Pipeline Frequency Boosting: Hiding Dual-Ported Block RAM Latency using Intentional Clock Skew

a place of mind
THE UNIVERSITY OF BRITISH COLUMBIA

Alexander Brant
alexb@ece.ubc.ca

Ameer Abdelhadi
ameer@ece.ubc.ca

Aaron Severance
aaronsev@ece.ubc.ca

Guy G.F. Lemieux
lemieux@ece.ubc.ca

Department of Electrical and Computer Engineering
University of British Columbia
Vancouver, Canada

Objectives

Hide dual-ported block RAM access latencies without additional pipeline stages or architectural changes.

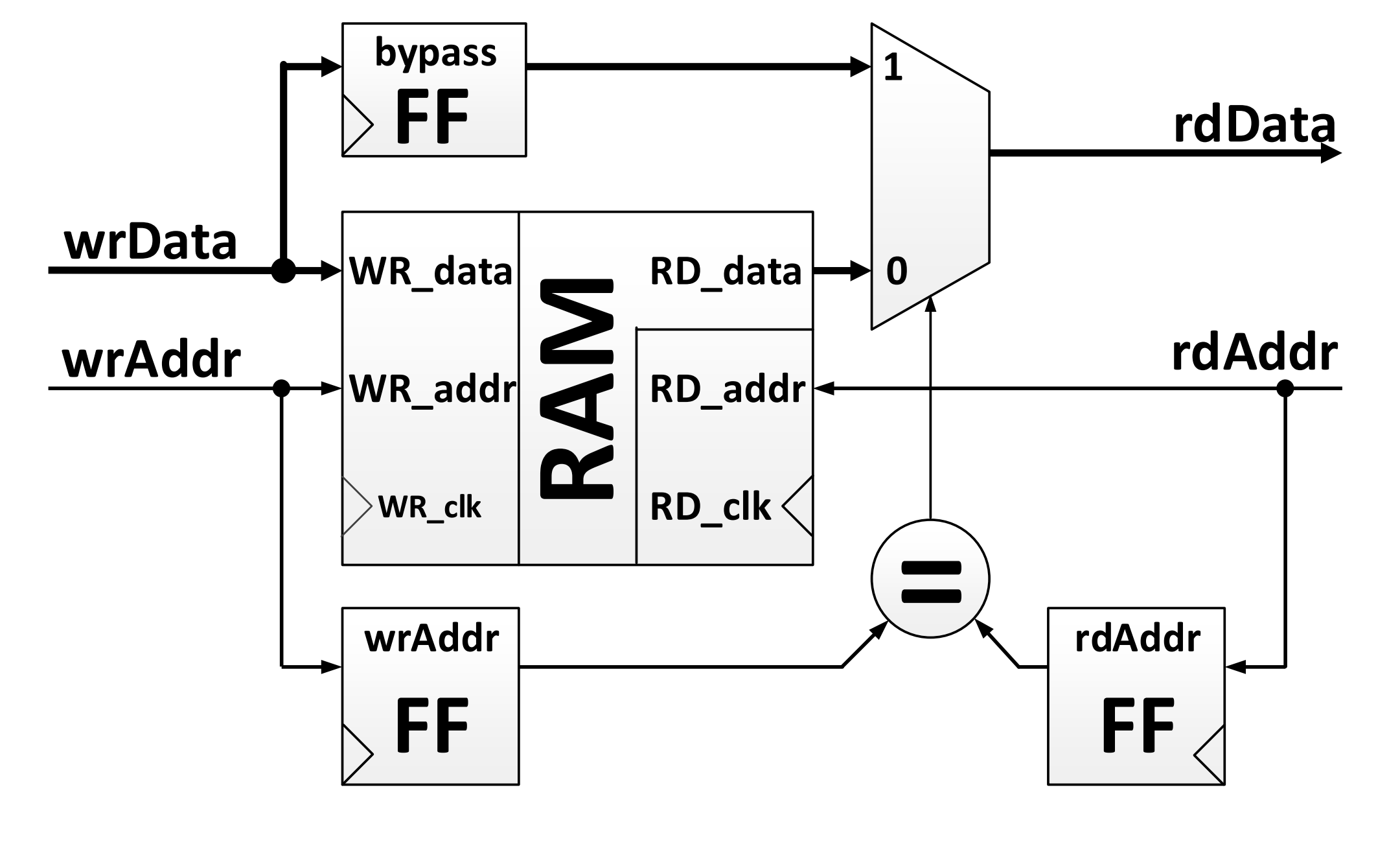
Method

Clock skewing is employed to effectively eliminate the read and write latency of memories, while preserving functionality, and using fewer resources than conventional bypass designs.

