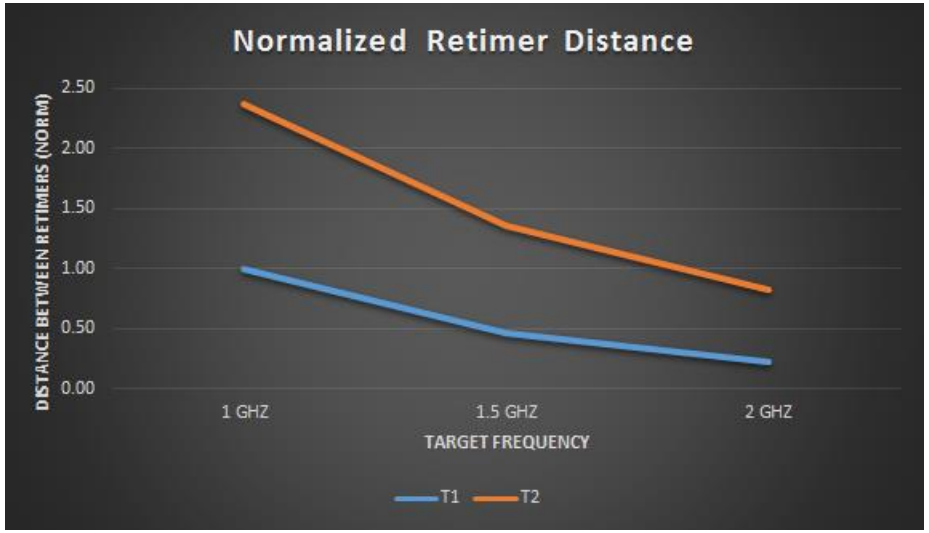
RECONFIGURABLE LINKS FOR SELF-TIMED ON-CHIP COMMUNICATION

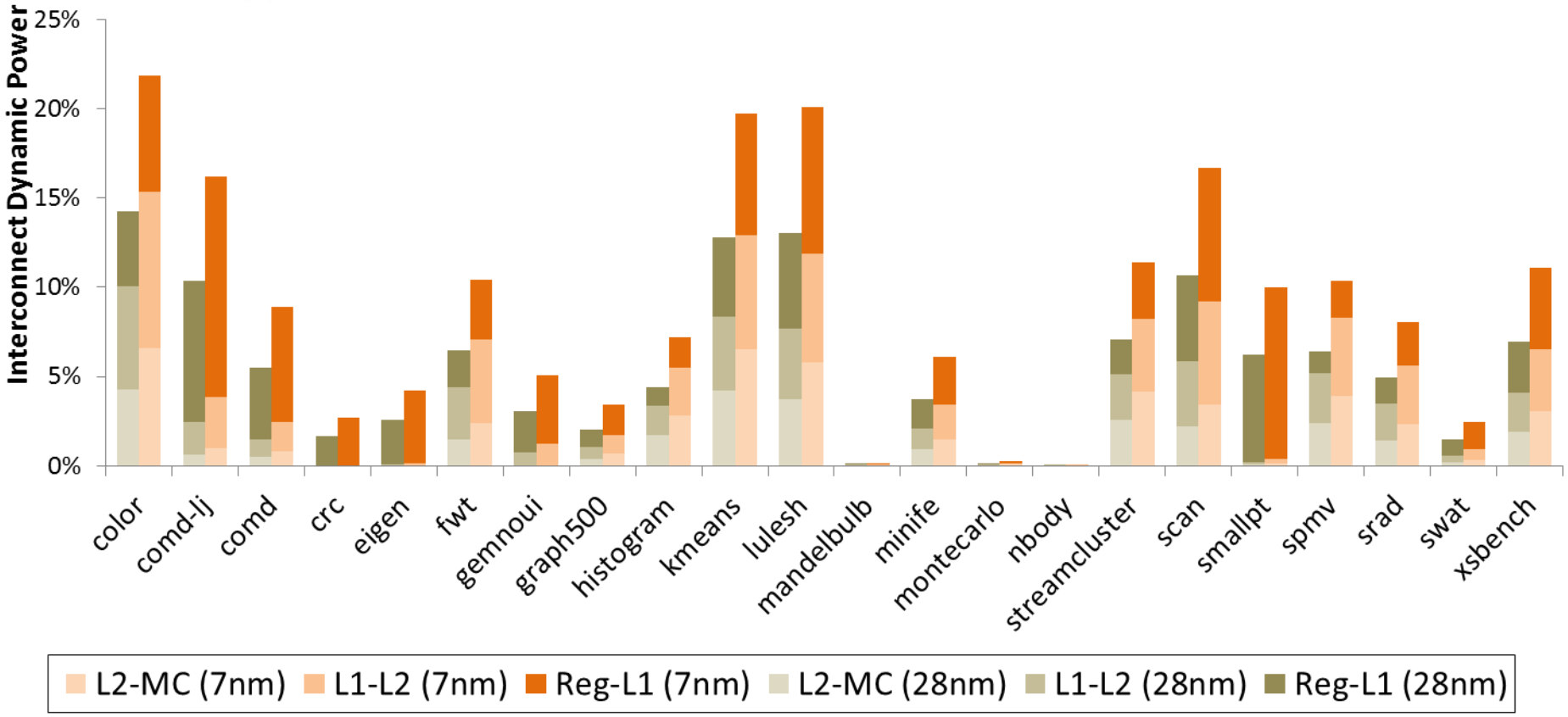
SHOMIT DAS, GREG SADOWSKI
AMD RESEARCH

TALKING POINTS

- Data movement energy reduction is a major challenge
- Timing heterogeneity is the enabler for a GALS world
- Self-timed pipeline controller and variations
- Speculative handshaking
- Real-time link bandwidth detection and dynamic reconfiguration



DATA MOVEMENT ENERGY



- Physical wire properties and scaling lead to increased communication energy
- Up to 14% dynamic power spent on interconnects in modern chips*
- Research in techniques to minimize toggling, compress data, PIM

