) avoiding challenges due to the rigidity of global timing:
Supporting new computation paradigms (continuous-time DSP's, cellular nano-arrays, nano-magnetics, energy harvesting, flexible electronics), extreme micro-parallelism, dynamic computational adaptivity, and ease of large-scale system integration;

(iv) robustness to process, voltage and temperature variation:
Allowing flexible accommodation of dynamic timing variations; and

(v) "on-demand," i.e. event-driven, operation:
Highly energy-proportional computing, without the need for extensive instrumentation of clock-gating at multiple design levels.

While the above themes capture promising opportunities, and the current industrial uptake indicates increasing commercial viability and interest, there remain open issues and challenges that asynchronous design must overcome to gain wider industry adoption. These include the development of automated CAD tool flows, test methodologies, robust design techniques, foundational components and architectures, and viable demonstrators. Our work is aimed to address these issues.

My Key Research Projects

My goal is to advance the viability of asynchronous and GALS design, as a paradigm to address the current critical engineering bottlenecks of scalability, power, and performance. There are currently eight main projects in my group. **Items (i), (iv) and (viii) cover some key application areas**, **while items (ii), (iii) and (v)-(vii) are on developing foundations and infrastructure.**

(i) developing low-power and high-performance **on-chip interconnection networks,** for high-performance parallel computers, as well as moderate-performance low-power embedded systems for consumer electronics, which efficiently and easily integrate heterogeneous components into a single system, and provide significant cost benefits over leading synchronous designs

(ii) designing robust **mixed-timing interface circuits,** to flexibly accommodate mixed-clock and clocked/asynchronous timing domains

(iii) developing practical **high-speed pipelines,** to support high-performance systems

(iv) developing a promising new class of signal processors**,** called **continuous-time digital signal processors (CT-DSPs),** whose operation adapts dynamically to the input sample rate, and which can accommodate a wide variety of input formats and sample rates with no design change, and which provides a signal-to-error ratio for some applications which exceeds that of clocked systems, *for use with a wide range of speech and audio applications*

(v) developing **ultra-low energy and reliable digital design,** using sub- and near-threshold circuits, to support critical applications requiring long battery lifetime and ultra-reliable operation, such as *bio-medical implants, sensors for environmental and infrastructure monitoring,* and *space applications*

(vi) developing robust **channel encoding for global asynchronous communication,** to support reliable and low-power systems

(vii) developing **computer-aided software design (CAD) tools and optimization algorithms,** for digital system design in each of the above areas, including techniques to perform design-space exploration on tradeoffs between hardware complexity, power, performance and reliability

(viii) promoting **technology transfer** of my asynchronous CAD tools and circuit design styles to industry

The following is a summary of some highlights of my research since the year 2000 (when I received tenure). In addition to the research overview, it also includes a summary of recent grants.

1 ON-CHIP INTERCONNECTION NETWORKS

Introduction and Motivation

Structured on-chip interconnection networks, also called *"networks-on-chip" (NoCs)*, have become the de facto standard in the last decade for system integration. These communication fabrics support the cost-effective and modular incorporation of multiple cores, accelerators, and caches/memories, for both large-scale parallel architectures and heterogeneous embedded systems.

Unlike classic bus-based architectures (with their significant costs in performance and resource contention) as well as full point-to-point connection networks (with their complexity and over-design), these networks borrow ideas from the macro-level networking community to systematically create a communication block that provides efficient resource sharing, quality-of-service, low-power, high-performance and scalability.

Given the modularity of these communication blocks, and the huge heterogeneity that must be supported when attaching multiple processing elements (which may operate at different or varying frequencies and voltages, or with distinct protocols from different vendors), with dynamically complex traffic patterns, there has been a huge interest in the use of asynchronous NoCs as a flexible integration medium. The asynchronous NoCs can be used to support either fully asynchronous systems, or to "glue" together diverse synchronous components in a GALS system. The ability to form an elastic asynchronous integrative communication medium, without any requirements of global clock distribution, and without the complexity of managing clock-gating at fine granularity and with highly-variable traffic, becomes highly attractive.

In practice, though, while there has been significant promise from several of the proposed asynchronous NoCs in the last 15 years, many of these designs suffer from significant overheads.

Most asynchronous NoCs have been based on highly robust asynchronous design styles, using *delay-insensitive (DI) codes* **on channels**, such as dual-rail or 1-of-4 codes. Such channel encoding provides great resiliency to uncertainty in packet arrival times, as well as tolerating arbitrary inter-bit timing skews. This approach also greatly simplifies physical design, since the correctness of transmission is invariant to delays of individual wires in channels. However, the cost of most DI codes is a significant degradation in coding efficiency, for example, using double the number of wires as in synchronous encoding.

In addition, many existing asynchronous NoCs use a highly robust *quasi-delay-insensitive (QDI) style* **in switches.** This style simplifies asynchronous timing validation, which only needs to check that each wire fanout point has roughly equal fanout delays. However, the cost of this design style is typically large area and power overhead, and the requirement of use of non-standard components.

Research Overview

The goal of our research is to support the design and optimization of high-throughput, low latency, flexible and low-power digital interconnection networks (i.e. *networks-on-chip [NoC's]*) using asynchronous design. Overall, this work aims to demonstrate that asynchronous and GALS NoC's can provide significant advantages in system power, area, latency, and support for heterogeneous interfaces, while still maintaining high throughput – in direct comparison to leading synchronous designs in

identical technology. The work has been funded by 3 NSF grants and 1 competitive university seed grant (RISE).

The key technical strategy is to use asynchronous design techniques with modest timing constraints, to enable much more efficient design.

In particular, a *single-rail bundled-data approach* is used **on channels**, which allows coding efficiency nearly identical to synchronous design. Only one extra forward "request" wire, or bundling signal, is added, which uses a worst-case delay to strobe when transmitted data is valid. This signal is only transmitted on demand, with data. This approach not only supports high coding efficiency on channels, but also allows "glitchy" (i.e. hazardous) datapath blocks to be used safely, as long as the bundling timing constraint is met; hence synchronous combinational blocks can be directly used, without alteration. In addition, *one-sided (or relative) timing constraints* are used **in switches**, which require that certain paths be always shorter than other paths.

Finally, while many asynchronous designs use "four-phase handshaking communication protocols," i.e. return-to-zero, which allow ease-of-design, since all signals are reset to zero between operations, these protocols have significant throughput overheads, since they require two round-trip communications per channel for each transmission (active and reset phases). In contrast, **we use "two-phase handshaking communication protocols"** (i.e. non-return-to-zero), which can be more complex to support efficiently in hardware, but have much better throughput, since they require only one round-trip communication per channel for each transmission.

The overall asynchronous NoC switch architecture is built with micro-level **asynchronous pipelining,** based on our Mousetrap pipelines (see "High-Speed Digital Pipelines" section below). These pipelines use single-rail bundled data, moderate relative timing constraints, and two-phase communication -- which together enable low area and power, and high performance. In addition, *a key feature is that these pipelines use only single-latch registers* – unlike most synchronous designs. Each architectural register is a bank of only a single layer of transparent D-latches, yet each can hold a distinct data item. In contrast, most synchronous designs require either flipflop-based or double-latch registers, or else use single-latch registers with complex two-sided timing constraints (i.e. "pulse-mode"). In contrast, Mousetrap uses single-latch architectural registers with only simple one-sided timing constraints (i.e. min constraints). This unusual feature results in the low area, low latency and high performance, underlying the asynchronous NoC router designs.

Summary of Contributions

While others have used the above techniques in asynchronous NoC's, no previous published work has demonstrated similar cost advantages (performance, area, power) over synchronous design, at both the switch- and network-level. We have also developed a complete computer-aided design (CAD) tool flow for these NoC's leveraging synchronous commercial CAD tools; and demonstrated highly-parametrizable automated synthesis (varying switch radix, layout, aspect ratio) with extensive experimental evaluation.

We have also developed and evaluated techniques to support efficient use of virtual channels (VCs). To support these NoC switch designs, we have also developed a new scalable family of asynchronous tree arbiters, which are highly efficient, fair and balanced.

To further extend the capabilities and utility of these asynchronous NoCs, we have developed two new techniques: *(i) latency acceleration* and *(ii) support for multicast.* The latter is the first general-purpose solution to supporting multicast in asynchronous 2D-mesh routers. (Testability and test techniques have recently been developed by my colleague, Davide Bertozzi; see below.) Very recently -- almost complete, not yet published --, we have developed a fully automated mapping and design procedure to commercial Xilinx FPGAs.

