ISCAS

Vancouver • 2004

Tutorial Notes

Tutorial 6:
Clocking and Synchronization issues in sub-100nm System on Chip (SoC) Designs

Presented by:
Ramalingam Sridhar, State University of New York at Buffalo
Ram Krishnamurthy, Intel Corporation
Sanu K. Mathew, Intel Corporation

Sunday Afternoon, May 23, 13:15 - 16:15

ISCAS

Vancouver • 2004

Tutorial Notes

Tutorial 6:
Clocking and Synchronization issues in sub-100nm System on Chip (SoC) Designs

Presented by:
Ramalingam Sridhar, State University of New York at Buffalo
Ram Krishnamurthy, Intel Corporation
Sanu K. Mathew, Intel Corporation

Sunday Afternoon, May 23, 13:15 - 16:15

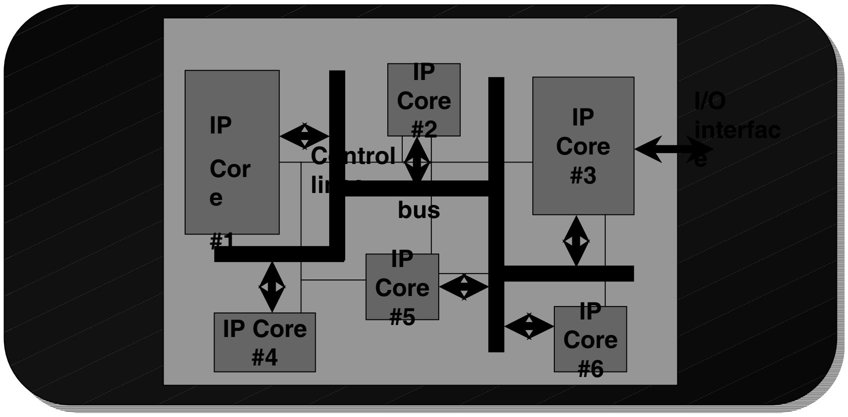
Clocking and Synchronization in UDSM SOCs

Professor Ramalingam Sridhar
University at Buffalo, SUNY
Buffalo, NY, USA

Outline

- UDSM SoC – An Overview
- Design Challenges for SoCs
- Clocking and Synchronization Problems
- Potential Solutions
  - Asynchronous Design
  - GALS
  - Current-mode Interconnects
  - Other Approaches
- Delay Variations and Wave-pipelining
- Summary

What is SoC?



- Heterogeneous integration of components makes SoCs powerful, flexible and versatile

UDSM SoCs Characteristics

- **Faster gates allow for systems operating at very high frequencies (multi-GHz)**
- **Smaller transistors result in**
  - **Higher Performance**
  - **Higher Density**
  - **More functionality**
- **Supply voltage and voltage swings reduced to keep power under control**
- **Threshold voltages scaled for high drive current and performance maintenance**

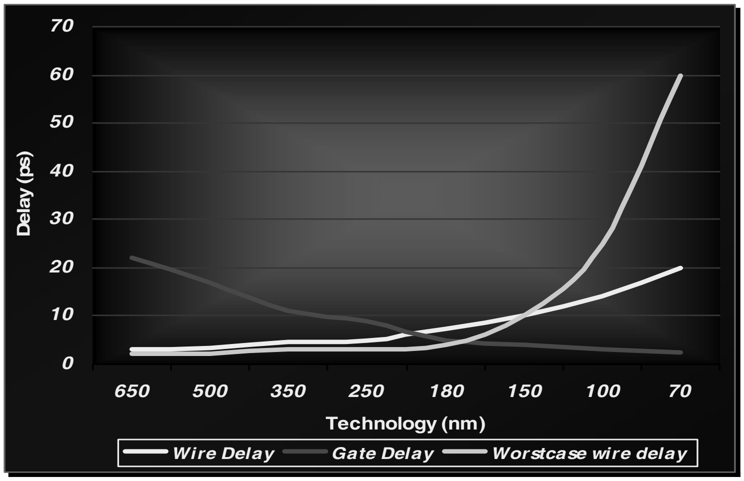
SoC Requirements

- To ensure seamless integration and operation of SoCs, different IPs have to be properly clocked and synchronized with one another
- Design of efficient clock distribution network and synchronization circuitry dictates proper functioning of the SoC

Challenges of SoC design in the UDSM domain

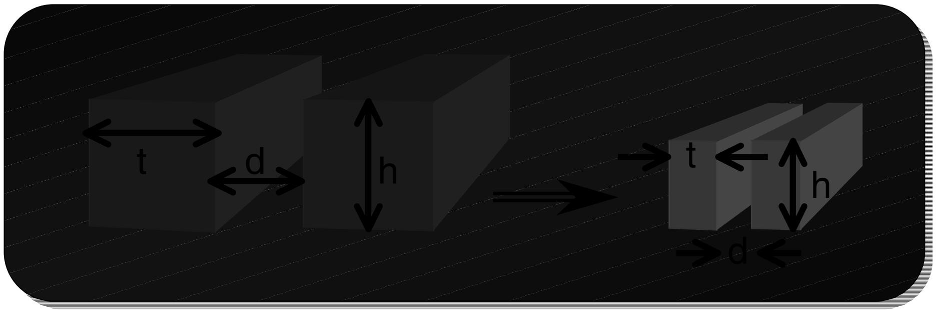
- Higher power consumption due to increasing sub-threshold and gate-oxide leakage currents
- Increasing interconnect delays limiting system throughput
- Synchronization problems due to unpredictable delay variations
- Environmental and process variations-induced timing uncertainties
- Signal Integrity problems due to closer integration and higher frequencies
- Susceptibility to soft-error (single-event upsets)

Scaling of gate and wire delays



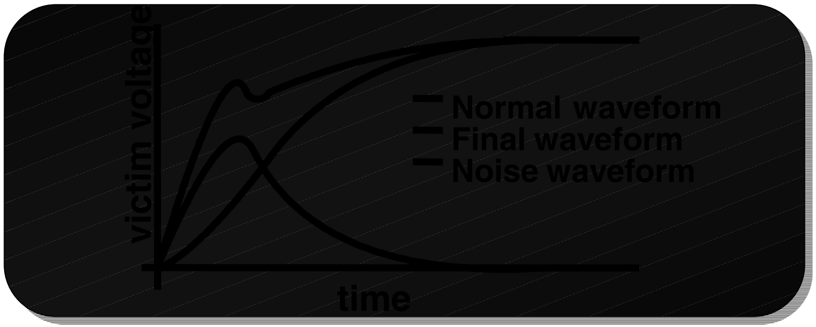
- **Interconnect delays dominate system performance in sub-90nm designs**

Interconnect Scaling



- **Wire resistance *R = ρl/A***
- **With scaling, *t* and *h* ↓ ⇒ *A* ↓ and *R* ↑ ⇒ delay↑**
- **Hence aspect ratio (*h)* ↑ to reduce delay**
- **However, high aspect ratios (~3) and reduced *d* lead to perfect formation of a parallel plate capacitor, inducing crosstalk between wires**

Delay Variations in Interconnects

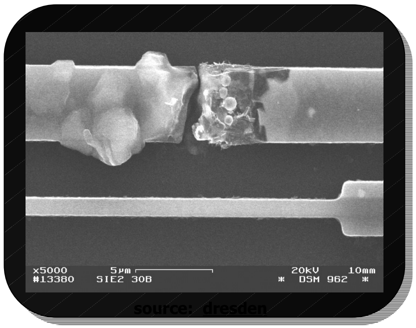


- Crosstalk induces noise voltage in neighboring wires that can aid/degrade signal integrity
- Thus wire delay can vary widely and unpredictably in UDSM technologies
- Techniques that reduce delay and its variations necessary

Signal Integrity - Noise

- In spite of Cu interconnects and low-K devices, higher layer interconnects are thin
- No low-K insulators between wires on same layers and cross-cap percentage is higher
- Hence, crosstalk induced delay variation is a serious challenge in UDSM
- Shielding, buffer insertions & net-reorder can still play crucial roles

Temperature Effects



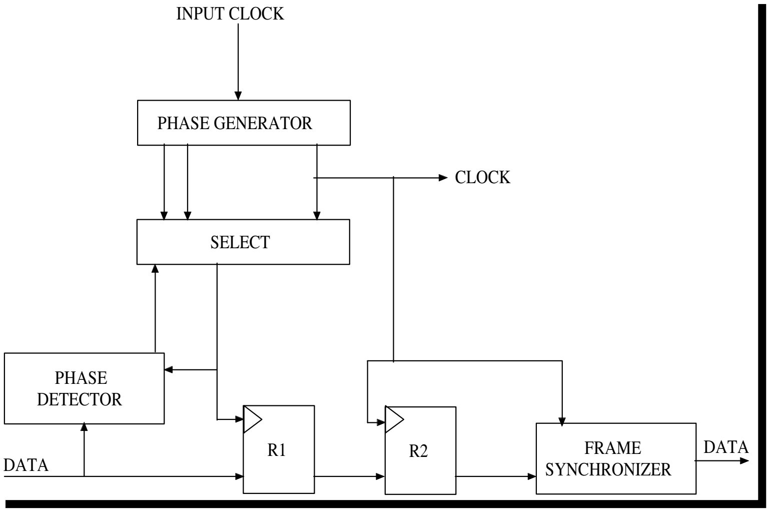
source: dresden

- Electromigration of the interconnects is accelerated due to increased resistance and hence heat generation in the wires
- Using low-k dielectrics for intra-level gap fill can cause significant increase in thermal effects owing to their lower thermal conductivity than SiO_2

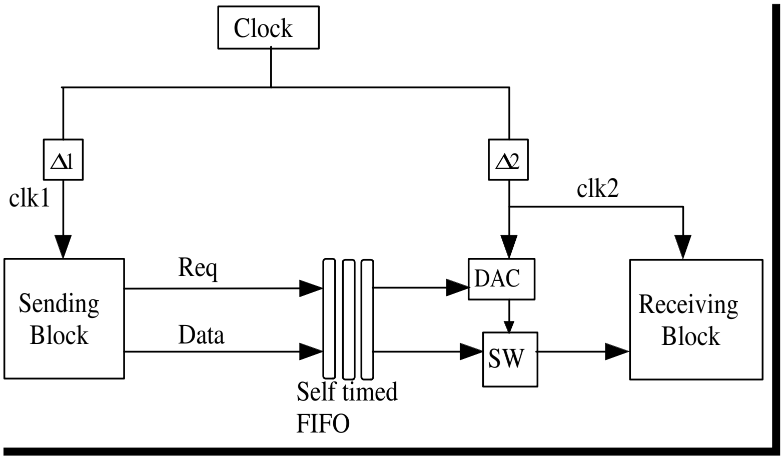
Clocking and Synchronization of IP Cores

- Popular synchronization techniques of the IPs
  - Mesochronous clocking
  - Plesiochronous clocking
- Other schemes such as Pausable clocking are also existent

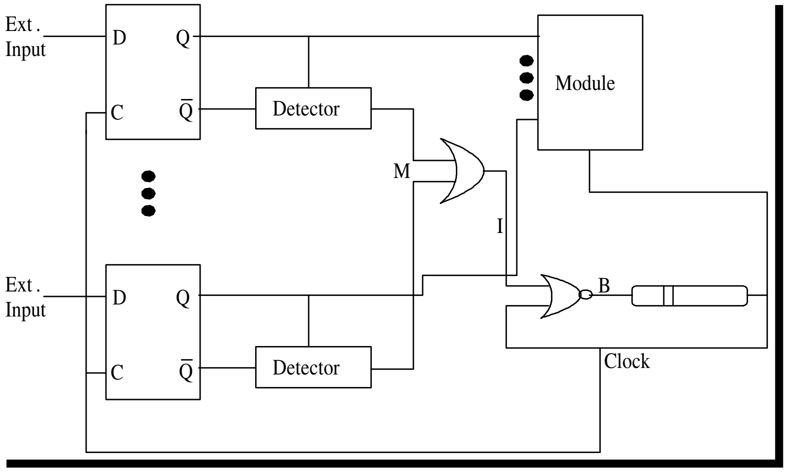
Block Diagram of a Re-synchronizer



Mesochronous Clocking

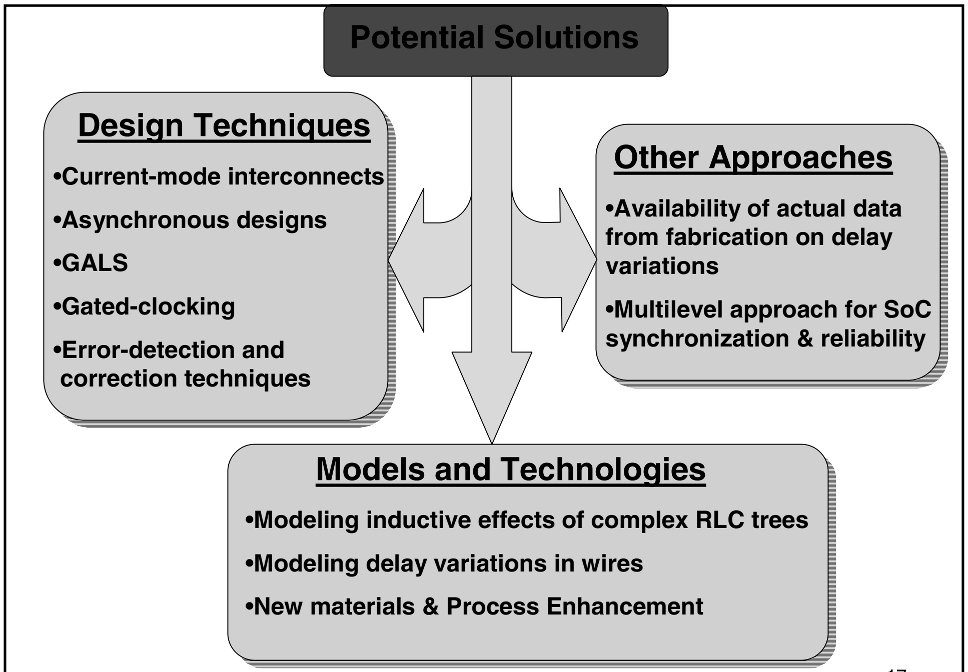


Pausable Clocking



Clocking of different IPs

- Clock signals should reach the respective components at appropriate times for successful operation of SoC
- Different Clock distribution schemes, such as, symmetric H tree, X tree and buffered clock tree schemes used to minimize skew



Potential Solutions

Design Techniques

- Current-mode interconnects
- Asynchronous designs
- GALS
- Gated-clocking
- Error-detection and correction techniques

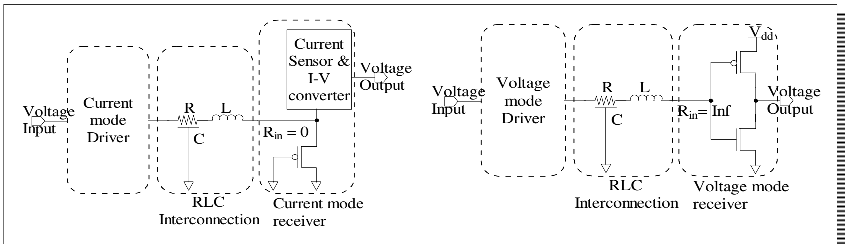
Other Approaches

- Availability of actual data from fabrication on delay variations
- Multilevel approach for SoC synchronization & reliability

Models and Technologies

- Modeling inductive effects of complex RLC trees
- Modeling delay variations in wires
- New materials & Process Enhancement

Current-mode Signaling



Current-mode scheme

Voltage-mode scheme

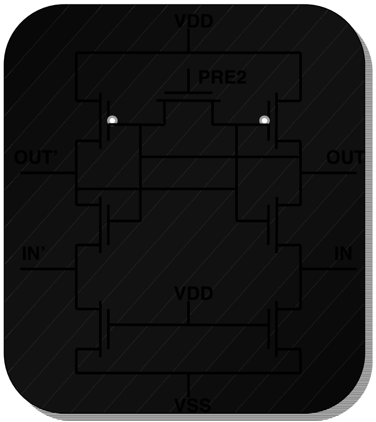
- Current-mode interconnects terminated with low input impedance. Voltage swing reduced on the interconnect
- Due to the short path to ground, voltage in current-mode interconnect need not charge/discharge large wire capacitances in a short time, thus reducing power
- This increases the bandwidth of current-mode scheme
- Since current-mode doesn’t use power rails as reference it is immune to power-supply noise. It is comparatively immune to crosstalk noise

Types of Current-mode Signaling

- Differential and single-ended signaling
  - Single-ended signaling uses single wire per signal thus reducing the area and associated power overhead in differential signaling
  - Differential signaling transmits a differential current per signal (using 2 wires)
- Noise immunity of differential signaling significantly high compared to single-ended

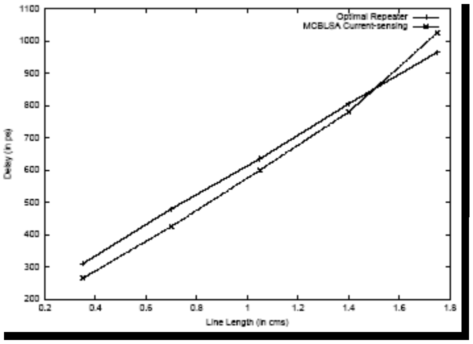
Differential method implementation

- MCBLSA interconnect delay is less compared to optimal repeater insertion (avg. 6%)
- Power consumed less than that in optimal repeater insertion
- Power consumption remains almost constant with increase in wire length



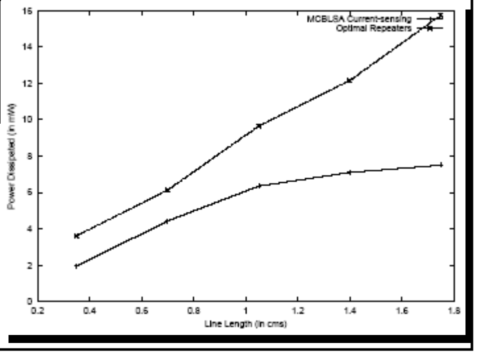
Modified Clamped Bit Line Sense Amplifier (MCBLSA) scheme, A. Maheswari *et.al., 2001*

Differential method implementation



Delay of optimal repeater method vs. MCBLSA scheme

Power consumed in optimal repeater method vs. MCBLSA scheme

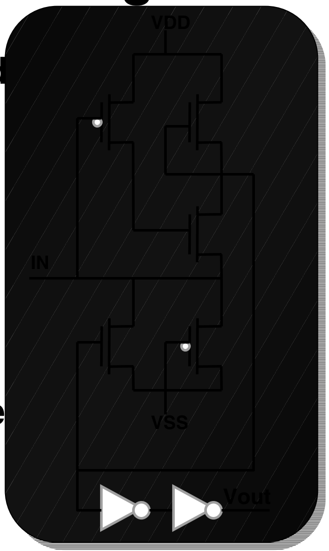


Differential method - disadvantages

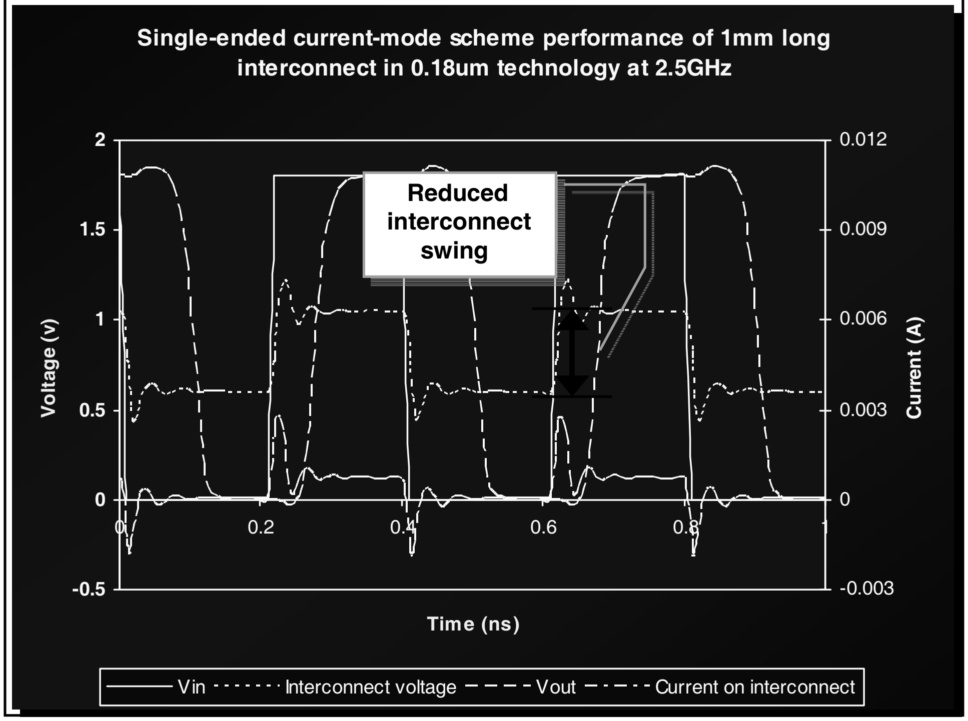
- Area overhead: Consumes almost double the routing area due to additional wire per signal
- Scope for reduction in power consumed if the additional wire is eliminated
- Single-ended design overcomes these disadvantages at the expense of noise immunity

Single-ended signaling

- Less area and routing overhead compared to MCBLSA scheme
- Delay and power consumption improved by 11% over the MCBLSA scheme in 0.18µm technology
- Delay variations due to cross-talk increased from 28% in MCBLSA to 32% in this scheme
- This scheme could operate successfully up to 10GHz in 0.18µm technology



Single-ended current sensing circuit, R. Sridhar *et.al.*, 2004

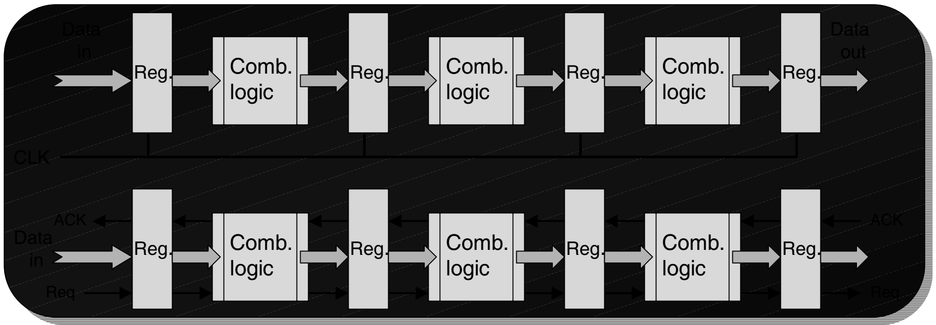


Current-mode signaling - Summary

- Current-mode mostly used in inter-chip signaling. On-chip use limited to CMOS SRAM circuits
- However its favorable characteristics in terms of bandwidth, noise immunity and power consumption makes it a good choice for on-chip global interconnects
- Voltage-mode circuits expected to saturate in the 5-10GHz domain. Current-mode circuits provide a promising alternative in this regard

25

Asynchronous Circuits



- Explicit local synchronization between blocks
- Provides modularity for SoC design – plug and play compatibility
- Robust with regard to wire delay, temperature and process variations
- Helps reduce power supply noise by reducing current peaks around clock edges

26

Asynchronous Design – Hurdles

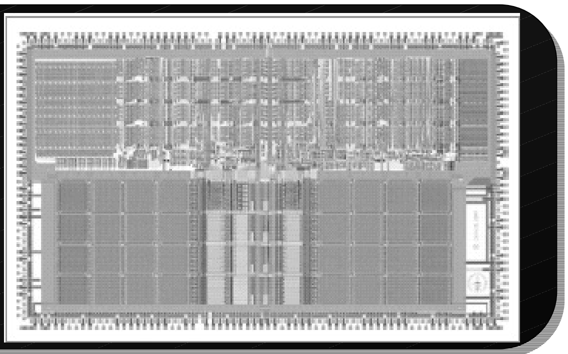
- EDA tools
  - Capable of complex timing analysis
- Difficult to test

Asynchronous Implementations

- Amulet Microprocessors (asynchronous ARM, Univ. of Manchester, 2000)
- Caltech Asynchronous Microprocessors (asynchronous MIPS, 1998)
- Titac2 (Univ. of Tokyo, 1997)
- Intel’s RAPPID Instruction Length Decoder (2001)

The miniMIPS processor

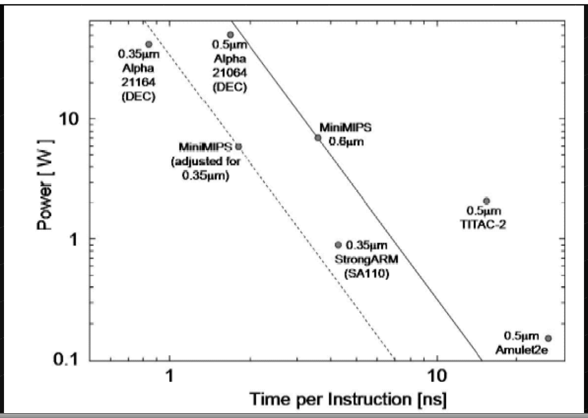
Asynchronous version of MIPS microprocessor



A. J. Martin *et.al.*
Caltech group, 1998

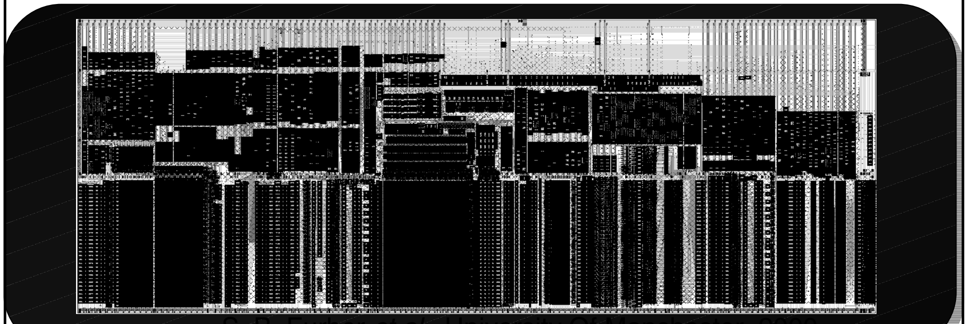
- Fully asynchronous 32-bit RISC µP similar to MIPS R3000
- Implements most of MIPS-I ISA
- Has two 4-KB on-chip caches, an instruction cache and a direct-mapped write-through data cache

miniMIPS – Results



- 180MIPS, 4W @ 3.3v; 100MIPS, 850mW @ 2v; 60MIPS, 220mW at 1.5v
- Fabricated chip reported to be up to 4 times faster than comparable synchronous commercial µPs

Amulet3 - Results



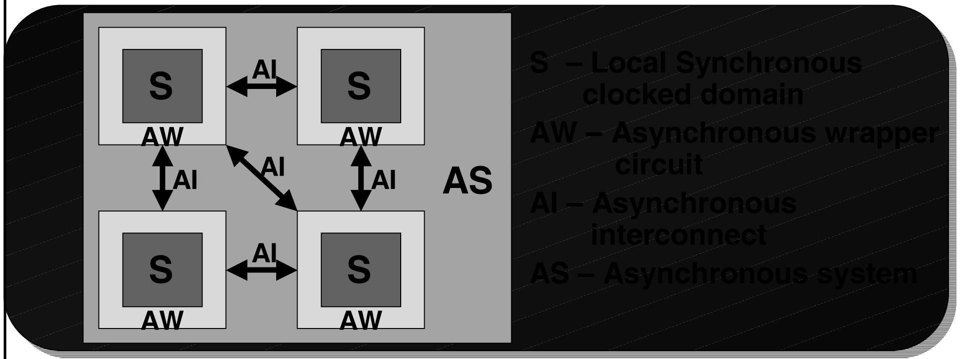
S. B. Furber *et.al.*, University Of Manchester, 2000

- ARM9TDMI is the synchronous implementation closest to Amulet3
- ARM9 – operates up to 120MHz with 1.1MIPS/MHz, 1.8mW/MHz ⇒ energy per instruction 610 MIPS/W
- Amulet3 – 220mW at 85 Dhrystone MIPS, energy per instruction 620 MIPS/W

Summary of Asynchronous Design

- Asynchronous design provides a promising alternative to clocking and synchronization problems of synchronous designs
- Offers significantly improved performance in terms of throughput and performance while maintaining minimal power overhead

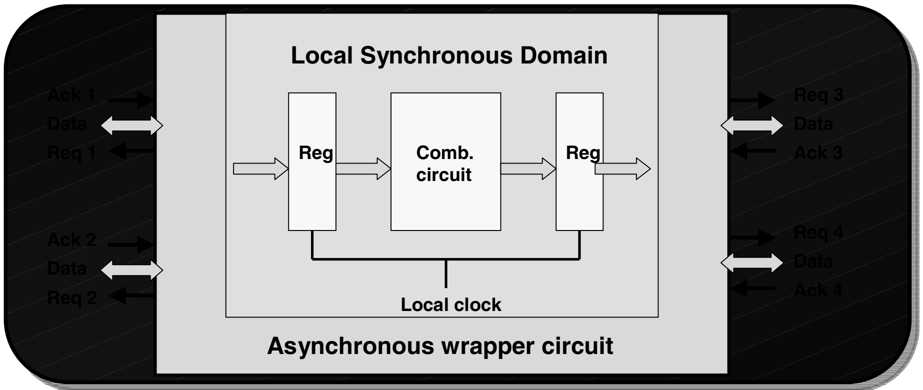
Globally Asynchronous Locally Synchronous (GALS) Systems



S – Local Synchronous clocked domain
AW – Asynchronous wrapper circuit
AI – Asynchronous interconnect
AS – Asynchronous system

- Large asynchronous design difficult to realize
- GALS combines advantages of synchronous and asynchronous designs

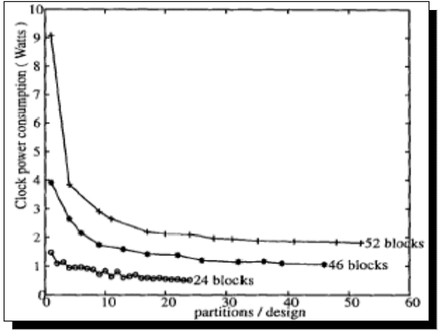
Globally Asynchronous Locally Synchronous (GALS) Systems



- GALS immune to global clock skew which degrades performance significantly in synchronous systems
- Tolerant to global wire delays and its variations

GALS Implementation - Results

- The partitioning of GALS is an important factor that affects final system peformance
- Some important observations of GALS implementation by J. Oberg *et.al.*
- Partitioning the system into synchronous (SB) and asynchronous blocks plays a significant role in determining power and performance
- Allows the SB s to run at different clock speeds and still be synchronized
- Results show upto 70% reduction in power
- Overheads in GALS increases with number of partitions



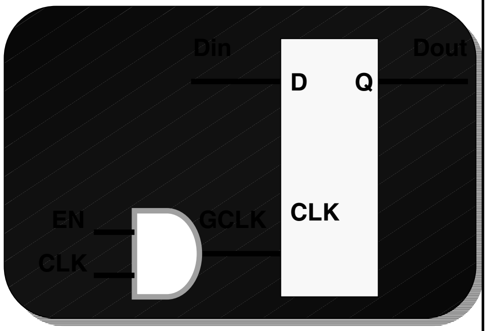
GALS Implementation - Results

- Some important observations of GALS implementation by D. Marculescu *et.al.*
- A 5-clock domain GALS processor was implemented in 0.25µm technology
- Power consumption in GALS 10% lower compared to synchronous equivalent
- Drop in performance ranging between 5-15%

36

Gated Clocking

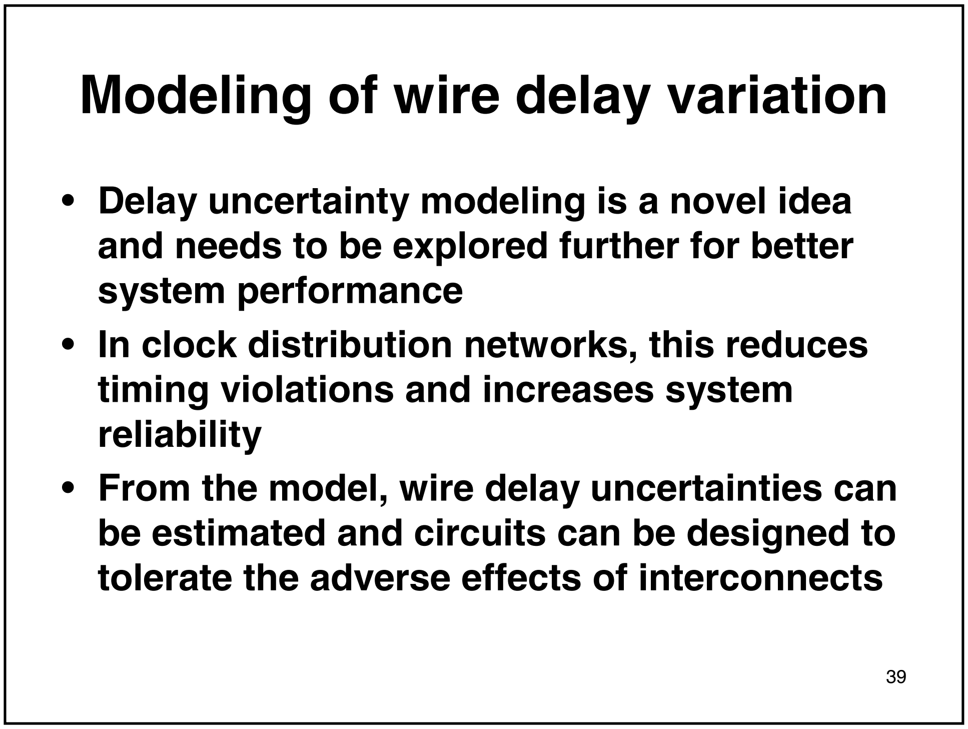
- The clock input to a synchronous block is turned off when inactive
- This saves power dissipated in the block when inactive