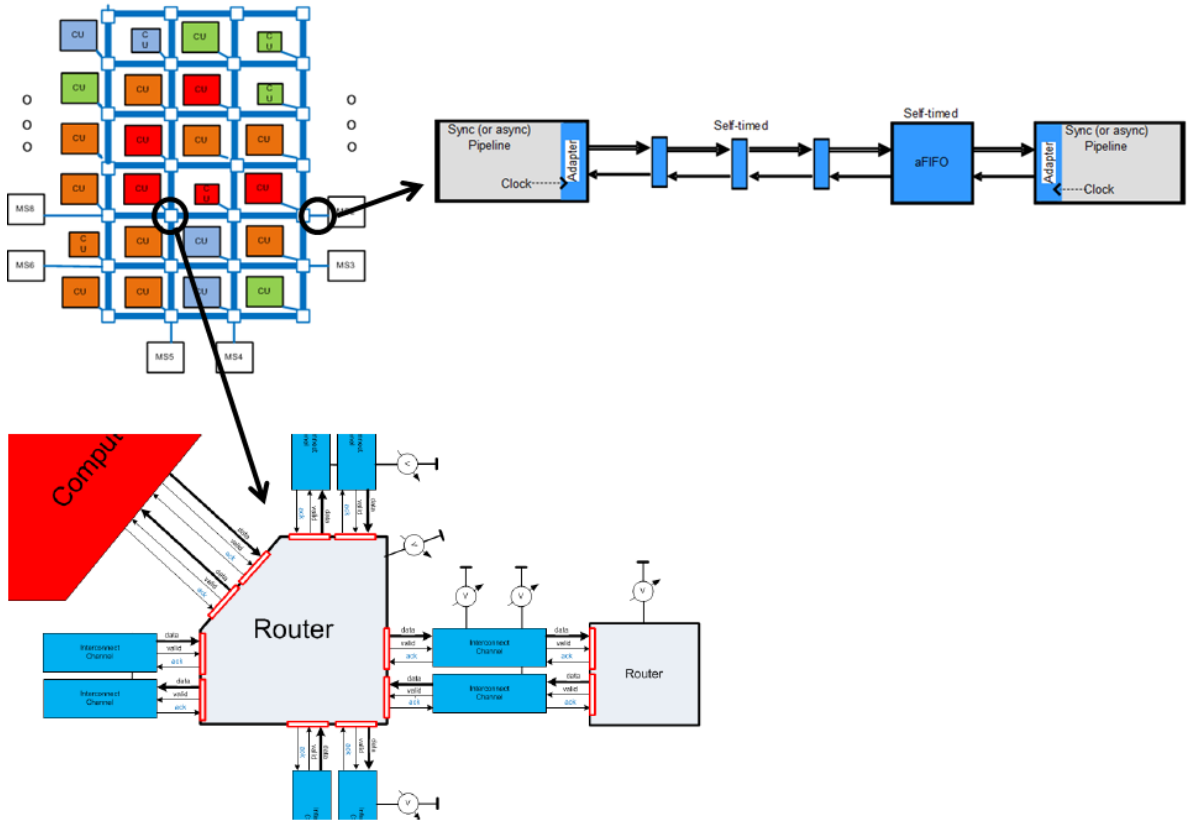
* “Measuring and modeling on-chip interconnect power on real hardware”, V. Adhinarayanan, IISWC 2016

GALS SYSTEMS

- Presence of several timing domains facilitates average-case performance in presence of PVT variation
- Functional blocks in a heterogeneous SoC are not constrained by global timing reference
- Power supply noise and spatio-temporal workload variations create the need for even more fine-grained timing domains
- Ring oscillator based clock generation provides better correlation with delay variability
- Asynchronous communication among these domains in a multi-frequency regime

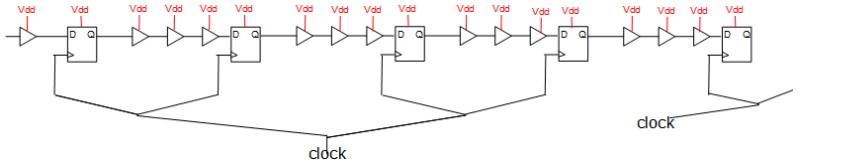
SELF-TIMED ON-CHIP COMMUNICATION

- Structured on-chip networks are replacing bus-based communication systems
- Built-in “clock” gating for lower energy costs
- Separating communication from computation allows us to avoid several challenges associated with asynchronous design and timing closure
- Routers, synchronizers, FIFOs, **data links**

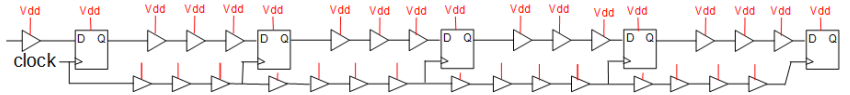


DATA LINKS

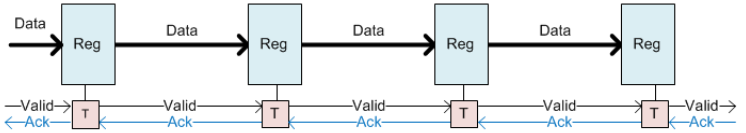
- Regular synchronous interconnect
  - requires clock distribution to every retiming flop = expensive in NoC of heterogeneous system
  - clock gating not easy to implement in long connections

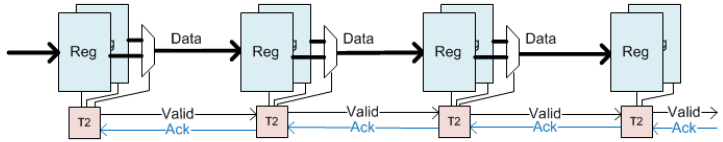


- Source synchronous bus interconnect
  - data-clock bundle, no global clock distribution
  - no retimer level flow control (once data is sent is has to flow thru) = requires a data buffer at the receiver



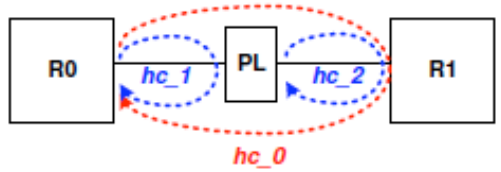
- For self-timed NoC we need self-timed data links
  - Basic handshake
    - connection is not used during Ack propagation time = not full usage of bandwidth
    - register could be just a latch, so area and power savings
    - the series of registers are acting as a FIFO
  - Full bandwidth
    - fixes the problem of transmission gaps due to Ack propagation
    - need to measure the Valid-Ack propagation time in order to know when to send the next data

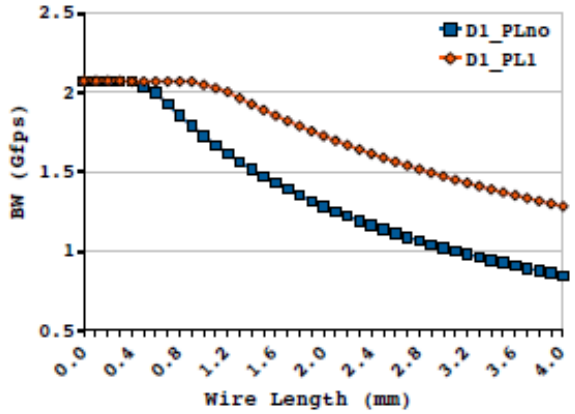


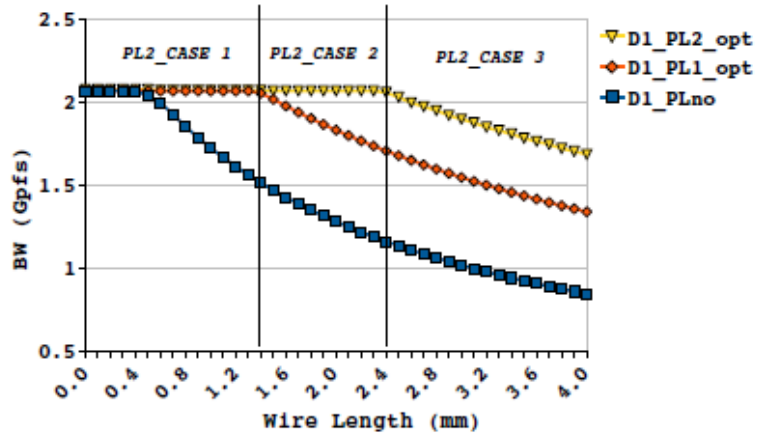


LINK PIPELINING

- Long wire delays can be the limiting factor to NoC bandwidth
- Adding intermediate pipeline latches on the links preserves/enhances available bandwidth while adding some latency
- Higher number of retiming latches provides higher bandwidth
- Placement of latches must be to balance the newly formed handshake cycles so as to eliminate bottlenecks
- Selective pipelining to optimize power*delay metric