Finally, by invitation, we have validated our work at AMD Research, experimentally migrating our technology, and completing a head-to-head experimental comparison in advanced 14nm FinFET industrial technology, and demonstrating major benefits in switch area, latency and energy, over a commercial synchronous NoC design.

Some of this work has been completed in collaboration with colleagues at University of Ferrara, Italy (Prof. Davide Bertozzi's group), and other projects only at Columbia University.

Taken together, this body of our recent research has substantially advanced the state-of-art of asynchronous and GALS NoC's, and demonstrated its viability for use in a variety of commercial applications.

Technical Highlights

This work is divided into two sub-projects: we target two distinct network topologies, *(i) 2D Mesh,* and *(ii) variant Mesh-of-Trees (MoT)*. The 2D Mesh is our main and current focus: it is widely used, provides good resource sharing, and is highly scalable. The variant MoT is a lightweight tree topology, with indirect structure, with small binary router nodes, which is especially useful for forming networks in chip multiprocessors (CMPs) for last-level cache or memory access with light traffic, where latency is critical. Much of our earliest work was on this variant MoT, and then migrated recently to the more complex and challenging 2D Mesh networks. Much of the discussion below focuses on 2D Mesh networks, which is our main focus at this time.

Basic 5-ported router switch architecture [DATE-13]

The 2D Mesh networks use 5-ported routers. We introduced a basic 5-ported router design in this paper. A post-layout evaluation of the new switch design, in comparison with a highly-efficient synchronous implementation, *xpipes Lite*, demonstrated: a *reduction in overall power of 85%/73% (vs. synchronous without/with clock gating), a 71% reduction in switch area, and a 44% reduction in average energy/flit, while maintaining nearly comparable throughput (903 ps/cycle) in a 45nm low power technology.*

The above design is almost entirely standard-cell based, making it practical for commercial application. Our solution provides a unique direct comparison with a state-of-the-art synchronous design (*xpipes Lite*), and demonstrates significant overall cost benefits – including highlighting that high-performance asynchronous designs can have significantly lower area than synchronous designs, and can provide much lower average power even compared to synchronous clock-gated designs.

This paper was a Best Paper Finalist at DATE-13, and voted the best paper of the "network-on-chip" track out of more than 60 track submissions.

Automated synthesis flow using synchronous CAD tools (Synopsys DC/IC Compiler)
[DATE-13, Async-17, NGCAS-17]
We presented a basic semi-automated synthesis strategy in [DATE-13], using commercial synchronous CAD tools. The key challenges are to support two classes of timing constraints: *(i) "bundling constraints" on the datapath* (i.e. the bundled worst-case "request" signal must always be slower than the corresponding data, but ideally with tight margins), and *(ii) relative timing constraints on the control* (i.e. there are race conditions identified, where certain paths must always be slower than other paths). An initial heuristic methodology was introduced to use synchronous CAD tools to enforce these constraints. In addition, certain control blocks which must be hazard-free, are identified and given "don't touch" requirements to avoid arbitrary and hazardous restructuring of the logic.

More recently, we have developed a complete automated tool flow, which allows heuristic iterative gradual convergence, to meet the timing constraints. The tool flow has been presented in [Async-17, NGCAS-17]. Input is at the RTL level. The flow has been validated down to layout level, and shown to support significant parametrizability: a wide range of switch designs has been implemented, with varied radix (i.e. # of ports), NoC topology and aspect ratio.

Extensions to support virtual channels (VCs) [VLSI-SoC-14, DATE-17]
For asynchronous VC's, we demonstrated that a fully replicated switch with distinct VC control on links is the best solution [VLSI-SoC-14], rather than a single switch with internal sharing between VCs. In our recent collaboration with AMD [DATE-17], an improved technique for credit-based VC control is introduced, using a "lazy credit update," which avoids performance contention between a credit increment and a critical credit decrement request (i.e. pending transmission). A patent between AMD and my PhD student Weiwei Jiang has been filed on this innovation.

Development of a new family of scalable N-way asynchronous arbiters [Async-15]
The use of asynchronous arbiters is a critical component when designing asynchronous NoCs. These mediate between competing requests in continuous time, and hence are much more challenging than synchronous clock-based arbiters to design. A scalable family is needed, to support varied radices, i.e. number of ports. Previous work has demonstrated that tree arbiters can provide low latency and modular design. However, we have determined that this earlier work suffered from severe asymmetries: unequal latencies and unbalanced win rates for different clients (i.e. input requests). In addition, transient correctness problems were observed: two grants could be asserted simultaneously, with one heading low only after a new grant has been asserted high, called the "grant overlapping problem." In our work, we developed a new scalable family of asynchronous tree arbiters, that provides low latency, and much better equalization of path delays (i.e. access time) and win rates between clients. It also, in practice, eliminates the correctness problems. *When evaluated across multiple cost objectives, this new family achieves overall the best state-of-art in designing low-latency N-way asynchronous arbiters.*

This paper was a Best Paper Finalist at the IEEE Async-15 Symposium.

System latency acceleration techniques [ASPDAC-14, DAC-15]
System latency is a critical cost metric in NoC design, for many applications. We developed lightweight techniques to accelerate system latency, using advance monitoring and pre-arbitration and pre-allocation of channels. These use of a small "shadow" lightweight monitoring network, which tracks the structure

of the data network but with only a few extra wires and small control cells. When a packet is injected into the network, an advance notification token is rapidly transmitted across the shadow network, following the same source-to-destination path as the actual data transmission, and ideally sets up both the arbitration and the channel allocations along the path, in advance. The protocol requires no global control, and gracefully adapts to local congestion. These papers target both a variant MoT topology [ASPDAC-14] and a 2D Mesh topology [DAC-15].

The DAC-15 paper demonstrated a uniform improvement in system latency in a 2D mesh topology, across all benchmarks in moderate traffic, ranging from 34.4-37.9%. Interestingly, the approach also enabled throughput gains for most benchmarks, ranging from 14.7-27.1%. *These benefits are further improvements over the strong results already reported for our basic 2D mesh design in [DATE-13].*

Support for multicast communication [DAC-16, NOCS-17]

Multicast is defined as sending the same packet from a single source to an arbitrary subset of destinations. It is a critical communication paradigm, which is used to support a number of scenarios in modern architectures: in cache coherence protocols, to send write invalidates to multiple cores; in shared operand networks, to deliver operands to multiple instructions; and in multi-threaded applications, to notify barrier synchronization to multiple processors. Multicast is also gaining importance with new technologies, such as wireless, photonic, and CDMA, as well as in neuromorphic computing.

The goal of this work is to support efficient parallel multicast in asynchronous NoCs. *No general-purpose solutions to this problem have been previously proposed*, only narrow solutions customized for specialized applications (CDMA, neuromorphic computing). *Our NOCS-17 paper presents the first general-purpose asynchronous NoC to support multicast communication in a 2D Mesh topology.* In DAC-16, our approach is the first to support asynchronous multicast in variant MoT topologies.

The multicast strategy, for the 2D Mesh topology, is to include a highly concurrent "continuous-time multi-way read buffer." This unique new asynchronous component allows the write of a packet's flits into a single shared input buffer, followed by multiple independent reads, almost entirely decoupled, that operate at their own rates, for all requesting output ports to read the packet. Coupling is only enforced at the end of packet, i.e. tail flit. The approach exploits the lack of discretizing global clock, and supports fully separable and continuous reads from the various requesting output ports, thus scavenging read access whenever needed for the body flits of a packet, regardless of the different rates or congestion of the various output ports.

This paper was a Best Paper Finalist at the ACM NOCS-17 Symposium.

Industrial technology transfer: AMD Research [DATE-17]

By invitation, we experimentally transferred our basic 2D mesh NoC to AMD Research in Boxborough, MA. A switch design was implemented in advanced commercial 14nm FinFET technology, and compared directly to a recent AMD commercial NoC. *This was the first direct "apples-to-apples" comparison of an asynchronous NoC and a commercial synchronous NoC in identical advanced technology.*

Our design included a dual-plane switch, with each plane supporting 2 VCs. A semi-automated tool flow was used at AMD, to support the asynchronous design (see details on the design and flow in [DATE-17]).