- Clock combined with other signals using combinational logic to control clock input to a synchronous block
- This strategy fits well in particular for SoCs to reduce power consumption

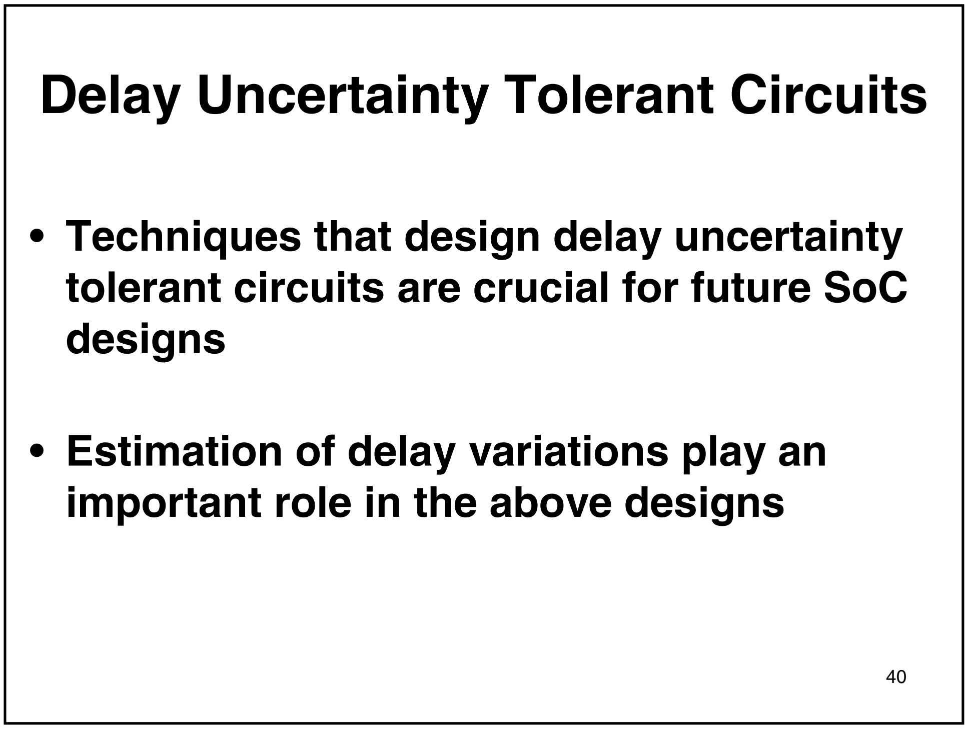
Gated Clocking …

- The effectiveness of this scheme depends on the frequency of operation of the various synchronous blocks
- The power saved in gated-clocked circuits is 30% lower than buffered clock circuits on an average
- The disadvantages of this scheme include
  - Area overhead
  - Requires complex timing analysis tools
  - Introduces additional clock skew due to the combinational elements



Modeling of wire delay variation

- Delay uncertainty modeling is a novel idea and needs to be explored further for better system performance
- In clock distribution networks, this reduces timing violations and increases system reliability
- From the model, wire delay uncertainties can be estimated and circuits can be designed to tolerate the adverse effects of interconnects



Delay Uncertainty Tolerant Circuits

- Techniques that design delay uncertainty tolerant circuits are crucial for future SoC designs
- Estimation of delay variations play an important role in the above designs

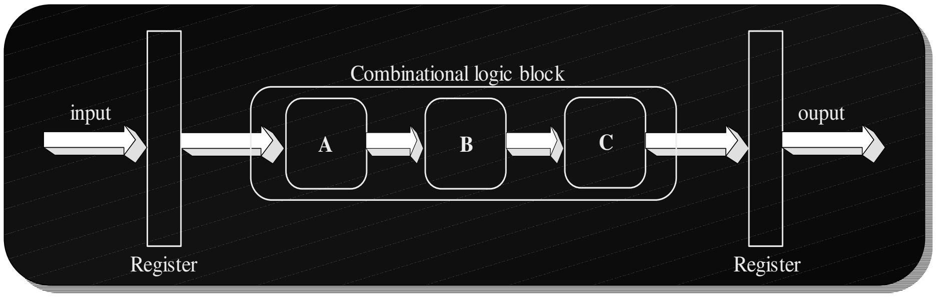
Delay Uncertainty Tolerant Circuits

- Specialized circuits and techniques to counteract the UDSM impacts are necessary
  - Wave pipelining and other circuit level asynchronous and pipelining techniques provide an insight into this problem
  - Incorporating self-checks and some tolerance circuitry
- Validation on adders and multipliers

Introduction to Wave Pipelining

- WP is a design method that uses gate capacitance to hold information between successive stages
- Allows multiple sets of data to coexist, and thus increases the throughput
- Boosts the throughput of a system without additional registers

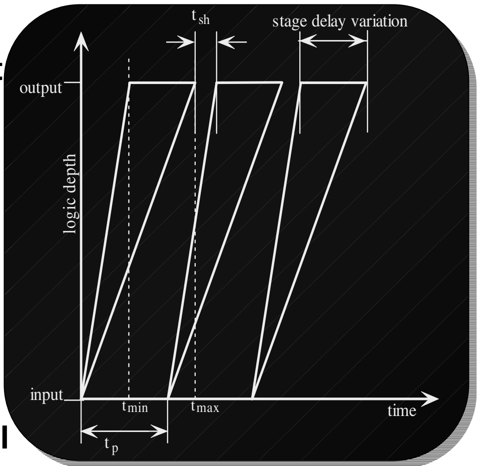
Clocking Wave pipelined circuits



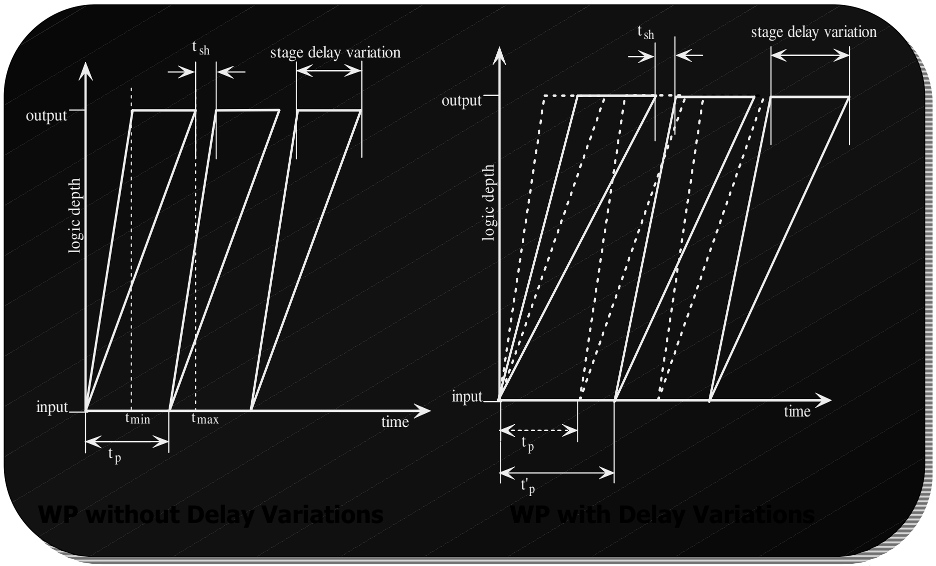
- Input applied to *A* at t=0
- Output of *A* appears at t=t_A
- Similarly for *B* and *C* – t_B & t_C respectively

Wave Pipelining – Timing Constraints

- First input is applied to *A* @ time t=0 and held for at least t_{SH}
- Delays of *A* - t_{min} & t_{max}
- New inputs applied at *A* such that new outputs appear not before t_{max}
- Hence new inputs at *A* should be applied only after time $t_p > (t_{max} – t_{mix}) + t_{SH} + \Delta$ where Δ is any unconditional clock skew



How Delay Variations affect WP



WP without Delay Variations

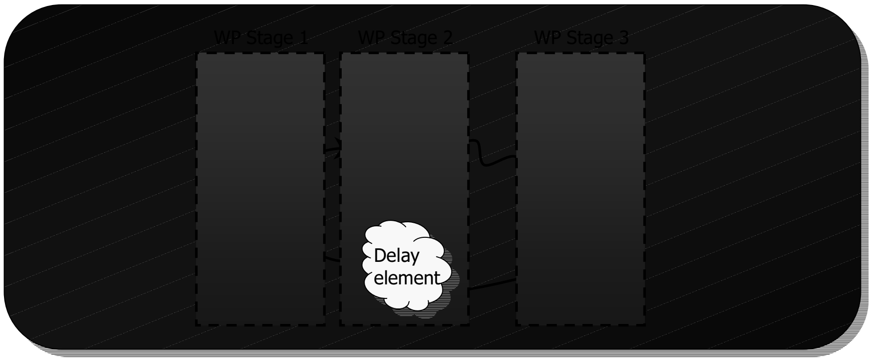
WP with Delay Variations

Wave Pipelining Design Principle

- The basic design of WP involves equalizing delays between successive stages

- The cause of the delay variations could be:
  - Difference in propagation paths
  - Data dependent delays
  - Temperature and process variations
  - Power supply drift and noise
  - Signal noise such as crosstalk

Difference in Propagation Path

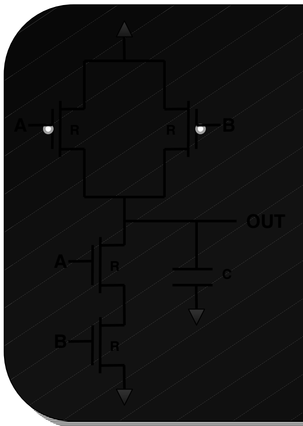


- Can be adjusted by inserting delay elements and gate sizing

Temperature and Process Variations

- In WP, incorporating Nowka's delay variation model results in higher number of data waves as compared to the fixed frequency clocking
- This is achieved by generating clock signal whose frequency depends on the delay in the logic network
- Such a design takes into consideration the process, temperature and any local variations that affect delay

Data Dependent Delays



Input A B	Delay (τ)
0 0	(R/2)*C
0 1	(R)*C
1 0	(R)*C
1 1	(2R)*C

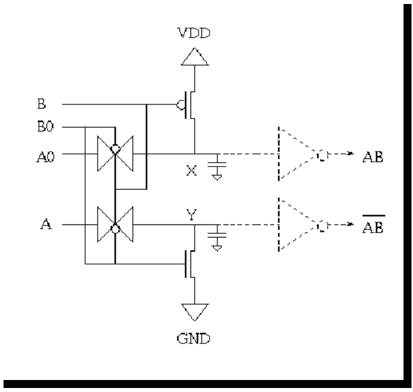
- Data dependent delays can be compensated by appropriate choice of design style

Logic Design Styles

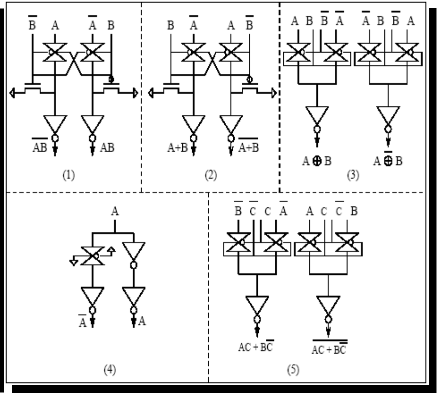
- A suitable logic design style should have the following features:
  - Equal rise and fall times
  - Gate delay independent of current/previous input patterns
  - Gate speed adjustment with predictable effects
  - High noise immunity, low power, high speed

Wave-pipelined Transmission Gate Logic

-



WTGL Nand Cell


$\overline{B}$ A $\overline{A}$ B
$\overline{AB}$
AB
(1)
B $\overline{A}$ A $\overline{B}$
A+B
$\overline{A+B}$
(2)
A B $\overline{B}$ $\overline{A}$ $\overline{A}$ B $\overline{B}$ A
$A \oplus B$
$A \overline{\oplus} B$
(3)
A
$\overline{A}$
A
(4)
$\overline{B}$ $\overline{C}$ C $\overline{A}$ A C $\overline{C}$ B
$AC + B\overline{C}$
$\overline{AC + B\overline{C}}$
(5)

WTGL Library

51

Advantages of WTGL Cells

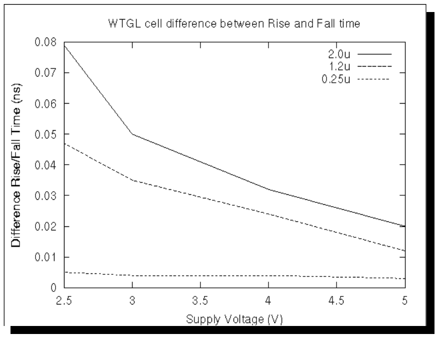
- Dual-rail logic – less prone to noise and error
- Has higher noise immunity
- Produces rail-to-rail full swing
- Delay easily adjusted by varying output inverter sizes
- Every WTGL has equal delay hence balancing easier
- High throughput and power efficient

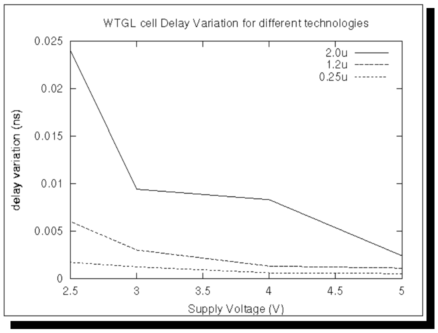
52

WTGL in UDSM Technologies

- WTGL has linear delay variation as a function of supply voltage
  - Reducing supply voltage results in quadratic power saving
- Rise/fall time also same in different supply voltage in deep sub-micron
- Output inverter in WTGL can be sized to have more current driving capability and do fine tuning
  - In bigger scale, size of the inverter has large effect on power dissipation

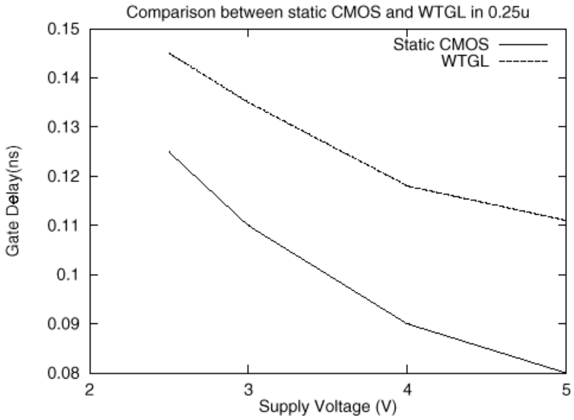
WTGL in UDSM Technologies





•Performance of WTGL cells is seen to improve as we scale down technology

WTGL in UDSM Technologies



55

Comparison of Logic Styles

Logic Style	T_{HL}(ps)	T_{LH}(ps)	T_{fall}(ps)	T_{rise}(ps)	Power in 2um (uW)	Power in 0.25um (uW)
Balanced NAND	**100**	**126**	**230**	**256**	**89**	**7.5**
NPCPL	**78**	**99**	**NA**	**NA**	**266**	**33**
WTGL	**109**	**99**	**70**	**83**	**250**	**8.9**
Cross coupled	**64**	**64**	**95**	**85**	**780**	**140**

56

Summary of Different Logic Styles

- Logic Styles compared.
  - Balanced NAND: Good power characteristics
  - WTGL: Faster with less delay variation.
- Logic styles not considered
  - NPCPL: Reduced output voltage swing
  - Cross Coupled: Large power dissipation

Implementation Results for a 4:2 Compressor

	Delay	Delay Variation	Power (mW)
Balanced NAND w/interconnect effect	0.98ns	0.032ns	1.72
Balanced NAND w/o interconnect effect	0.75ns	0.04ns	1.33
WTGL w/ interconnect effect	0.19ns	0.01ns	1.27
WTGL w/o interconnect effect	0.16ns	0.018ns	1.15

Results from WP Multiplier circuit

- Multiplier has 7.8ns propagation delay and delay variation is 1.32ns.
- Best clock cycle is 2.3ns.
- Wave pipelined multiplier is 3.5 times faster than non pipelined multiplier.
- Conventional pipelined multiplier which has 5 stage has 6.4 ns propagation delay.
  - Each stage is 1.5n and clocking overhead is 0.9ns.
  - Clock period of conventional pipelined multiplier is 3.15ns.
  - Wave pipelined multiplier is 35% faster than conventional multiplier.

Results from WP Adder circuit

- A 4-bit Brent-Kung WP adder was laid out in 180nm and simulated
- Simulations showed a latency of 1.9ns with delay variation of 12%
- The layout was enhanced using the above techniques and the new latency was 1.83ns with a delay variation of 6%

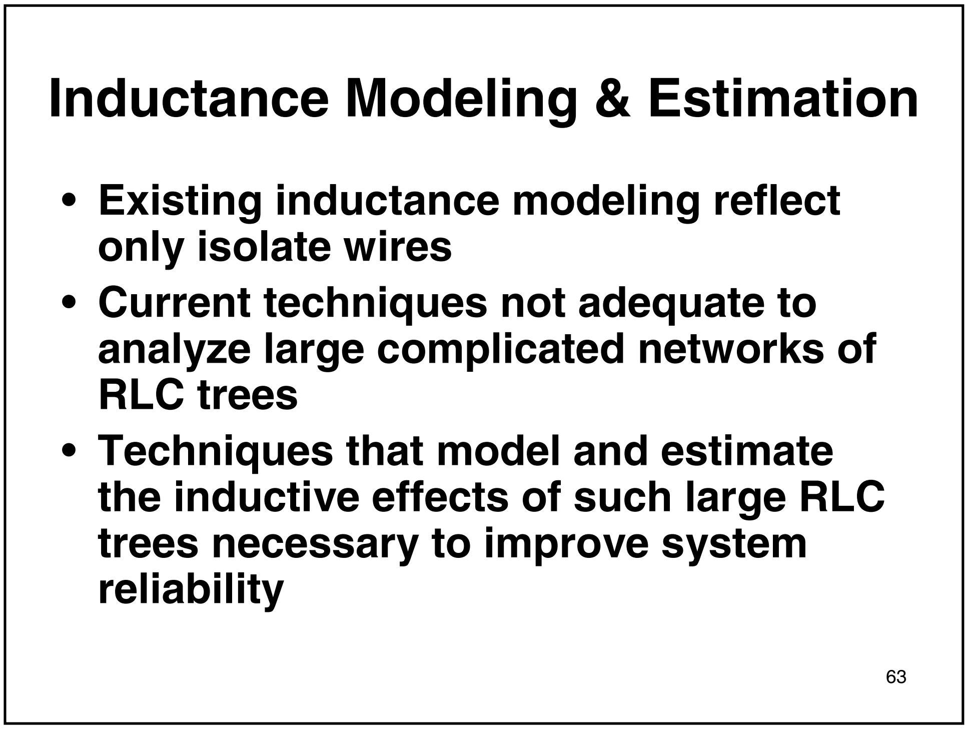
Wave Pipelining Summary

- WP is a technique that can achieve maximum clock cycle without using additional registers.
  - In deep sub-micron design, it helps to reduce clock skew and power dissipation by clock distribution.
- For WP, WTGL cell is proper logic style in deep sub-micron design.
  - Less input dependent delay variation.
  - Less power consumption.
  - Same rise and fall time.
  - Functionality.
  - Less affected by temperature variation.

WP and SoC design

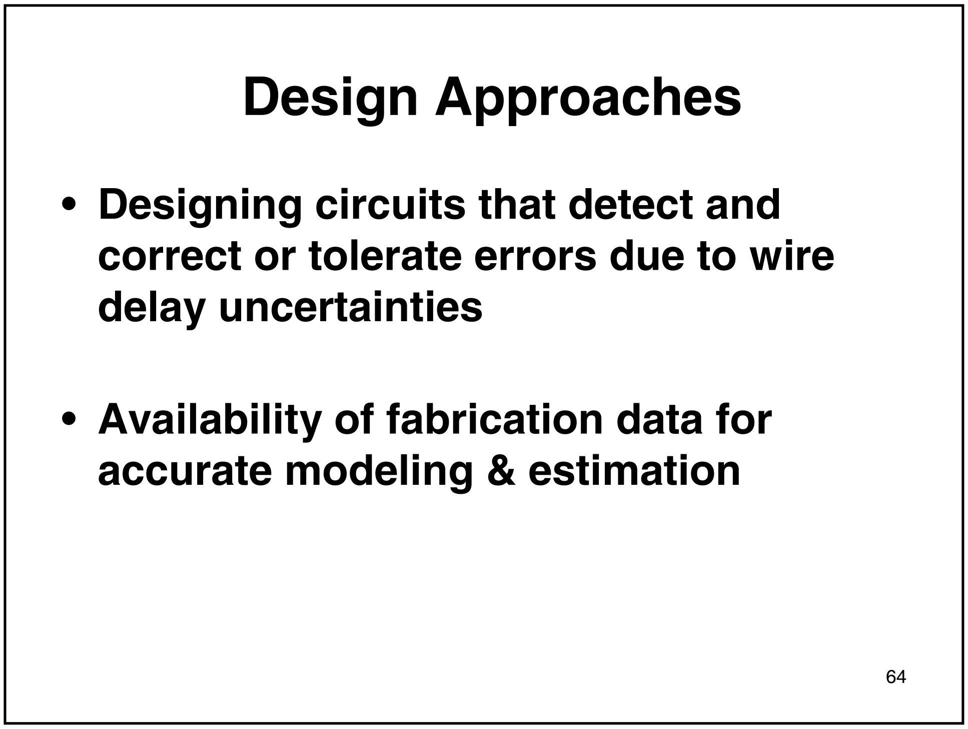
- The ideas learnt from WP technique in terms of minimizing delay variations could be applied to different IP cores of SoCs for minimum delay variation between the components

- Helps in improving reliability and performance of the overall system



Inductance Modeling & Estimation

- Existing inductance modeling reflect only isolate wires
- Current techniques not adequate to analyze large complicated networks of RLC trees
- Techniques that model and estimate the inductive effects of such large RLC trees necessary to improve system reliability



Design Approaches

- Designing circuits that detect and correct or tolerate errors due to wire delay uncertainties
- Availability of fabrication data for accurate modeling & estimation

Design Approaches…

- Network-on-Chips (NoC) by Luca Benini and DeMichelli's group
  - Attempts to model the SoC into different layers similar to the network OSI model
  - The interconnects form the lowest layer – the physical layer
  - One of ideas behind this concept is routing data rather than routing wires

Development of New Materials

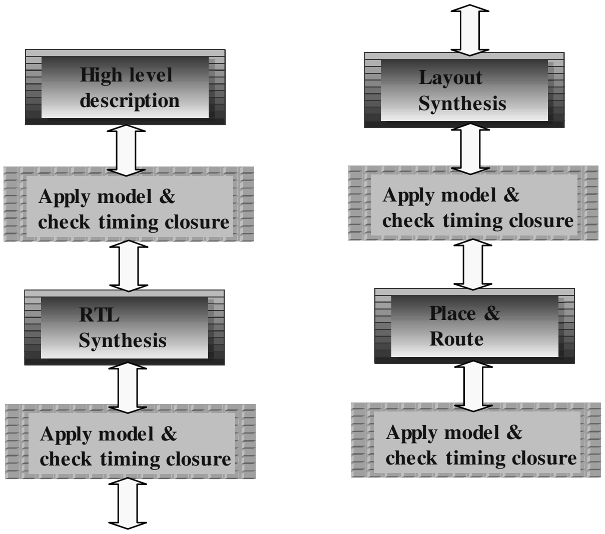
- In addition to design approaches, new materials and processes that overcome the problems of inductance, noise and delay uncertainties need to be explored

- This helps in conforming to Moore's Law for a longer time

Using a Multi-level Approach

- A consorted effort from all design perspectives should be adopted to achieve reliable performance in UDSM SoC designs
- A multi level approach that applies the model developed to check the inductive and delay unpredictability effects should be followed at each level of design flow

SoC Design Flow



High level description

Apply model & check timing closure

RTL Synthesis

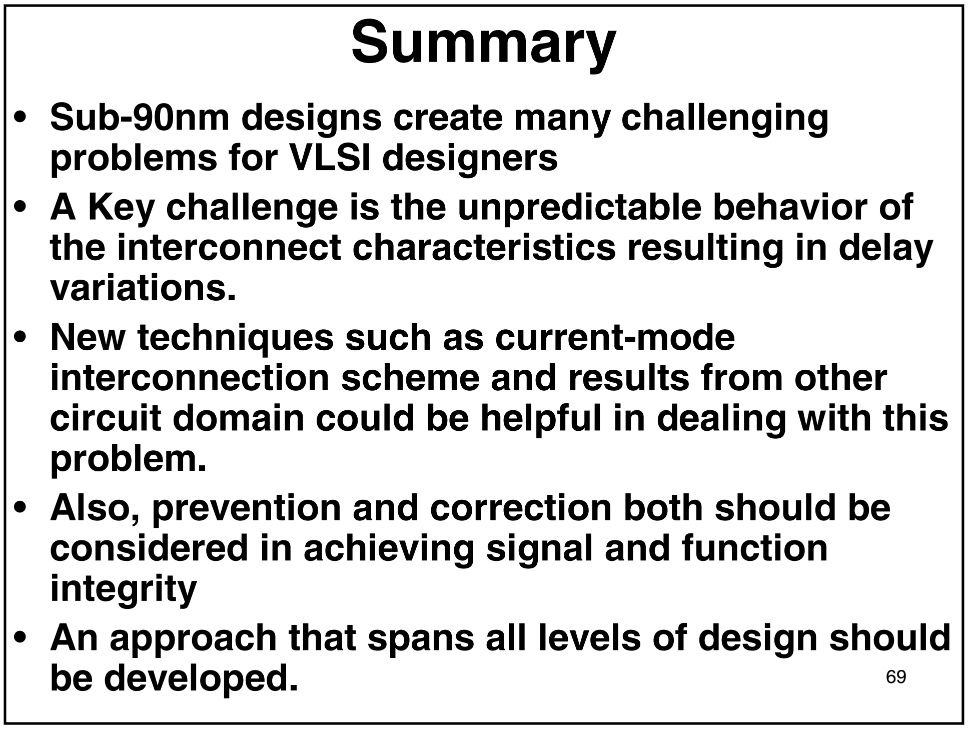
Apply model & check timing closure

Layout Synthesis

Apply model & check timing closure

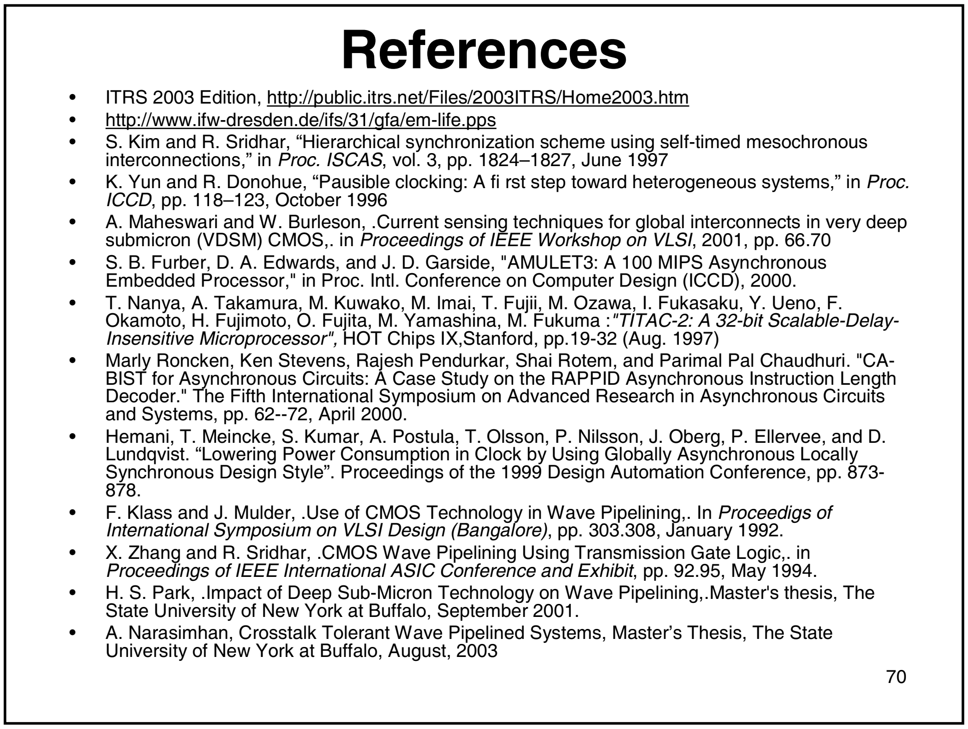
Place & Route

Apply model & check timing closure



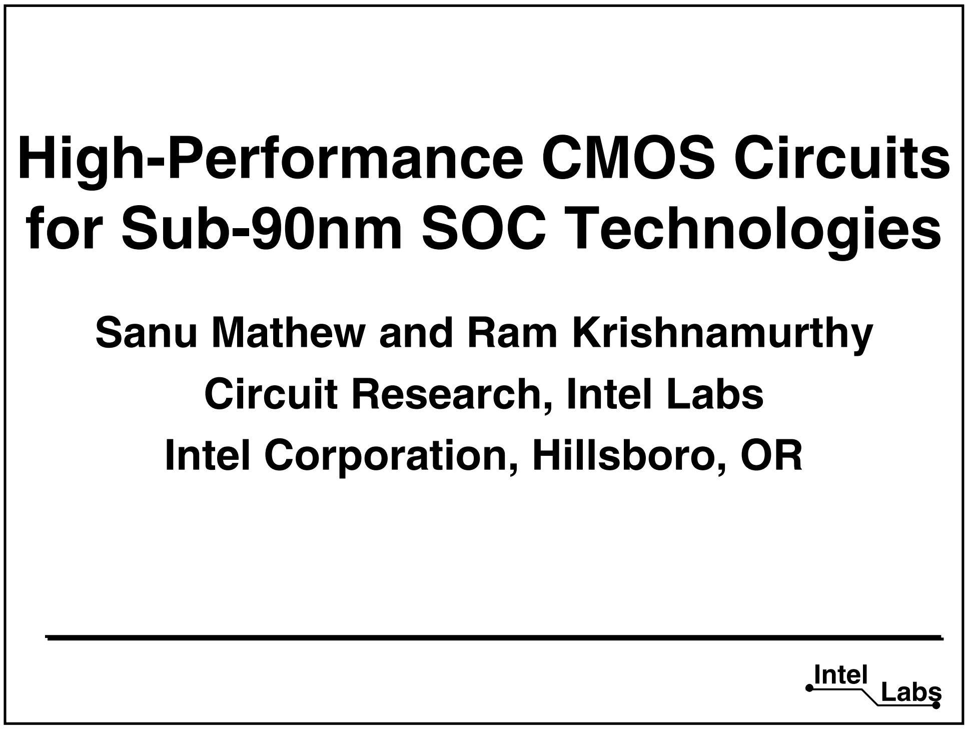
Summary

- Sub-90nm designs create many challenging problems for VLSI designers
- A Key challenge is the unpredictable behavior of the interconnect characteristics resulting in delay variations.
- New techniques such as current-mode interconnection scheme and results from other circuit domain could be helpful in dealing with this problem.
- Also, prevention and correction both should be considered in achieving signal and function integrity
- An approach that spans all levels of design should be developed.