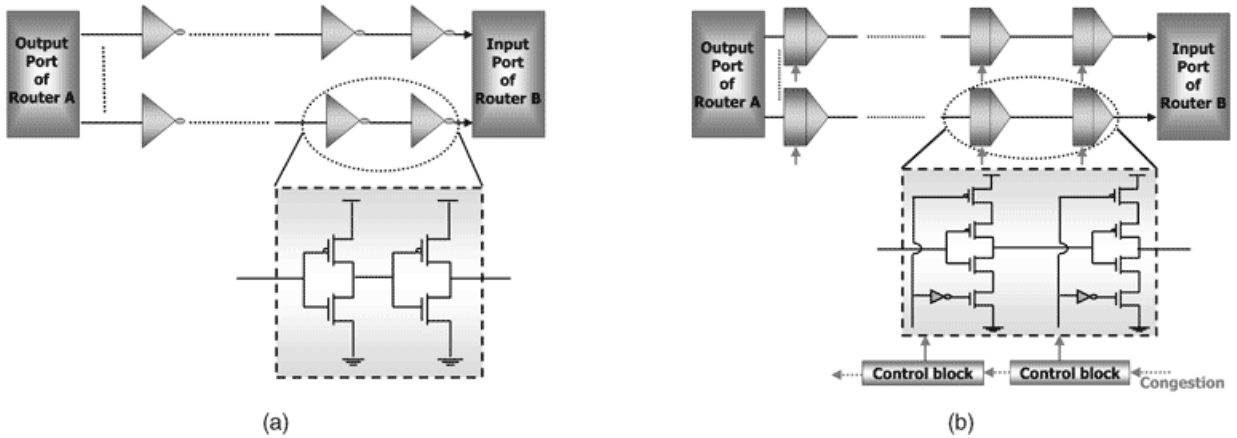




*Figures from "Asynchronous Network-on-Chip Design and Evaluation", J. You 2011

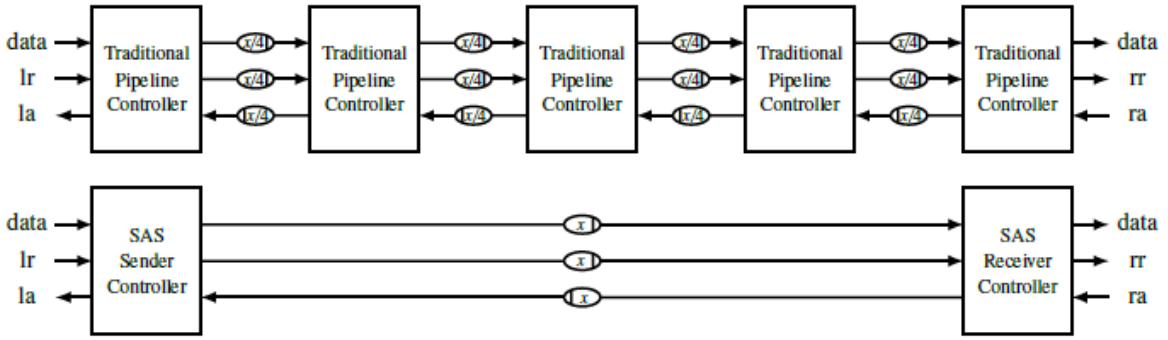
OTHER EFFORTS

- Store data on tri-state repeaters
- During high network load, enable storage functionality on repeaters
- Packet-level solution



"Adaptive channel buffers in on-chip networks", A. Kodi, IEEE Trans on Computers 2008

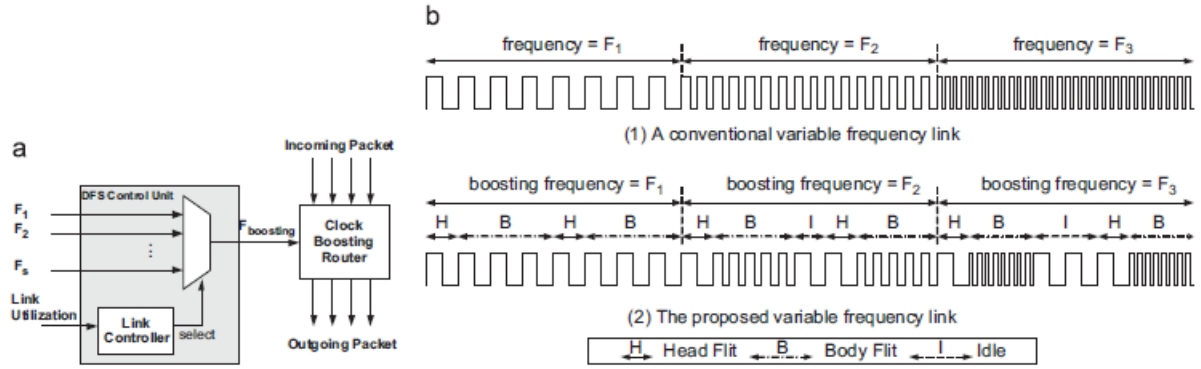
- Eliminate back-propagation delay for asynchronous bundled data links
- Token counting at sender and sufficient storage at receiver (SAS)
- Discussed more in detail as we go along



"SAS: Source Asynchronous Signaling protocol for asynchronous handshake communication free from wire delay overhead", S. Das, ASYNC 2013

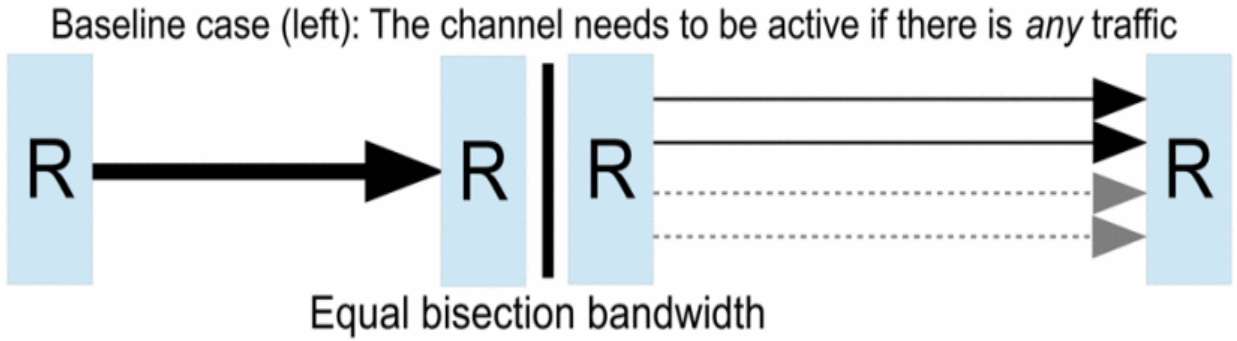
OTHER EFFORTS

- Dynamic Frequency Scaling (DFS) for links
- Use separate clocks for header and body flits
- Since path is already reserved by header, body flits can be “clock boosted”



“A variable frequency link for a power-aware NoC”, S. Lee, Integration VLSI Journal 2009