Initial results were quite promising: *the asynchronous router had 55% lower area, 28% latency improvement, and 88% and 58% savings in idle and active power, respectively.*

Results were also extrapolated to two other configurations: (i) a 7-ported router with 2 VCs, and (ii) a 5-ported router with 8 VCs. The former is important for 3D stacking, and the latter represents a more realistic configuration. While some parameters increased in cost, the relative asynchronous area and power benefits are largely maintained, though latency improvements are somewhat reduced for the 7-ported configuration.

Details of this work are presented below, in the "Technology Transfer" section.

Basic asynchronous NoC switch design: targeting a variant MoT topology

[NOCS-10, TCAD-11, NOCS-11]

This earlier work (2008-2013) was led by my group, but in collaboration with a parallel architecture group at the University of Maryland, which is developing the synchronous shared-memory processor environment (called XMT, a PRAM-based massively-parallel architecture) and simulation tools. My group led on the design of the new asynchronous network. It is aimed at medium-to-high end multi-processors, yet uses robust handshaking protocols, with mainly standard-cell design techniques and portable design flows.

Experimental Results: In direct comparison to a fabricated synchronous chip (Balkan et al., *Hot Interconnects-07*) in the same 90nm technology, our new asynchronous NoC exhibited: *82-91% lower energy/packet, 64-84% less area, and up to 1.7x better system latency than an 800 MHz synchronous network* (for up to 73% of maximum synchronous traffic injection rate), but with some performance degradation at very high traffic rates. Initial simulations of a GALS NoC, in direct comparison with a synchronous architecture, on realistic parallel kernels (array summation, matrix multiplication, breadth-first search, array increment) show promising results.

Additional considerations: testing and testability techniques

Testing and design-for-testability are critical components of a viable design methodology. Test techniques for the underlying Mousetrap pipelines in our NoC switches are discussed below (see "High-Performance Digital Pipelines" section). For testing of our asynchronous NoC switches, recent work from our colleagues at the University of Ferrara, Profs. Davide Bertozzi and Michele Favalli and their students, has demonstrated a systematic approach for built-in self-testing and signature-based analysis:

G. Miorandi, A. Celin, M. Favalli and D. Bertozzi, "A Built-in Self-Testing Framework for Asynchronous Bundled-Data NoC Switches Resilient to Delay Variations," ACM Int. Symposium on Networks-on-Chip *(NOCS-16)*, pp. 1-8, 2016.

Selected Publications and Patents:

K. Bhardwaj, P. Mantovani, L. Carloni S.M. Nowick, "Towards a Comprehensive Methodology for Implementing Asynchronous NoCs on FPGAs," ACM/IEEE Design, Automation and Test in Europe Conference *(DATE-19),* Florence, Italy (deadline Sept. 2018) ***(in preparation).***

K. Bhardwaj, W. Jiang and S.M. Nowick, "Achieving Lightweight Multicast in Asynchronous NoCs Using a Continuous-Time Multi-Way Read Buffer," ACM Int. Symposium on Networks-on-Chip *(NOCS-17)*, pp. 6:1-6:8, 2017 ***(Best Paper Finalist)***.

D. Bertozzi, G. Miorandi, M. Tala and S.M. Nowick, "Cost-Effective and Flexible Asynchronous Interconnect Technology for GALS Networks-on-Chip," IEEE New Generation of Circuits and Systems Conference *(NGCAS-17),* Genoa, Italy (September 2017).

G. Miorandi, M. Balboni, S.M. Nowick and D. Bertozzi, "Accurate Assessment of Bundled-Data Asynchronous NoCs Enabled by a Predictable and Efficient Hierarchical Synthesis Flow", IEEE Int. Symposium on Asynchronous Circuits and Systems *(Async-17)*, pp. 10-17, 2017.

W. Jiang, D. Bertozzi, G. Miorandi, S.M. Nowick, W.P. Burleson and G. Sadowski, "An Asynchronous NoC router in a 14nm FinFET library: Comparison to an Industrial Synchronous Counterpart," ACM/IEEE Design Automation and Test in Europe Conference *(DATE-17)*, pp. 732-733, 2017.

K. Bhardwaj and S.M. Nowick, "Achieving Lightweight Multicast in Asynchronous Networks-on- Chip Using Local Speculation," ACM/IEEE Design Automation Conference *(DAC-16),* Austin, TX (June 2016).

W. Jiang, K. Bhardwaj, G. Lacourba and S.M. Nowick, "A Lightweight Early Arbitration Method for Low-Latency Asynchronous 2D-Mesh NoC's," ACM/IEEE Design Automation Conference *(DAC-15),* San Francisco, CA (June 2015).

G. Miorandi, D. Bertozzi and S.M. Nowick, "Increasing Impartiality and Robustness in High-Performance N-Way Asynchronous Arbiters," IEEE International Symposium on Asynchronous Circuits and Systems *(Async-15),* Mountain View, CA (May 2015) ***(Best Paper Finalist)***.

G. Miorandi, A. Ghiribaldi, S. Nowick and D. Bertozzi, "Crossbar Replication vs. Sharing for Virtual Channel Flow Control in Asynchronous NoCs: a Comparative Study," IFIP/IEEE International Conference on Very Large Scale Integration and System-on-Chip *(VLSI-SoC-14),* Playa del Carmen, Mexico (October 2014).

A. Ghiribaldi, D. Bertozzi and S.M. Nowick, "A Transition-Signaling Bundled Data NoC Switch Architecture for Cost-Effective GALS Multicore Systems," ACM/IEEE Design, Automation and Test in Europe Conference *(DATE-13),* Grenoble, France (March 2013) ***(Best Paper Finalist)***.

G. Gill, S.S. Attarde, G. Lacourba and S.M. Nowick, "A Low-Latency Adaptive Asynchronous Interconnection Network Using Bi-Modal Router Nodes," ACM Int. Symposium on Networks-on-Chip *(NOCS-11)*, Pittsburgh, PA (May 2011).

M.N. Horak, S.M. Nowick, M. Carlberg and U. Vishkin, "A Low-Overhead Asynchronous Interconnection Network for GALS Chip Multiprocessors," *IEEE Transactions on CAD*, vol. 30:4, pp. 494-507 (April 2011) ***(selected for special section on networks-on-chip).***

M.N. Horak, S.M. Nowick, M. Carlberg and U. Vishkin, "A Low-Overhead Asynchronous Interconnection Network for GALS Chip Multiprocessors," ACM Int. Symposium on Networks-on-Chip *(NOCS-10)*, Grenoble, France (May 2010).

S.M. Nowick, G.D. Gill and S. Attarde, *"Bi-Modal Arbitration Nodes for a Low-Latency Adaptive Asynchronous Interconnection Network and Methods for Using the Same."* ***U.S. Patent #9,537,679*** (January 3, 2017).

S.M. Nowick, M.N. Horak and M. Carlberg, "*Asynchronous Digital Circuits Including Arbitration and Routing Primatives for Asynchronous and Mixed-Timing Networks."* ***U.S. Patent #8,766,667*** (July 1, 2014).

S.M. Nowick, M.N. Horak and M. Carlberg, *"Asynchronous Digital Circuits Including Arbitration and Routing Primitives for Asynchronous and Mixed-Timing Networks."* ***U.S. Patent #8,362,802*** (January 29, 2013).

2 MIXED-TIMING INTERFACES

This aim of this research is to develop a practical and flexible set of interface circuits for mixed-timing digital systems. Heteregeneous systems, which combine asynchronous and/or multiple clock domains, are becoming increasingly common. Support for reliable and efficient interfacing between these domains is a critical capability.

Our work is the first to define a complete and modular family of robust and efficient circuits that can connect any combination of digital interfaces in different timing domains: *clocked-clocked* (with different clock rates), *clocked-async*, and *async-clocked*. Most future digital systems are expected to have mixed timing domains, yet few practical solutions have been previously proposed to adequately support their interfaces. These designs avoid the significant performance overheads of "pausible" or "stoppable" clocking schemes, as well as their modularity issues (pausible/stoppable clocking requires engineering changes inside of distinct clocked blocks).