References

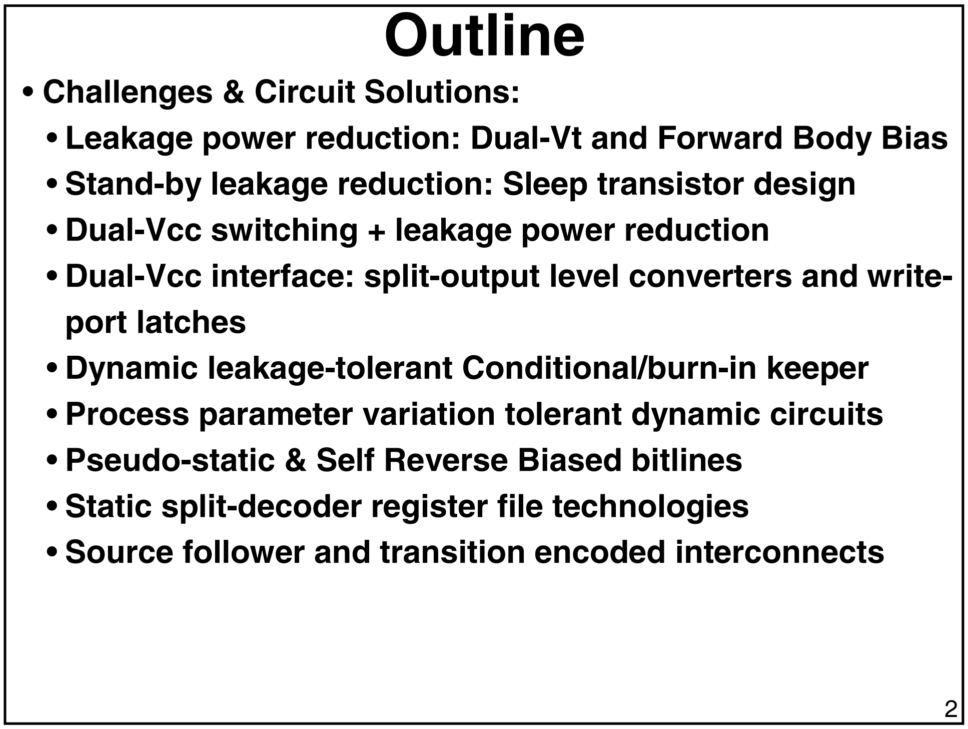
- ITRS 2003 Edition, http://public.itrs.net/Files/2003ITRS/Home2003.htm
- http://www.ifw-dresden.de/ifs/31/gfa/em-life.pps
- S. Kim and R. Sridhar, “Hierarchical synchronization scheme using self-timed mesochronous interconnections,” in *Proc. ISCAS*, vol. 3, pp. 1824–1827, June 1997
- K. Yun and R. Donohue, “Pausible clocking: A fi rst step toward heterogeneous systems,” in *Proc. ICCD*, pp. 118–123, October 1996
- A. Maheswari and W. Burleson, .Current sensing techniques for global interconnects in very deep submicron (VDSM) CMOS,. in *Proceedings of IEEE Workshop on VLSI*, 2001, pp. 66.70
- S. B. Furber, D. A. Edwards, and J. D. Garside, "AMULET3: A 100 MIPS Asynchronous Embedded Processor," in Proc. Intl. Conference on Computer Design (ICCD), 2000.
- T. Nanya, A. Takamura, M. Kuwako, M. Imai, T. Fujii, M. Ozawa, I. Fukasaku, Y. Ueno, F. Okamoto, H. Fujimoto, O. Fujita, M. Yamashina, M. Fukuma :*"TITAC-2: A 32-bit Scalable-Delay-Insensitive Microprocessor",* HOT Chips IX,Stanford, pp.19-32 (Aug. 1997)
- Marly Roncken, Ken Stevens, Rajesh Pendurkar, Shai Rotem, and Parimal Pal Chaudhuri. "CA-BIST for Asynchronous Circuits: A Case Study on the RAPPID Asynchronous Instruction Length Decoder." The Fifth International Symposium on Advanced Research in Asynchronous Circuits and Systems, pp. 62--72, April 2000.
- Hemani, T. Meincke, S. Kumar, A. Postula, T. Olsson, P. Nilsson, J. Oberg, P. Ellervee, and D. Lundqvist. “Lowering Power Consumption in Clock by Using Globally Asynchronous Locally Synchronous Design Style”. Proceedings of the 1999 Design Automation Conference, pp. 873-878.
- F. Klass and J. Mulder, .Use of CMOS Technology in Wave Pipelining,. In *Proceedigs of International Symposium on VLSI Design (Bangalore)*, pp. 303.308, January 1992.
- X. Zhang and R. Sridhar, .CMOS Wave Pipelining Using Transmission Gate Logic,. in *Proceedings of IEEE International ASIC Conference and Exhibit*, pp. 92.95, May 1994.
- H. S. Park, .Impact of Deep Sub-Micron Technology on Wave Pipelining,.Master's thesis, The State University of New York at Buffalo, September 2001.
- A. Narasimhan, Crosstalk Tolerant Wave Pipelined Systems, Master’s Thesis, The State University of New York at Buffalo, August, 2003



High-Performance CMOS Circuits for Sub-90nm SOC Technologies

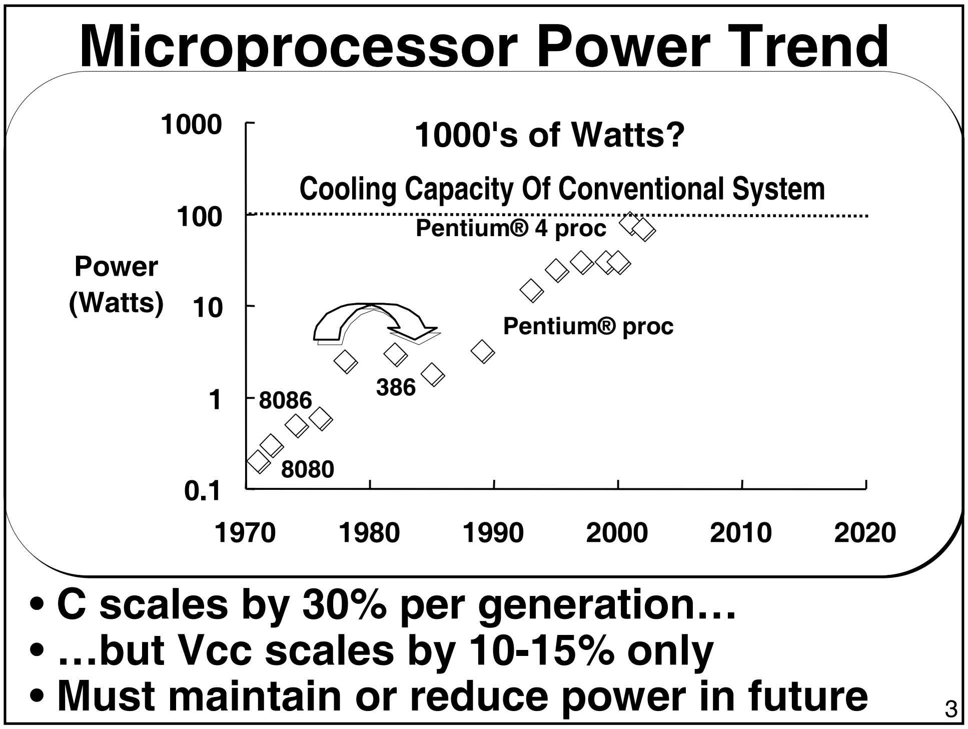
Sanu Mathew and Ram Krishnamurthy
Circuit Research, Intel Labs
Intel Corporation, Hillsboro, OR

Intel Labs



Outline

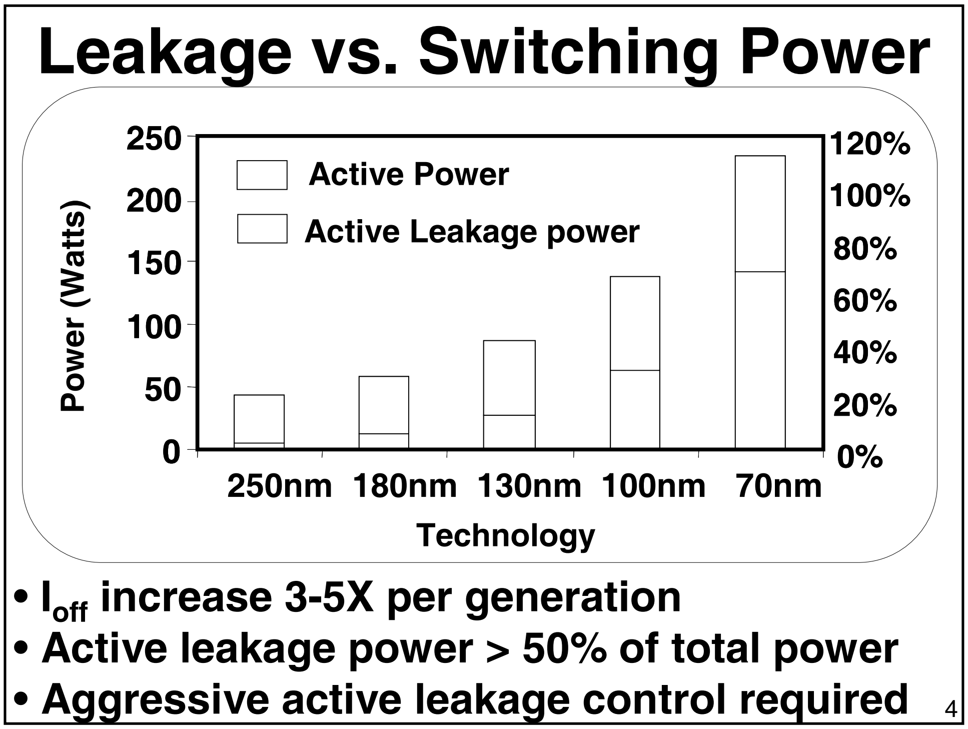
- Challenges & Circuit Solutions:
  - Leakage power reduction: Dual-Vt and Forward Body Bias
  - Stand-by leakage reduction: Sleep transistor design
  - Dual-Vcc switching + leakage power reduction
  - Dual-Vcc interface: split-output level converters and write-port latches
  - Dynamic leakage-tolerant Conditional/burn-in keeper
  - Process parameter variation tolerant dynamic circuits
  - Pseudo-static & Self Reverse Biased bitlines
  - Static split-decoder register file technologies
  - Source follower and transition encoded interconnects



Microprocessor Power Trend

- C scales by 30% per generation…
- …but Vcc scales by 10-15% only
- Must maintain or reduce power in future

3

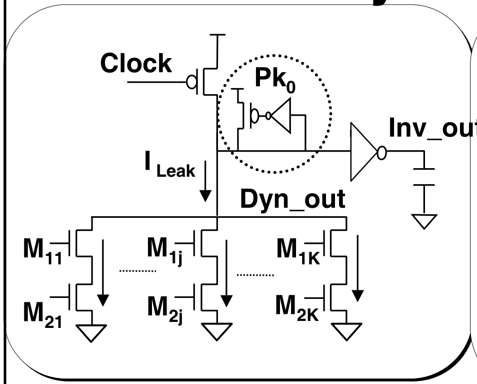


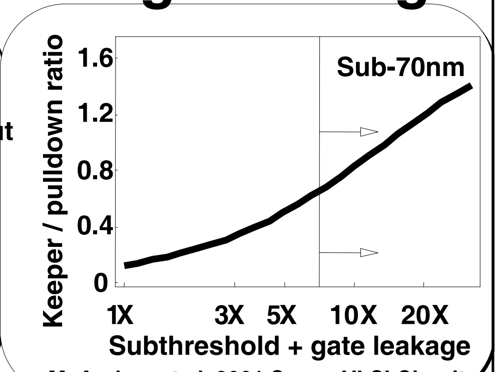
Leakage vs. Switching Power

- I_{off} increase 3-5X per generation
- Active leakage power > 50% of total power
- Aggressive active leakage control required

4

Functionality with High Leakage



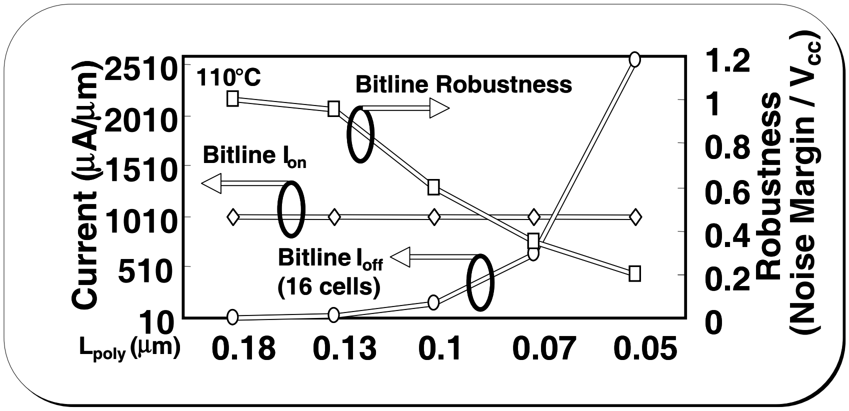


M. Anders et al, 2001 Symp. VLSI Circuits

Sub-70nm Dynamic Circuit Active Leakage Tolerance:

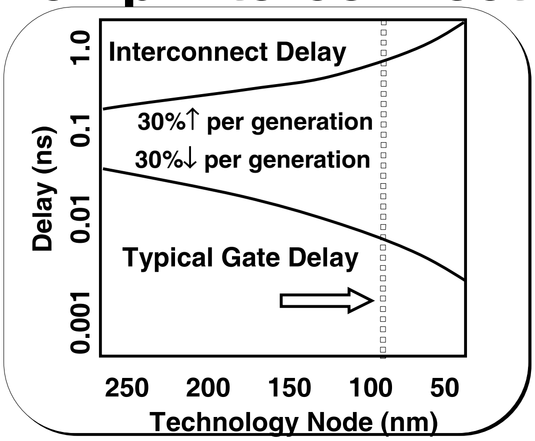
- Cache, RF, Arrays, Bitlines most affected
- Keeper sizes > 50% of pulldown strength
- High contention ⇒ degraded performance
- Slow keeper shutoff ⇒ high short-circuit power

Bitline Leakage Tolerance



- Bitline I_{on}/I_{off}: 60% ↓ per generation
- Leakage tolerant bitline techniques required

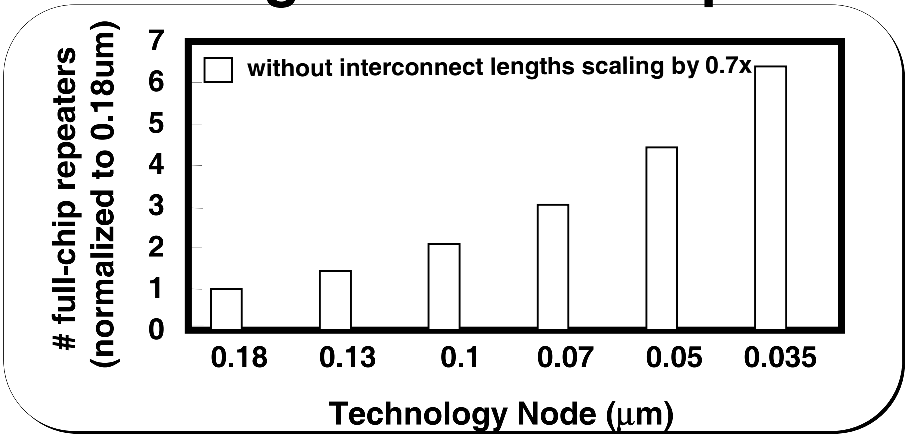
On-chip Interconnect RC



- RC/μm increases 40-60% per generation
- Local inter-gate wires dominate critical-path delays
- Global wire lengths not scaling by 0.7x
- Copper, low-K ILD: modest benefit

7

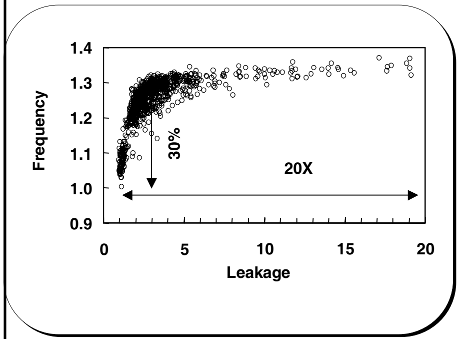
Increasing Number of Repeaters

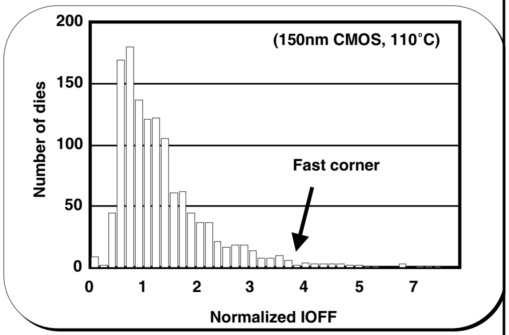


- RC/μm is only one side of the story…
- Finer pipeline global buses ⇒ more flop-repeaters
- Exponential increase in bus repeaters aggravates power problem

8

Process Parameter Variation Tolerance





- Significant variation in IOFF (hence F_{max} spread)
- Worsening with process scaling
- Excess leakage dies: lack in robustness
- Low leakage dies: over-designed for robustness

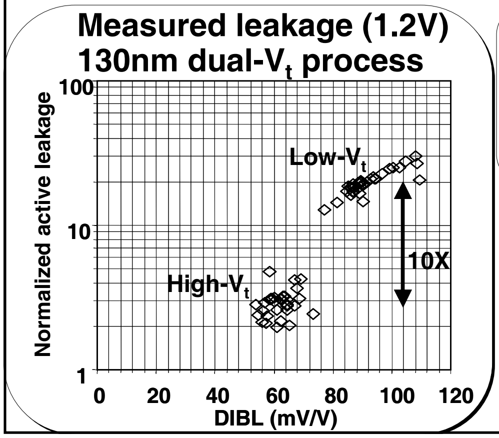
Process parameter variation tolerant circuit techniques

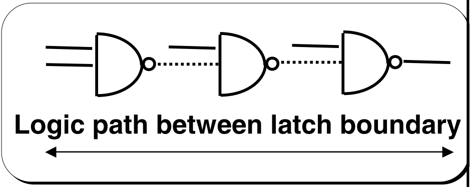
Outline

- Challenges & Circuit Solutions:
  - Leakage power reduction: Dual-Vt and Forward Body Bias
  - Stand-by leakage reduction: Sleep transistor design
  - Dual-Vcc switching + leakage power reduction
  - Dual-Vcc interface: split-output level converters and write-port latches
  - Dynamic leakage-tolerant Conditional/burn-in keeper
  - Process parameter variation tolerant dynamic circuits
  - Pseudo-static & Self Reverse Biased bitlines
  - Static split-decoder register file technologies
  - Source follower and transition encoded interconnects

Active Leakage Reduction: Dual V_t design

- Motivation: Exploit two V_t's provided by process
  - High-V_t with nominal I_{off}
  - Low-V_t with 10X higher I_{off} (15% better delay)

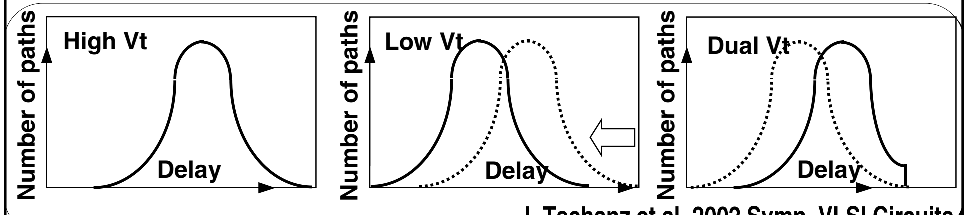




Goal: selective high-V_t usage for leakage power reduction

11

Active Leakage Reduction: Dual V_t design



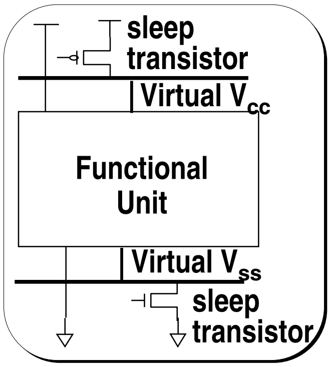
J. Tschanz et al, 2002 Symp. VLSI Circuits

- Methodology:
  - Low-V_t on critical paths (best delay)
  - High-V_t on non-critical paths: 10X lower leakage
- Selective low-V_t and high-V_t insertion enables >1X and <<10X active leakage power
- Challenges: Tool-flow, accurate slack estimation

12

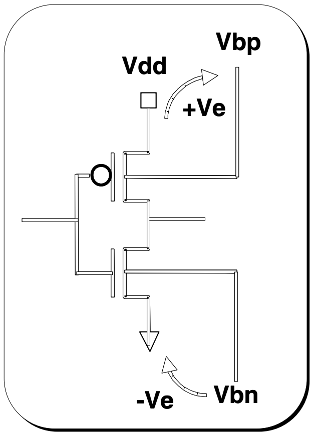
Standby Leakage Reduction: Sleep Transistor design

- Motivation: Cut off power supply in sleep-mode
  - Insert “sleep” transistor between main supply and functional unit’s supply rails
  - Latches tied to main supply rails: retain state



Standby leakage benefit for 5% delay penalty

Switching + Leakage Reduction: Forward Body Bias

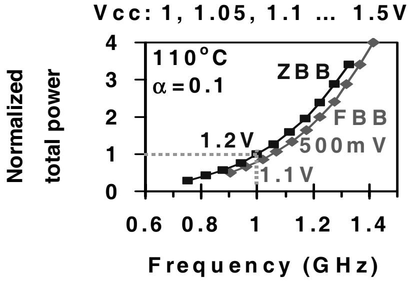


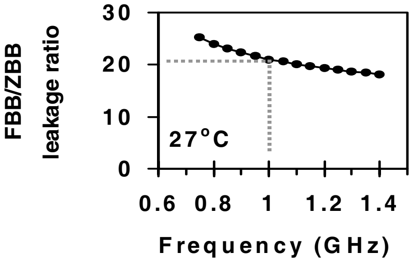
A. Keshavarzi et al, 2002 Symp. VLSI Circuits

20% power reduction at 1GHz

8% ↑ frequency at iso-power

20X ↓ idle-mode leakage





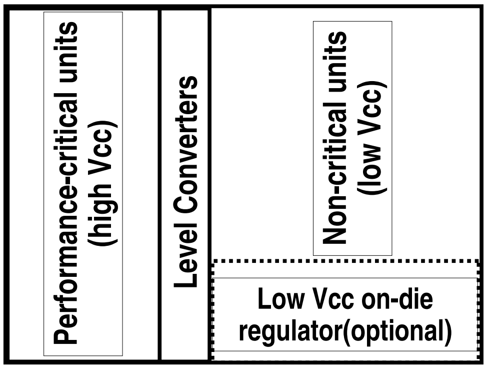
Outline

- Challenges & Circuit Solutions:
  - Leakage power reduction: Dual-Vt and Forward Body Bias
  - Stand-by leakage reduction: Sleep transistor design
  - Dual-Vcc switching + leakage power reduction
  - Dual-Vcc interface: split-output level converters and write-port latches
  - Dynamic leakage-tolerant Conditional/burn-in keeper
  - Process parameter variation tolerant dynamic circuits
  - Pseudo-static & Self Reverse Biased bitlines
  - Static split-decoder register file technologies
  - Source follower and transition encoded interconnects

15

Switching + Leakage Reduction: Dual Supply Design

- Motivation: power-optimized performance
- High (regular) supply for critical units
- Lower supply for non-critical units
- Lower V_{cc} generated off-chip or regulated on-die



Performance-critical units (high Vcc)

Level Converters

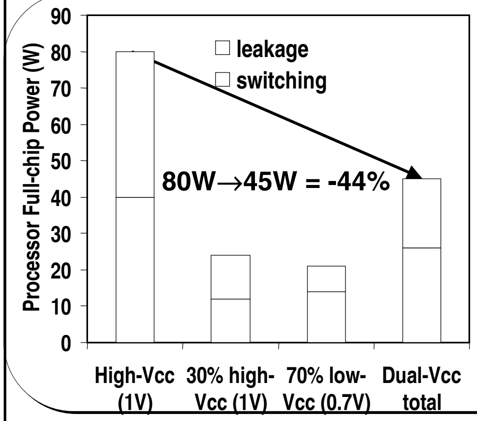
Non-critical units (low Vcc)

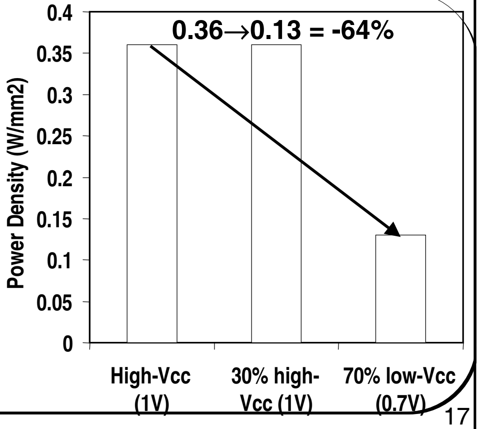
Low Vcc on-die regulator(optional)

16

Switching + Leakage Reduction: Dual Supply Design

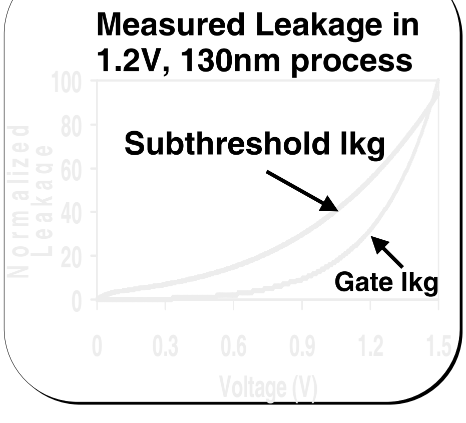
- Cumulative processor-level power and power density benefit
- Challenges: level converters, low-V_{cc} distribution

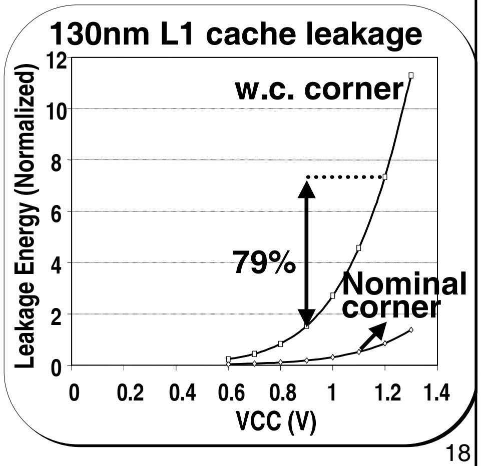




Switching + Leakage Reduction: Dual Supply Design

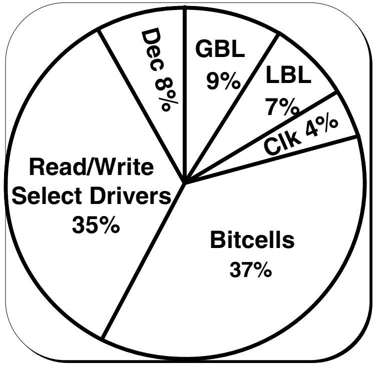
- Active leakage benefit with lower supply voltage
- Exponential subthreshold and gate leakage reduction





Register File Energy Breakup

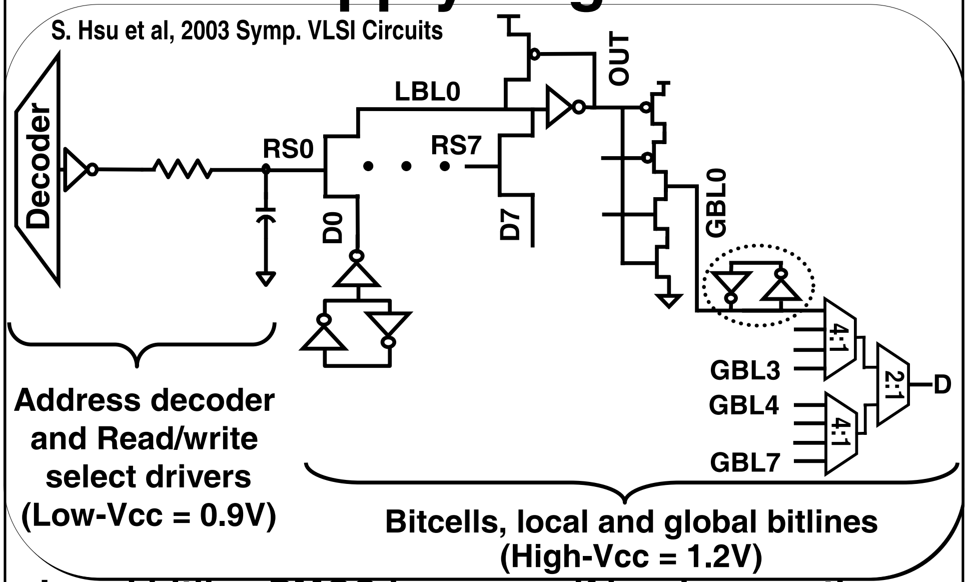
90nm CMOS, 1.2V, 110°C simulation



- Active leakage = 83% of total energy
- Address decoder and read/write select drivers contribution = 43% of total energy
- Goal: Lower supply voltage on decoder and read/write select drivers to reduce total energy

Dual Supply Register File

S. Hsu et al, 2003 Symp. VLSI Circuits



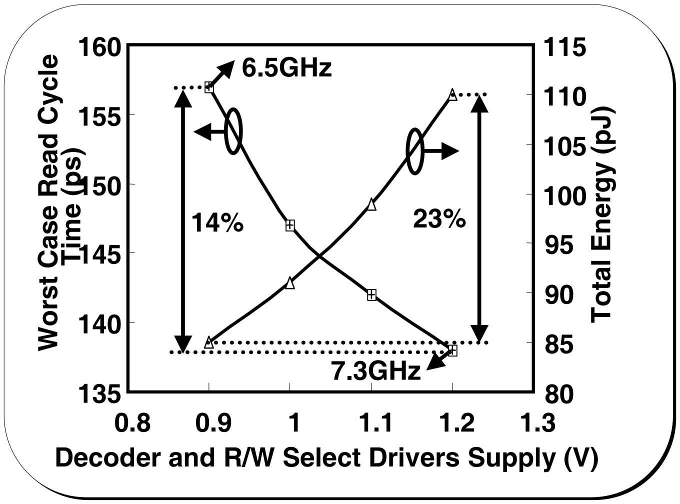
Address decoder and Read/write select drivers (Low-Vcc = 0.9V)

Bitcells, local and global bitlines (High-Vcc = 1.2V)

- Local bitline PMOS keeper: self level converting
- Bitcells stability unaffected: operating on high-Vcc

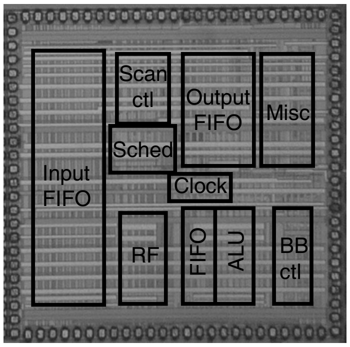
Dual-Vcc RF Energy-Delay

90nm CMOS, 110°C simulation (high-Vcc = 1.2V)

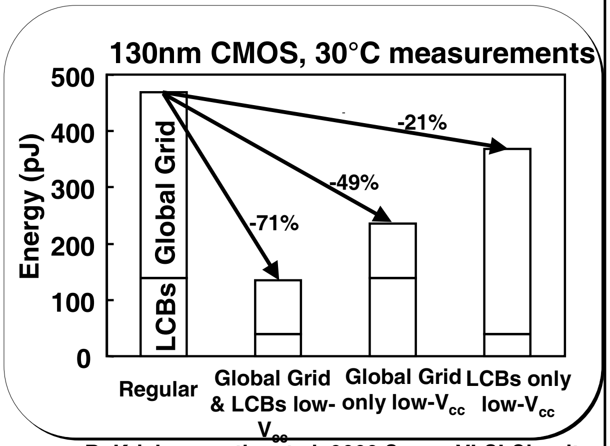


- 6.5GHz dual supply operation (7.3GHz at 1.2V)
- 23% total energy reduction with dual supply design

Dual Vcc Clocking



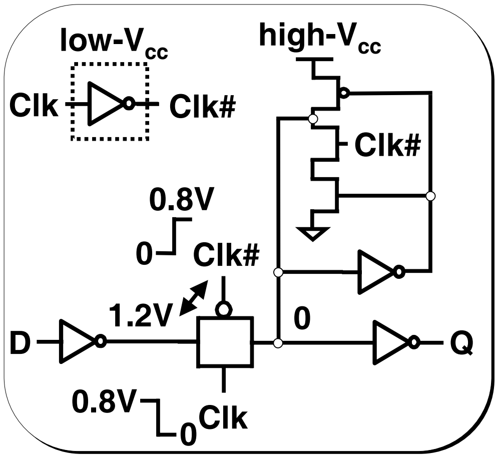
5GHz 130nm Integer Execution Core

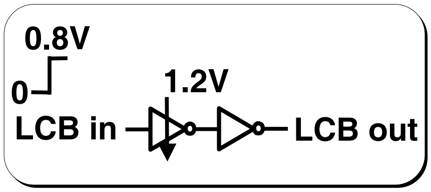


R. Krishnamurthy et al, 2002 Symp. VLSI Circuits

- Goal: Combat increasing clock power in μP's
- High-V_{cc} (1.2V) on datapath, Low-V_{cc} (0.8V) on clock
- 21-71% clock energy reduction

Dual-Vcc DC Power Impact



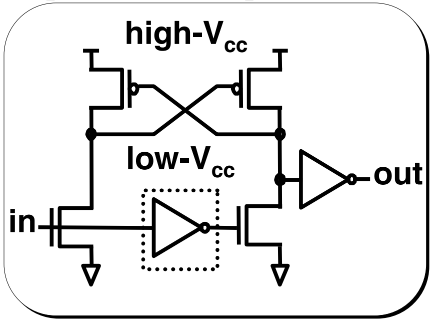


LCB DC power = 350µW

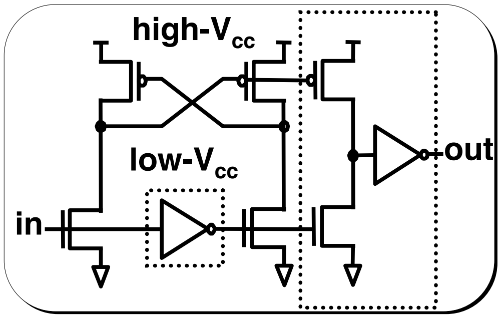
32-bit processor pass-gate latch:
DC power = 290µW

- DC power free latches and LCBs required to enable practical dual-Vcc core/cache interface or low-Vcc clocking