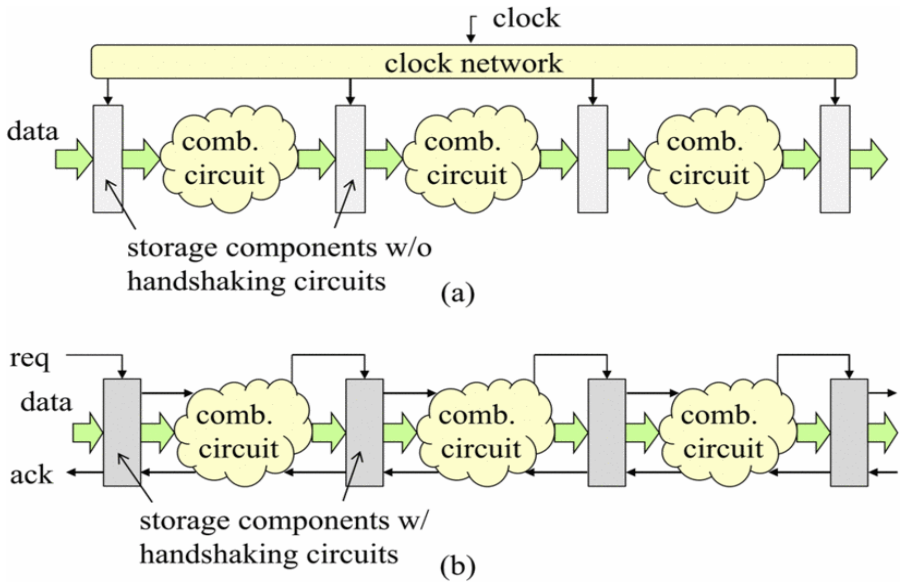
- Break down channels into lanes
- Each lane has independent power-gating capability
- Activate VCs individually

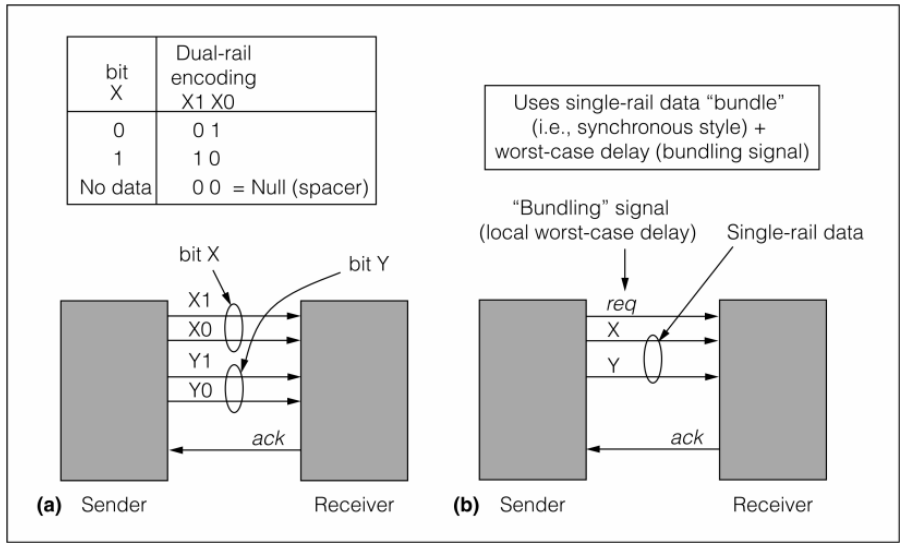


“Variable width datapath for on-chip static power reduction”, G. Michelogiannakis, NOCS 2014

BUNDLED-DATA PIPELINES

- No clock network in asynchronous communication
- Dual-rail encoding employs 2 wires per bit and completion detection to identify valid codeword
- Single-rail bundled data protocol is better suited in the wire-constrained environment
- Better power and area overheads for bundled data
- Matched delay element in control logic to provide latching guidance

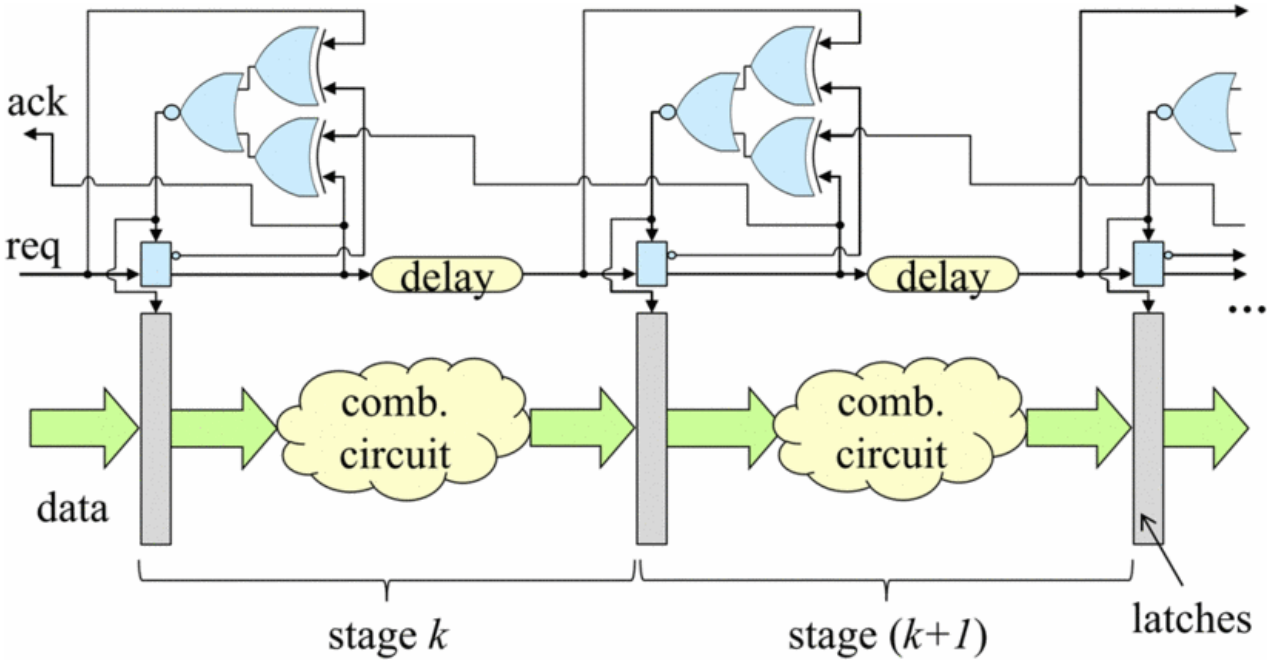




"High performance asynchronous pipelines: an overview", S. Nowick IEEE Design and Test 2011

PIPELINE CONTROLLER

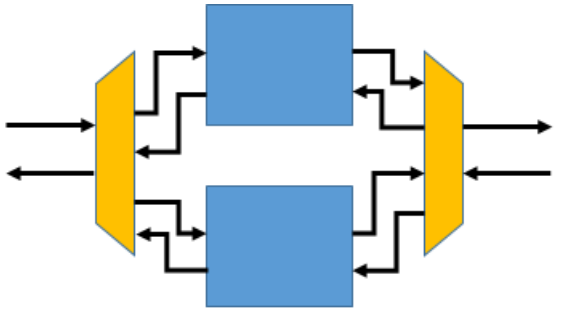
- 2-phase bundled data protocol
- Normally opaque latches
- Low glitch power, easy synthesis, verification