Our proposed circuits combine several benefits: (i) they can be built using standard gates (no custom circuits required); (ii) they can robustly handle arbitrary timing discrepancies between interfaces (no restriction on the relationship between the different clock rates); (iii) they have high throughput (no synchronization overhead during steady-state operation); (iv) they only require a small constant number of synchronizer circuits (only 3) for the entire FIFO, independent of the number of cells (unlike several previous approaches, such as from Intel, where the # of synchronizers grows linearly with FIFO capacity) and (v) they are easily scalable (i.e. built using a token ring architecture, with replicated cells).

This work has been influential, with a number of leading researchers building on, or using, this approach in published work: D. Marculescu/CMU [ISCA-02] (our FIFO's are modeled and used in their superscalar simulator); D. Albonesi/Cornell [Micro-02]; S. Moore/Cambridge [Async-02]; and many others.

Our publications on this research have received **535 total cites**. These include 3 conference papers where the initial ideas were introduced, followed by an IEEE Transactions on VLSI Systems journal paper that combines them together, and 2 patents (see below).

Selected Publications and Patents:

T. Chelcea and S.M. Nowick, "Robust Interfaces for Mixed-Timing Systems," *IEEE Transactions on VLSI Systems*, vol. 12:8, pp. 857-873 (August 2004).

T. Chelcea and S.M. Nowick, "Robust Interfaces for Mixed-Timing Systems with Application to Latency-Insensitive Protocols," ACM/IEEE Design Automation Conference *(DAC-01),* Las Vegas, NV (June 2001).

T.Chelcea and S.M.Nowick, "Low-Latency FIFO's for Mixed-Clock Systems," IEEE Computer Society Annual Workshop on VLSI *(WVLSI-00),* Orlando, FL (April 2000).

T. Chelcea and S.M. Nowick, "Low-Latency Asynchronous FIFO's using Token Rings," IEEE International Symposium on Advanced Research in Asynchronous Circuits and Systems *(Async-00)*, Eilat, Israel (April 2000).

T. Chelcea and S.M. Nowick, "Low Latency FIFO Circuit for Mixed Clock Systems," ***U.S. Patent #7,197,582*** (March 27, 2007).

T. Chelcea and S.M. Nowick, "Low Latency FIFO Circuits for Mixed Asynchronous and Synchronous Systems," ***U.S. Patent #6,850,092*** (February 1, 2005).

3 HIGH-SPEED DIGITAL PIPELINES

The goal of this research is to develop a set of practical asynchronous pipeline circuit structures to support the design of high-performance systems. In synchronous systems, pipelining is the foundation of most high-performance design. Therefore, the development of effective asynchronous pipeline structures is critical to support wide use of clockless design.

We have developed three new efficient and widely used asynchronous pipeline structures: ***(i) MOUSETRAP pipelines, (ii) lookahead pipelines,*** and ***(iii) high-capacity pipelines.*** *MOUSETRAP* uses static logic, and *lookahead* and *high-capacity* pipelines use dynamic logic. All share common goals: high performance implementations, low area, and use of simple local one-sided timing constraints that are easy to satisfy. MOUSETRAP pipelines can be built using existing standard cell libraries. All have demonstrated multi-GigaHertz performance, with low power and area overheads.

This work has been influential.

MOUSETRAP pipelines were adopted by the DARPA CLASS project, led by Boeing Corporation (2005-2007), for use in an experimental asynchronous chip (see "Grant Highlights" below). Philips Semiconductors, through their Handshake Solutions incubated startup, incorporated them into an experimental version of their asynchronous CAD tool flow. A variant of Mousetrap pipelines was developed by NXP/Philips Semiconductors for use in their experimental Aetherial network-on-chip in the early 2000's, and included in their design library. Our Mousetrap publications (ICCD-01, TVLSI-07) and patent have received **316 total cites**.

Our first paper on *lookahead pipelines* received a Best Paper award at the IEEE Async-00 Symposium.

IBM Research (2001-2002) adopted our *high-capacity pipelines* for use in a fabricated experimental FIR filter chip for disk drive reads (see "Technology Transfer" below).

We have also developed systematic and lightweight techniques for *test generation and testability* for our asynchronous pipelines, for both stuck-at and delay faults, and demonstrated their applicability to Mousetrap (see below, Shi et al., ITC-05). Extensions for minimally-intrusive at-speed testing of delay faults have been developed by my former PhD student, Montek Singh, and his group at UNC, and applied to both Mousetrap and HC pipelines (Gill et al., "Low-overhead testing of delay faults in high-speed asynchronous pipelines," IEEE Async-06 Symposium).

Four patents have been issued for these pipeline styles. Taken together, our papers introducing these 3 pipeline structures (conference, journal), and associated patents, have received **771 total cites**.

Selected Publications and Patents:

S.M. Nowick and M. Singh, "High-Performance Asynchronous Pipelines: an Overview." IEEE Design & Test of Computers, vol. 28:5, pp. 8-22 (September/October 2011).

M. Singh and S.M. Nowick, "MOUSETRAP: High-Speed Transition-Signaling Asynchronous Pipelines," *IEEE Transactions on VLSI Systems*, vol. 15:6 (June 2007).

M. Singh and S.M. Nowick, "The Design of High-Performance Dynamic Asynchronous Pipelines: Lookahead Style," *IEEE Transactions on VLSI Systems*, vol. 15:11 (November 2007).

M. Singh and S.M. Nowick, “The Design of High-Performance Dynamic Asynchronous Pipelines: High-Capacity Style," *IEEE Transactions on VLSI Systems*, vol. 15:11 (November 2007).

F. Shi, Y. Makris, S.M. Nowick and M. Singh, “Test Generation for Ultra-High-Speed Asynchronous Pipelines,” IEEE International Test Conference *(ITC-05*), Washington, D.C. (October 2005).

M. Singh and S.M. Nowick, “High-Throughput Asynchronous Pipelines for Fine-Grain Dynamic Datapaths," ***Best Paper Award,*** IEEE International Symposium on Advanced Research in Asynchronous Circuits and Systems *(Async-00)*, Eilat, Israel (April 2000).

M. Singh and S.M. Nowick, “Circuits and Methods for High-Capacity Asynchronous Pipeline Processing,” ***U.S. Patent #7,053,665*** (5/31/06).

M. Singh and S.M. Nowick, “Asynchronous Pipeline with Latch Controllers,” ***U.S. Patent #6,958,627*** (10/25/05).

M. Singh and S.M. Nowick, “Circuits and Methods for High-Capacity Asynchronous Pipeline,” ***U.S. Patent #6,867,620*** (3/15/2005).

M. Singh and S.M. Nowick, “High-Throughput Asynchronous Dynamic Pipelines,” ***U.S. Patent #6,590,424*** (7/8/03).

4 CONTINUOUS-TIME DIGITAL SIGNAL PROCESSORS (CT-DSP's)

The goal of this research is to develop a promising new class of DSP's, *using variable-rate, i.e. amplitude-based, sampling,* called continuous-time DSP's. This work is in collaboration with Prof. Yannis Tsividis (Columbia EE) and joint PhD student Christos Vezyrtzis.

The work grows out of early theory, from Prof. Tsividis, on how DSP's can avoid classic quantization at fixed time intervals (x-axis), and instead adapt their quantized sample rate to conform to the rate of change of the input signal (y-axis).

Potential benefits are significant: (i) complete elimination of all aliasing (unlike classic quantization), (ii) substantial reduction in overall power (periods of quiescence or low activity inherently result in no or few samples), (iii) much greater flexibility in accommodating a wide range of input formats, with no design change, and (iv) the capability of designing high-resolution DSP's with lower cost and sample bit-width than classical DSP's.

While a number of prior continuous-time ADC's/DAC's have been developed, ***no flexible and general-purpose CT-DSP has previously been designed.***

The design problems are challenging, since the DSP core must preserve time spacing of input events on its outputs, and also must observe and react in continuous (i.e. not discrete) time to sample arrival. As a result, the CT-DSP's involve an unusual amalgam of (*i) real-time delay-lines and computation*, which translates input time intervals to the outputs precisely, and *(ii) asynchronous digital components and control*, which can handle irregular and unpredictable signal arrival rates, and correctly initiate operation. Only a small handful of prior CT-DSP cores have been implemented, but these are either very early prototypes (e.g. 1-bit sample width, limited sample format support, with timing race conditions not resolved) or specialized for very high throughput (with short delay lines and limited functionality).