Split-output Level Converter LCB

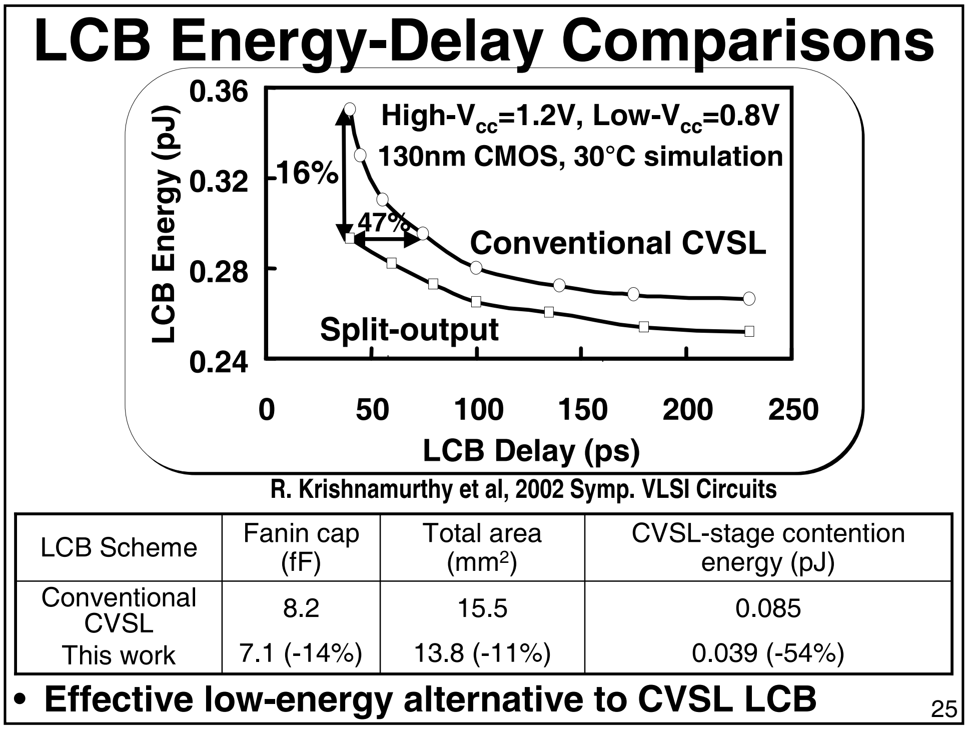


Conventional CVSL LCB



Split-output LCB

- Contention in CVSL LCB degrades delay
- Split-output LCB decouples CVSL stage from output driver stage
  - Fast level conversion due to low contention
  - Reduced fanin load on clock grid

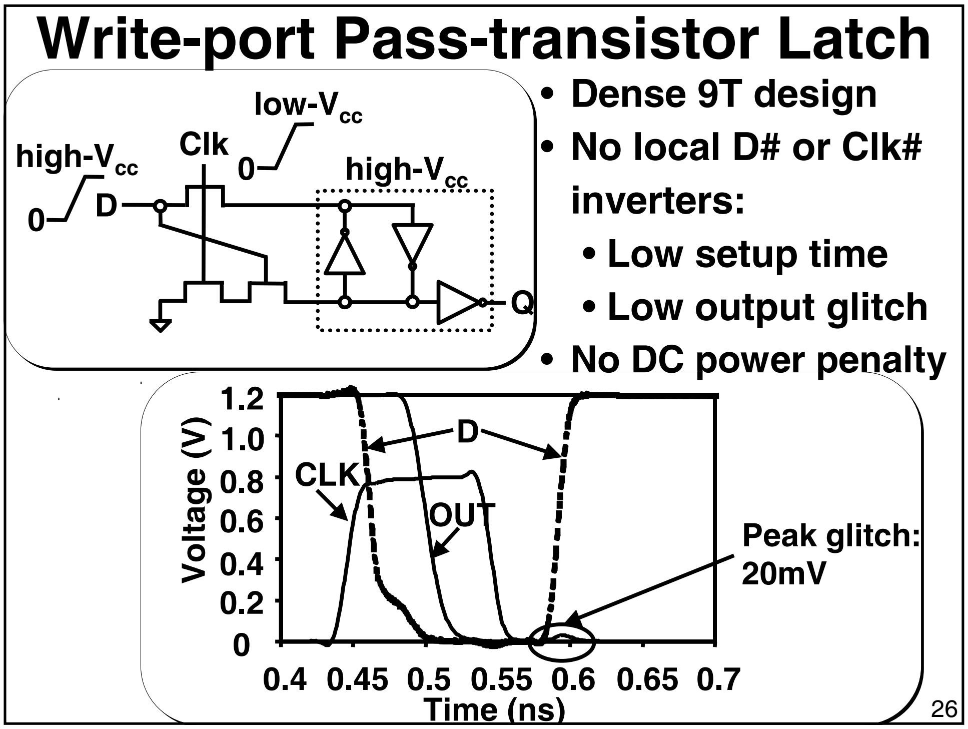


LCB Energy-Delay Comparisons

R. Krishnamurthy et al, 2002 Symp. VLSI Circuits

LCB Scheme	Fanin cap (fF)	Total area (mm^2)	CVSL-stage contention energy (pJ)
Conventional CVSL	8.2	15.5	0.085
This work	7.1 (-14%)	13.8 (-11%)	0.039 (-54%)

- **Effective low-energy alternative to CVSL LCB**



Write-port Pass-transistor Latch

- **Dense 9T design**
- **No local D# or Clk# inverters:**
  - **Low setup time**
  - **Low output glitch**
- **No DC power penalty**

Write-port Latch Comparisons

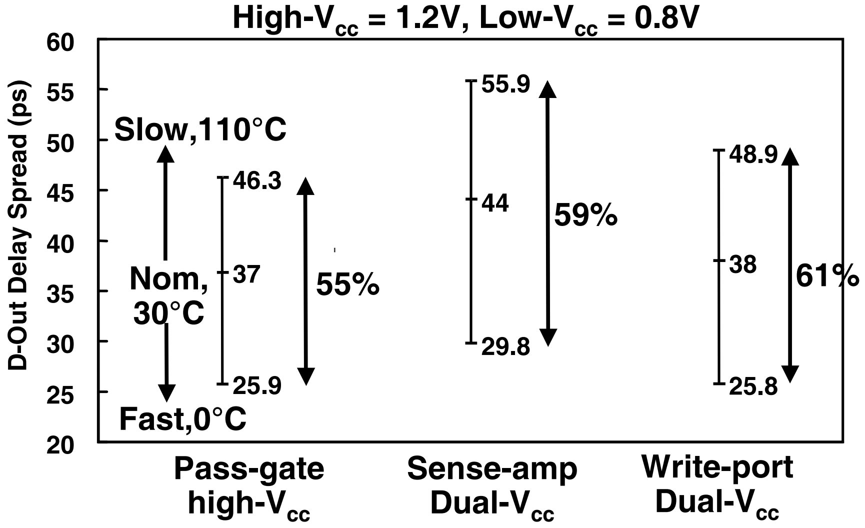
130nm CMOS, 30°C simulation

Latch Scheme	Number of Transistors	D→Q (ps)	O/P glitch (mV)	Energy (pJ)
Pass-gate high-V_{cc}	11	37	18mV	0.76
Sense-amp dual-V_{cc}*	11	44	60mV	0.59
Write-port dual-V_{cc}	9	38	20mV	0.66

* H. Kawaguchi et al, JSSC, May 1998

- Data activity = 0.1, Clock activity = 1.0
- Optimized for constant fanin and fanout load
- 12% energy reduction vs. pass-gate latch
- 14% delay reduction vs. sense-amp latch

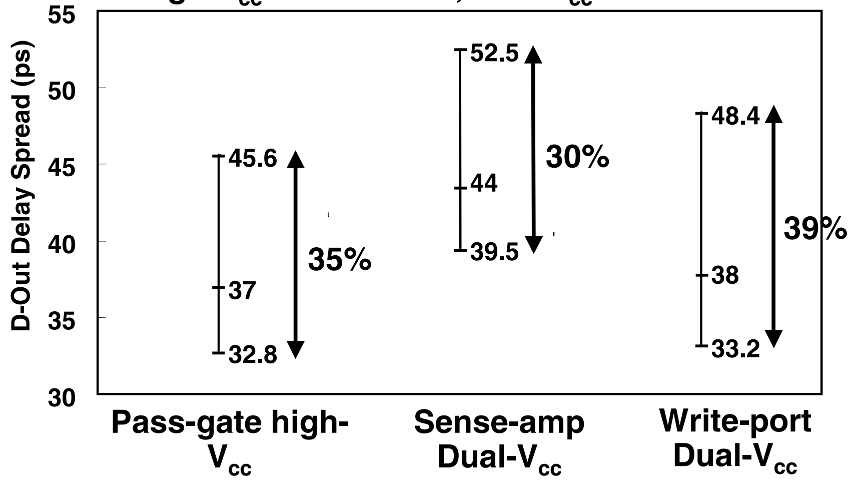
Process Skew Sensitivity



- Comparable delay spread across fast/slow corners

Supply Variation Sensitivity

High-V_{cc} = 1.2V ±15%, Low-V_{cc} = 0.8V ±15%

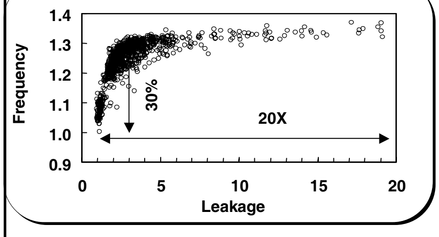


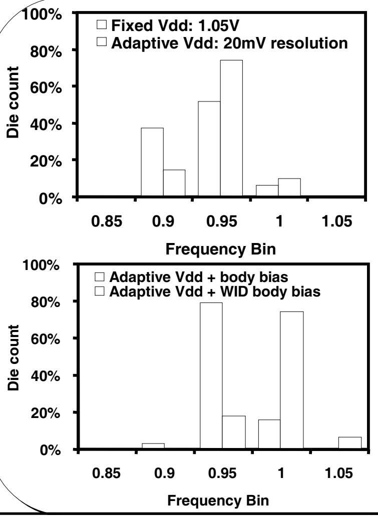
- Comparable delay spread for supply variations

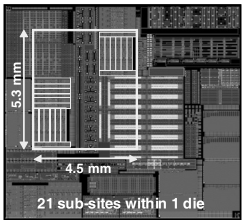
29

Adaptive Vcc: Variation-tolerant Circuits

- Motive: change Vcc adaptively to reduce impact of parameter variations
  - Large Fmax vs. Isb spread (worsening with scaling)
  - Lower Vcc on leakage-limited circuits (subject to stability limits)
  - Higher Vcc on speed-limited circuits (subject to reliability limits)







J. Tschanz et al, 2002 Symp. VLSI Circuits

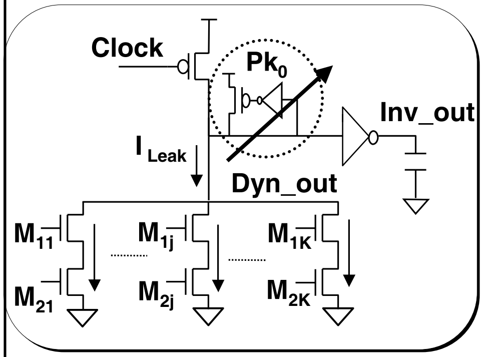
30

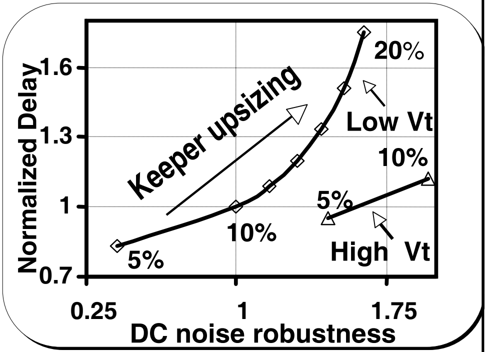
Outline

- Challenges & Circuit Solutions:
  - Leakage power reduction: Dual-Vt and Forward Body Bias
  - Stand-by leakage reduction: Sleep transistor design
  - Dual-Vcc switching + leakage power reduction
  - Dual-Vcc interface: split-output level converters and write-port latches
  - Dynamic leakage-tolerant Conditional/burn-in keeper
  - Process parameter variation tolerant dynamic circuits
  - Source follower and transition encoded interconnects

31

Improving Dynamic Leakage Tolerance: Keeper Upsizing



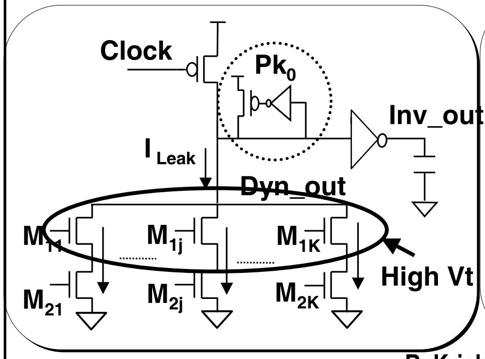


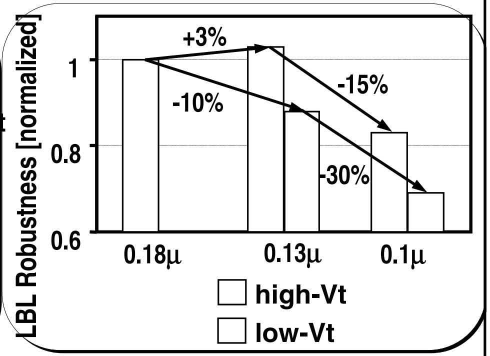
A. Alvandpour et al, 2001 Symp. VLSI Circuits

- Robustness = DC Noise Margin / Vcc
- Traditional noise engineering ⇒ diminishing ROI

32

Dual Vt Scaling Trends

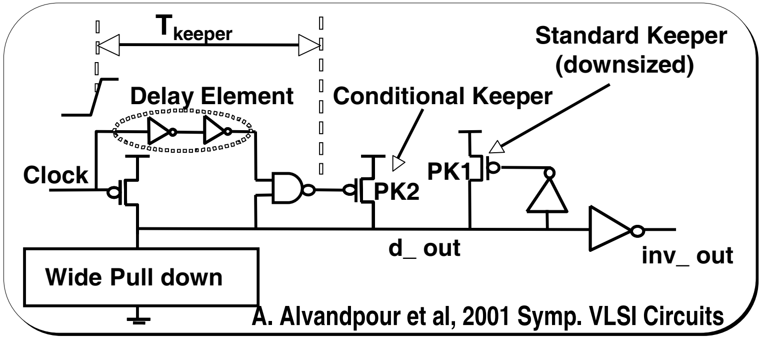




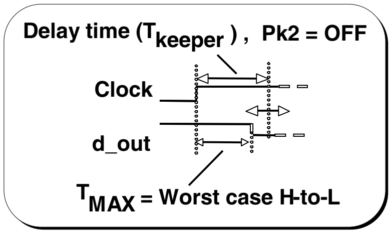
R. Krishnamurthy et al, 2001 Great Lakes VLSI Symp.

- Replace NMOS pulldowns with high-V_t
- Good one-time solution for 130nm node
- 15-30% degradation for both high- and low-V_t in sub-130nm
- Dual-Vt bitlines don’t scale well beyond 130nm

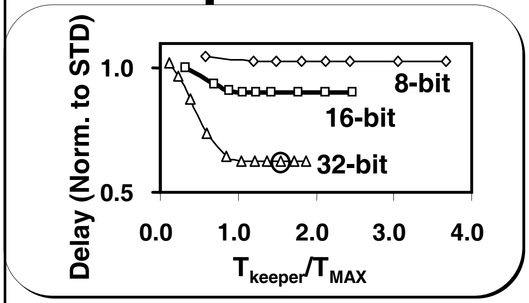
Leakage-tolerant Conditional Keeper Domino Technology

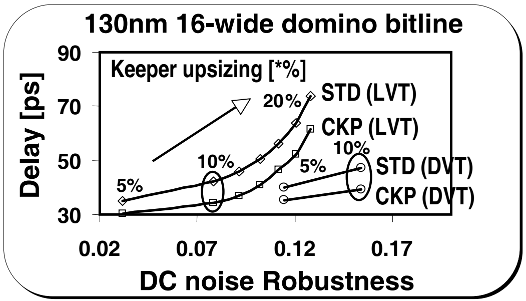


A. Alvandpour et al, 2001 Symp. VLSI Circuits


Delay time (T$_{keeper}$) , Pk2 = OFF
Clock
d_out
T$_{MAX}$ = Worst case H-to-L

Leakage-tolerant Conditional Keeper Domino Technology

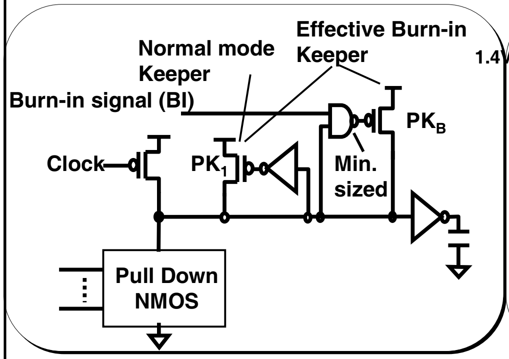


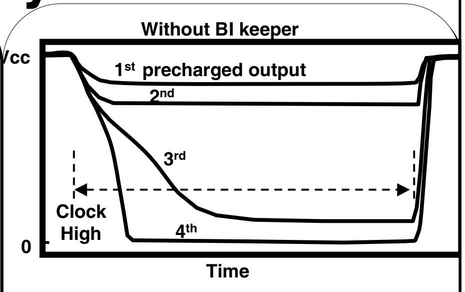


- Motivation:
- Weak keeper (low contention) during evaluation window
- Strong keeper activated only if dynamic node "high"
- 20% delay reduction at same robustness
- High-performance "dual-V_t enabler"

35

Burn-in Tolerant Dynamic Circuits



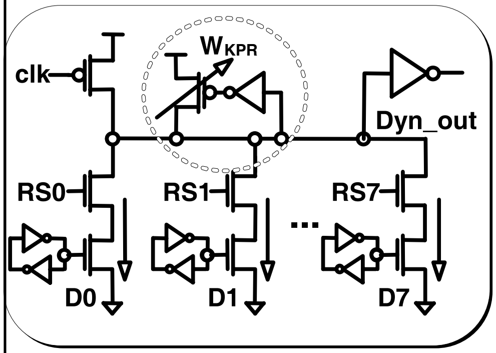


A. Alvandpour et al, 2002 CICC

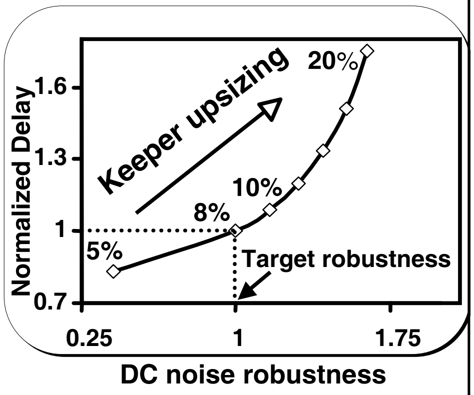
- Leakage sensitive circuits not functional at burn-in
  - Elevated supply and temperature
- Larger keepers increase delay at "normal" condition
- Conditional keeper enables functional burn-in testing
- 2X lower noise during burn-in
- 50% better delay than upsizing BKM keeper

36

Leakage Variations Impact

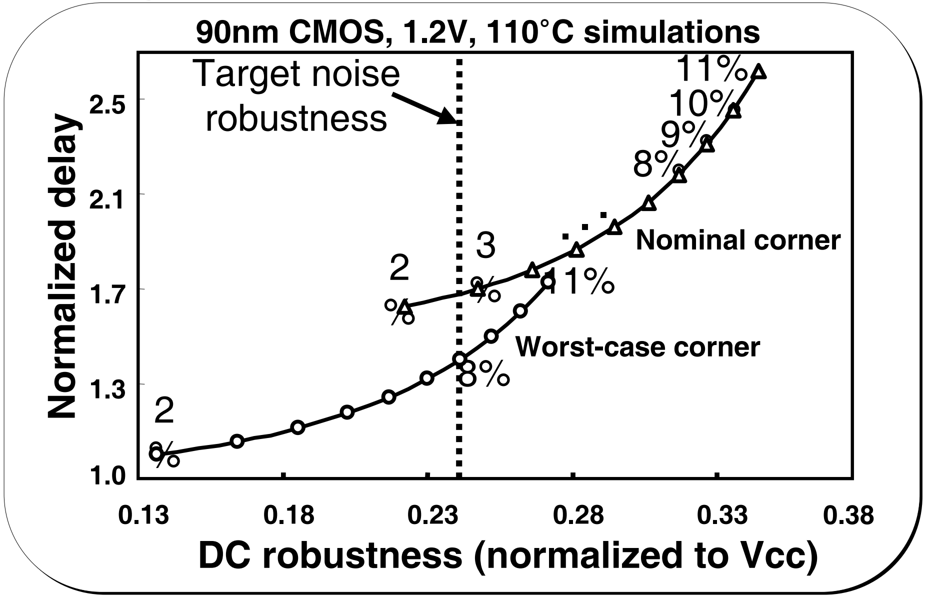


Dynamic 8-way Bitline

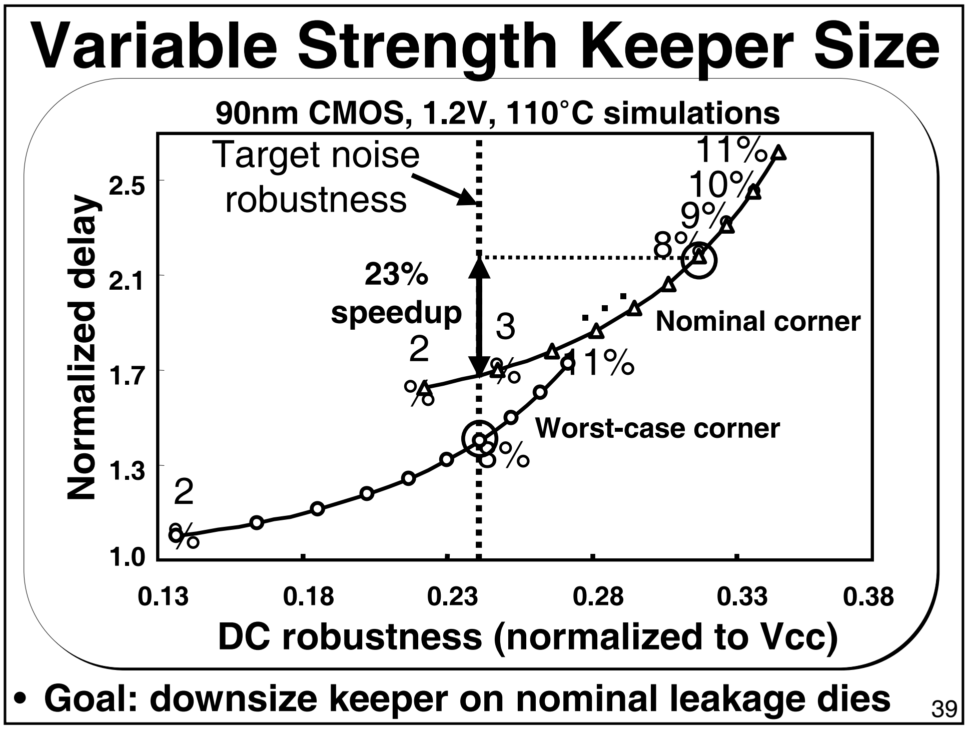


- Dynamic circuit NMOS pulldown leakage variation:
  - Keeper size determined for target robustness at worst-case leakage corner
  - Excess leakage dies: fail to meet target robustness
  - Lower leakage dies: over-designed for robustness

Delay and Robustness Spread

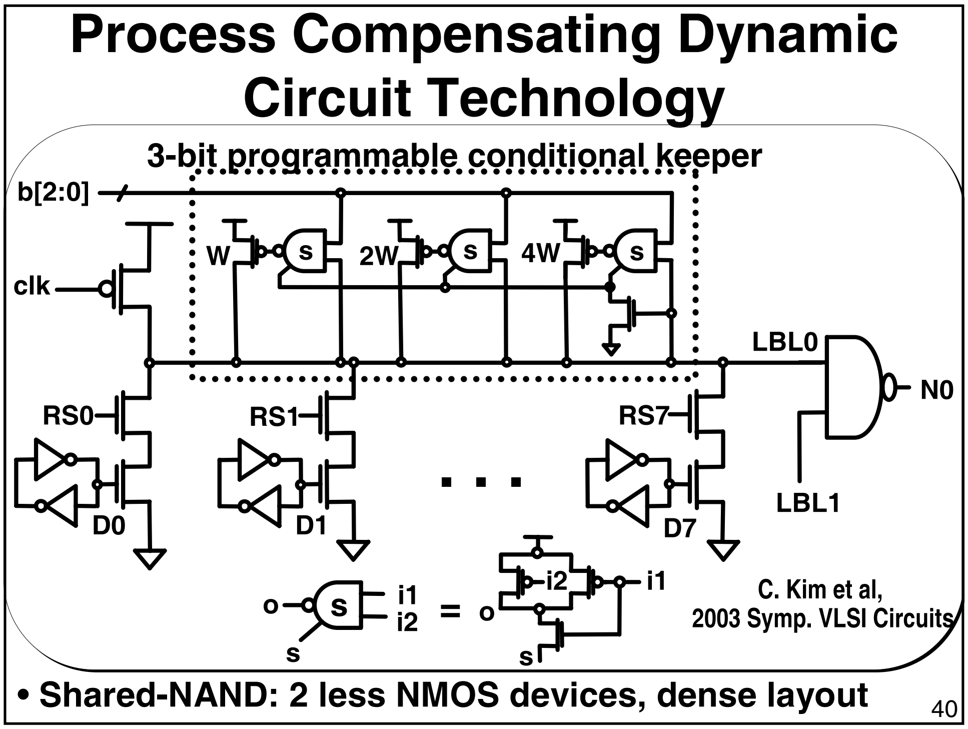


- Fast corner keeper sizing is sub-optimal for delay



Variable Strength Keeper Size

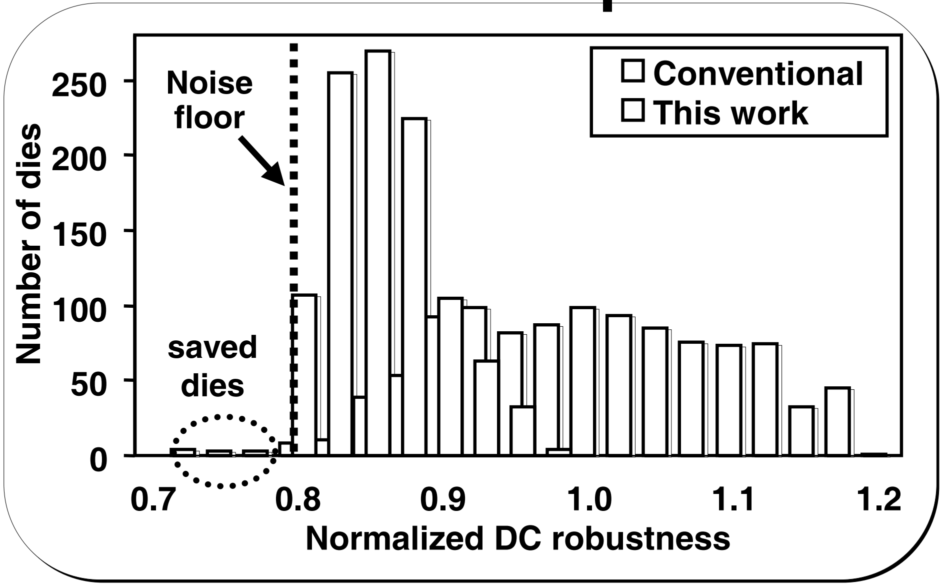
- Goal: downsize keeper on nominal leakage dies



Process Compensating Dynamic Circuit Technology

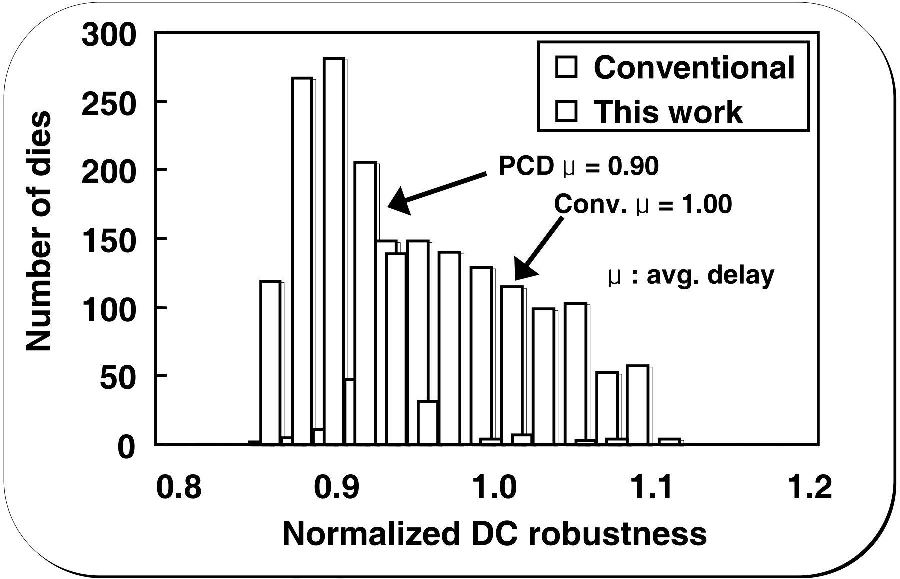
- Shared-NAND: 2 less NMOS devices, dense layout

Robustness Squeeze



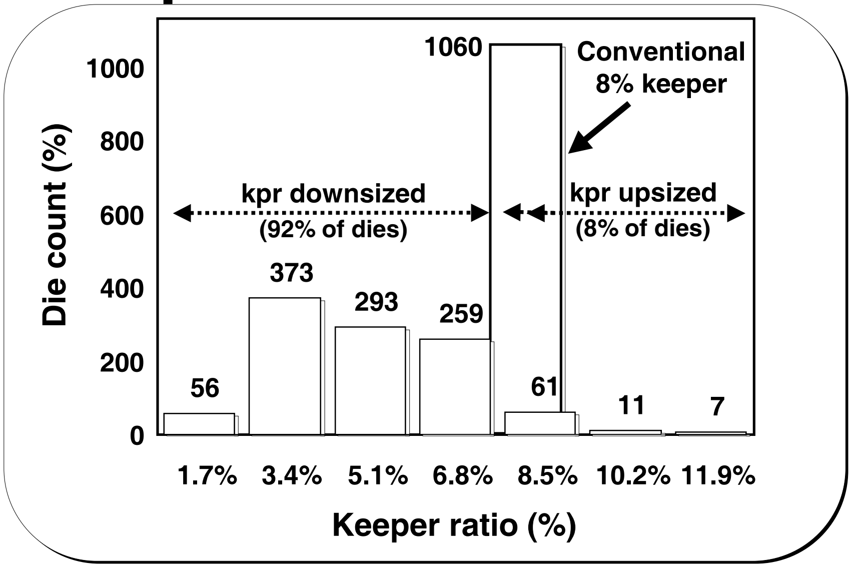
- 5X reduction in robustness failing dies

Delay Squeeze



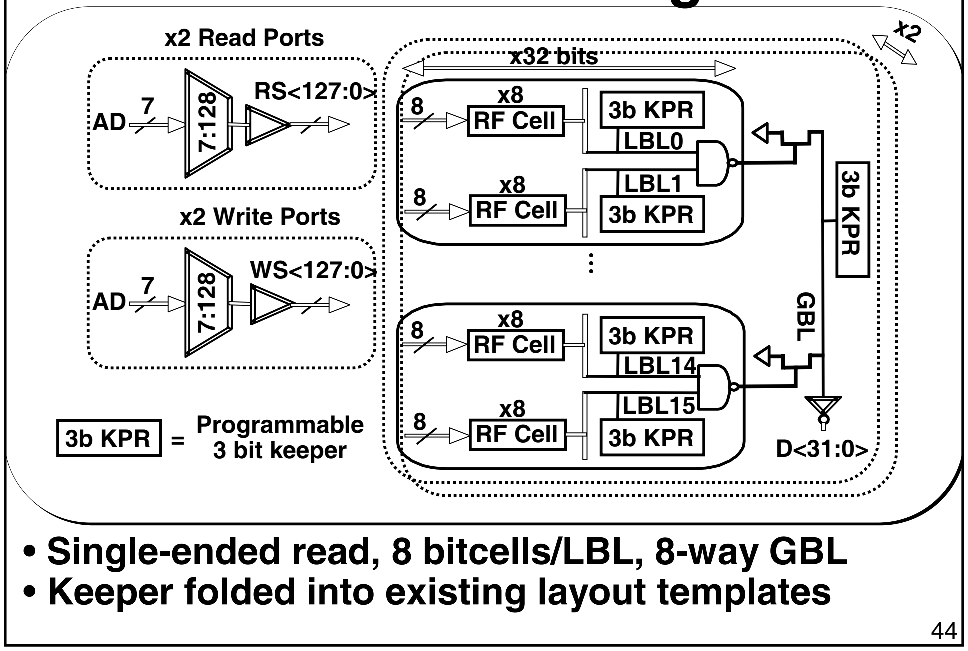
- 10% opportunistic speedup

Keeper Ratio Distribution



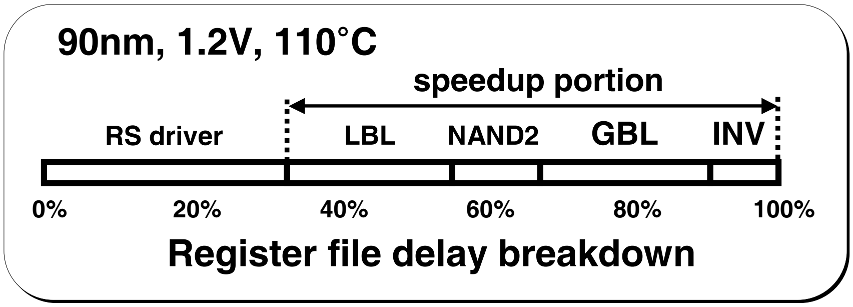
- Keeper downsized in 92% of dies

128x32b 2R2W PCD Register File



- Single-ended read, 8 bitcells/LBL, 8-way GBL
- Keeper folded into existing layout templates