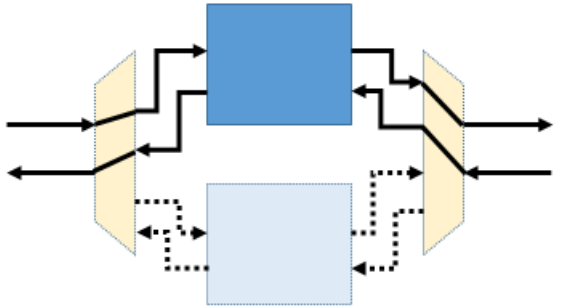


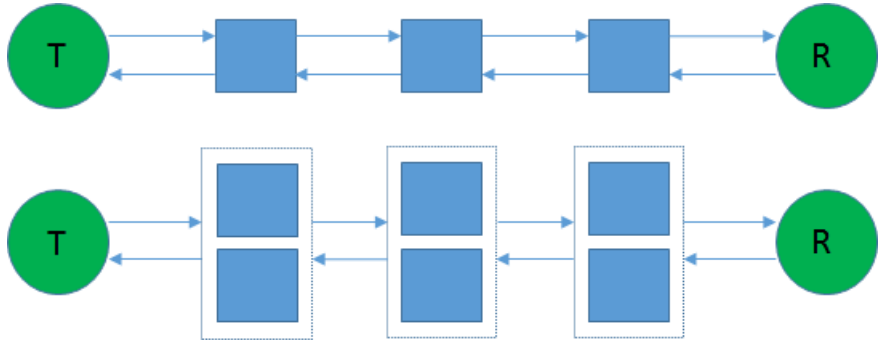
"A new asynchronous pipeline template for power and performance optimization", K. H. Ho, DAC 2014

T2 ELEMENT

- In handshaking protocols, round-trip cycle time determines throughput; long channel bandwidth is dominated by wire delay
- We propose storing 2 data tokens at each retiming stage and employing credit-based token flow
- Same number of wires as before, and channel protocol is still the same; only additions are merge-join logic, extra set of latches, and pipeline controller at each stage
- Latches, not flops; therefore, similar storage as clocked techniques



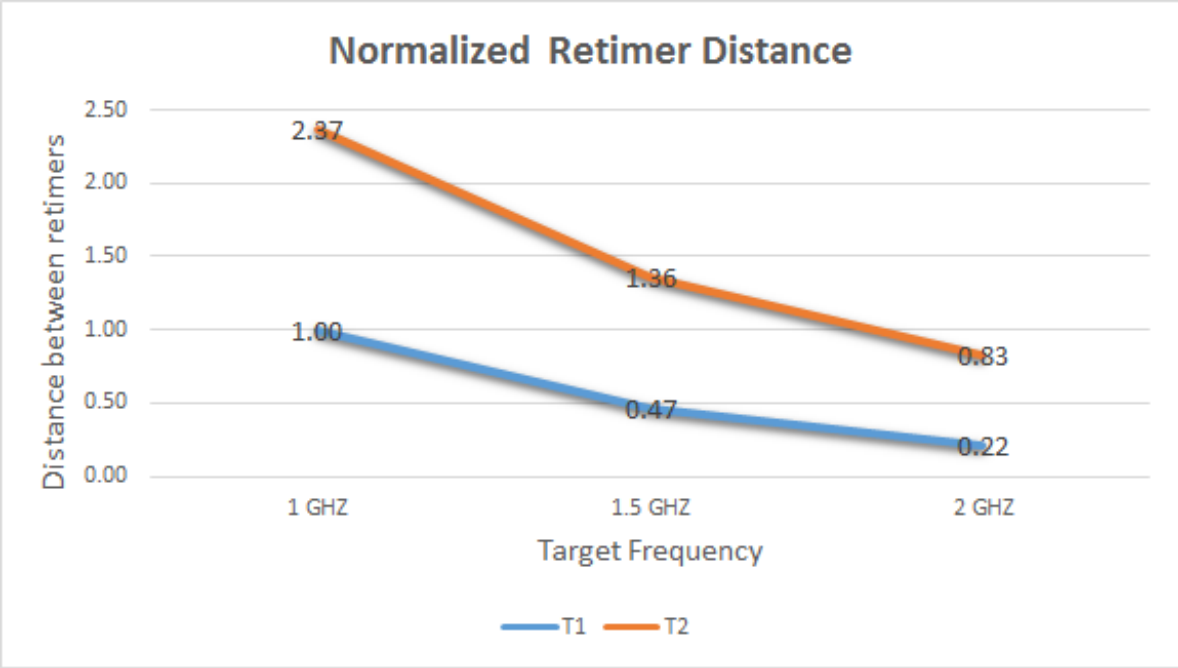


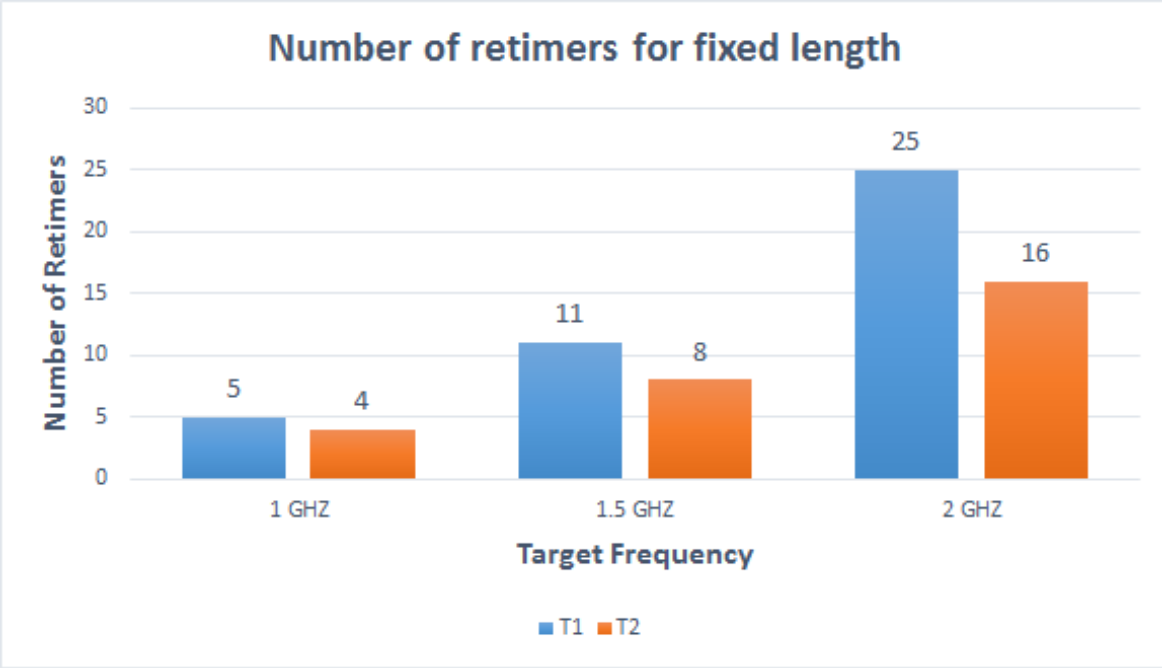


T2 ELEMENT

- Pipelined data link can be abstracted as linear flow-through FIFO
- SAS technique adds extra storage in the form of (non-linear) logarithmic FIFOs at the terminals
- That technique is challenged by how many tokens can exist on the link at a given time; i.e. signal integrity for wavepipelining
- This method allows for intermediate pipeline stages, which turns the long link into a series connection of logarithmic FIFOs with capacity = 2
- Depending on the overhead of implementation, T2 method allows for close to twice the bandwidth as compared to the conventional handshaking (T1) method