In ESSCIRC-13 and JSSC-14 papers, ***we demonstrate the first fully-functional general-purpose CT-DSP chip.*** The design, including asynchronous logic and a calibrated delay line, is shown capable of handling a wide variety of input formats (sync PCM, sync/async PWM, sync/async sigma-delta), arbitrary bit widths, and time-varying input rates, with no changes in the DSP design. This capability is not possible in any synchronous DSP without reprogramming. In addition, for certain inputs, it has a signal-to-error ratio that exceeds that of clocked systems. Moreover, the frequency response remains intact for any type and rate of its input. The chip also is the first CT-DSP to include on-chip tuning.

Finally, in our ICCD-12 paper, we propose techniques to reduce power in the calibrated delay line. A novel "adaptive granularity" approach is used, with asynchronous circuitry monitoring the current input sample rate, and changing the granularity (i.e. dynamically varying the number of pipeline stages) of the delay line accordingly, to reduce overall power (i.e. pipeline depth) for less dense traffic.

Overall, the CT-DSP chip not only makes an important advance in the DSP field; it also highlights the novel and unique benefits that asynchronous design can provide, to enable this architecture.

Selected Publications and Patents:

C. Vezyrtzis, W. Jiang, S.M. Nowick and Y. Tsividis, "A Flexible, Event-Driven Digital Filter with Frequency Response Independent of Input Sample Rate," *IEEE Journal of Solid State Circuits,* vol. 49:10, pp. 2292-2304 (October 2014).
C. Vezyrtzis, Y. Tsividis and S.M. Nowick, "Improving the Energy Efficiency of Pipelined Delay Lines Through Adaptive Granularity," *IEEE Transactions on VLSI Systems,* vol. 23:10, pp. 2009-2022 (October 2015).

Y. Tsividis, M. Kurchuk, S.M. Nowick, B. Schell and C. Vezyrtzis, "Event-Based Data Acquisition and Digital Signal Processing in Continuous Time." Chapter in: *Event-Based Control and Signal Processing* (ed., M. Miskowicz), CRC/Taylor & Francis (2015).

C. Vezyrtzis, S.M. Nowick and Y. Tsividis, "A Flexible, Clockless Digital Filter," European Solid State Circuits Conference *(ESSCIRC-13),* Bucharest, Romania (September 2013).

C. Vezyrtzis, Y. Tsividis and S.M. Nowick, "Designing Pipelined Delay Lines with Dynamically-Adaptive Granularity for Low-Energy Applications," IEEE International Conference on Computer Design *(ICCD-12),* Montreal, Canada (October 2012) (***Best Paper Award, Logic and Circuit Design Track).***

5 ULTRA-LOW ENERGY AND RELIABLE DIGITAL DESIGN

The goal of this research is to develop extremely low-energy digital circuits, using sub-threshold or near-threshold voltage levels, which at the same time provide good performance and high reliability. This work was pursued in collaboration with Prof. Mingoo Seok (Columbia EE).

The challenge of low-voltage low-energy digital design is to handle the extremes of device variability and vulnerability to external noise which come with reduced Vdd. Both issues have been heavily studied in the synchronous domain, with much instrumentation and calibration required to mitigate these issues, or else requiring severe margining which limits performance.

Asynchronous design is highly promising for this application, since several design styles have great robustness to timing variability, i.e. using few timing assumptions and delay-insensitive data encoding, and all eliminate the need for synchronization with a fixed-rate global clock. In this project, in our ISPLED-13 paper, we identify two leading high-performance and highly-robust asynchronous pipeline styles, both using dynamic logic: (i) PS0 (Dean/Horowitz, Stanford), and (ii) PCHB (Lines, Caltech). The former was used in commercial HaL processors in the late 1990's, and the latter is currently used at Intel's Switch & Router Division (formerly Fulcrum Microsystems) for commercial high-performance Ethernet switch chips.

We identify a key bottleneck for use of these dynamic pipelines in subthreshold design: "keeper fighting" issues at low voltages. We then propose a novel solution: *adding lightweight monitoring and control, to avoid keeper fighting at each pipeline stage before new data is evaluated.* ***To the best of our knowledge, this is the first approach to demonstrate that these robust asynchronous pipelines can operate safely at Vdd of 0.3V.*** Our Async-13 paper demonstrates mitigation techniques for correct operation of a leading static asynchronous pipeline, Mousetrap.

Taken together, this initial work is advancing the ability to use robust and high-performance asynchronous pipelines, for extreme low-energy applications.

Selected Publications and Patents:

Y. Chen, M. Seok and S.M. Nowick, "Robust and Energy-Efficient Asynchronous Dynamic Pipelines for Ultra-Low-Voltage Operation Using Adaptive Keeper Control," IEEE International Symposium on Low Power Electronics and Design *(ISLPED-13),* Beijing, China (September 2013).

J. Liu, S.M. Nowick and M. Seok, "Soft MOUSETRAP: a Bundled-Data Asynchronous Pipeline Scheme Tolerant to Random Variations at Ultra-Low Supply Voltages," IEEE International Symposium on Asynchronous Circuits and Systems *(Async-13),* Santa Monica, CA (May 2013).

6 ROBUST ENCODING FOR GLOBAL ASYNCHRONOUS COMMUNICATION

This aim of this research is to develop practical codes, and hardware support, for robust asynchronous communication.

These codes can be used for communication either in fully asynchronous systems, or in GALS systems which allow use of synchronous cores, memories and function units integrated through a flexible asynchronous network. This work is centered on ***delay-insensitive (DI) codes***, which are also known as "unordered codes" in the synchronous community. The benefit of these codes is that they tolerate arbitrary and unknown skew in the arrival of individual bits in an asynchronous transmission, and hence support *timing-robust* communication.

The research includes the design of several new practical DI codes: **(i) Zero-Sum**, which also provides error correction (i.e. *an error-correcting unordered [ECU] code*); **(ii) DI Bus-Invert**, which provides low power through use of selective bit inversion; and **(iii) Level-Encoded Transition-Signalling (LETS)**, which provides both high-throughput, using a two-phase protocol and low switching activity. The work on Zero-Sum codes also includes heuristic extensions to provide good coverage of 2-bit correction while guaranteeing 100% coverage of 1-bit correction. In addition, supporting hardware blocks for protocol conversion, encode/decode and completion detection have been developed. Protocol conversion allows efficient encodings for global channels to interface to alternative efficient encodings for local computation nodes. Completion detectors are critical components of asynchronous systems: they determine when a valid DI codeword has been received.

LETS codes have recently been successfully used in the Stanford University "Neurogrid" project (led by Prof. Kwabena Boahen), to build a leading large-scale neuromorphic system, where low-power, high-throughput and robust inter-neuron communication is facilitated using asynchronous DI communication (see Sec. V.B.: *Proceedings of the IEEE*, vol. 102:5, pp. 699-716, April 2014).

Selected Publications and Patents:

M.Y. Agyekum and S.M. Nowick, "Error-Correcting Unordered Codes and Hardware Support for Robust Global Communication," *IEEE Transactions on Computer-Aided Design,* vol. 31:1, pp. 75-88 (January 2012).

M.Y. Agyekum and S.M. Nowick, "A Delay-Insensitive Bus-Invert Code and Hardware Support for Robust Global Communication," ACM/IEEE Design, Automation and Test in Europe Conference *(DATE-11),* Grenoble, France (May 2011).

M. Cannizzaro, W. Jiang and S.M. Nowick, "Practical Completion Detection for 2-of-N Delay-Insensitive Codes," IEEE International Conference on Computer Design *(ICCD-10),* Amsterdam, The Netherlands (October 2010).

M.Y. Agyekum and S.M. Nowick, "An Error-Correcting Unordered Code and Hardware Support for Robust Global Communication," ACM/IEEE Design, Automation and Test in Europe Conference *(DATE-10),* Dresden, Germany (March 2010).

A. Mitra, W.F. McLaughlin and S.M. Nowick, "Efficient Asynchronous Protocol Converters for Two-Phase Delay-Insensitive Global Communication," *IEEE Transactions on VLSI Systems*, vol. 18:7, pp. 1043-1056 (July 2009).

P.B. McGee, M.Y. Agyekum, M.A. Mohamed and S.M. Nowick, "A Level-Encoded Transition Signaling Protocol for High-Throughput Asynchronous Global Communication," IEEE Int. Symposium on Asynchronous Circuits and Systems *(Async-08)*, Newcastle-upon-Tyne, UK (April 2008).