Pipeline Bypassing/Forwarding

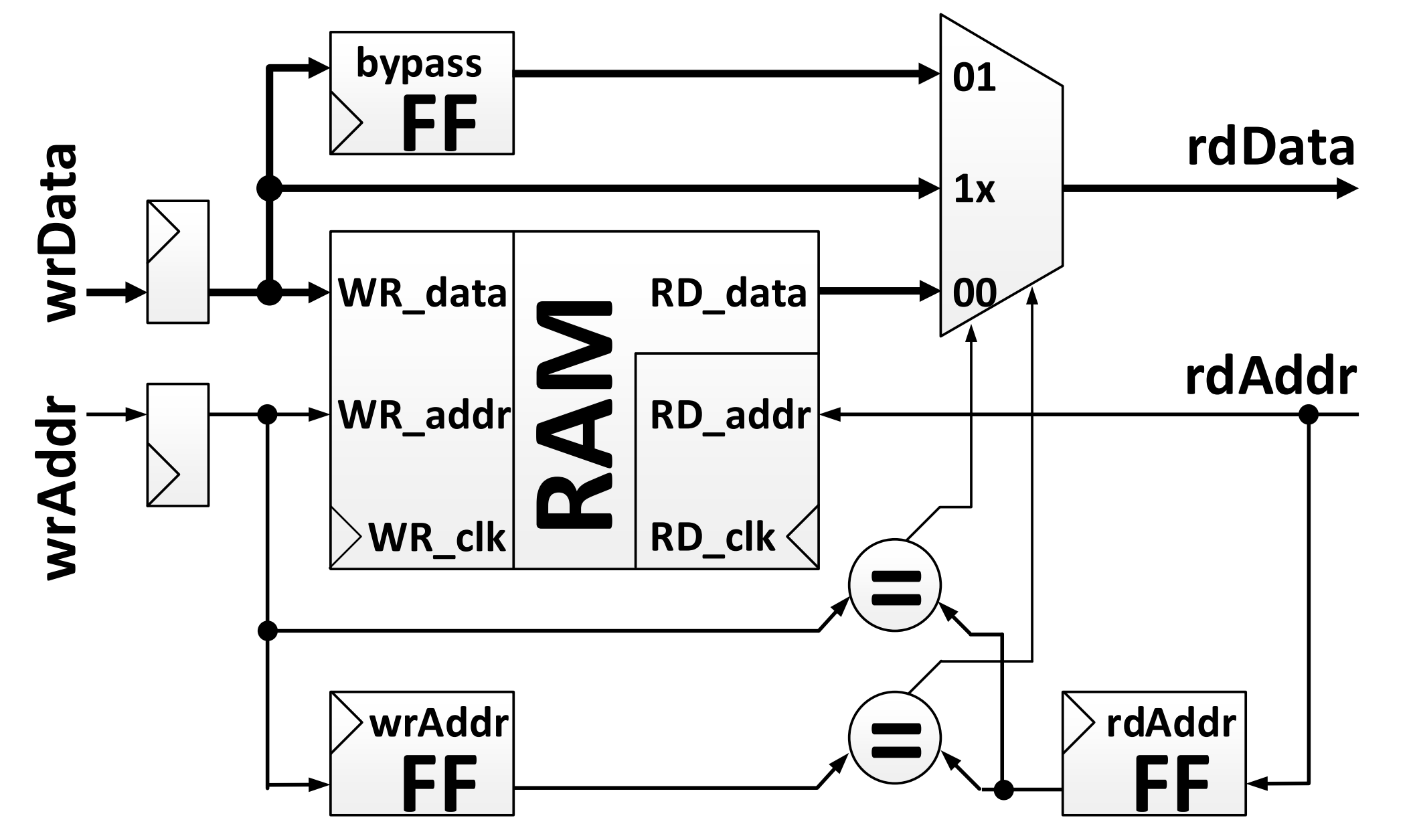
Written data is passed forward through a bypass register, skipping any additional cycles incurred by RAM writing process.