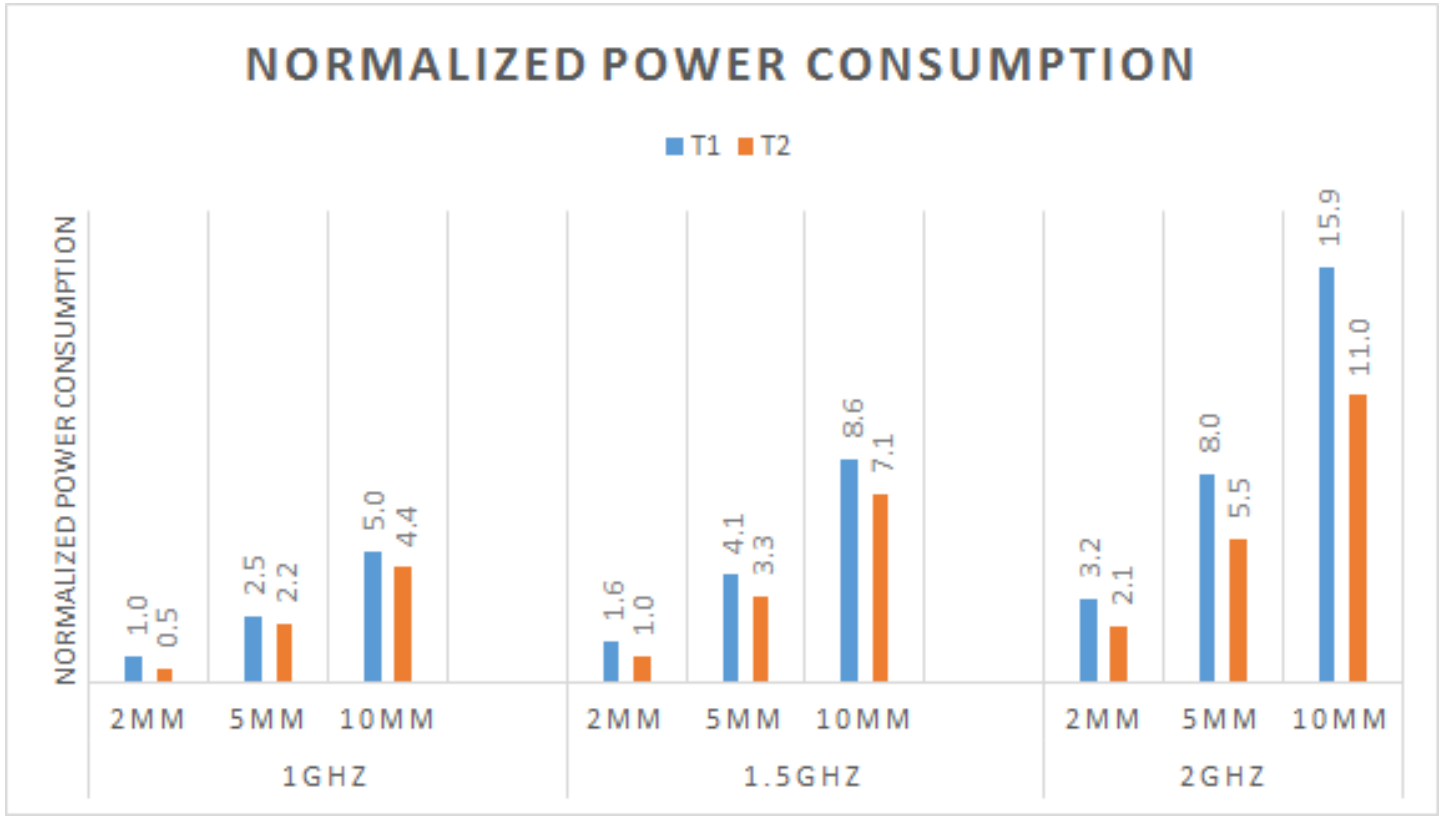
T2 ELEMENT





- Speculative handshaking allows for relaxed handshake timing constraints
- For a target frequency, fewer retimers are required
- Conversely, for a fixed number of retimers, much higher throughput is possible

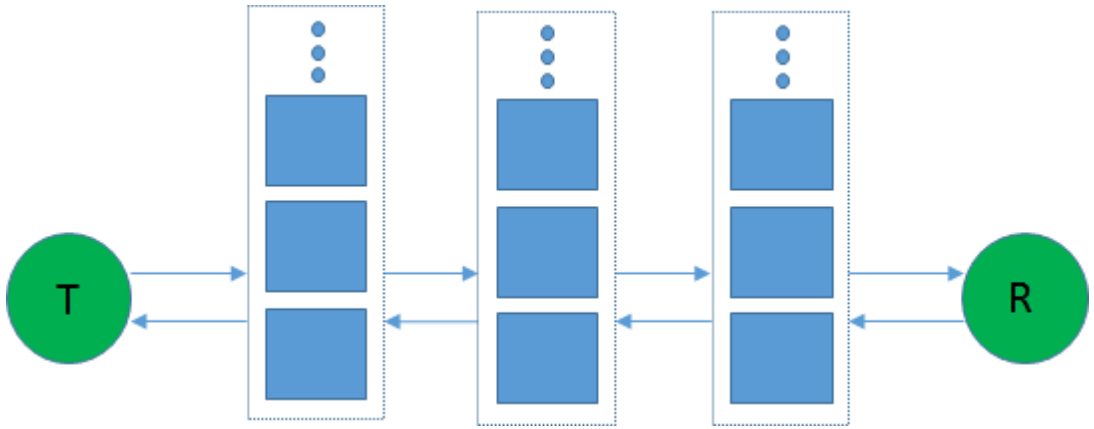
T2 ELEMENT



- Smaller logic footprint in the T2 scheme
- Up to 53% power savings
- Higher throughput target = larger savings when compared to T1