7 COMPUTER-AIDED DESIGN TOOLS

The goal of this work is to develop practical CAD tools for asynchronous design. CAD tools are critical to the widespread adoption of asynchronous circuits.

My PhD thesis research, and accompanying publications, presented some of the earliest solutions for asynchronous CAD design and optimization techniques and tool development, targeting asynchronous controllers, called ***"locally-clocked" burst-mode machines***. These asynchronous state machines were the first to guarantee correct hazard-free operation when simultaneously targeting: implementations with realistic gate-level mapping, supporting a fairly general specification style, yet still providing very low latency and area overhead. The thesis also introduced the first general solution to the 2-level minimization problem for hazard-free logic.

Our publications on "locally-clocked" asynchronous controllers together have received **450 total cites**. These include 2 conference papers where initial ideas were introduced (ICCD-91, ICCAD-91) and my PhD thesis.

More recently, my focus has been on four key areas: (i) individual controllers, (ii) timing-robust circuit styles (i.e. threshold circuits), (iii) performance analysis and timing verification of concurrent systems, and (iv) synthesis and optimization of large-scale asynchronous systems.

I have released new versions of the first three of these tools under a comprehensive asynchronous design framework, called "**CaSCADE"** (Columbia University and USC Asynchronous Design Environment). Each tool is available for free download, and includes extensive tutorial support, setup documentation and benchmark examples. For download access, or just to view tutorial slides, see: http://www.cs.columbia.edu/~nowick/asynctools.

(i) MINIMALIST: a CAD Package for Synthesis & Optimization of Asynchronous Controllers

With several of my students, I have developed and released a large software CAD package for the synthesis of asynchronous burst-mode controllers, called MINIMALIST.

I introduced and formalized a new class of asynchronous controllers, with a Mealy-machine style, called "burst-mode" (which was based on earlier ad hoc, and not fully implementable, specifications used at HP Laboratories).

"Burst-mode" asynchronous controllers have a number of attractive features: simple timing constraints, very low latency (i.e., input-to-output paths) and power consumption, and targeting of existing (i.e. synchronous) standard-cell gate libraries for ease of implementation. They have been successfully applied to a variety of industrial and large-scale designs, e.g. cache and SCSI controllers, and controllers for an experimental low-power infrared communication chip from HP Laboratories ("Stetson" project). More recently, burst-mode controllers and the MINIMALIST package have been used at NASA Goddard Space Flight Center to design experimental space measurement chips (see "Technology Transfer" below).

The Minimalist CAD package provides a number of practical benefits for designers. It includes:

(a) a sophisticated set of optimization algorithms (both exact and heuristic) for several synthesis steps: two-level and multi-level hazard-free logic minimization, and optimal critical race-free state assignment (under a so-called 'input encoding' model) – where we have proposed the first complete solutions to these problems [ICCAD-92, TCAD-95, ICCAD-95];

(b) an extensive set of designer scripts, allowing users to target different cost functions (speed vs. area), typically providing dozens of alternative implementations synthesized under different user metrics, thus supporting design-space exploration;

(c) extensive practical designer support, including a Verilog backend, graphical interfaces, a command-line shell (for customized synthesis runs), and automatic insertion of initialization circuitry;

(d) a top-to-bottom verifier, to validate the final implementation against the initial specification.

The tool has been highly visible: it has been downloaded to over 100 sites in over 18 countries. It has also been used in a joint project between the nominee and NASA Goddard Space Flight Center (see below). Portions of the tool (hazard-free two-level logic minimization) have been used in experimental projects at HP Laboratories (Stetson project) and also incorporated into other asynchronous CAD tools (3D tool from Kenneth Yun).

Selected Publications and CAD Tools:

R.M. Fuhrer and S.M. Nowick, *Sequential Optimization of Asynchronous and Synchronous Finite-State Machines: Algorithms and Tools.* Kluwer Academic Publishers, Boston, MA (2001).

L. Lavagno and S.M. Nowick, "Asynchronous Control Circuits," Chapter 10 of *Logic Synthesis and Verification* (pp. 255-284) ***(invited contribution).*** S. Hassoun and T. Sasao, editors. Kluwer Academic Publishers, Boston, MA (2002).

M. Theobald and S.M. Nowick, "Fast Heuristic and Exact Algorithms for Two-Level Hazard-Free Logic Minimization," *IEEE Transactions on Computer-Aided Design,* vol. 17:11, pp. 1130-1147 (November 1998).

R.M. Fuhrer and S.M. Nowick, "Symbolic Hazard-Free Minimization and Encoding of Asynchronous Finite State Machines," ACM/IEEE International Conference on Computer-Aided Design *(ICCAD-95),* San Jose, CA (November 1995).

S.M. Nowick and D.L. Dill, "Exact Two-Level Minimization of Hazard-Free Logic with Multiple-Input Changes," *IEEE Transactions on Computer-Aided Design,* vol. 14:8, pp. 986-997 (August 1995).

See also CAD tool download information, and tutorial slides, available at
http://www.cs.columbia.edu/~nowick/asynctools.

(ii) Synthesis and Optimization of Asynchronous Threshold Circuits

Another interesting research area has been to develop CAD tools for a class of extremely timing-robust asynchronous circuits: ***dual-rail threshold circuits***.

These circuits are especially important for future-generation digital systems, because they gracefully tolerate wide variations and unpredictability due to process, temperature and voltage variability. However, there has been only limited previous work on developing systematic optimization techniques for these circuits. The goal is to support automated design of substantial digital systems (datapath + control) using this design style.

A complete tool package, called **ATN_OPT,** is available on the CaSCADE web site, with a tutorial; see: http://www.cs.columbia.edu/~nowick/asynctools.

In particular, my students and I have developed novel optimization algorithms and CAD tools for two synthesis steps: ***(i) technology mapping (atn_map)***, and ***(ii) multi-level optimization (atn_relax)***. The work demonstrates the feasibility of powerful CAD optimizers for robust asynchronous circuits, which can obtain *over 50% performance and area improvement* without any loss of their timing-robustness properties.

The technology mapper, ***atn_map***, is the *first systematic and general approach* for technology mapping of asynchronous threshold circuits. It is based loosely on synchronous techniques (such as those incorporated into Synopsys' Design Compiler), but with novel modifications to several steps to ensure that no hazards or timing constraints are introduced. A basic method has been developed to handle individual cost functions (area, delay). A recent extension has been developed to handle an important combined cost function: *area optimization under hard delay constraints*. The latter is the first asynchronous logic synthesis approach to systematically ensure that hard timing requirements are met.

The multi-level optimizer, ***atn_opt***, is based on the notion of "local relaxation". The strategy is to implement selected nodes in a given netlist using "eager evaluation", i.e. allowing these nodes to produce outputs *early (*before all inputs have arrived). Other nodes in the netlist are implemented in a conservative "input-complete" manner: only producing outputs after all inputs have arrived. If the relaxed nodes are carefully selected, this hybrid 'relaxation' approach *still* ensures a fully timing-robust overall circuit.

These tools have been applied to substantial systems with thousands of gates and hundreds of inputs and outputs: a complete DES encryption circuit, a GCD circuit, and the largest MCNC combinational benchmarks. For technology mapping, *average delay improvements of 31.6% and area improvements of 9.5% were obtained.* For multi-level optimization, *average delay improvements of 16.1% and area improvements of 34.9% were obtained.* Using our recent technology mapping approach targeting delay-area tradeoffs, as a post-processing step, *further area reduction by 10.7% on average can be recovered*, with no degradation of delay. The resulting circuits still remain hazard-free and timing-robust. This approach has also extended this work to handle simple *combinational* dual-rail circuits, i.e. dual-rail circuits that do no have hysteresis. This circuit family is widely used, and though less timing-robust than sequential threshold circuits, provides higher-performance circuits. The optimization methods can apply directly to this larger class.

Publications and CAD Tools:

C. Jeong and S.M. Nowick, "Technology Mapping and Cell Merger for Asynchronous Threshold Networks," *IEEE Transactions on Computer-Aided Design*, vol. 27:4, pp. 659-672 (April 2008).

C. Jeong and S.M. Nowick, "Optimization for Timing-Robust Asynchronous Circuits based on Eager Evaluation," IEEE International Symposium on Asynchronous Circuits and Systems *(Async-08),* Newcastle-upon-Tyne, UK (April 2008).