Single-Stage Bypass



Load bypass register if read-after-write

Fully Pipelined Bypass

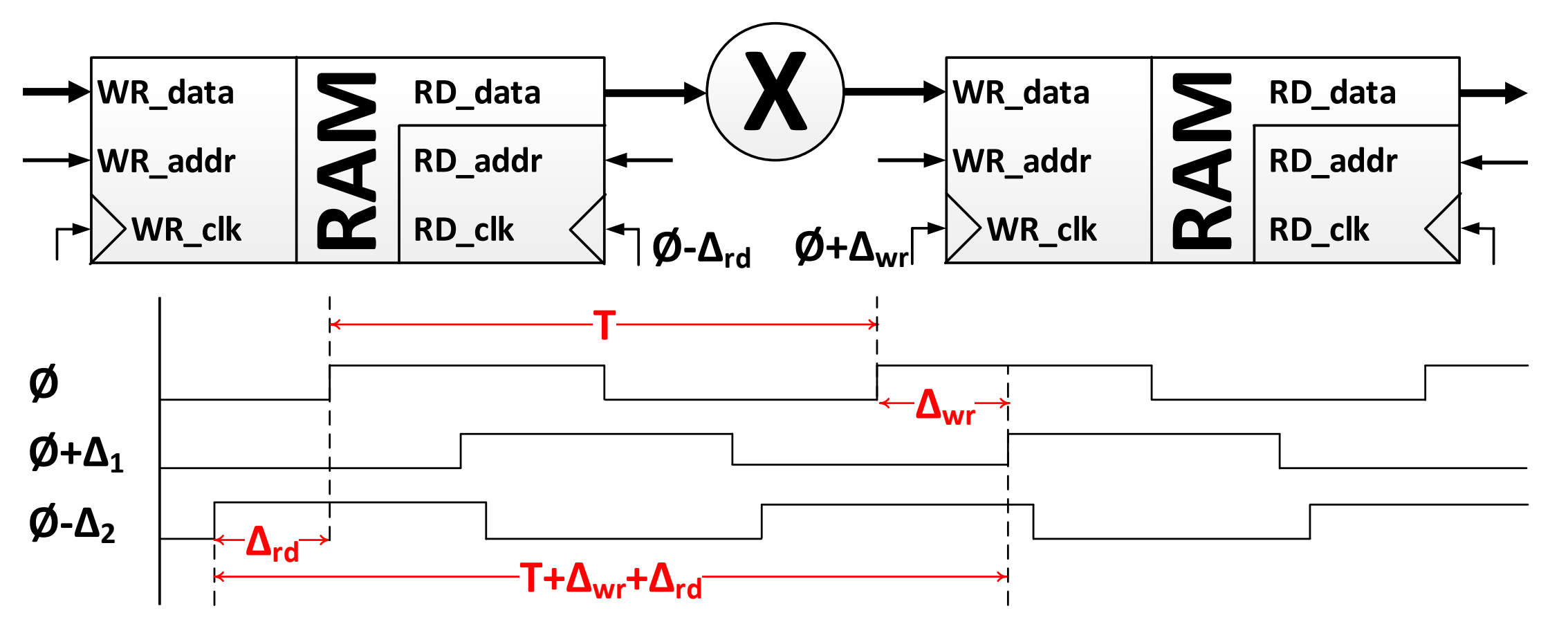


Earlier read-after-write is also bypassed