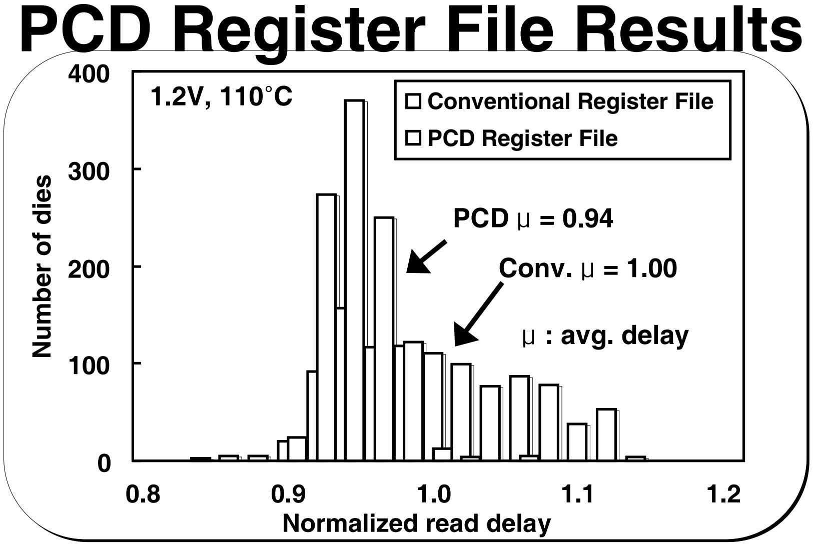
PCD Register File Delay, Energy



Register file delay breakdown

- Speeds up 67% of RF critical path delay
- 2% worst-case total energy overhead

PCD Register File Results

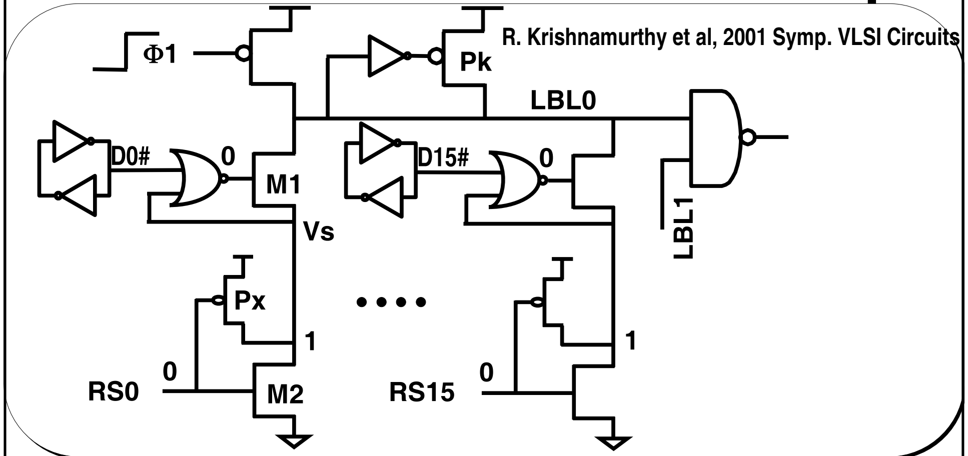


Read Delay Benefit	5.5%
Robustness Failing Dies	0.2% (5X ▼)
Read Delay Variation: σ/µ	6.1%→2.3% (2.7X ▼)

Outline

- Challenges & Circuit Solutions:
  - Leakage power reduction: Dual-Vt and Forward Body Bias
  - Stand-by leakage reduction: Sleep transistor design
  - Dual-Vcc switching + leakage power reduction
  - Dual-Vcc interface: split-output level converters and write-port latches
  - Dynamic leakage-tolerant Conditional/burn-in keeper
  - Process parameter variation tolerant dynamic circuits
  - Pseudo-static & Self Reverse Biased bitlines
  - Static split-decoder register file technologies
  - Source follower and transition encoded interconnects

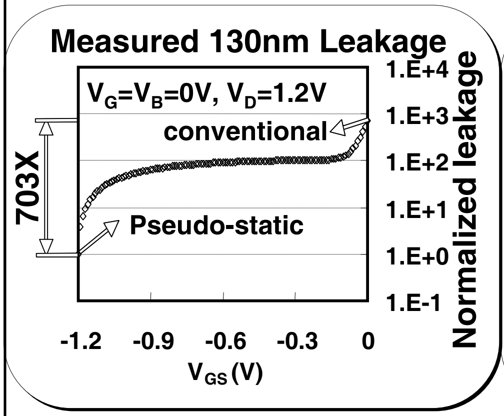
Pseudo-static Bitline Technique

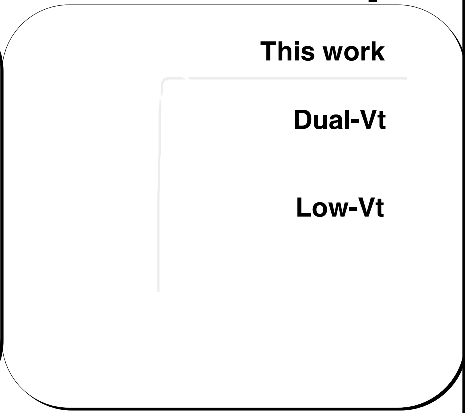


R. Krishnamurthy et al, 2001 Symp. VLSI Circuits

- Goal: $V_{GS} = -V_{cc}$ and $V_{DS} = 0$ on deselected bitline's access transistors
- No oxide stress or additional bias voltages

Pseudo-static Bitline Technique





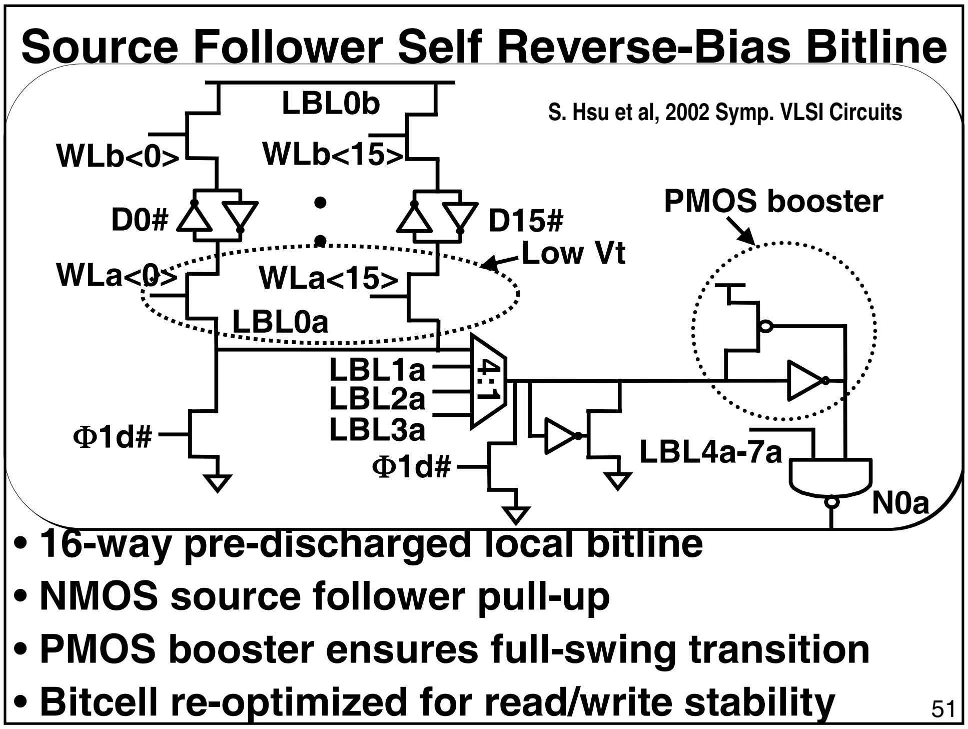
- 130nm measurement: 703X reduction in bitline leakage (~4 process generations)
- Scalable replacement for dual-Vt bitlines

6GHz 130nm Pseudo-static RF

130nm, 1.2V, 110C

LBL Scheme	Read Delay	DC robustness (DC noise margin/Vcc)	Energy/transition (normalized)
Low-Vt	158ps	0.072	1.0
Dual-Vt	178ps	0.157	0.95
This work	165ps	0.214	1.02

- 256-entryx32-bit 4-read, 4-write ported register file
- Single-cycle latency & throughput: performance critical
- 6GHz operation (8% read delay improvement) with simultaneous 36% robustness benefit over dual-V_t
- Scalable to sub-130nm technologies

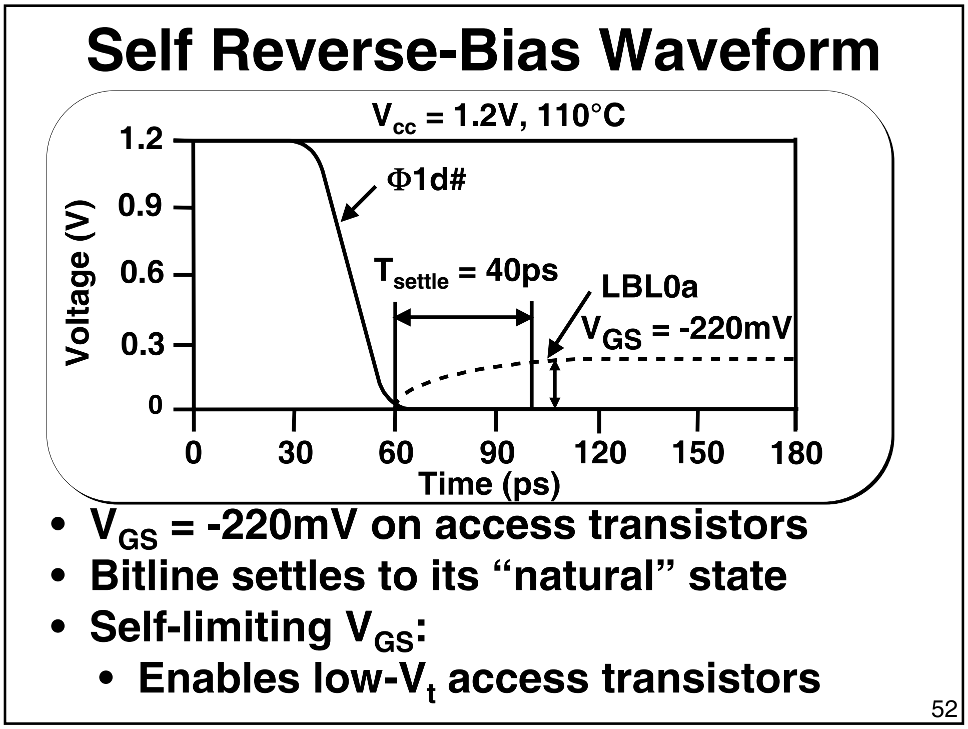


Source Follower Self Reverse-Bias Bitline

S. Hsu et al, 2002 Symp. VLSI Circuits

- 16-way pre-discharged local bitline
- NMOS source follower pull-up
- PMOS booster ensures full-swing transition
- Bitcell re-optimized for read/write stability

51

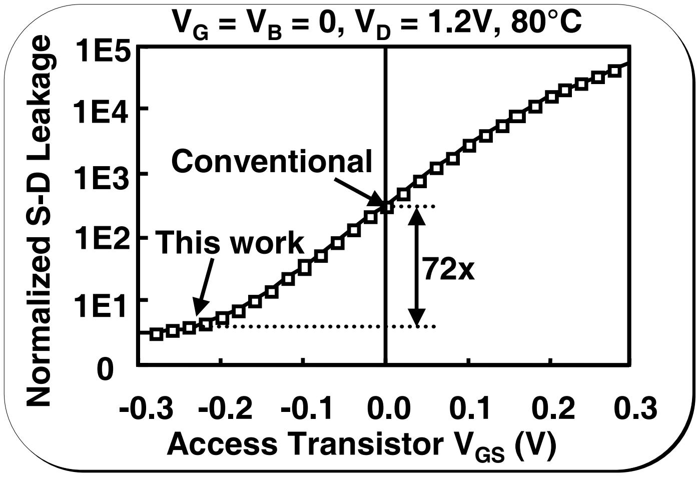


Self Reverse-Bias Waveform

- V_{GS} = -220mV on access transistors
- Bitline settles to its “natural” state
- Self-limiting V_{GS}:
  - Enables low-V_t access transistors

52

Leakage Reduction Measurement



- 130nm low-Vt NMOS leakage measurement
- 72X bitline leakage reduction
- 7X leakage reduction over dual-V_t

53

4.5GHz Single-Cycle L0 Cache using SRB
130nm CMOS, 1.2V, 110°C simulation

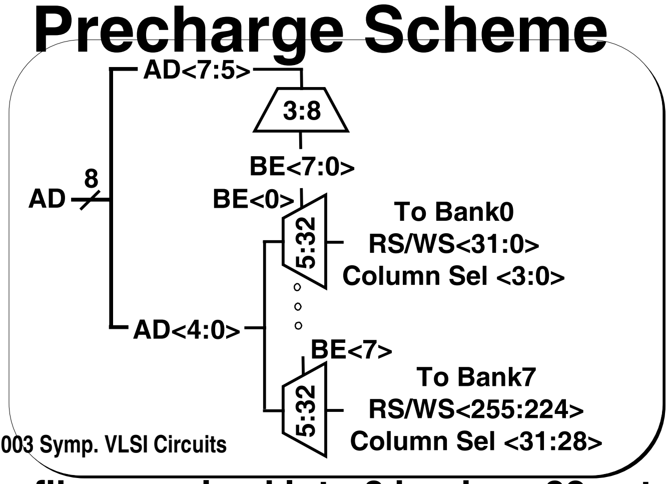
LBL Scheme	Read Delay	Energy/ Transition
Conventional	247ps	26.88pJ
This work	220ps (-11%)	28.14pJ (+5%)

LBL Scheme	DC robustness (DC noise margin/Vcc)	LBL droop/rise
Conventional	0.114	350mV
This work	0.233	220mV

- 11% total read path delay improvement
- 5% energy overhead due to PMOS boosters
- 2X simultaneous DC robustness increase
- 37% reduction in bitline droop / rise
- Scalable replacement for dual-Vt bitlines

54

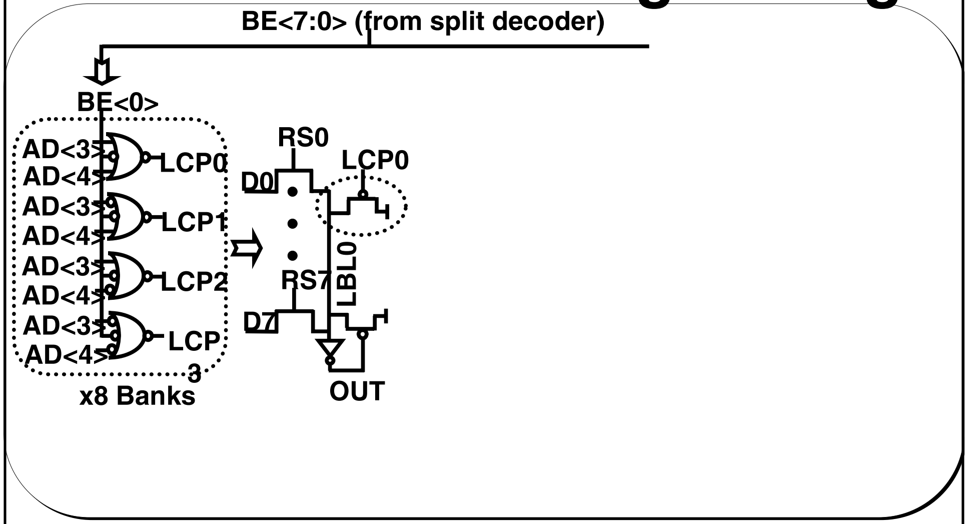
Split Decoder Conditional Precharge Scheme



S. Hsu et al, 2003 Symp. VLSI Circuits

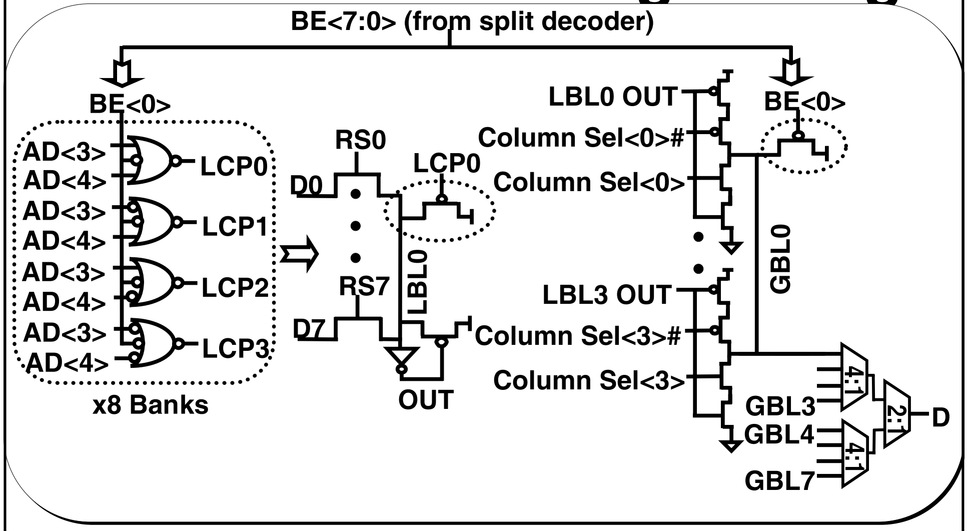
- Register file organized into 8 banks x 32 entries
- 1 first-level 3:8 address decoder:
  - Produces 8 Bank Enable signals BE<7:0>
- 8 second-level 5:32 address decoders (1 per bank)

Conditional Precharge Design



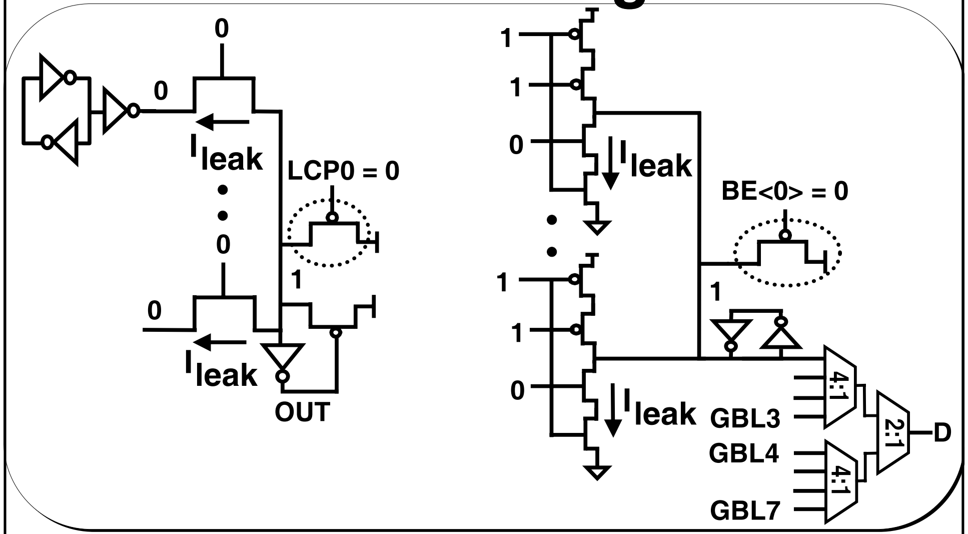
- LCP statically anchors deselected LBL

Conditional Precharge Design



- LCP statically anchors deselected LBL
- BE statically anchors deselected GBL
- Deselected LBLs and GBLs strongly held at "1"

Worst-case Leakage Vector



- Conditional precharge layout area overhead <1%
    - Extra device folded into keeper layout template
    - Precharge control signals routed on upper layer

6.5GHz Single-cycle Split Decoder RF

90nm CMOS, 1.2V, 110°C simulation

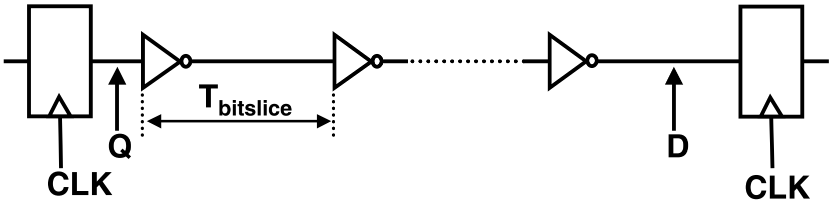
LBL/GBL Scheme	DC robustness (Deselected BL)	DC robustness (One entry enabled)	Noise recovery time
Conventional static	220mV	310mV	23ps
Conventional dynamic	192mV	196mV	Non-recoverable
This work	365mV	312mV	23ps

- DC robustness for deselected bitline:
  - 65% (90%) higher than static (dynamic) bitlines
- Same DC robustness as static with 1 entry enabled
- Full recovery from AC noise (same as static)
- 1% read delay penalty due to excess diffusion cap

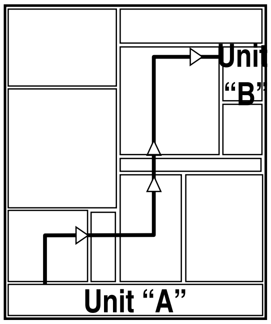
Outline

- Challenges & Circuit Solutions:
  - Leakage power reduction: Dual-Vt and Forward Body Bias
  - Stand-by leakage reduction: Sleep transistor design
  - Dual-Vcc switching + leakage power reduction
  - Dual-Vcc interface: split-output level converters and write-port latches
  - Dynamic leakage-tolerant Conditional/burn-in keeper
  - Process parameter variation tolerant dynamic circuits
  - Pseudo-static & Self Reverse Biased bitlines
  - Static split-decoder register file technologies
  - Source follower and transition encoded interconnects

Interconnect Reality

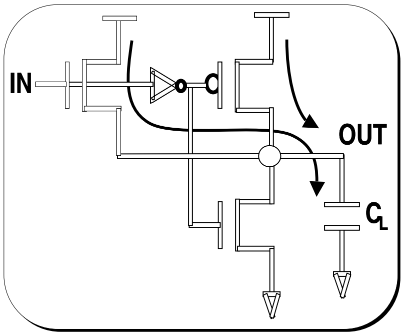


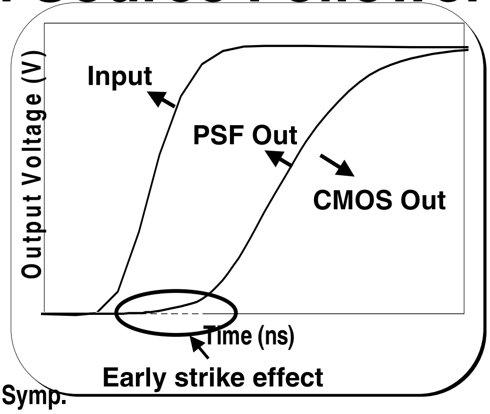
$$T_{clk} = T_{clk\text{-}Q} + N^*T_{bitslice} + T_{setup} + T_{skew}$$



- Microprocessor global bus limitations:
- Ideal world: $T_{repeater} = T_{RCsegment} = T_{bitslice}/2$
- Real world:
- Don't get repeaters where you want
- Floorplan decides repeater locations
- Performance << ideal T_{clk}

PMOS-Boosted Source Follower



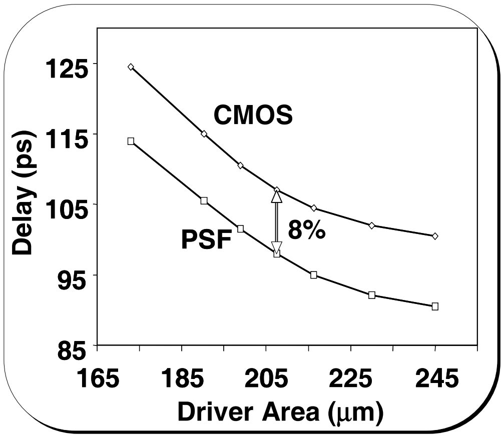


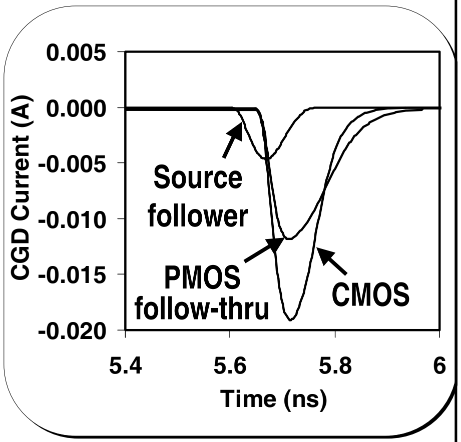
Early strike effect

R. Krishnamurthy et al, 2001 VLSI Circuits Symp.

- **Source follower NMOS to accelerate driver strength**
  - **NMOS begins fast pullup ⇒ "early" strike effect**
  - **PMOS follow-through completes full-swing transition**
- **Full-swing CMOS driver robustness**

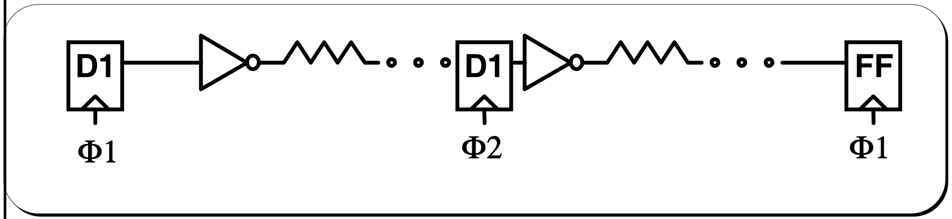
L1 Cache Bus Results





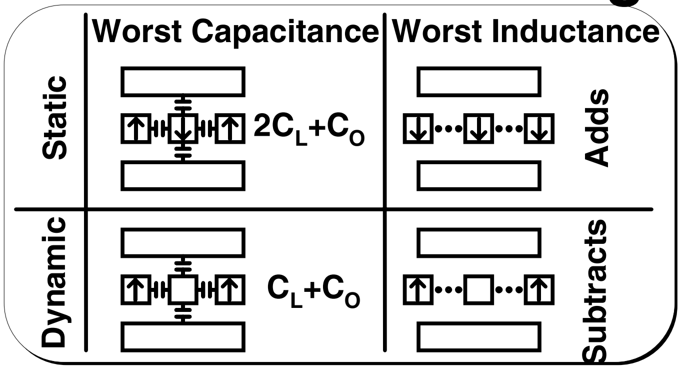
- 8% driver delay reduction for same fanin and area
- 10% simultaneous peak current reduction (7% lower decoupling capacitance)
- Effective alternative to upsizing CMOS drivers

Background: Conventional Dynamic Bus



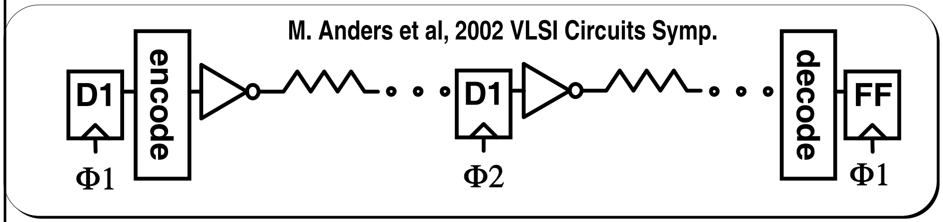
- Domino timing applied to interconnect
- Monotonic transitions
  - –**Reduced collinear capacitance**
    - **Static (worst case) = 2X**
    - **Dynamic (worst case) = 1X**
  - – **Φ2 repeater required – susceptible to noise**
- Higher transition activity when input = 1
- Static CMOS inverters drive all segments

Dynamic Bus Advantages



- Capacitance effects reduced
  - **Collinear capacitance reduced 2X**
  - **Orthogonal capacitance unchanged**
- Inductance effects reduced
  - **Can oppose transition for static bus**
  - **Can reduce capacitive effects for dynamic bus**

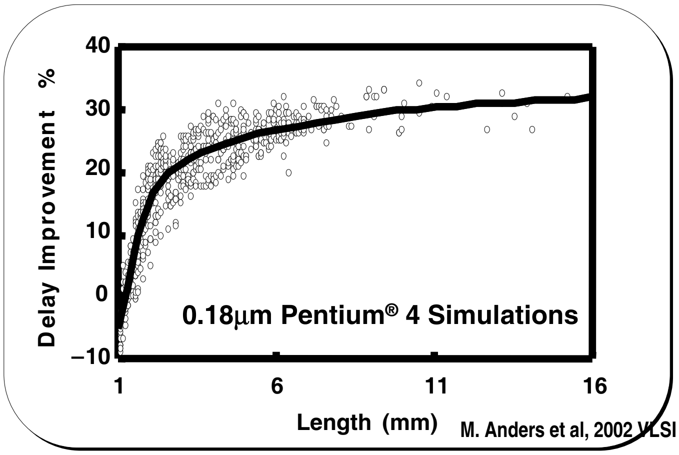
Transition-Encoded Bus



M. Anders et al, 2002 VLSI Circuits Symp.

- Encoder circuit
  - **XOR of previous and current input**
  - **Domino compatible output**
- Decoder circuit
  - **XOR of previous output and bus state**

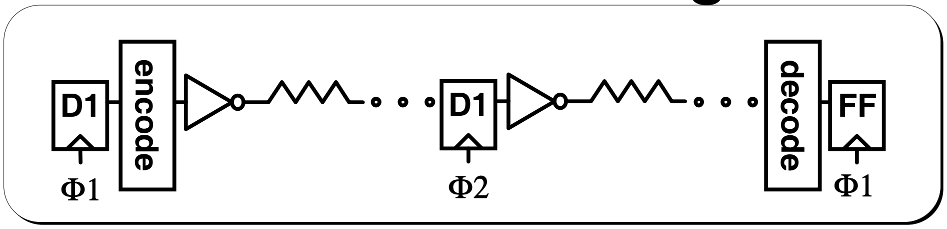
Transition-Encoded Bus



M. Anders et al, 2002 VLSI Circuits Symp.

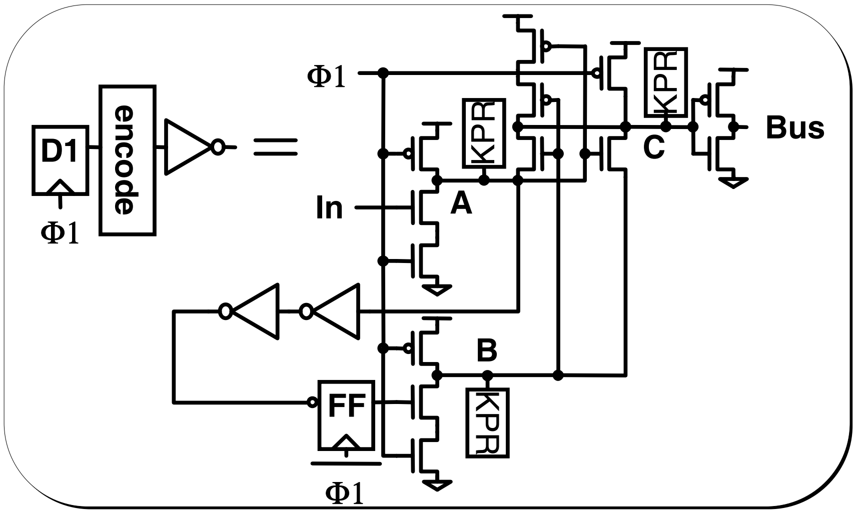
- Transition only when current input != previous input
- Dynamic bus performance but energy profile of static bus
- Energy scales linearly with input switching activity
- 79% of full-chip buses: 10%-35% delay improvement

TED Bus Advantages



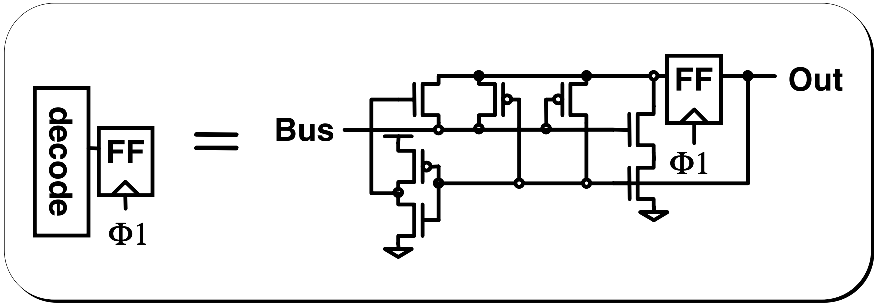
- Dynamic bus performance improvement
  - **Collinear capacitance reduction**
- Static bus energy
  - **Transition dependent switching activity**
- Noise-insensitive Φ2 repeater required
  - **Regains noise immunity of CMOS inverter**