C. Jeong and S.M. Nowick, "Optimization of Robust Asynchronous Circuits by Local Input Completeness Relaxation," IEEE Asia-South Pacific Design Automation Conference *(ASPDAC-07)*, Yokohama, Japan (January 2007).

C. Jeong and S.M. Nowick, "Optimal Technology Mapping and Cell Merger for Asynchronous Threshold Networks," IEEE International Symposium on Asynchronous Circuits and Systems *(Async-06)*, Grenoble, France (March 2006).

C. Jeong and S.M. Nowick, "Methods, Media and Means for Forming Asynchronous Logic Networks," ***U.S. Patent #7,729,892*** (June 1, 2010).

See tutorial slides and CAD tool download information available at
http://www.cs.columbia.edu/~nowick/asynctools.

(iii) Performance Analysis and Timing Verification of Concurrent Systems

An important component of an effective design methodology is to have efficient performance analysis and timing verification tools, both to evaluate a system's operation and also as a basis for driving further optimization.

Performance analysis and timing verification are especially challenging for asynchronous systems: since there is no global clock, and system-level operation is often highly decoupled and concurrent, it is difficult to adapt existing synchronous analytical or simulation approaches.

In this work, we have developed two novel approaches to performance analysis: ***(a) using stochastic delay models*****,** and ***(b) using min/max delay models***.

A complete tool package, called **DES (Discrete Event System) Analyzer,** is available on the CaSCADE web site, with a tutorial; see: http://www.cs.columbia.edu/~nowick/asynctools.

Approach (a), called ***des_perf***, focuses on modeling stochastic delay distributions for each component in a concurrent system, and then using Markovian techniques to find key metrics: average-case throughput, latency, relative input arrival order and component utilization. Critical and slack paths can also be identified. These metrics can then be used either to rate the system's performance, or as a basis for performance-driven optimization of the system.

Approach (b), called ***des_tse***, does not consider a delay distribution, but rather extremes of behavior: modeling min/max delays of each component. It then evaluates the 'time-separation' between two successive events in the system. This approach can then be used to rapidly determine relative orderings of input arrivals and extremes of overall system behavior. It is therefore a basis for formal verification of feasible behaviors.

Selected Publications:

P.B. McGee and S.M. Nowick, "An Efficient Algorithm for Time Separation of Events in Concurrent Systems," ACM/IEEE International Conference on Computer-Aided Design *(ICCAD-07)*, San Jose, CA (November 2007).

P.B. McGee and S.M. Nowick, "Efficient Performance Analysis of Asynchronous Systems Based on Periodicity," IEEE/ACM International Conference on Hardware/Software Codesign and System Synthesis *(CODES-05),* Jersey City, NJ (September 2005).

See also tutorial slides and CAD tool download information available at http://www.cs.columbia.edu/~nowick/asynctools.

(iv) Synthesis and Optimization of Large-Scale Asynchronous Systems

A final area has been to develop CAD tools for the synthesis and optimization of entire asynchronous systems.

Two approaches have been developed. With my PhD student Tiberiu Chelcea and colleagues at the University of Manchester, we have developed a "back-end optimizer" suitable for improving the existing asynchronous commercial CAD tool flow at Philips Semiconductors. The Philips approach uses simple syntax-directed compilation, from a high-level concurrent specification language to asynchronous circuits. In this approach, each construct in the specification is replaced directly by a corresponding small "process" (i.e. a concurrent component) in hardware. Our proposed approach is to build on this unoptimized compiler flow, and insert a powerful back-end optimizer, which performs peephole and resynthesis transformations on the netlist of concurrent hardware components, and thereby further restructures and improves the circuit. Initial results of our tool indicate up to 50% performance improvements.

An alternative approach, developed with my PhD student Michael Theobald, is closer to classic high-level synthesis from the synchronous world, but introduces concurrency-enhancing system-level transformations, in a strict architecture where each function unit has a dedicated asynchronous controller.

Selected Publications:

T. Chelcea and S.M. Nowick, "Resynthesis and Peephole Transformations for the Optimization of Large-Scale Asynchronous Systems," IEEE/ACM Design Automation Conference *(DAC-02)*, New Orleans, LA (June 2002).

M. Theobald and S.M. Nowick, "Transformations for the Synthesis and Optimization of Asynchronous Distributed Control," IEEE/ACM Design Automation Conference *(DAC-01)*, Las Vegas, NV (June 2001).

8 TECHNOLOGY TRANSFER

Finally, I have been involved in several efforts to promote asynchronous technology transfer. (For other highlights, see DARPA CLASS project under "Grant Highlights" below.)

(i) AMD Research (2015-2017)

By invitation from Advanced Micro Devices (AMD), we experimentally migrated our recent high-performance and low-energy asynchronous network-on-chip (NoC) switch into their industrial design flow, and conducted the first direct ("apples-to-apples") comparison between an asynchronous NoC design and a recent commercial synchronous NoC chip in identical advanced technology (14nm FinFET).

The development of the asynchronous NoC switch was led by my group in collaboration with the University of Ferrara (Italy), and the transfer of our NoC technology to AMD was led by my PhD student, Weiwei Jiang. As part of this project, Jiang also developed a new efficient approach for asynchronous virtual channels, using a "lazy credit update", with a joint patent filed by AMD. The synchronous chip was manufactured by AMD and is currently used in several commercial products.

Experimental results, on a single network node, show the asynchronous NoC design performed significantly better when compared to the clock-gated synchronous chip: ***55% less area with 28% lower latency – along with a 58% reduction in active power and an 88% reduction in idle power***. The experiments were conducted at AMD Research, in collaboration with AMD Fellows Greg Sadowski and Wayne Burleson.

"This was a highly-productive and effective collaboration between Steve Nowick's group and AMD," said Sadowski. *"His group's asynchronous network-on-chip has great potential for our future systems, once the automated tool flow is developed."*

This evaluation was published in the 2017 ACM/IEEE Design, Automation and Test in Europe (DATE) Conference.

In recent developments, we also now provide a complete and automated synthesis CAD tool flow for asynchronous NoC's, leveraging synchronous commercial tools, which was published in the 2017 IEEE Async Symposium.

(ii) NASA Goddard Space Flight Center: Space Measurement Applications (2006-2008)

By invitation, I collaborated with NASA Goddard Space Center in the design of series of asynchronous chips for space measurement applications.

In Spring 2006, I was invited by a senior engineer at NASA/Goddard (now a manager) to collaborate in designing an asynchronous measurement circuit for space applications. The motivation for using asynchronous design are: the complete removal of a high-speed sampling clock, lower power, and

design flexibility in implementing a highly-concurrent and fine-grained micro-architecture which can handle varied input sampling rates.

The chip includes a number of my asynchronous burst-mode controllers, and the design made extensive use of my *Minimalist* CAD package. The asynchronous circuit was designed by myself and by my NASA colleague, Duane Armstrong, with additional support for simulation and physical design by other NASA engineers. The design has been included in a fabricated chip; it achieved desired performance targets but with significantly lower area and power than their previous synchronous designs. This work received positive interest from NASA scientists, and there is real interest in its use in future space missions. (However, Armstrong has since left for a management position at Stennis Space Center.)

(iii) IBM T.J. Watson: FIR Filter Chip for Disk Drive Reads (2001-2002)

My former PhD student (Montek Singh) and I transferred one of our high-speed asynchronous pipeline styles to IBM, for use in an experimental FIR filter chip for disk drive reads.

This work presents an important advance in the design of high-speed asynchronous circuits: it demonstrates that an *industry-standard FIR filter* could be designed by mixing asynchronous pipelines and synchronous interfaces, achieving *higher-throughput* and *significantly-lower latency* than the *best* comparable commercial synchronous design.

The chip was designed by my PhD student, Montek Singh, together with IBM engineers, including Dr. Jose Tierno and others, and used several of our group's design components and techniques. The entire core of the filter was asynchronous -- designed using our *high-capacity dynamic pipelines* --, while the external interfaces were synchronous. Hence, the chip appears to the environment as a fully synchronous system. The chip was designed to IBM commercial specifications.