Clock Skew Scheduling

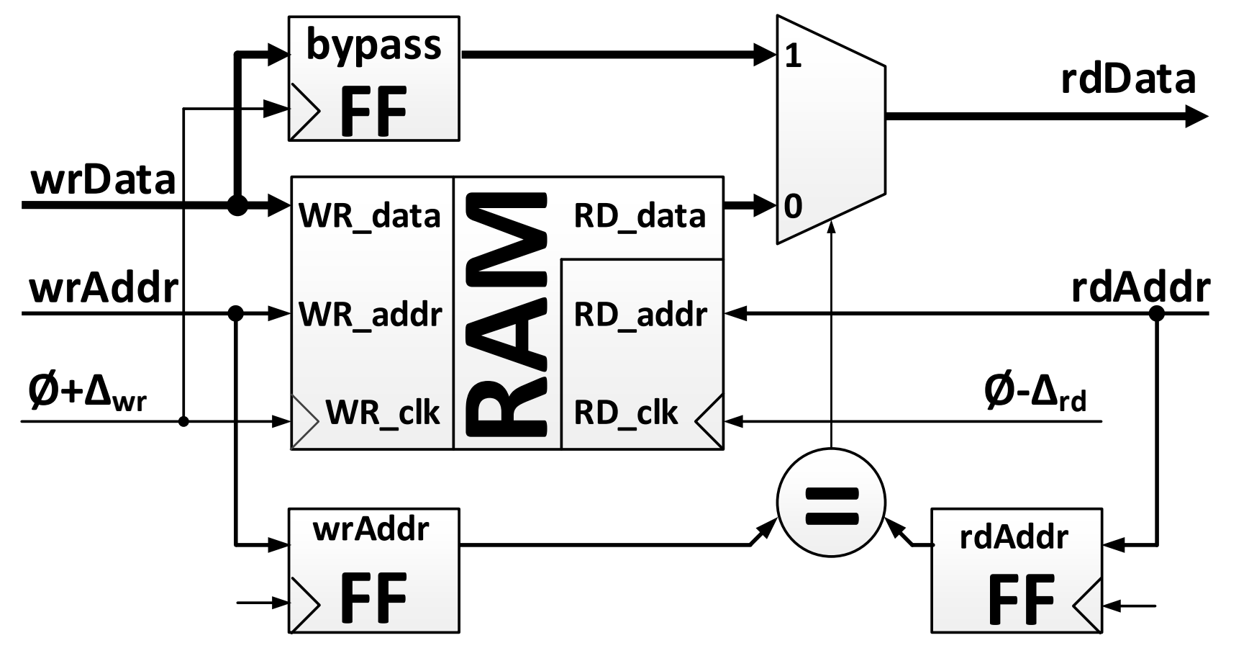
Useful or intentional skewing introduces skews to clocked elements allowing critical paths longer periods.