Encoder Circuit



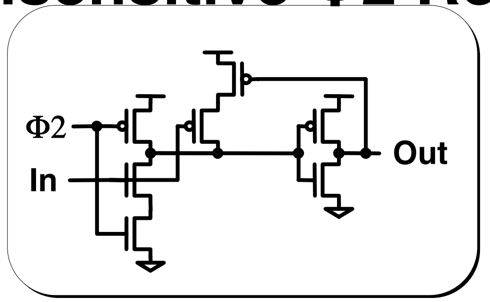
- Domino-compatible XOR of current and previous inputs
- Small FF holds previous state

Decoder Circuit



- Original data recovered
  - –**When Bus = 1, Out transitions**
  - –**When Bus = 0, Out does not change**
- Flip-flop must be initially reset
- No clock required for decode

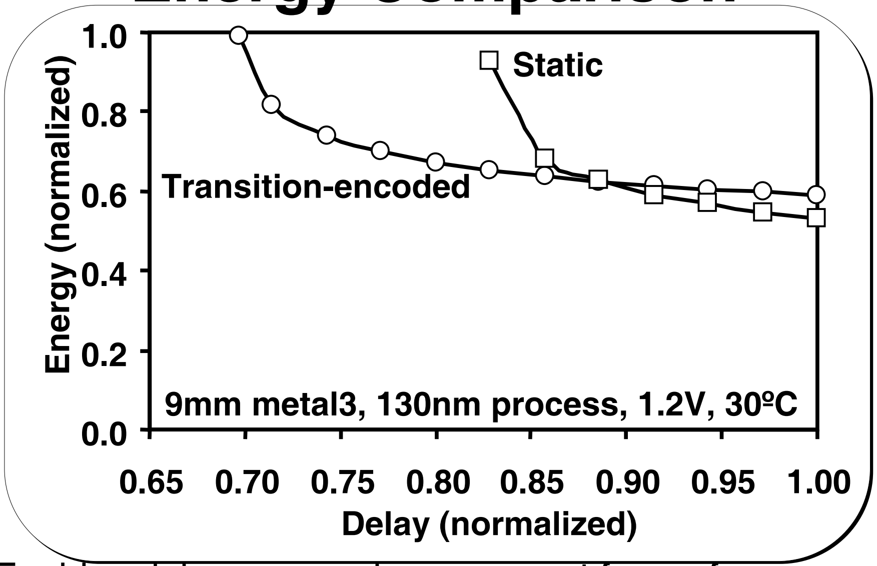
Noise Insensitive Φ2 Repeater



- Sized for $V_{CC}/2$ trip point
  - **Same as static CMOS inverter**
- Latching behavior of domino repeater
- Minimal delay penalty

71

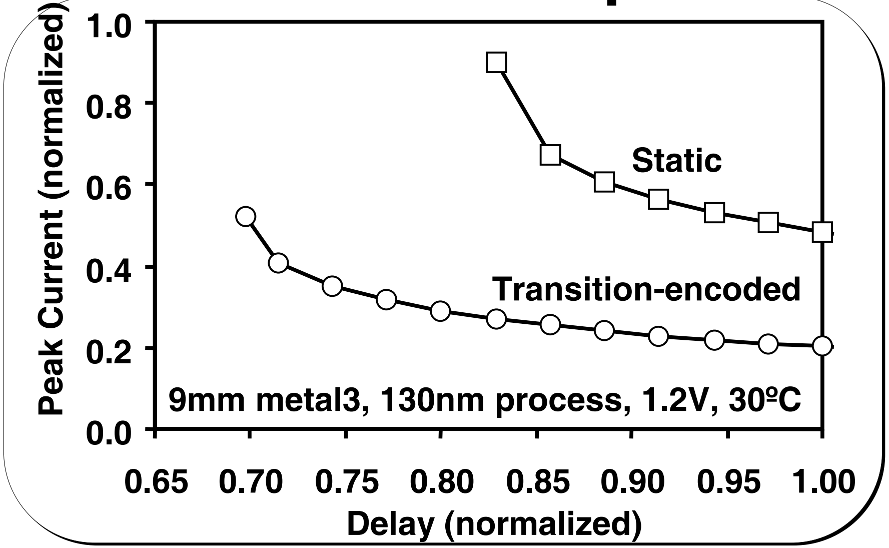
Energy Comparison



Enables delay, energy improvement for performance-critical buses

72

Peak Current Comparison



9mm metal3, 130nm process, 1.2V, 30ºC

Enables peak current reductions across all delays due to smaller transistors

Advantages Summary

	Equal delay	Equal transistor width	Equal driver size
Delay reduction	0%	19%	22%
Total transistor width reduction	32%	0%	-20%
Peak current reduction	49%	30%	17%
Energy increase	9%	16%	19%

- Averaged over 3-9mm buses
- Metal3 in 130nm technology, 1.2V, 30ºC

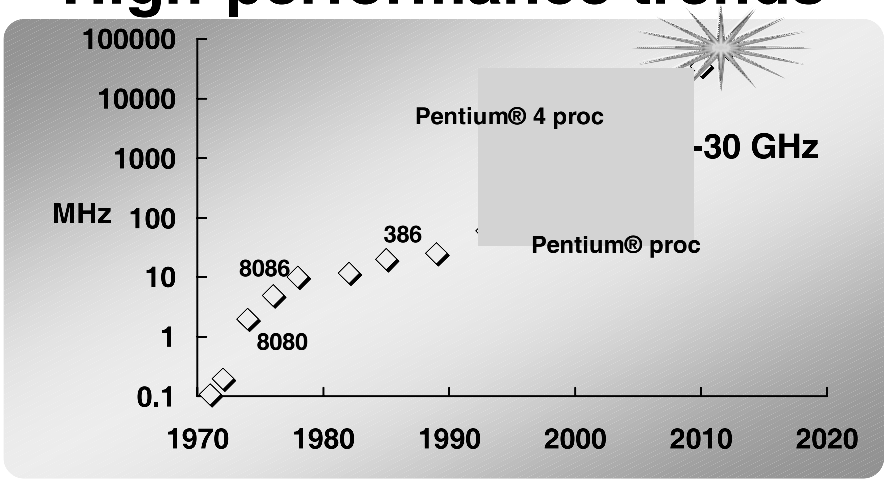
Summary

- Performance demand continues, barriers: Power, Leakage, Interconnect, Variations
- To get E.D = $(0.7)^4$, V_t has to scale aggressively
- Active & standby leakage reduction strategies
- Leakage-tolerant bitline technologies
- Process variation tolerant circuit technologies
- High-speed SF & transition encoded interconnects

Outline

- **Motivation**
- **High-performance Adders**
  - **- 6.5GHz 32-bit Han-Carlson ALU**
  - **- 4GHz 32-bit Address Generation Unit**
- **Low-power Multiplier**
  - **- 1 GHz 16x16 multiplier**

High-performance trends

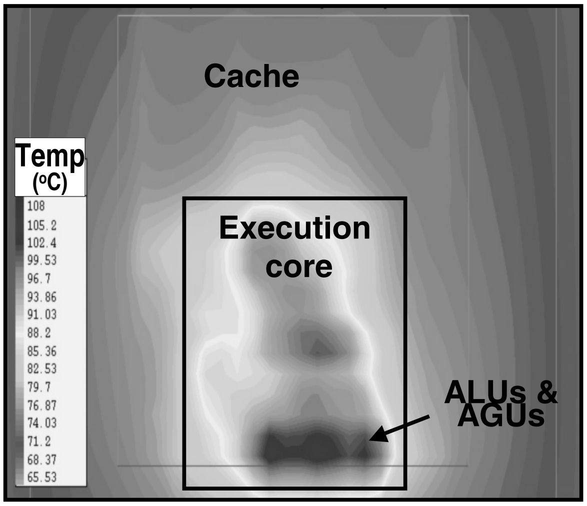


- Frequency doubles every generation
- Performance-critical units
  - ALUs & AGUs
  - Register files, L0 Caches

} Single-cycle throughput & latency

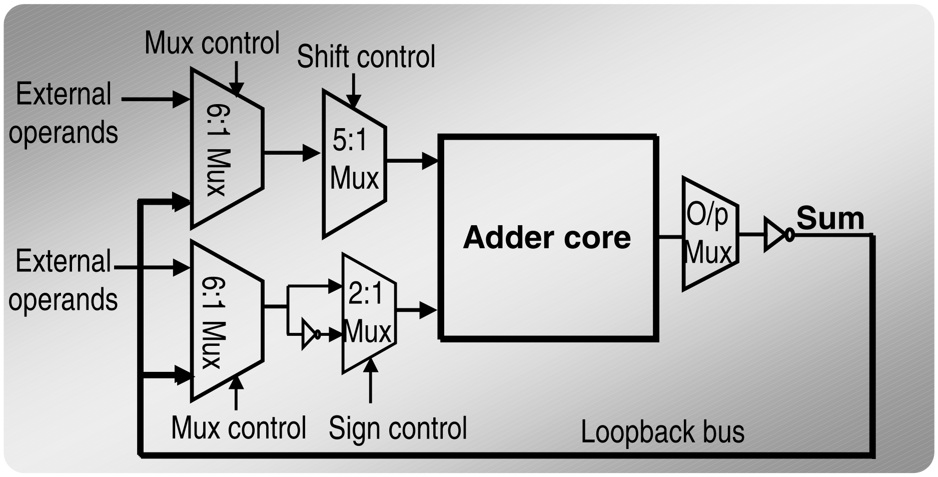
Motivation

Processor thermal map



- ALUs: performance and peak-current limiters
- High activity ⇒ thermal hotspot
- Goal: high-performance energy-efficient design

32-bit ALU architecture



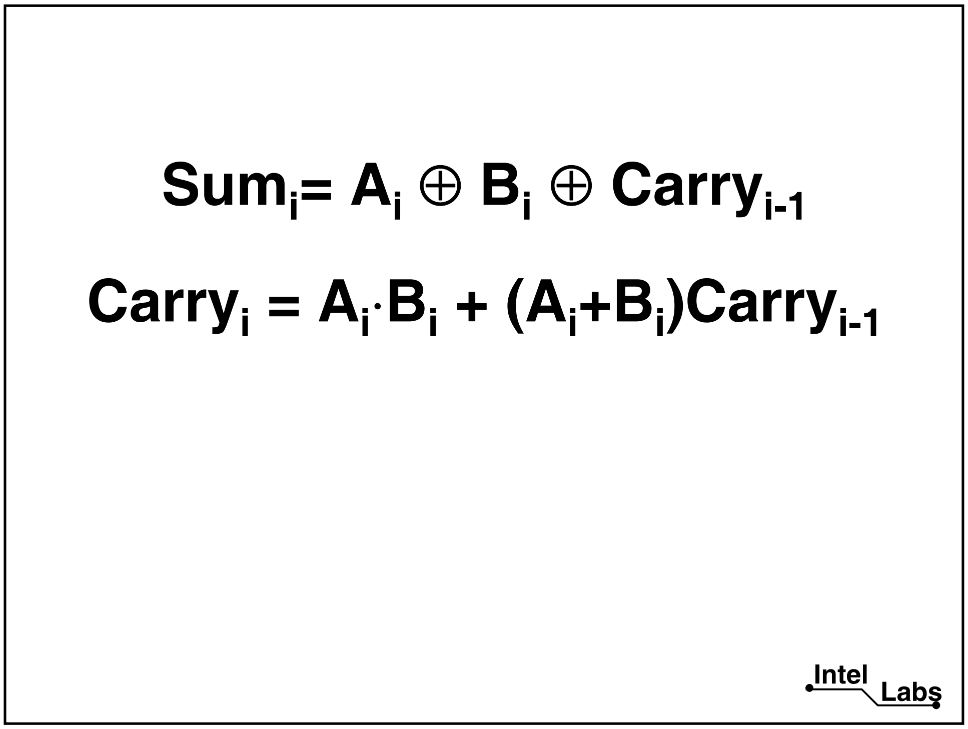
Multiple ALUs clustered together in the execution core ⇨High power density

79

A 6.5GHz, 130nm Single-ended Dynamic ALU

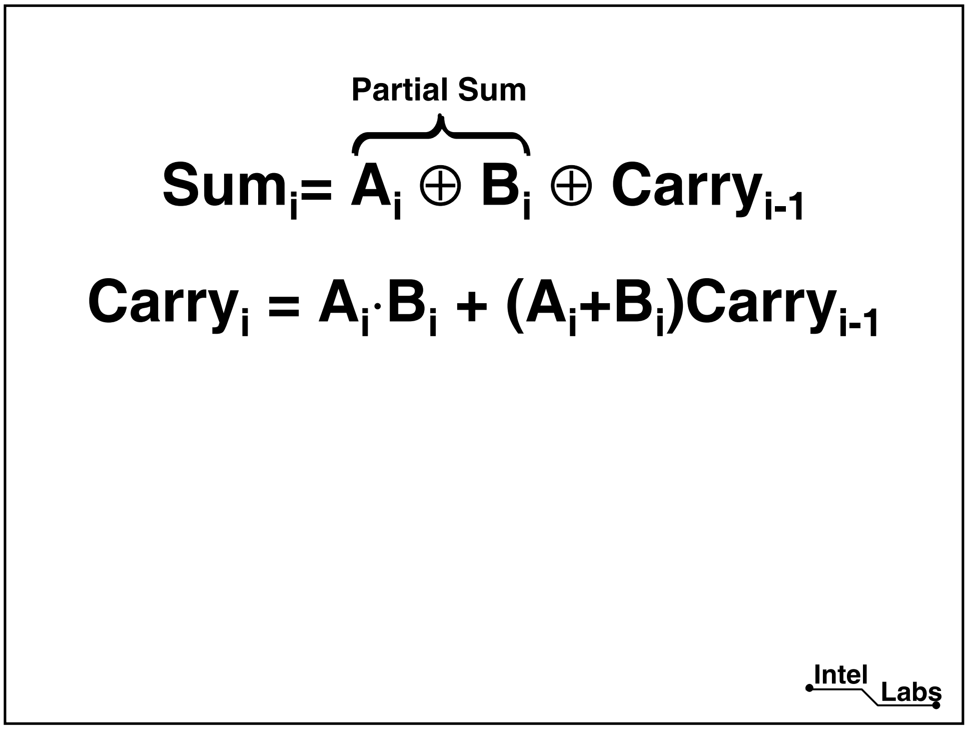
[M. Anders et al, ISSCC 2002]

Intel Labs



$$Sum_i = A_i \oplus B_i \oplus Carry_{i-1}$$

$$Carry_i = A_i \cdot B_i + (A_i + B_i)Carry_{i-1}$$



$$Sum_i = \overbrace{A_i \oplus B_i}^{\text{Partial Sum}} \oplus Carry_{i-1}$$

$$Carry_i = A_i \cdot B_i + (A_i + B_i)Carry_{i-1}$$

$$\mathrm{Sum}_i = \overbrace{A_i \oplus B_i}^{\text{Partial Sum}} \oplus \mathrm{Carry}_{i-1}$$

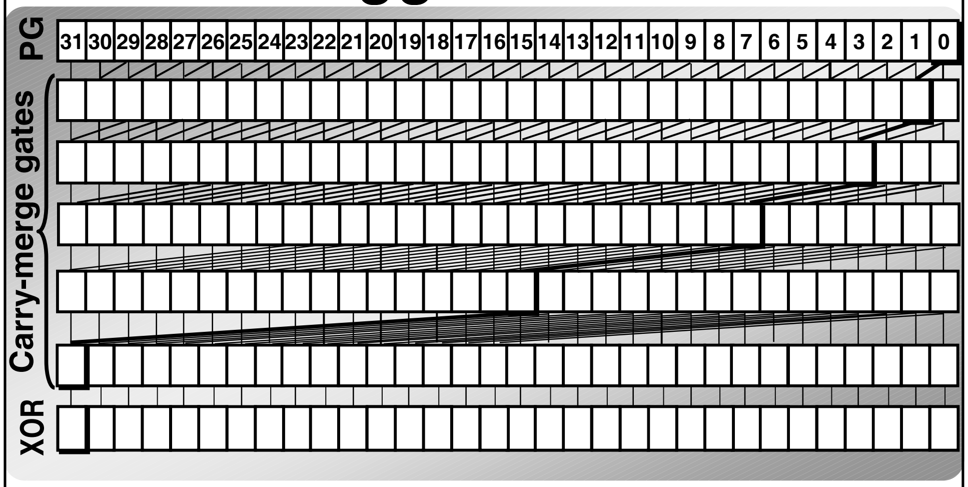
$$\mathrm{Carry}_i = \underbrace{A_i \cdot B_i}_{\text{Generate}} + \underbrace{(A_i + B_i)}_{\text{Propagate}}\mathrm{Carry}_{i-1}$$

$$\mathrm{Sum}_i = \overbrace{A_i \oplus B_i}^{\text{Partial Sum}} \oplus \mathrm{Carry}_{i-1}$$

$$\mathrm{Carry}_i = \underbrace{A_i \cdot B_i}_{\text{Generate}} + \underbrace{(A_i + B_i)}_{\text{Propagate}}\mathrm{Carry}_{i-1}$$

$$\mathrm{Carry}_i = \quad G_i \quad + \quad P_i \cdot \mathrm{Carry}_{i-1}$$

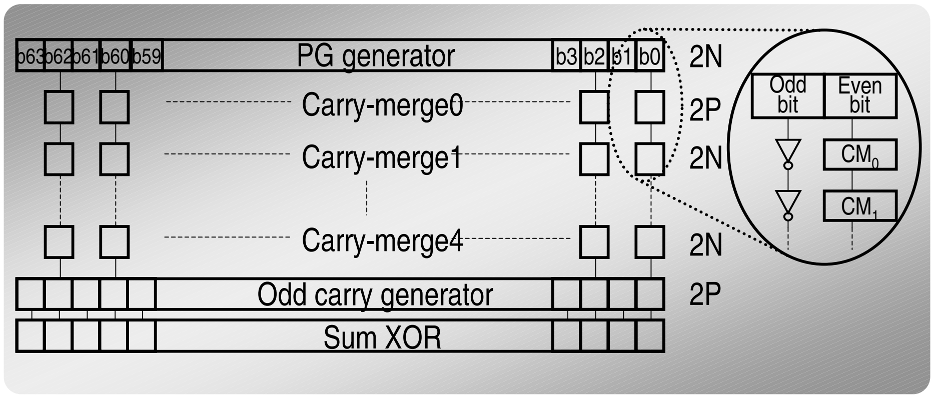
32-bit Kogge-Stone Adder



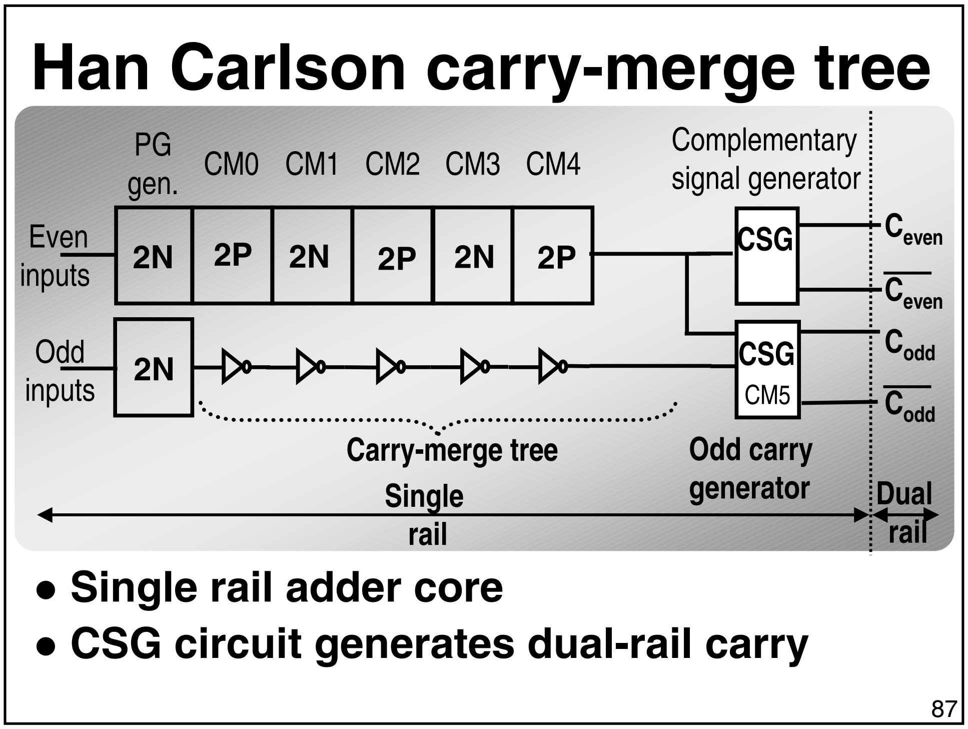
- Critical path = PG+5+XOR = 7 gate stages
- Generate,Propagate fanout of 2,3
- Maximum interconnect spans 16b

Energy inefficient

32-bit Han-Carlson adder core

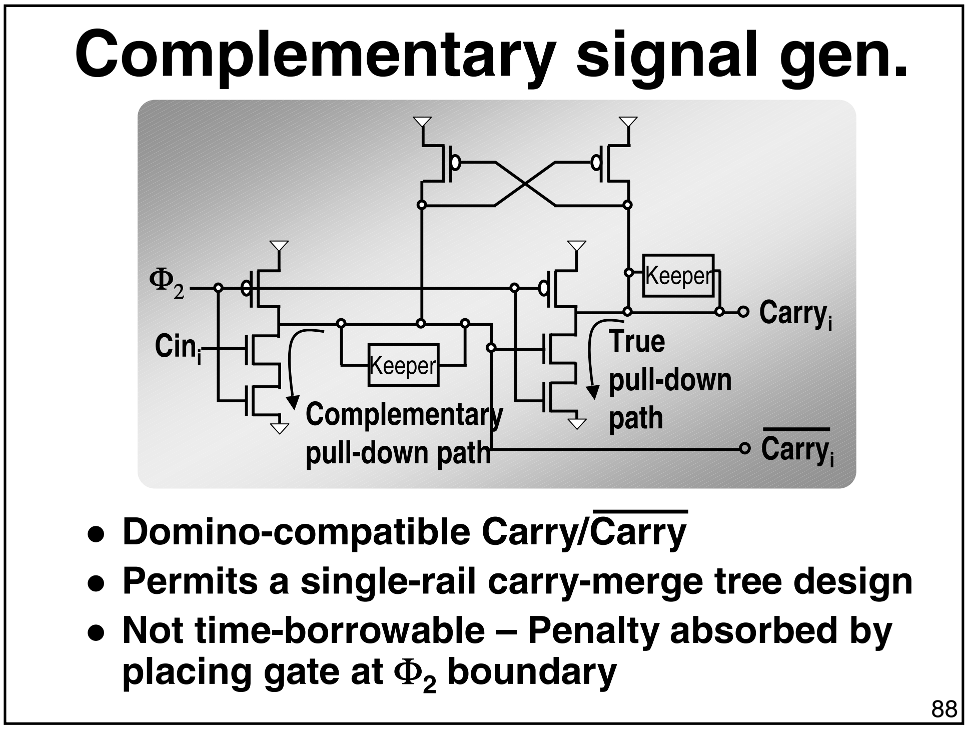


- Carry-merge done on even bitslices
- 50% fewer carry-merge gates vs Kogge-Stone
- Extra logic stage generates odd carries



Han Carlson carry-merge tree

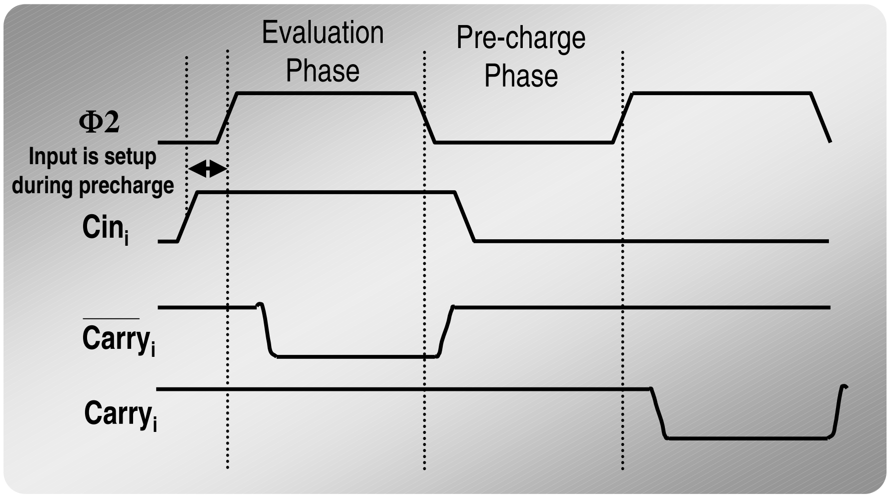
- Single rail adder core
- CSG circuit generates dual-rail carry



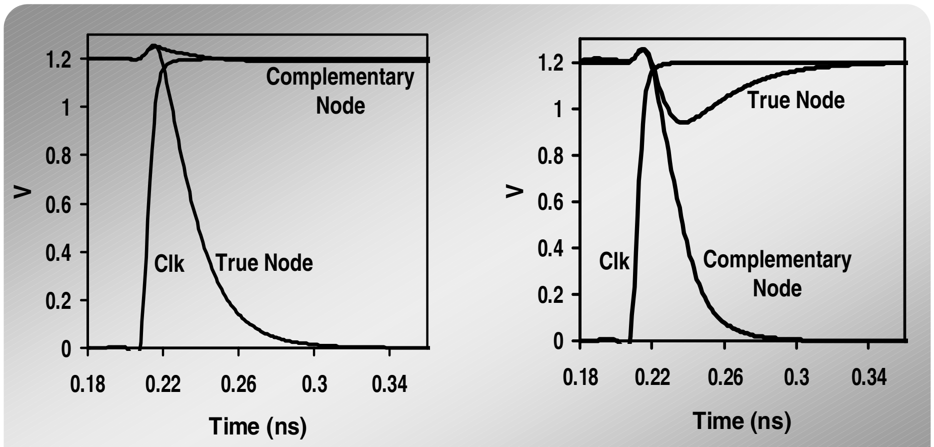
Complementary signal gen.

- Domino-compatible Carry/$\overline{\text{Carry}}$
- Permits a single-rail carry-merge tree design
- Not time-borrowable – Penalty absorbed by placing gate at Φ_2 boundary

CSG: Timing Diagram

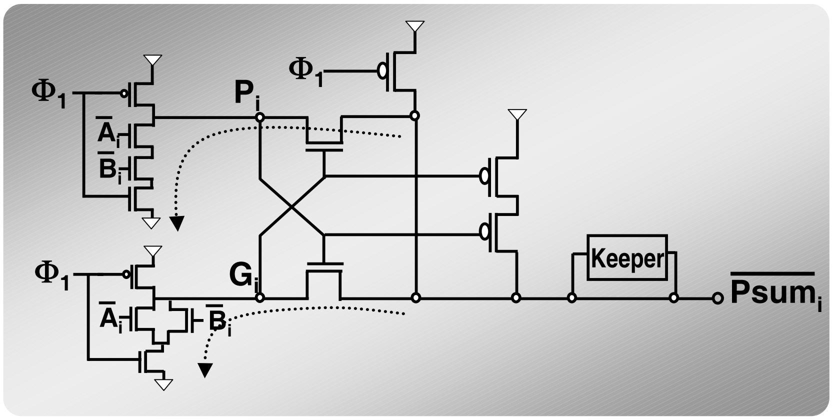


CSG: Simulation waveforms



(a) Cin=0 (b) Cin=1

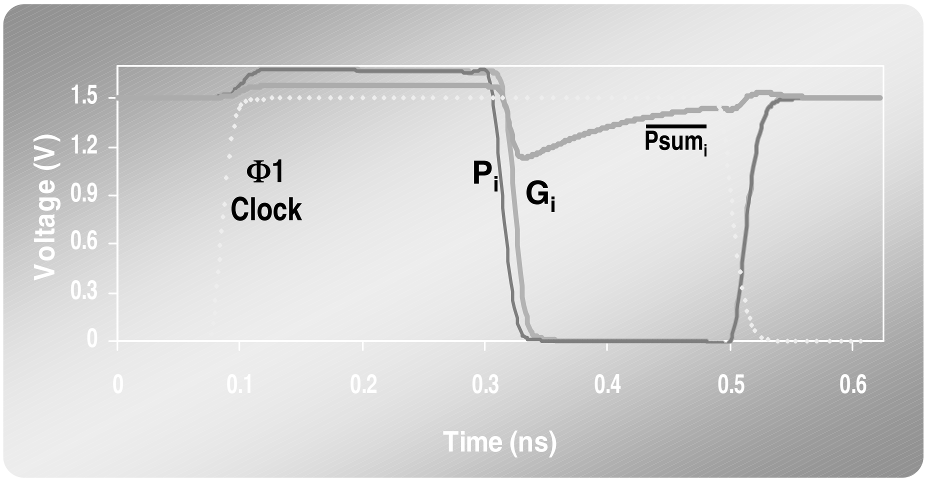
Partial sum generator



- Generates domino-compatible partial sum
- Placing the gate at Φ_1 boundary mitigates output noise-glitches

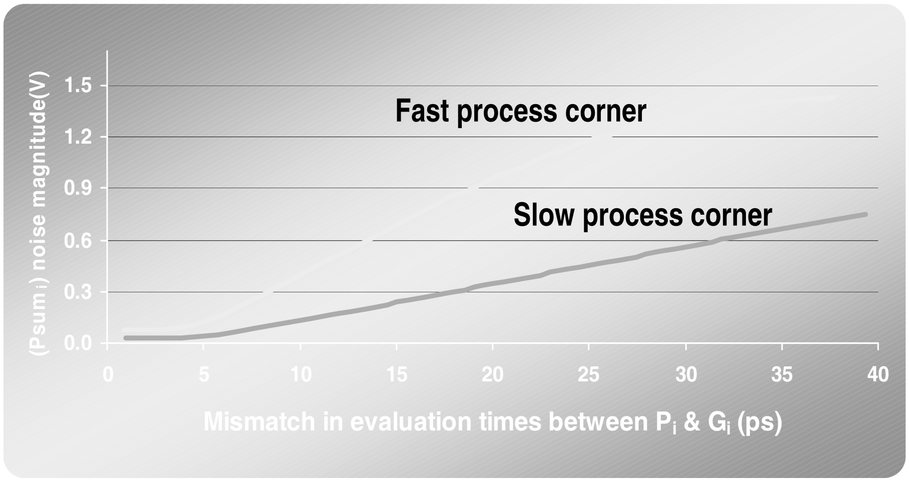
91

Dynamic XNOR: Simulation Waveforms



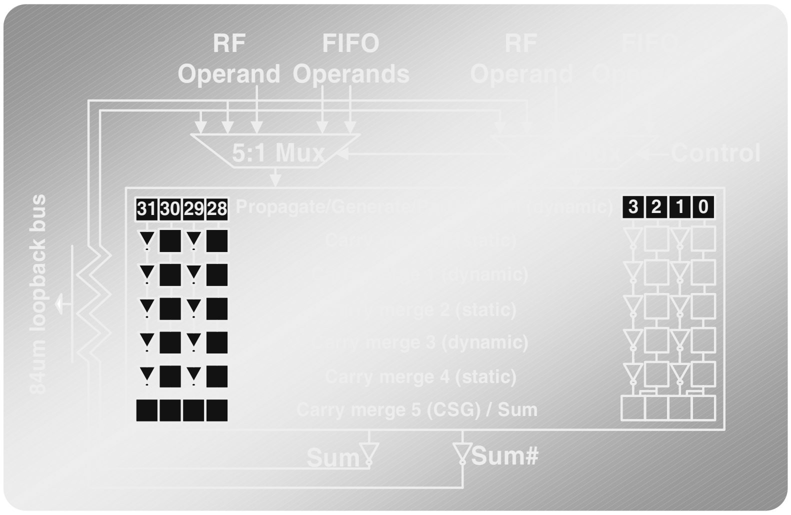
92

Dyn. XNOR: Noise Sensitivity



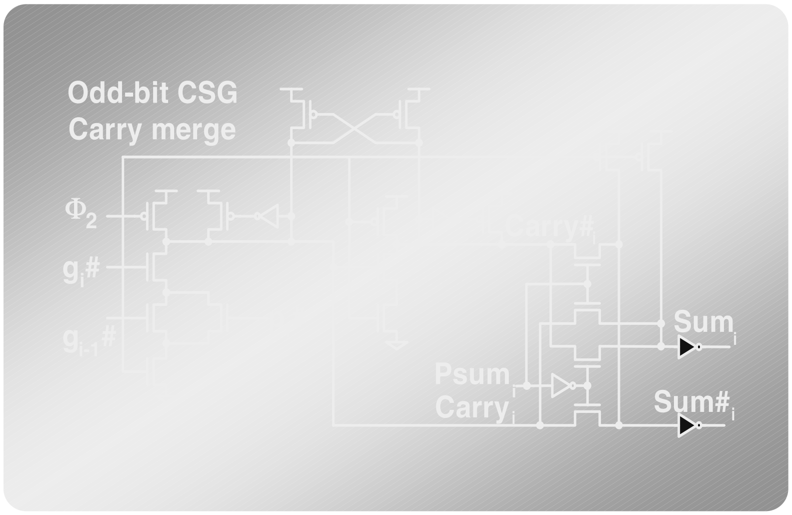
- Mismatch in input evaluation times can cause output noise