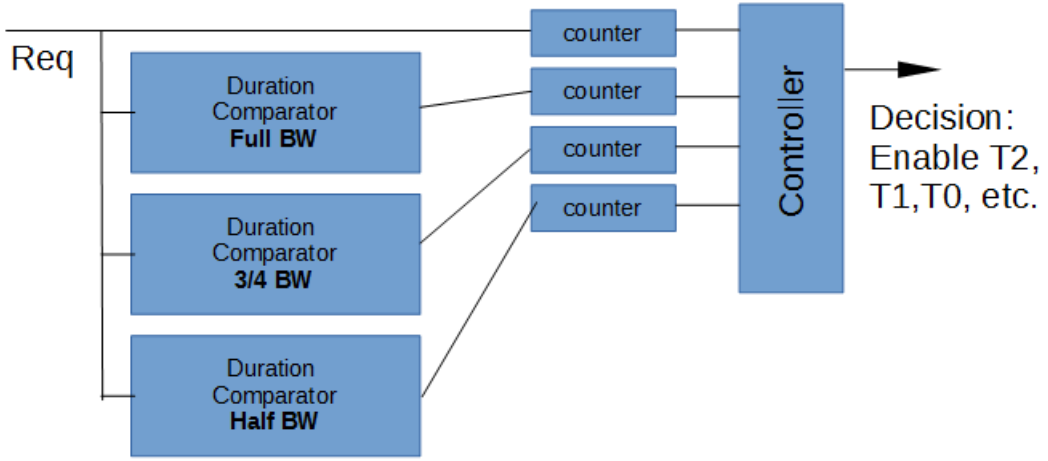
TN ELEMENT

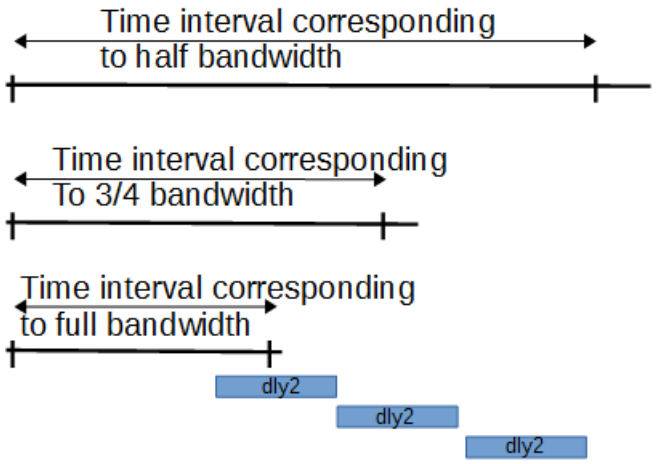
- Number of pipeline latches, and placement of retimer stages determine the available bandwidth on the link
- Speculative handshaking based credit-token techniques can allow for large number of tokens to be ‘in-flight’ on the link
- Similar to SAS, signal integrity concerns require minimum pulse separation
- Area overheads of the latches are also of concern; complex MUXing



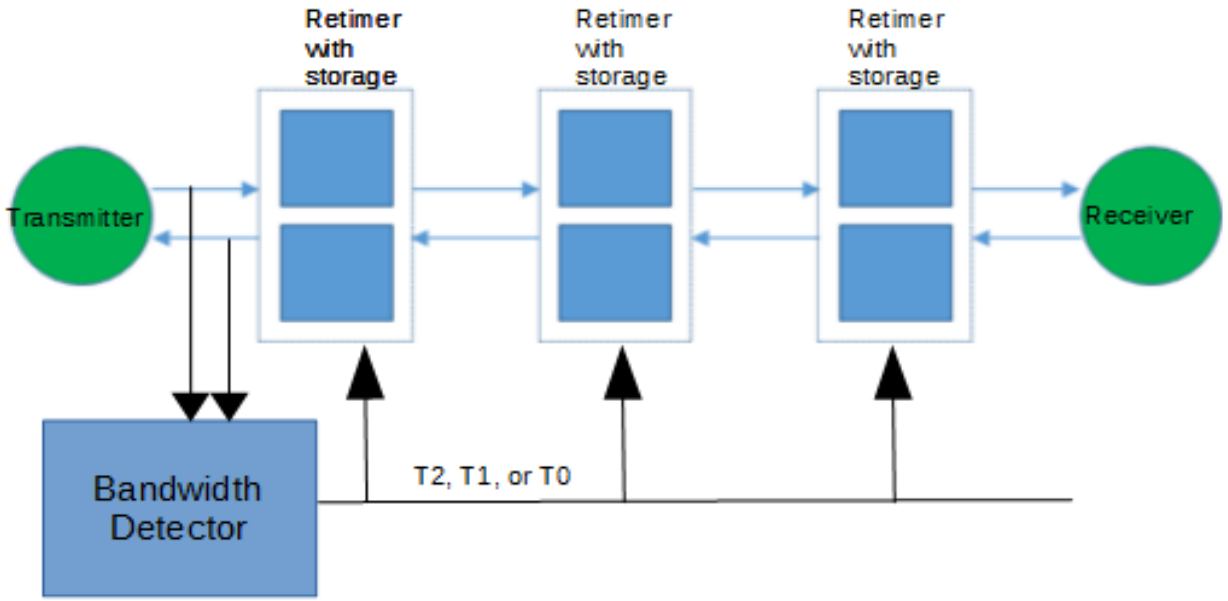
BANDWIDTH DETECTION

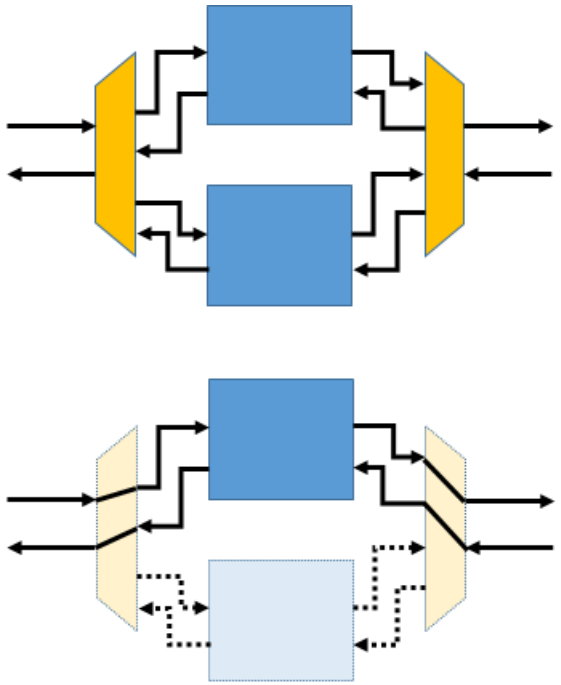
- We need to count req-req cycle times
- Fast, medium, and slow ‘buckets’
- Each request is sorted into these buckets
- Dominant bandwidth range is used to determine operating mode
- This apparatus can be shared among several links on a round-robin basis