Pipelined Block RAM



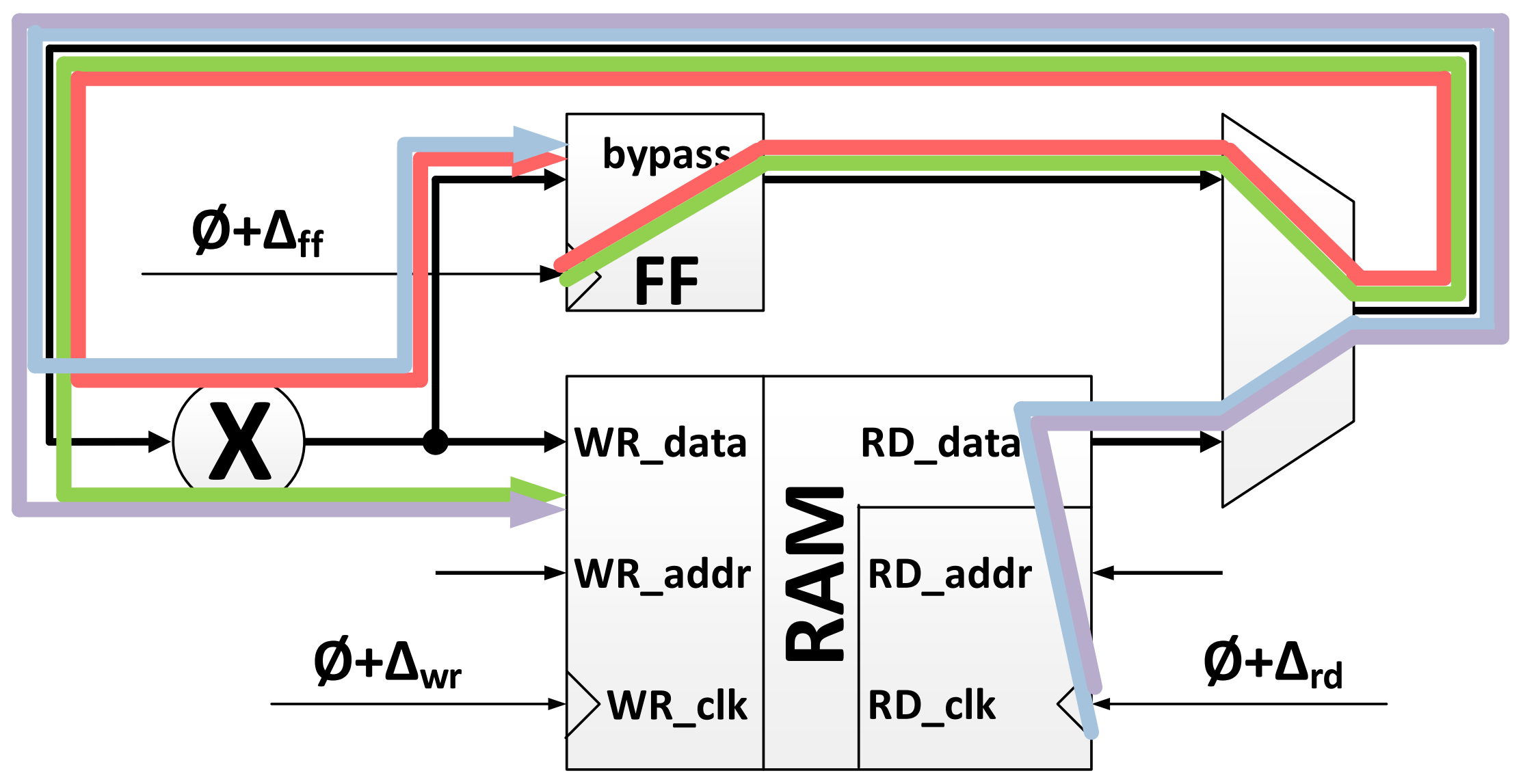
The write port is clocked late and the read port is clocked early to provide the processing logic a wider clock period.

Single-Stage Bypass Block RAM



Clock scheduling for a Single-stage Bypass: The write lags by Δ_{wr} and the read leads by Δ_{rd}.

Timing Paths Analysis



$T \geq \Delta_{ff} + t_{c\rightarrow o}(ff) + t_d(ff \rightarrow mux) + t_d(mux) + t_d(mul) + t_d(mul \rightarrow ff) + t_{su}(ff) - \Delta_{ff}$

$T \geq \Delta_{ff} + t_{c\rightarrow o}(ff) + t_d(ff \rightarrow mux) + t_d(mux) + t_d(mul) + t_d(mul \rightarrow ram_{wr}) + t_{su}(ram_{wr}) - \Delta_{wr}$

$T \geq \Delta_{rd} + t_{c\rightarrow o}(ram_{rd}) + t_d(ram_{rd} \rightarrow mux) + t_d(mux) + t_d(mul) + t_d(mul \rightarrow ff) + t_{su}(ff) - \Delta_{ff}$

$T \geq \Delta_{rd} + t_{c\rightarrow o}(ram_{rd}) + t_d(ram_{rd} \rightarrow mux) + t_d(mux) + t_d(mul) + t_d(mul \rightarrow ram_{wr}) + t_{su}(ram_{wr}) - \Delta_{wr}$