Han-Carlson ALU Organization



- Single-rail dynamic 9-stage low-V_t design

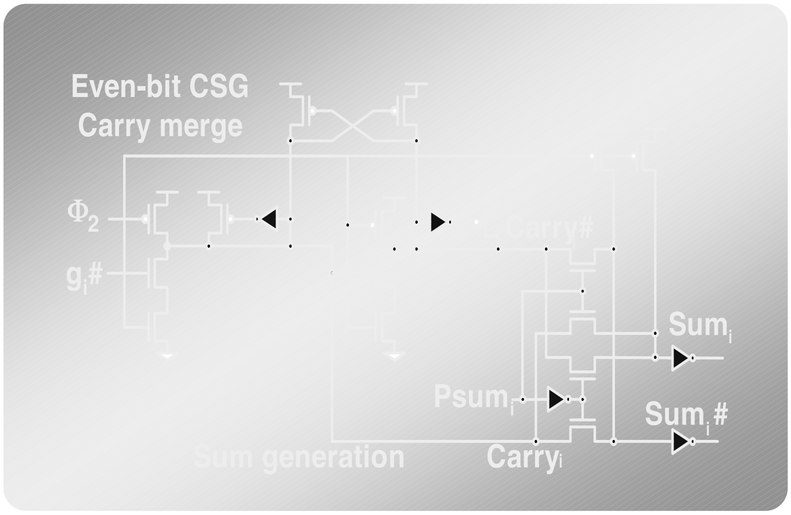
Odd-bits CSG Sum Generation



- Final carry-merge CSG(dual-rail carry output)
  → pass-transistor sum XOR

95

Even-bits CSG Sum Generation

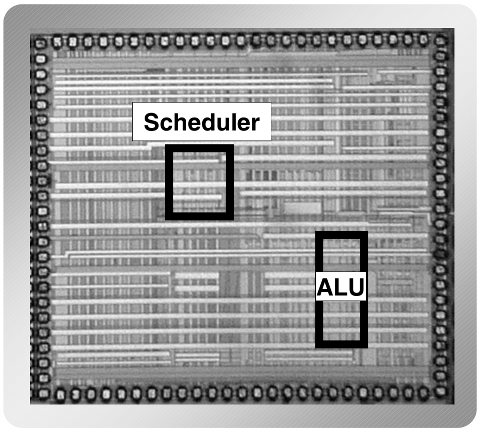


- Domino-compatible sum
- Dual-rail sum from single-ended g inputs

96

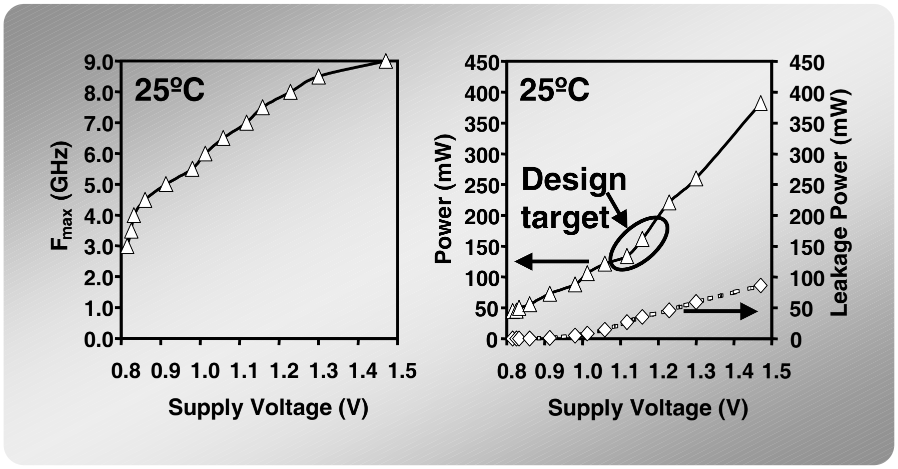
Die Micro-photograph

- 130nm 6-metal dual-Vt CMOS
- Scheduler:
  - 210µm x 210µm
- ALU:
  - 84µm x 336µm



97

Delay and Power Measurements



- 6.5GHz at 1.1V, 25ºC
- Power: 120mW total, 15mW leakage
- Scalable to 10GHz at 1.7V, 25ºC

98

Improvements Over Dual-rail Domino

Area	50%
Performance (Delay)	10%
Active Leakage	40%
Robustness	equal

- Leakage reduced by eliminating dual-rail logic
- Robustness not compromised
- CSG improves both area and performance

A 4GHz 130nm Address Generation Unit with 32-bit Sparse-tree Adder Core

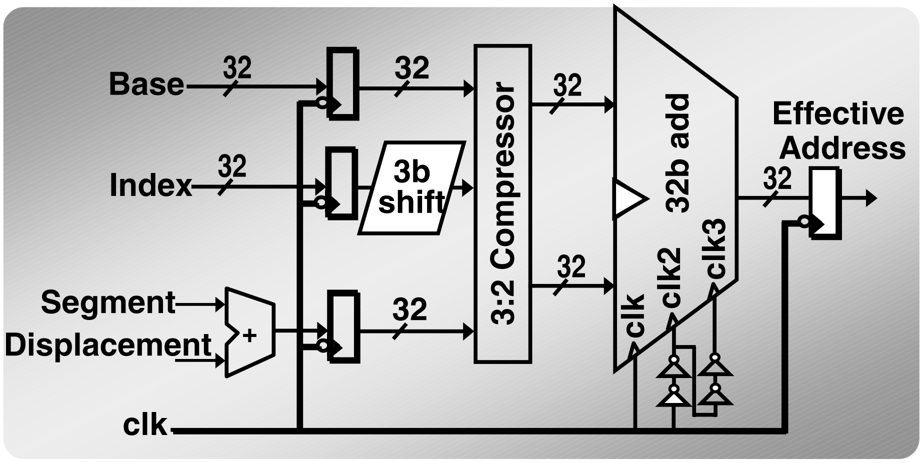
[S. Mathew et al, VLSI Symp. 2002]

Intel Labs

Outline

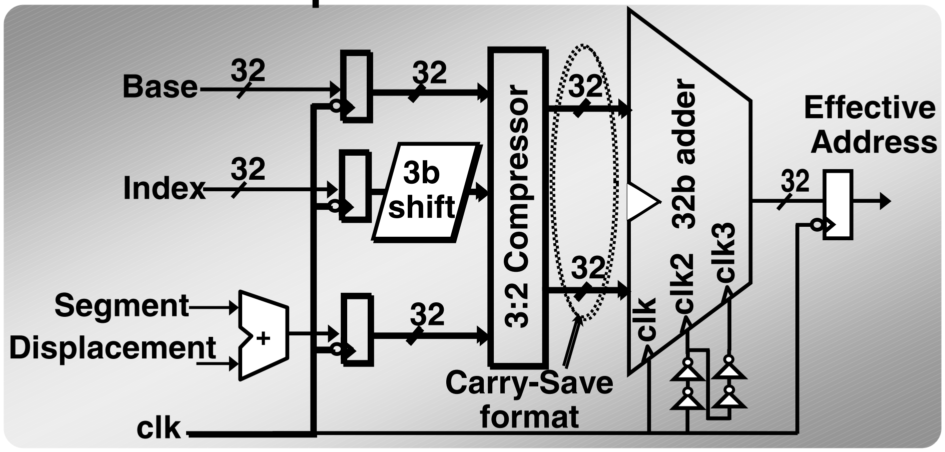
- Address Generation Unit (AGU) organization
- Sparse-tree adder core
- Dual-V_t semi-dynamic design
- Sub-130nm scaling trends
- Summary

AGU Architecture



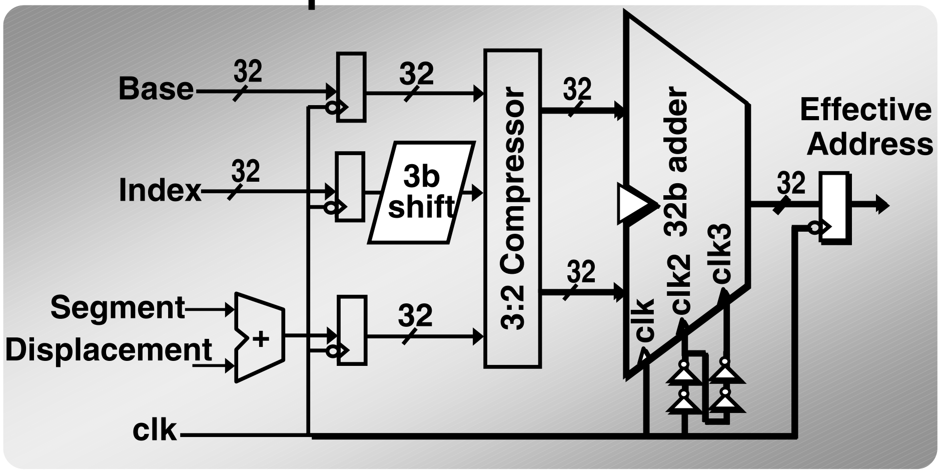
- Single-cycle latency and throughput
- Effective Address = Base + Index*Scale + (Segment +Displacement)
- 2-phase address computation

AGU Operation: Phase 1



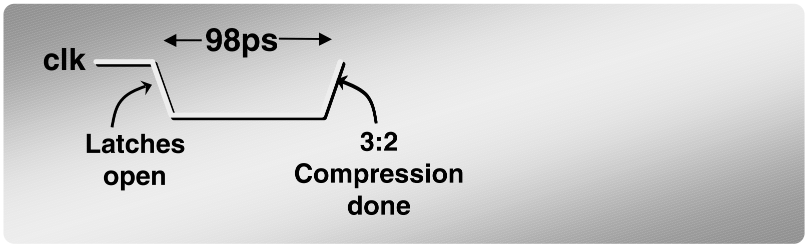
- Index pre-scaled via 3-bit barrel shifter
- 3:2 compressor renders partial address:
  - Carry-save format
- Adder in pre-charge state

AGU Operation: Phase 2

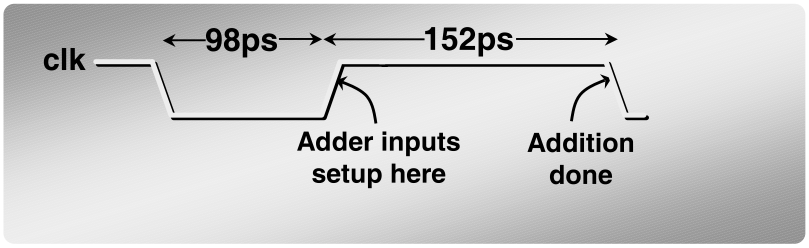


- Carry-save to binary format conversion:
  - 2’s complement parallel 32-bit adder

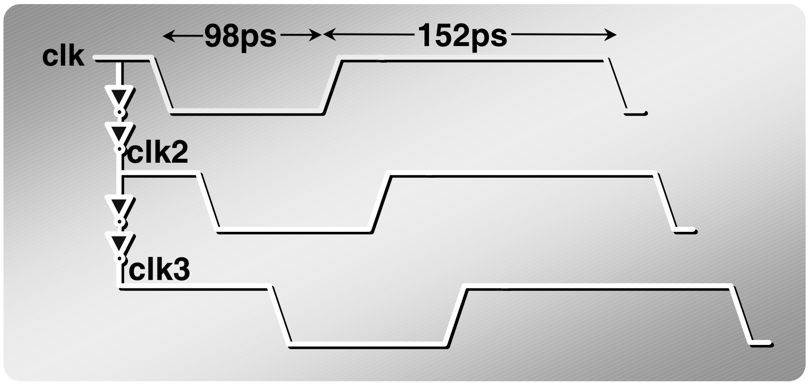
Timing Diagram



Timing Diagram

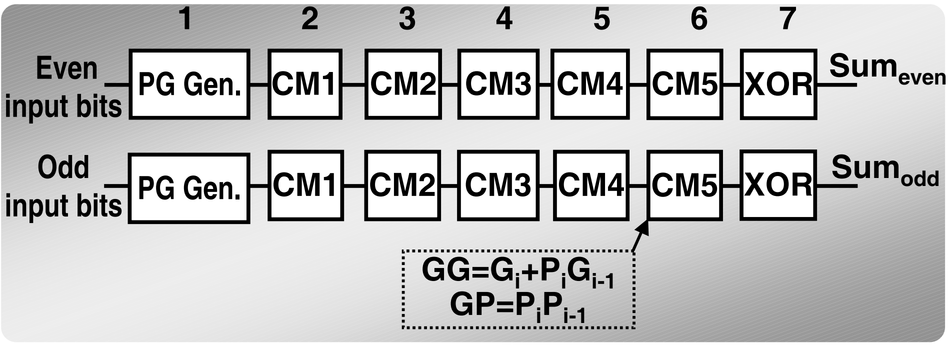


Timing Diagram



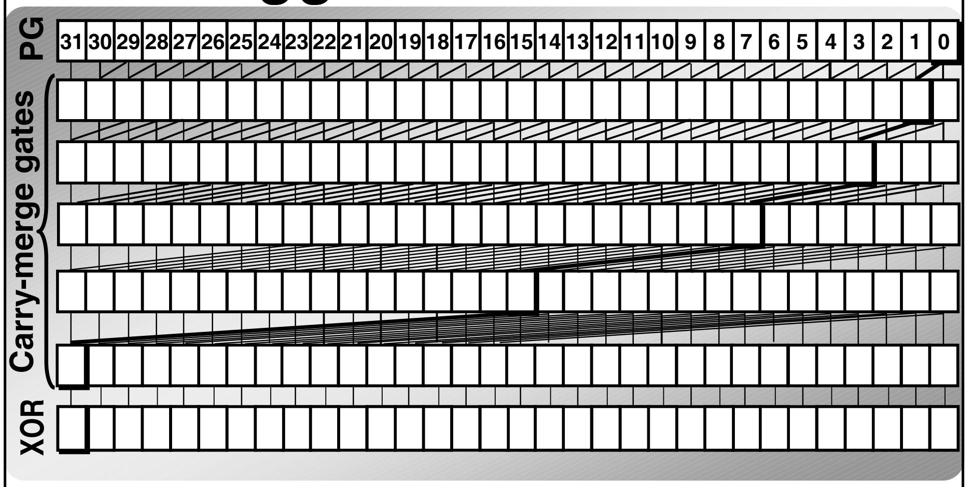
- Seamless time-borrowable clock boundaries
- 152ps (6.6GHz) 32-bit adder core required

High-performance Adders: Kogge Stone



- Generate all 32 carries:
  - Full-blown binary tree ⇒ energy-inefficient
- # Carry-merge stages = $\log_2(32)$ ⇒ 5 stages

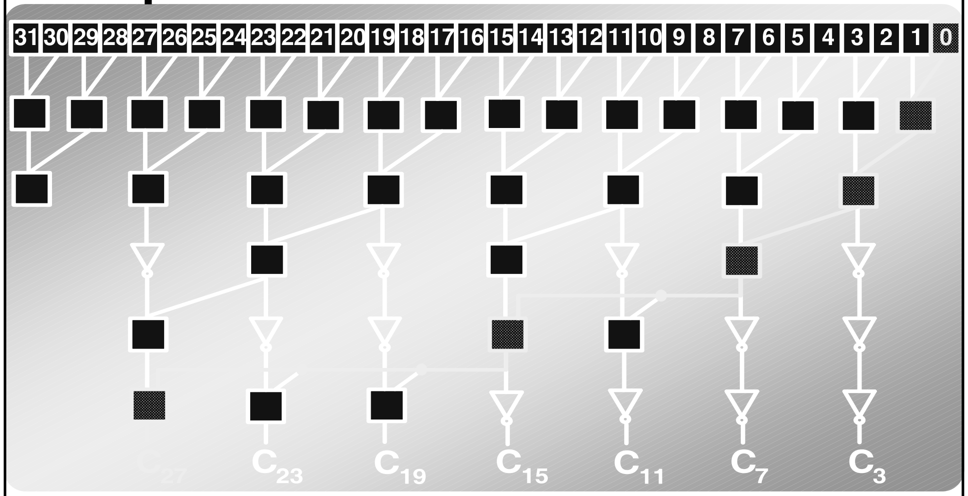
Kogge-Stone Adder



- Critical path = PG+5+XOR = 7 gate stages
- Generate,Propagate fanout of 2,3
- Maximum interconnect spans 16b

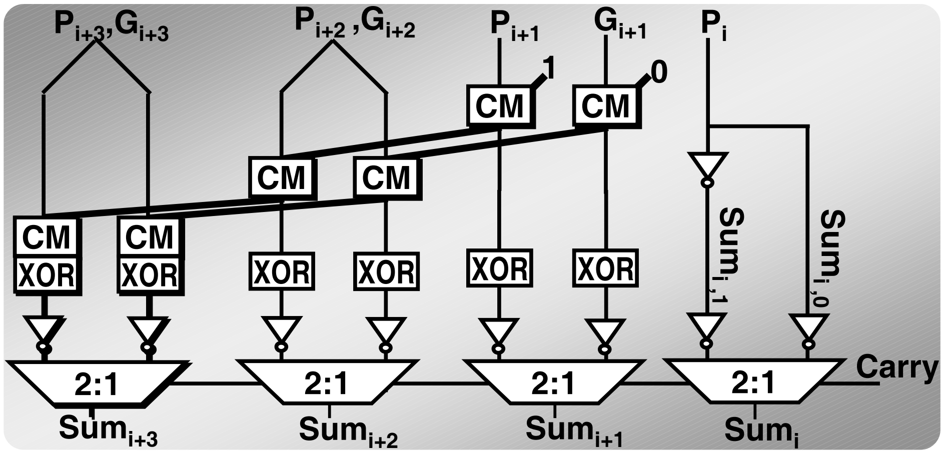
} Energy inefficient

Sparse-tree Adder Architecture



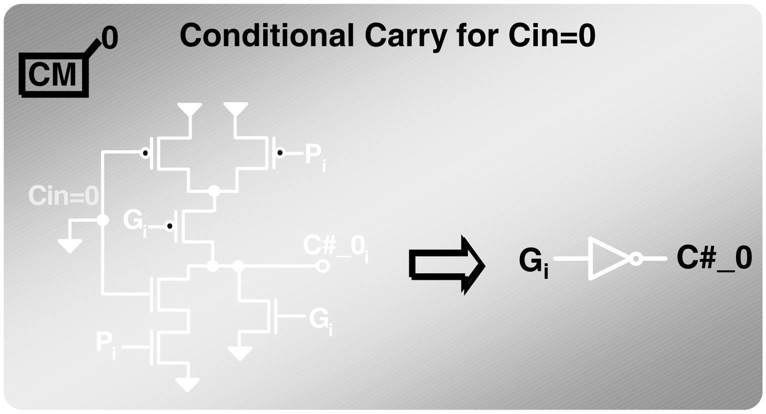
- Generate every 4th carry in parallel
- Side-path: 4-bit conditional sum generator
- 73% fewer carry-merge gates⇒energy-efficient

Non-critical Sum Generator

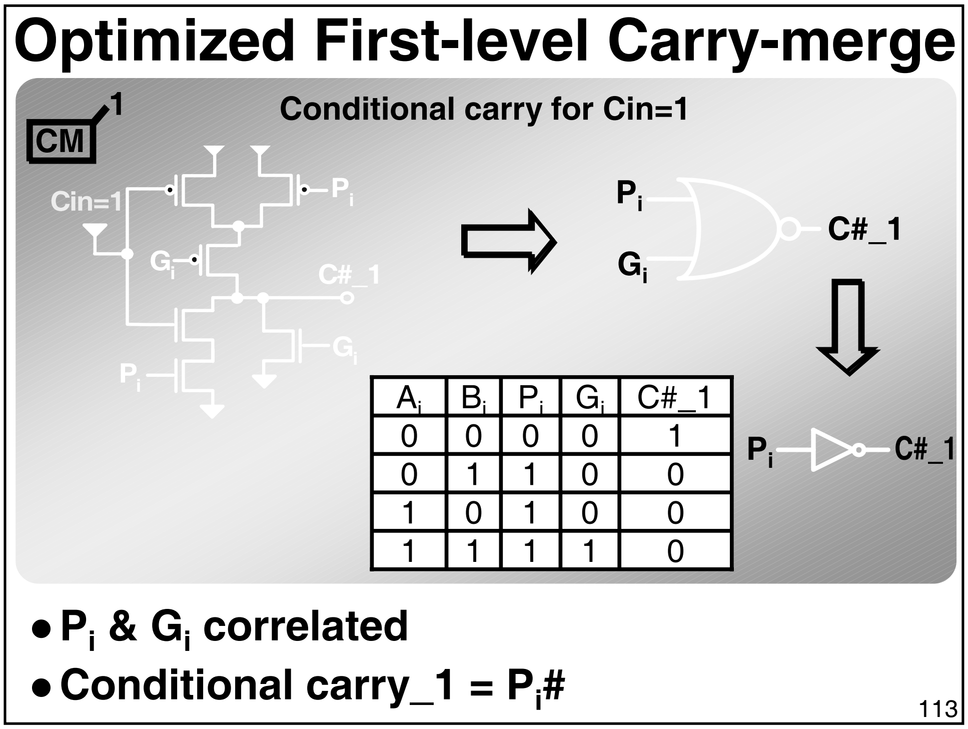


- Non-critical path: ripple carry chain
- Reduced area, energy consumption, leakage
- Generate conditional sums for each bit
- Sparse-tree carry selects appropriate sum

Optimized First-level Carry-merge



- Carry-merge stage reduces to inverter
- Conditional carry_0 = G_i#

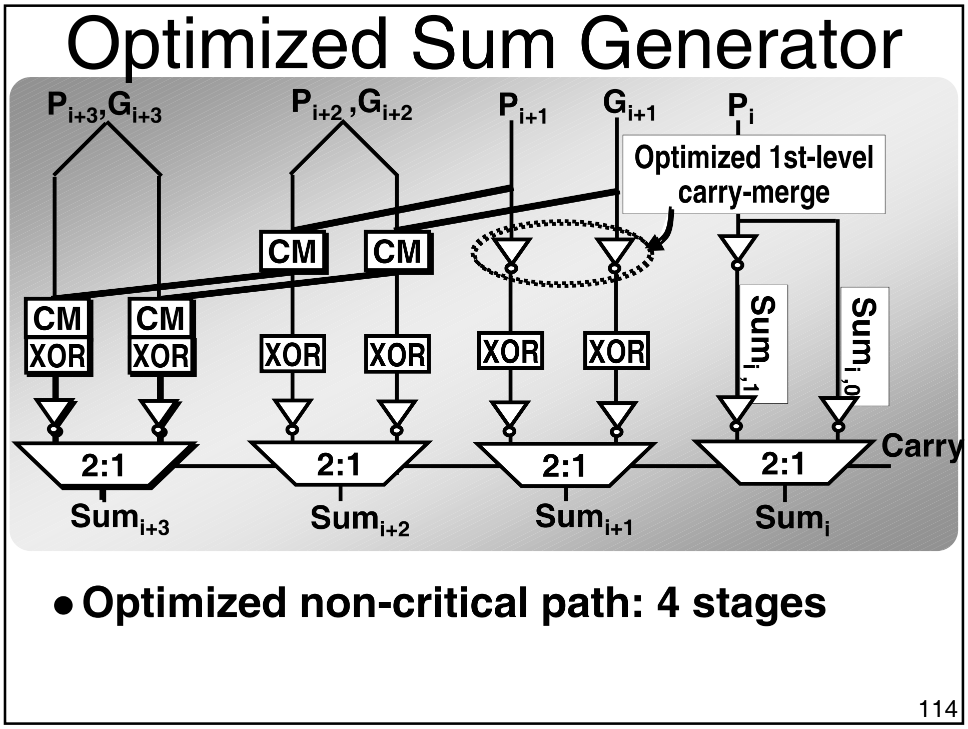


Optimized First-level Carry-merge

Conditional carry for Cin=1

A_i	B_i	P_i	G_i	C#_1
0	0	0	0	1
0	1	1	0	0
1	0	1	0	0
1	1	1	1	0

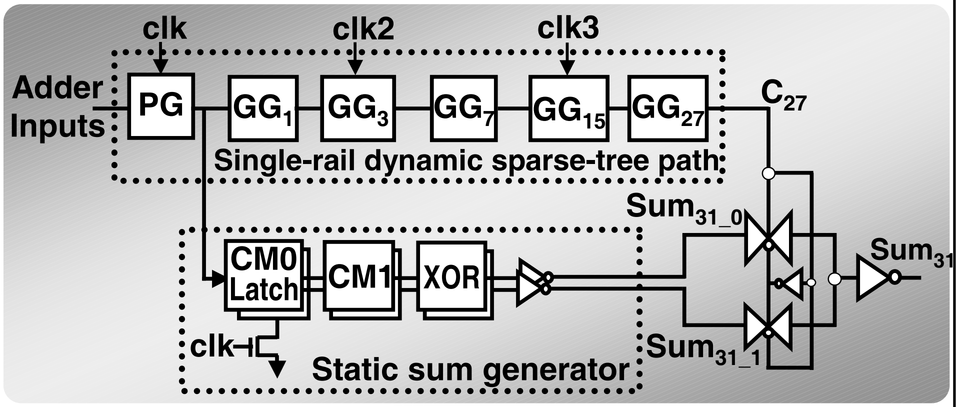
- **P_i & G_i correlated**
- **Conditional carry_1 = P_i#**



Optimized Sum Generator

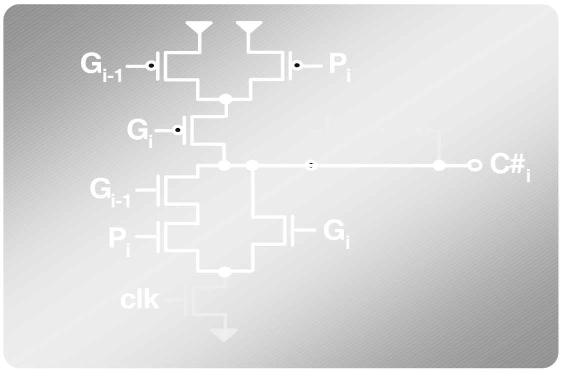
- **Optimized non-critical path: 4 stages**

Adder Core Critical Path



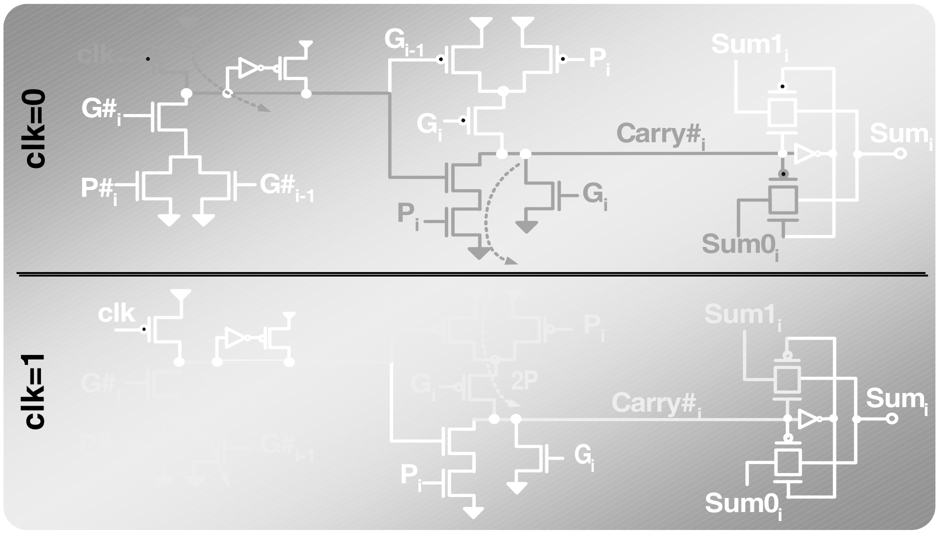
- Critical path: 7 gate stages ⇒ same as KS
- Sparse-tree: single-rail dynamic
- Exploit non-criticality of sum generator
- Convert to static logic⇒Semi-dynamic design

1st-level Carry-merge: Static Latch



- Holds state in pre-charge phase
- Prevents pre-charging of static stages

Domino-Static Interface



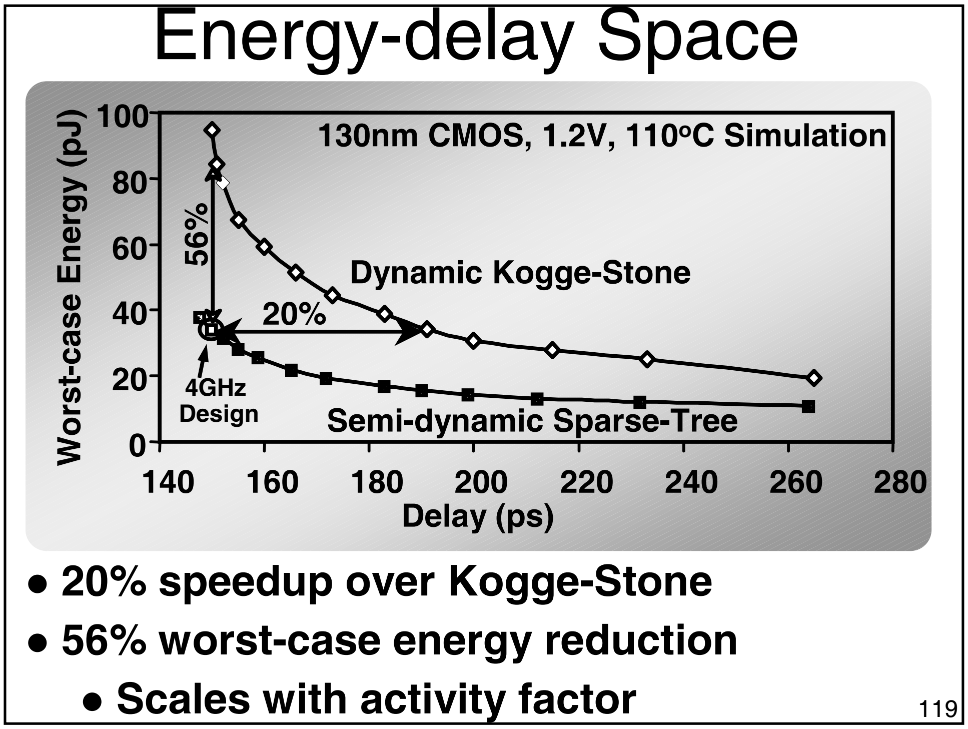
- Sum=Sum0 during pre-charge
- Mux output resolves during evaluation

117

Sparse-tree Architecture

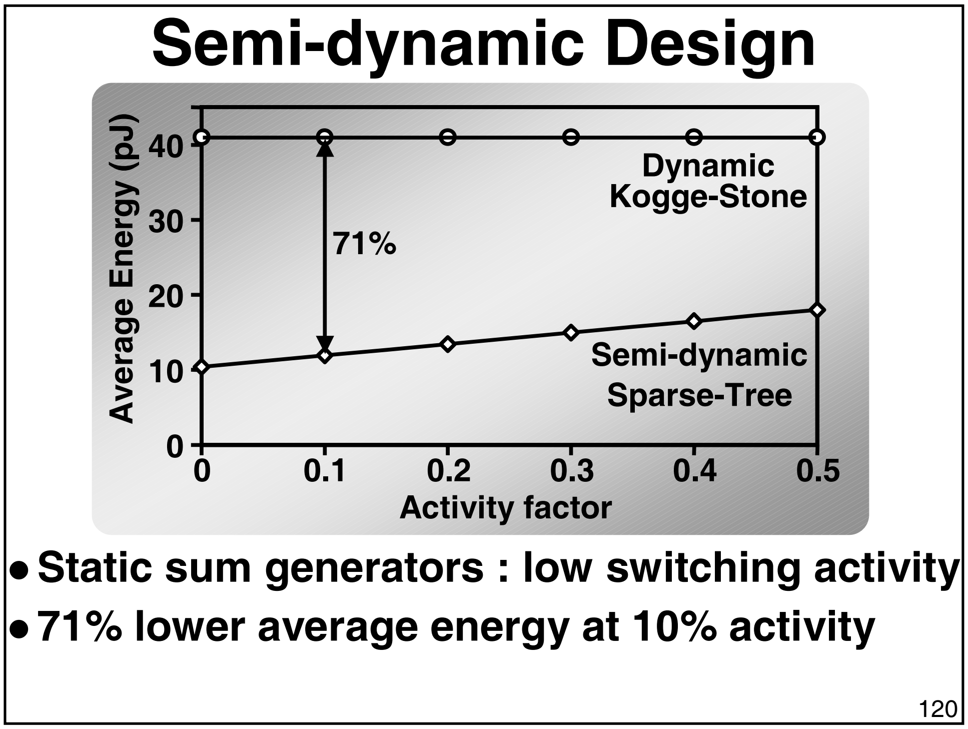
- **Performance impact: (20% speedup)**
  - **33-50% reduced G/P fanouts**
  - **80% reduced wiring complexity**
  - **30% reduction in maximum interconnect**
- **Power impact: (56% reduction)**
  - **73% fewer carry-merge gates**
  - **50% reduction in average transistor size**