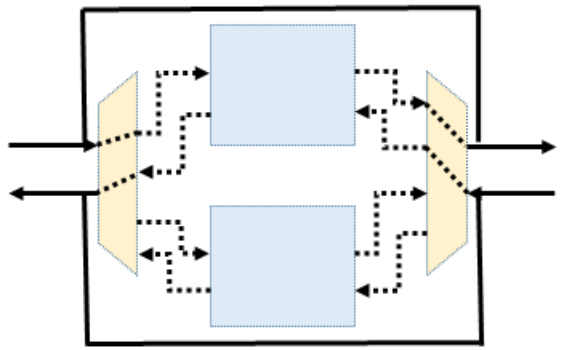


RECONFIGURABLE INTERCONNECT





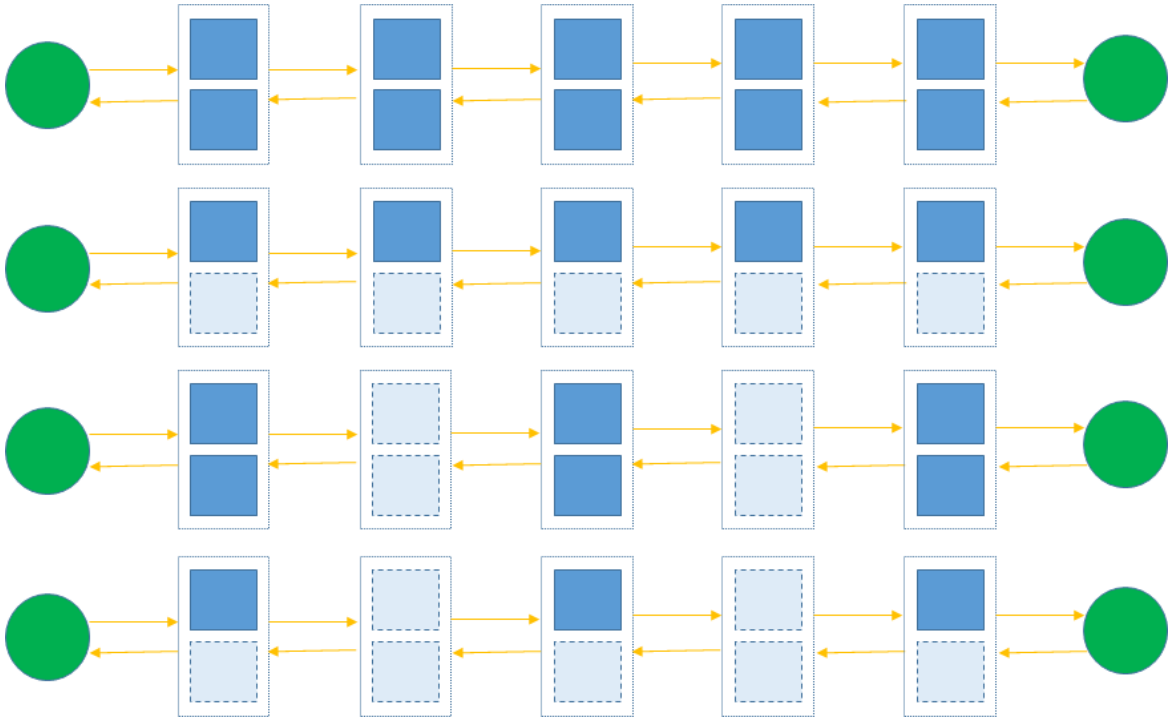
- Bandwidth detector provides guidance as to which mode will be most suitable (T2/T1/T0)
- Communicated to all retiming stages along the link
- Power gating + MUXing converts T2 stage to T1 or T0 (bypass)
- T0 stage still provides repeater functionality



T2/T1/T0

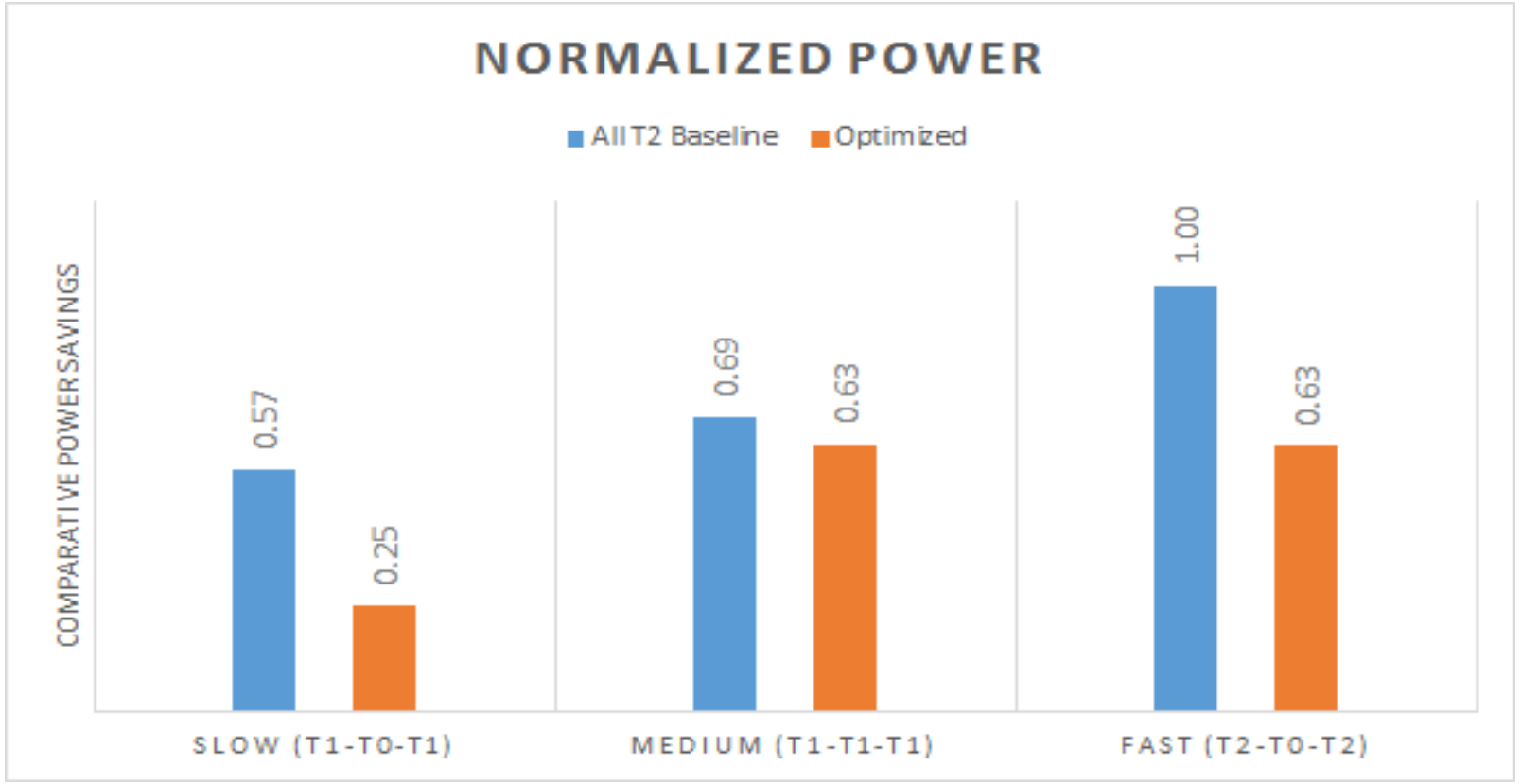
RECONFIGURABLE INTERCONNECT

- Available link bandwidth is dependent on the number of data tokens that can be stored on the link
- Channel is designed to support maximum burst of traffic; at other times, turn off extra stages to save static power
- Several configurations are possible, depending on the predicted link activity
- We evaluate 3 modes in this paper: fast (T2-T0-T2), medium (T1-T1-T1), slow (T1-T0-T1)



Different operating modes for the same link

RESULTS



- All-T2 baseline – 2mm link at 1 GHz
- Up to 55% savings in the 'SLOW' case
- Comparisons will look different for other link configurations and bandwidth targets

LOW VOLTAGE SIGNALING

- Another application is in links utilizing low voltage signaling
- When driver voltage is reduced, link latency increases, therefore reducing bandwidth
- Use default T1 element rated for normal operation
- Enable dormant T2 element to increase the bandwidth while saving power

FUTURE WORK

AMD

- Better traffic prediction techniques
- Improvement on pipeline latch controller (in preparation)
- Extension to synchronous links (in preparation)
- Study router-link balancing to determine retimer stage placement
- Apply to real-world designs and evaluate real workloads

SINCERE THANKS

AMD

- Wayne Burleson, Guoqing Chen for their inputs
- Joseph Greathouse for presenting
- Reviewers for their comments

DISCLAIMER & ATTRIBUTION

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.

The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION

BACKUP SLIDE - DURATION COMPARATOR

- Programmable delay lines measure time delays between successive input edges
- SR latches to create pulse separation between signal and its delayed value
- Output of circuit indicates the time both latches are ON
- Programmable delay calibrated to full/three quarters/half bandwidth delays