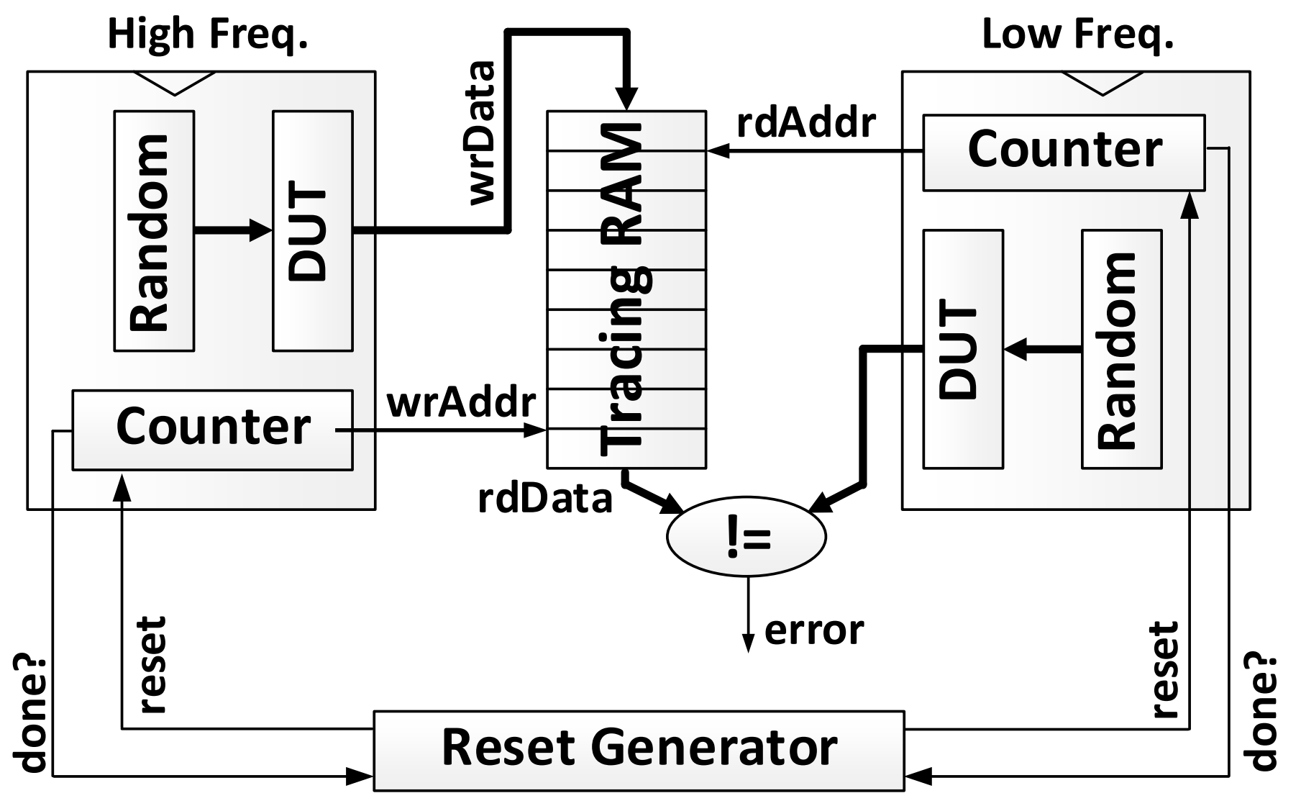
For minimal clock period, Δ_{wr} and Δ_{rd} are

$$\Delta_{wr} = t_d(mul \rightarrow ram_{wr}) + t_{su}(ram_{wr}) - t_d(mul \rightarrow ff) - t_{su}(ff) + \Delta_{ff}$$

$$\Delta_{rd} = t_{c\rightarrow o}(ff) + t_d(ff \rightarrow mux) - t_{c\rightarrow o}(ram_{rd}) - t_d(ram_{rd} \rightarrow mux) + \Delta_{ff}$$

Performance Testing Circuit

- Altera/Terasic DE2-115 board with a Cyclone IV-E
- Static timing analysis is derived using QuartusII
- A dual-clocked tracing RAM is written by a DUT at high freq. and read by a reference at low freq.
- Randomly initialized RAMs and random addresses
- Clock enables were avoided for high performance



Experimental Results

Design Method		Fmax (MHz) (QuartusII STA)	Fmax (MHz) (Tested)	#LUTs	#FFs
Fully Pipelined bypass		179	208	37	34
Single-Stage bypass	$\Delta_{wr}=\Delta_{rd}=0$	113	153	21	17
	Optimal Δ_{wr},Δ_{rd}	183 Δ_{wr}=3.9ns; Δ_{rd}=0	238 Δ_{wr}=1.9ns; Δ_{rd}=0	21	17

A pipeline with 16-bit data width and 8-bit address width

50% less FFs

43% less LUTs

14% performance improvement

a place of mind
THE UNIVERSITY OF BRITISH COLUMBIA