118



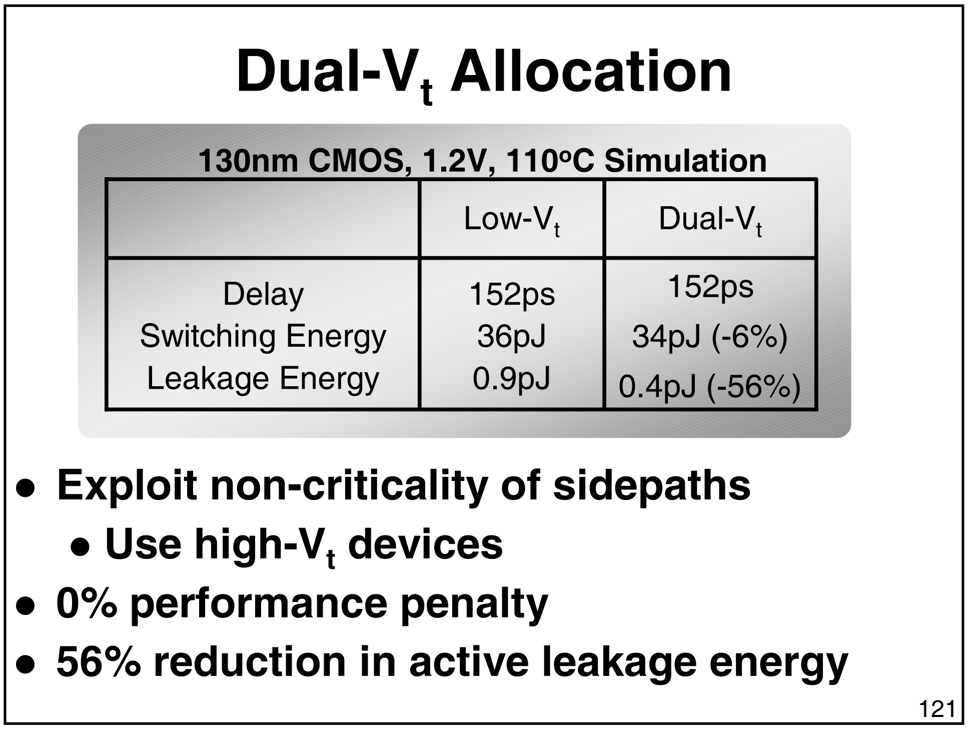
Energy-delay Space

- 20% speedup over Kogge-Stone
- 56% worst-case energy reduction
  - Scales with activity factor



Semi-dynamic Design

- Static sum generators : low switching activity
- 71% lower average energy at 10% activity

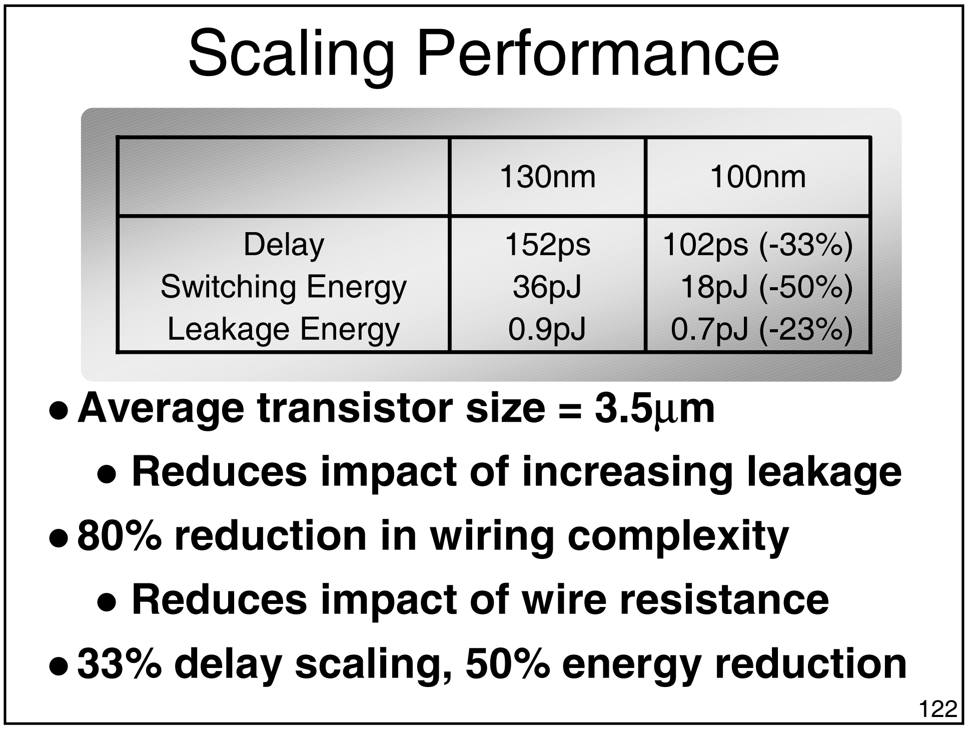


Dual-V_t Allocation

130nm CMOS, 1.2V, 110°C Simulation

	Low-V_t	Dual-V_t
Delay	152ps	152ps
Switching Energy	36pJ	34pJ (-6%)
Leakage Energy	0.9pJ	0.4pJ (-56%)

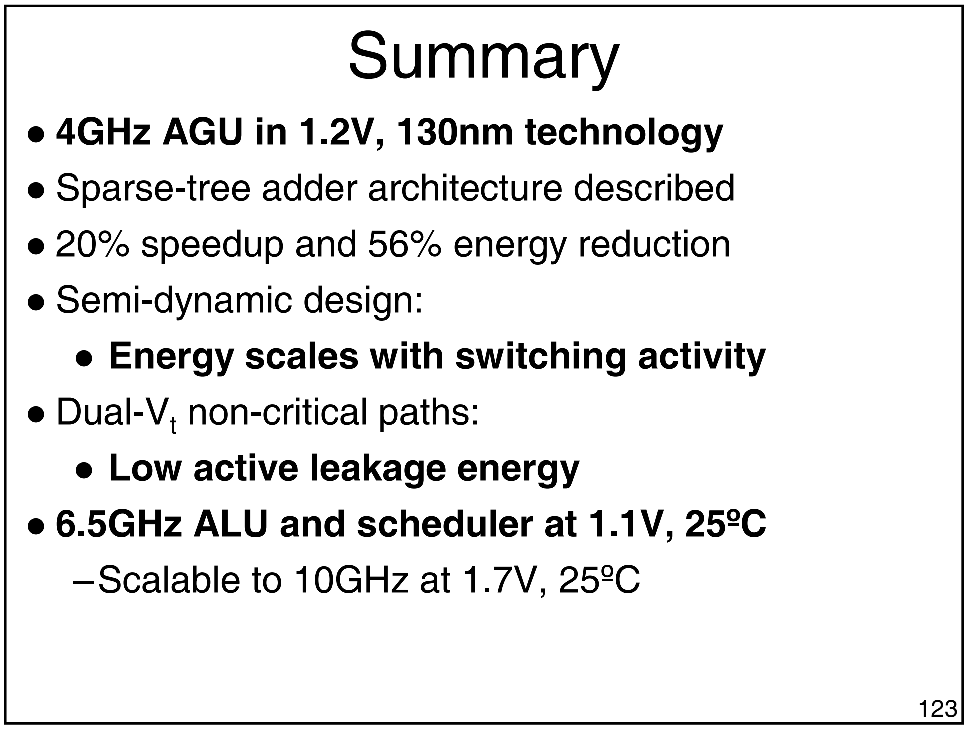
- **Exploit non-criticality of sidepaths**
  - **Use high-V_t devices**
- **0% performance penalty**
- **56% reduction in active leakage energy**



Scaling Performance

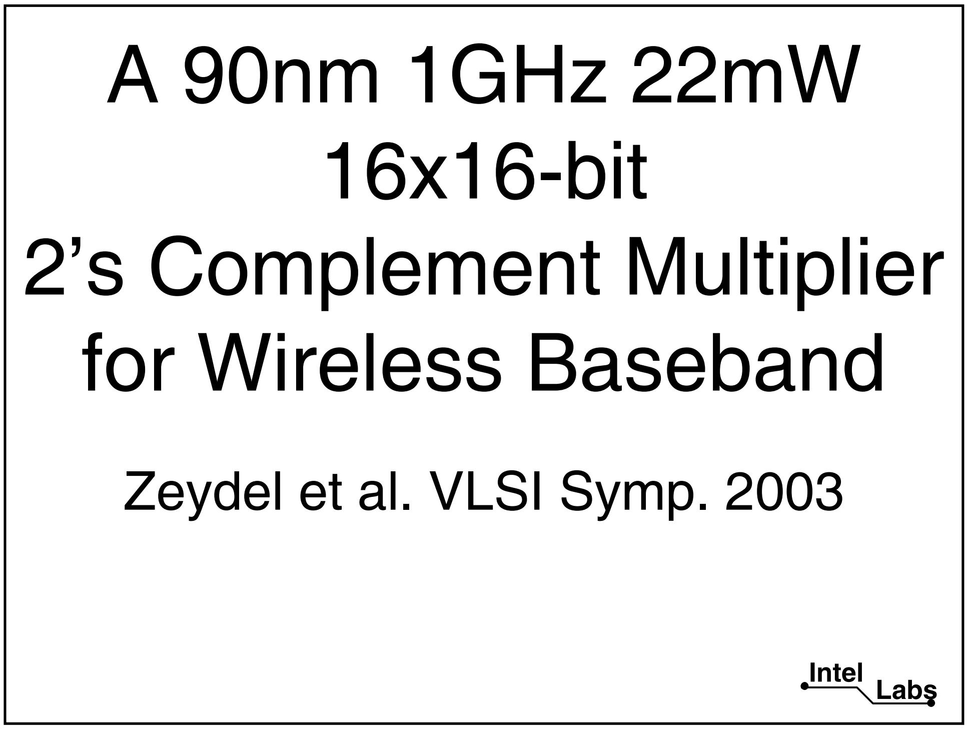
	130nm	100nm
Delay	152ps	102ps (-33%)
Switching Energy	36pJ	18pJ (-50%)
Leakage Energy	0.9pJ	0.7pJ (-23%)

- **Average transistor size = 3.5μm**
  - **Reduces impact of increasing leakage**
- **80% reduction in wiring complexity**
  - **Reduces impact of wire resistance**
- **33% delay scaling, 50% energy reduction**



Summary

- **4GHz AGU in 1.2V, 130nm technology**
- Sparse-tree adder architecture described
- 20% speedup and 56% energy reduction
- Semi-dynamic design:
  - **Energy scales with switching activity**
- Dual-V_t non-critical paths:
  - **Low active leakage energy**
- **6.5GHz ALU and scheduler at 1.1V, 25ºC**
  - –Scalable to 10GHz at 1.7V, 25ºC



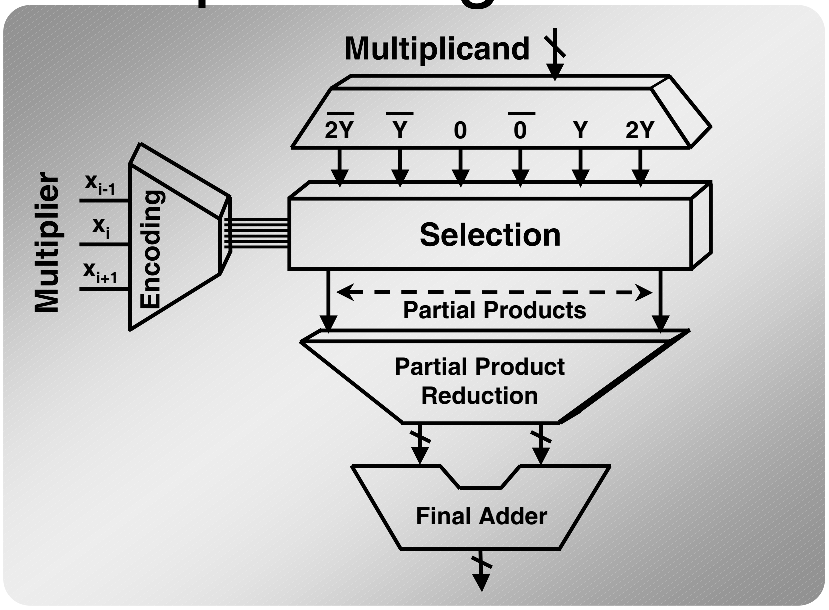
A 90nm 1GHz 22mW 16x16-bit 2’s Complement Multiplier for Wireless Baseband

Zeydel et al. VLSI Symp. 2003

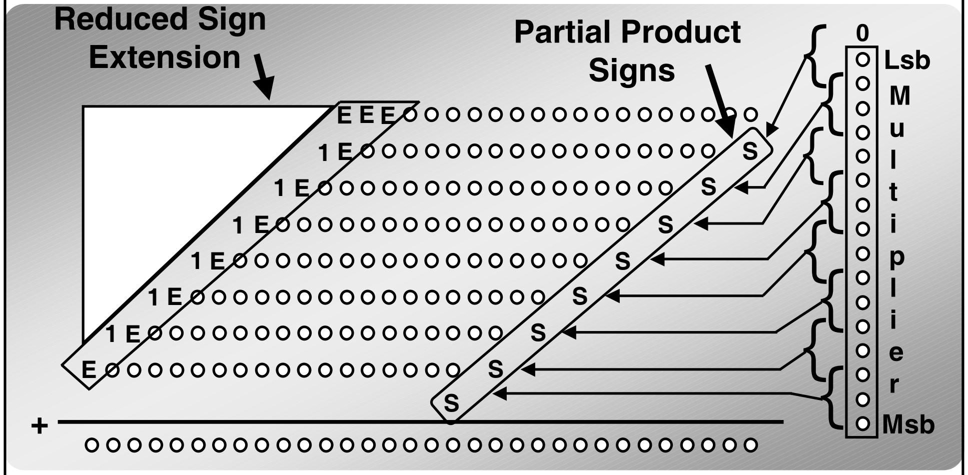
Outline

- Multiplier Block Diagram
- Booth Encoding and Select
- Optimized Partial Product Reduction
- Signal Arrival Optimized Final Adder
- 90nm Energy-Delay Results
- Conclusions

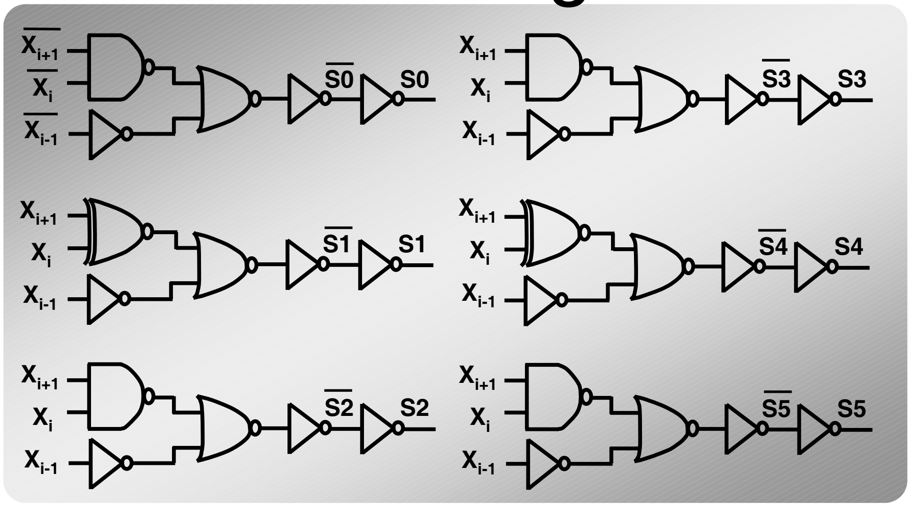
Multiplier Organization

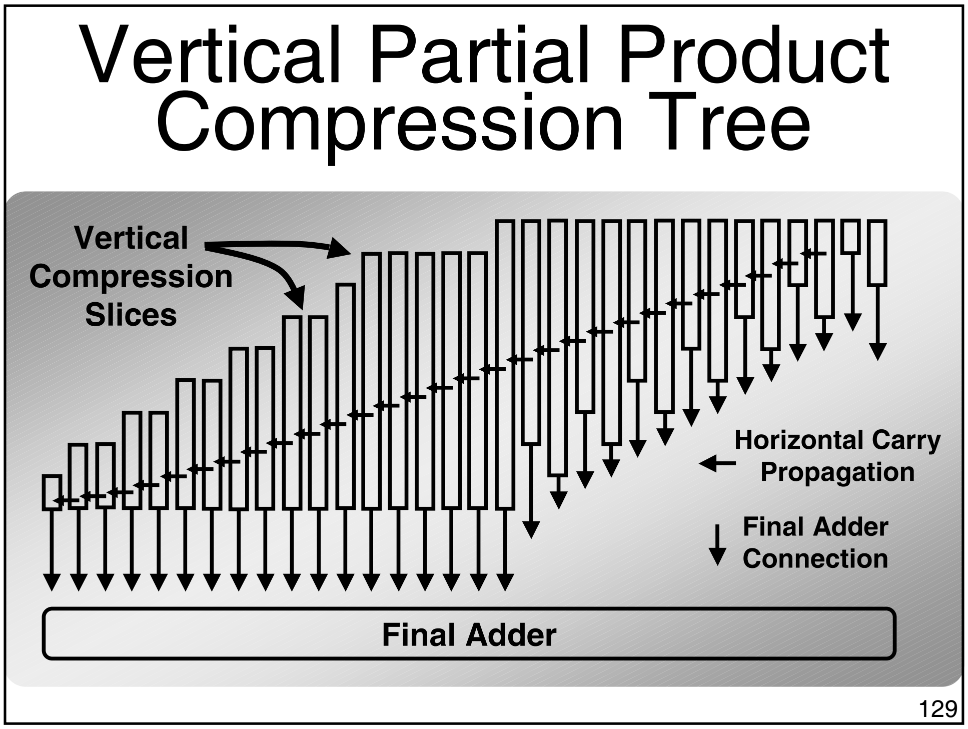


Booth Encoding and Selection



Booth Encoding Circuits





Vertical Partial Product Compression Tree

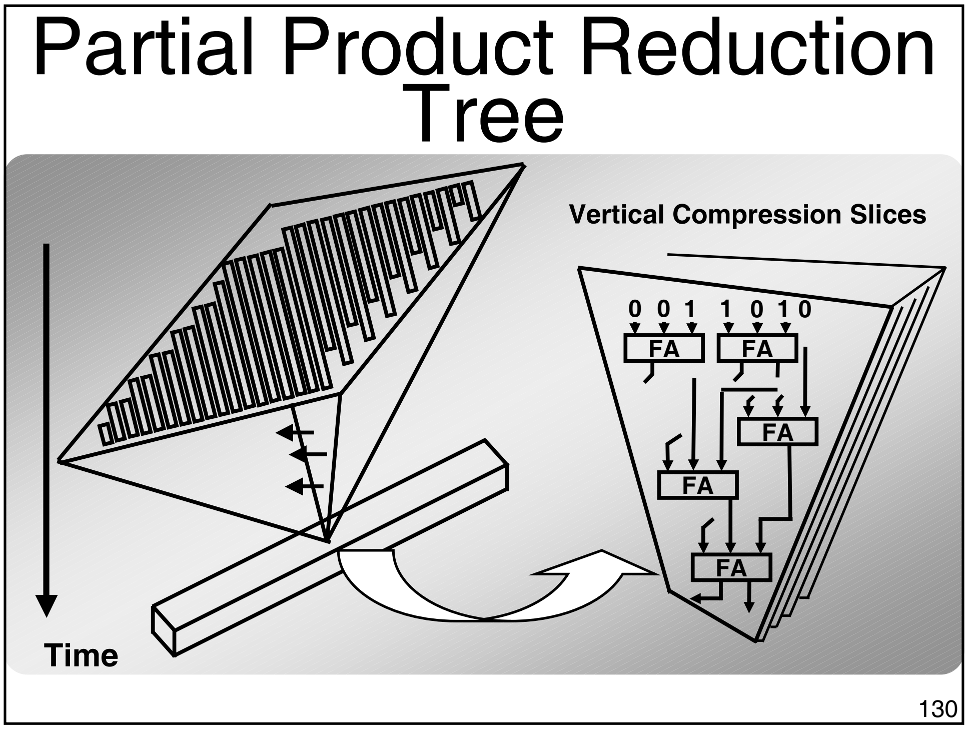
Vertical Compression Slices

Horizontal Carry Propagation

Final Adder Connection

Final Adder

129



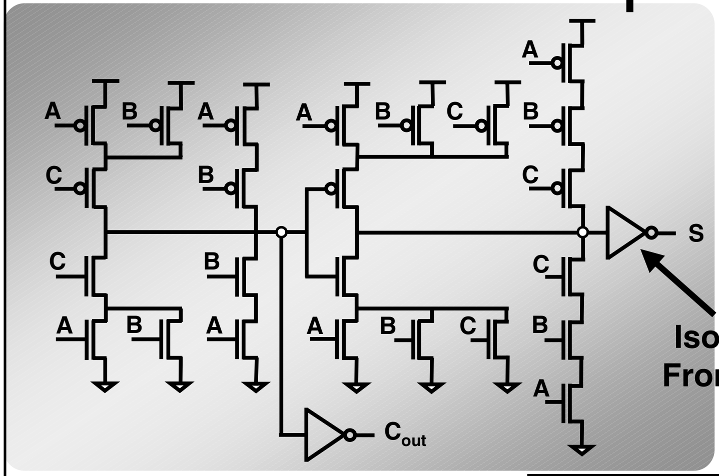
Partial Product Reduction Tree

Vertical Compression Slices

Time

130

PPRT 3:2 Compressor

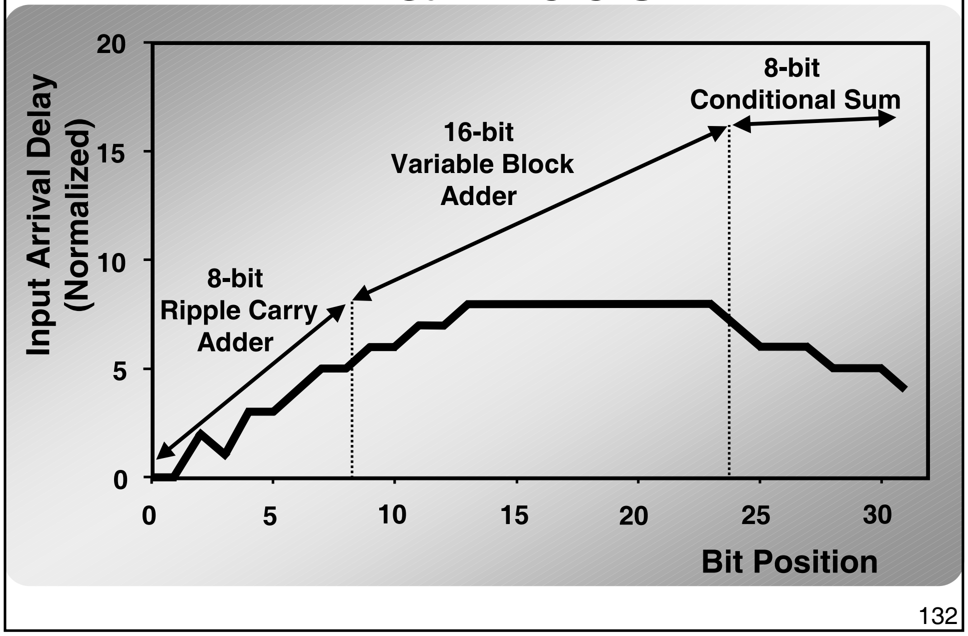


Isolates 3-stack From output load

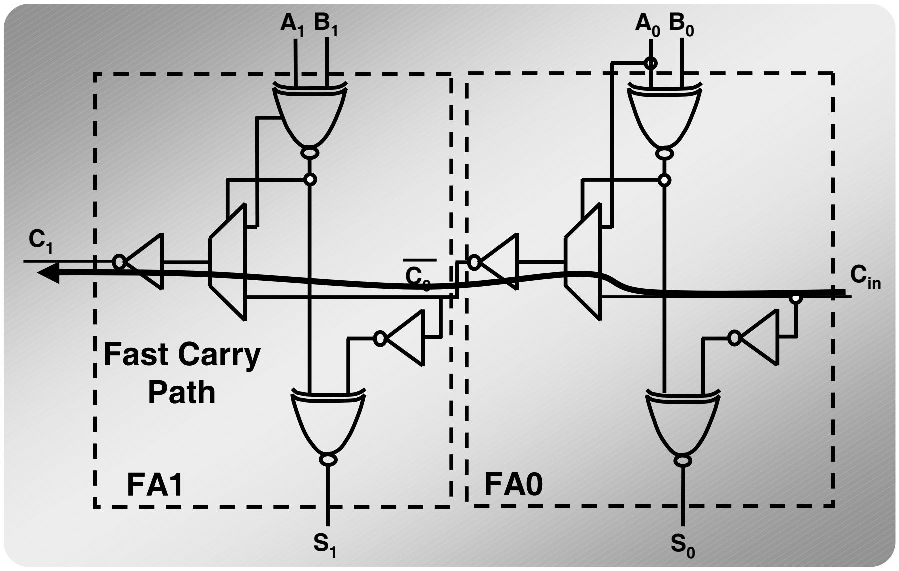
0.8V 110° C

Input	Worst-case Delay
Sum	240ps (1x)
Carry	165ps (0.69x)

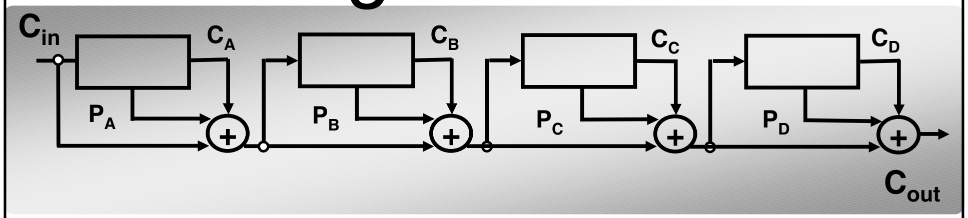
Final Adder

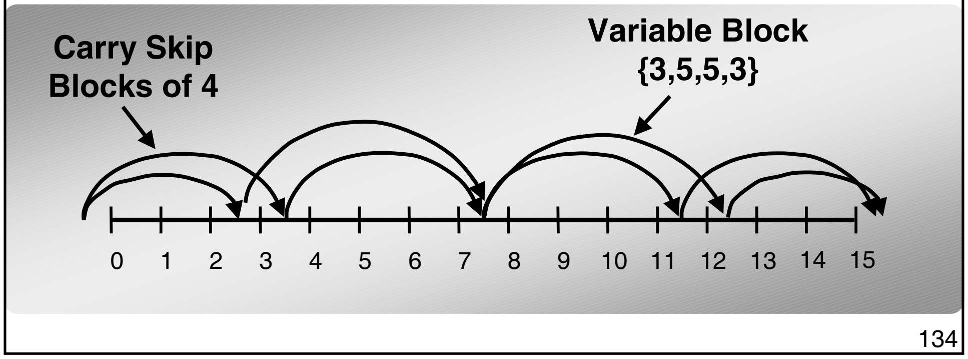


Fast Ripple Full Adders

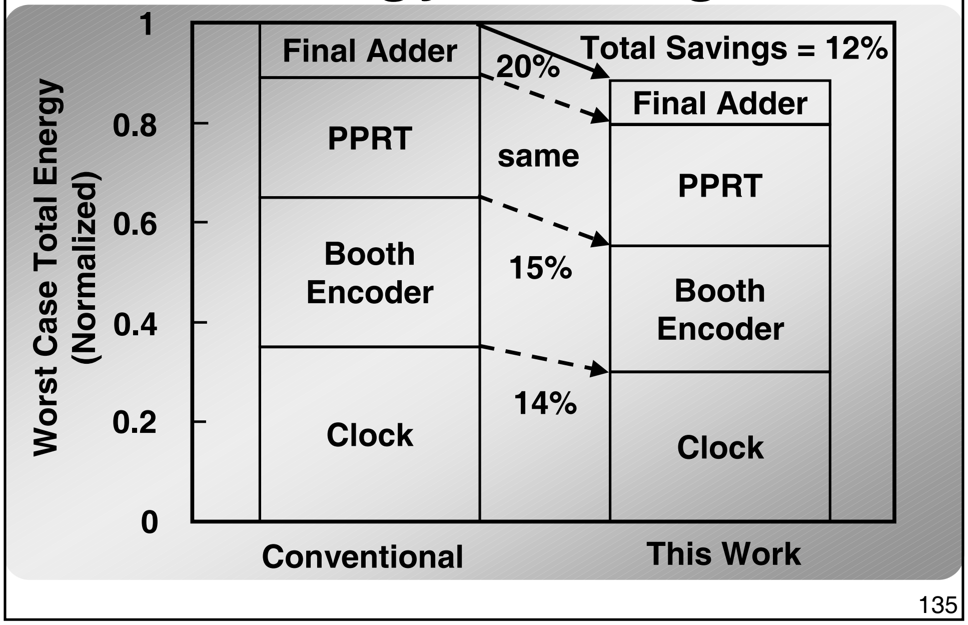


Carry Skip Adder Organization

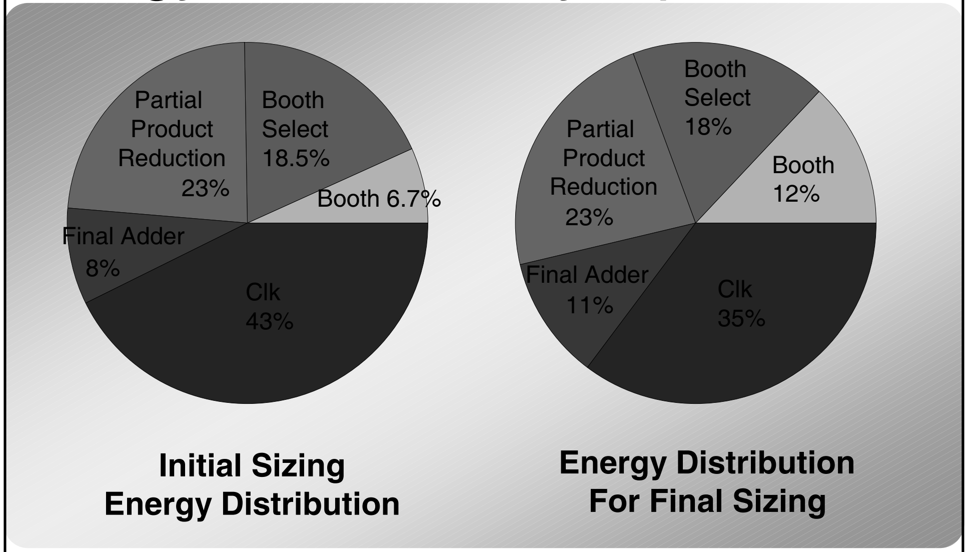




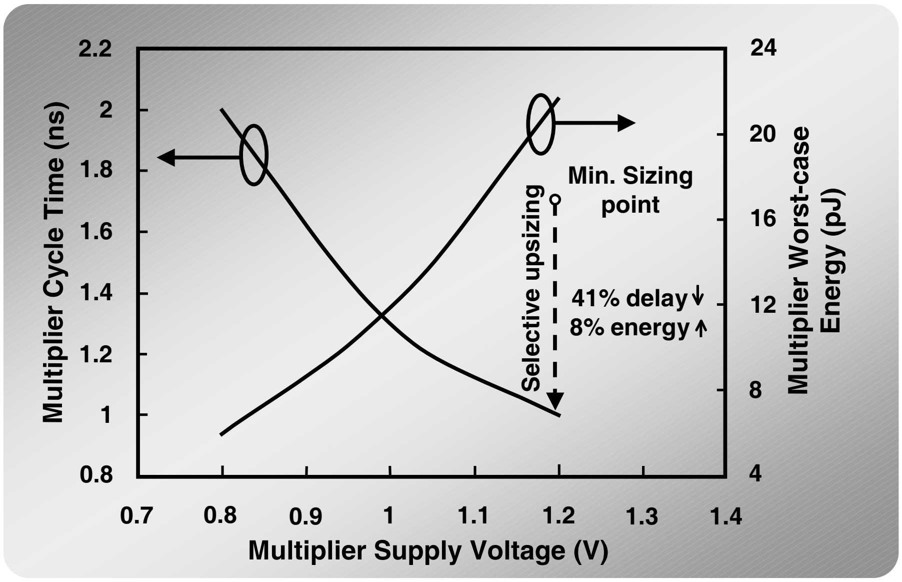
Energy Savings



Energy Based Delay Optimization



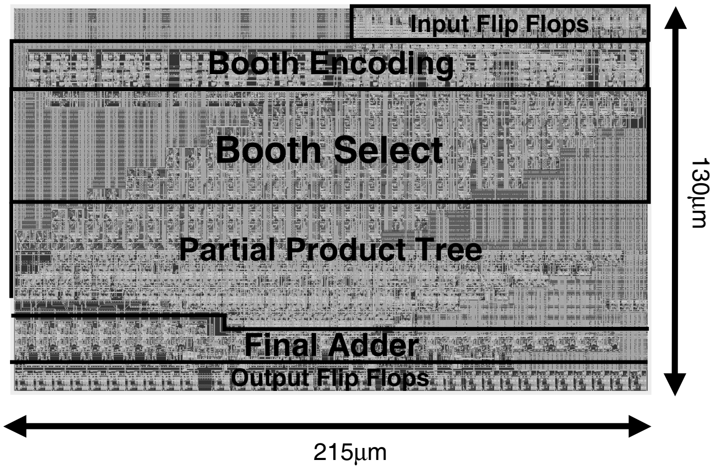
Simulation Results



Multiplier Layout

0.8V: 500MHz, 3mW

1.2V: 1GHz, 22mW



Summary

- 16-bit multiplier features:
  - Efficient Booth Encoding and Select
  - Delay and Area Optimized Partial Product Reduction Tree
  - Signal Profile Optimized Final Adder
  - Energy Optimized Sizing
- Enables multi-mode operation
  - **1GHz at 1.2V 22mW**
  - **500MHz at 0.8V 3mW**