The fabricated chip was evaluated in April-June 2001; it was fully functional. The synchronous interfaces operated at 1.3 Gigasample/sec in 0.18 micron technology, but the internal asynchronous pipelines (using our asynchronous dynamic pipelines) were shown to support a much higher rate: 1.8 Gigasample/sec. The chip met and exceeded the synchronous performance and power requirements even for the next-generation process at IBM; it also obtained 15% higher throughput than current-generation synchronous chips. Most significantly, the new hybrid design had the advantage over synchronous ones of *dynamically variable latency*, depending on input sampling rate. Hence, average latency was significantly lower than for the best comparable commercial synchronous IBM chips.

Selected Publications:

W. Jiang, D. Bertozzi, G. Miorandi, S.M. Nowick, W.P. Burleson and G. Sadowski, "An Asynchronous NoC router in a 14nm FinFET library: Comparison to an Industrial Synchronous Counterpart," ACM/IEEE Design Automation and Test in Europe Conference *(DATE-17),* pp. 732-733, 2017.

G. Miorandi, M. Balboni, S.M. Nowick and D. Bertozzi, "Accurate Assessment of Bundled-Data Asynchronous NoCs Enabled by a Predictable and Efficient Hierarchical Synthesis Flow", IEEE Int. Symposium on Asynchronous Circuits and Systems *(Async-17),* pp. 10-17, 2017.

M. Singh, J.A. Tierno, A. Rylyakov, S. Rylov, and S.M. Nowick, "An Adaptively-Pipelined Mixed Synchronous-Asynchronous Digital FIR Filter Chip Operating at 1.3 GigaHertz," *IEEE Transactions on VLSI Systems*, vol. 18:7, pp. 1043-1056 (July 2010).

J. Tierno, A. Rylyakov, S. Rylov, M. Singh, P. Ampadu, S.M. Nowick, M. Immediato, and S. Gowda, “A 1.3 GSample/s 10-tap Full-Rate Variable-Latency Self-Timed FIR with Clocked Interfaces,” International Solid State Circuits Conference *(ISSCC-02),* Monterey, CA (February 2002).

M. Singh, J.A. Tierno, A. Rylyakov, S. Rylov, and S.M. Nowick, “An Adaptively-Pipelined Mixed Synchronous-Asynchronous Digital FIR Filter Chip Operating at 1.3 GigaHertz" (***Best Paper Finalist)***, IEEE International Symposium on Advanced Research in Asynchronous Circuits and Systems *(Async-02)*, Manchester, UK (April 2002).

9 GRANT HIGHLIGHTS (2000-2015)

Below are some highlights of recent larger grants (see CV for additional smaller grants).

(i) NSF Award (medium-scale)

Title: ``***SHF:Medium: Power-Adaptive, Event-Driven Data Conversion and Signal Processing Using Asynchronous Digital Techniques"***

Time Period: 7/1/10-1/31/15
Total Grant Amount: $1,062,605
My Grant Portion: $500,000 *(approx.)*

Principal Investigator: Prof. Yannis Tsividis, EE Dept., Columbia U. (joint with co-PI S. Nowick)

This medium-scale NSF award targets ultra low-power microelectronic systems, through the development of a new approach to digital signal processing and conversion, called *continuous-time digital signal processing (CT-DSP)*. This work is aimed at applications that require continuous monitoring and processing of information, which may arrive infrequently, or at irregular or unpredictable intervals. Application areas include environmental sensors, and implantable or ingestible biomedical devices. Traditional synchronous processing is a poor match for these applications, because the regular sampling and clock control result in excessive power and aliasing. The event-based nature of the information calls for a drastic re-thinking of how these signals are monitored and processed. This research aims instead to provide a system controlled not by the clock, but by the actual arrival of each event. Asynchronous digital logic techniques, which are ideally suited for this work, are combined with continuous-time data conversion and digital signal processing. This new "event-based" approach promises significant power and energy reduction, as well as higher-quality sampling, in a fully programmable chip.

(ii) NSF Award (medium-scale)

Title: ``***CPA-DA-T: Design and Tools for Easy-to-Program Massively Parallel On-Chip Systems: Deriving Scalability Through Asynchrony"***

Time Period: 8/1/08-7/31/14
Total Grant Amount: $921,686
My Grant Portion: $461,000

Principal Investigator: Prof. Steven Nowick (joint with co-PI Prof. Uzi Vishkin, U. of Maryland)

This medium-scale NSF award is to support the design and optimization a high-throughput, flexible and low-power digital interconnection network for future desktop parallel processors. The asynchrony will facilitate lower power, handling of heterogeneous interfaces, and high access rates (with fine-grained pipelining). This work is in collaboration with the parallel processing architecture group at the University of Maryland, which is developing the synchronous shared-memory processor environment

(called XMT) and simulation tools. My group is leading the design of the novel asynchronous interconnection network. The goal is a *globally-asynchronous locally-synchronous (GALS) network*, that can integrate multiple synchronous cores and caches operating at unrelated clock rates. Most recent GALS networks have focused on low- to medium-performance embedded systems, or involve advanced circuit techniques for high-performance systems. In contrast, this proposal is aimed at medium-to-high end multi-processors, but using largely standard-cell design techniques, and portable design flows.

(iii) DARPA "CLASS" Project (MTO)

Principal Investigator/Lead: Boeing Corporation

Time Period: Fall 2005 through March 2007
Total Contract Amount: $14,000,000
My Subcontract Portion: $502,000

This DARPA program is the largest US government research funding for asynchronous digital design in the last 30 years. Its goal is to make asynchronous digital design viable for the commercial and military sectors. There were approximately 20 large-scale proposals submitted, and only 1 contract funded, headed by Boeing Corporation (PI), with participation of Philips Semiconductors (via its incubated asynchronous startup company, called Handshake Solutions), two asynchronous startups (Theseus Logic, Codetronix) and three key academic groups (Columbia, UNC, U. of Washington).

The two goals of the project are: (i) building a large-scale asynchronous demonstration chip (for Boeing military applications), and compare its performance and cost to an equivalent synchronous chip; and (ii) provide a "legacy asynchronous CAD tool" for future asynchronous designs.

I was brought onto the project at the start of Phase 2, to provide expertise in CAD tool development and optimization techniques. My role was to provide three key components: (i) CAD optimization techniques for robust asynchronous threshold circuits (to improve the unoptimized tool flow provided by Theseus Logic); (ii) CAD optimization techniques for the Philips-based asynchronous tool flow (collaborating with a team from its incubated startup company, Handshake Solutions); and (iii) migration our high-speed MOUSETRAP asynchronous pipelines into the tool flow.

(iv) NSF ITR Award (medium-scale)

Title: ***"A CAD Framework for the Design and Optimization of Large-Scale Asynchronous Digital Systems"***

Time Period: 9/1/00-8/31/07
Total Grant Amount: $1,600,000
My Grant Portion: $806,000

Principal Investigator: Prof. Steven Nowick (joint with co-PI Prof. Peter Beerel, USC).

(v) NSF ITR Award (medium-scale)

Title: ***"Asynchronous Digital Signal Processing for the Software Radio"***

Time Period: 9/1/00-8/31/03
Total Grant Amount: $969,227
My Grant Portion: $400,000 *(approx.)*

Principal Investigator: Prof. Ken Shepard, EE Dept., Columbia U. (joint with co-PI S. Nowick)

In grant (iv), we have developed a comprehensive CAD tool framework for the synthesis and optimization of asynchronous systems, targeting different asynchronous circuit styles (from high-speed/less robust to moderate-speed/highly-robust). This work encompasses the entire ***CaSCADE Tool Environment*** free public-domain tool release, available on the web site: www.cs.columbia.edu/~nowick/asynctools. Columbia tools include: (i) *MINIMALIST,* for synthesis and optimization of asynchronous burst-mode controllers; (ii) *ATN_OPT,* for asynchronous threshold networks; and (iii) *DES Analyzer,* for performance analysis and verification of concurrent systems. (See descriptions above.)

In grant (v), we explored the applicability of high-speed asynchronous pipelines to designing voltage-controlled asynchronous micro-architectures for software radio.

Note: The ITR initiative is an outgrowth of President Clinton's "PITAC" advisory committee for national information technology, to fund "long-term risk-taking research" in information technology. In 2000, only 62 medium-scale NSF ITR awards were granted, out of 920 submitted proposals, across all areas of information technology.

In 2000, I was 1 of only 4 investigators nationally, in all research areas, to receive 2 medium-scale NSF ITR awards.