Supply & Threshold Voltage Scaling

Prashant Agrawal

Department of Computer Sc & Engg

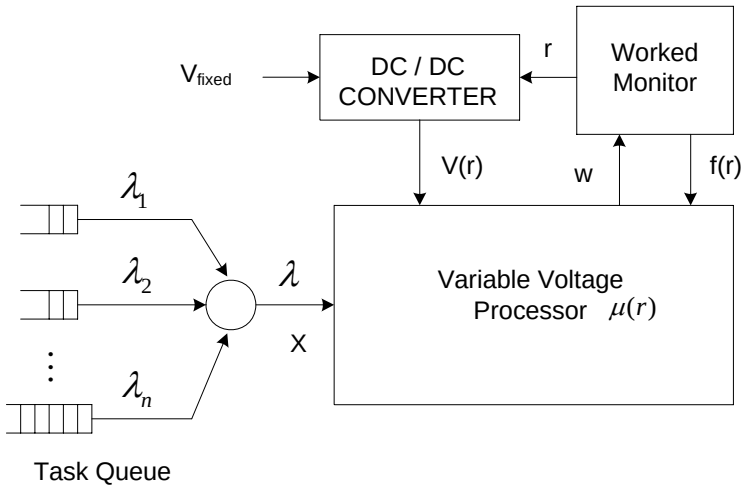
IIT Kharagpur

Outline

- Introduction
  - Dynamic Voltage Scaling
  - Body Biasing
- Previous Works

Dynamic Voltage Scaling

- Supply voltage can be dynamically changed during system operation
  - Cubic energy savings
  - Circuit slowdown
- Just –In-Time Computation
  - Stretch execution time to the maximum tolerable



Body Biasing

- Body Terminal Bias in Bulk CMOS
  - Dynamically alter MOSFET Vt
- Convenient Circuit-Level Technique
  - Reverse Body Bias (Vsb<0) to reduce sub-threshold leakage in standby mode
  - Forward Body Bias (Vsb>0) to improve performance in active mode
  - Adaptive Body Bias to compensate for leakage and performance spread

Combined dynamic voltage scaling and adaptive body biasing for lower power microprocessors under dynamic workloads

Authors

Martin, S.M.; Flautner, K.; Mudge, T.; Blaauw, D.

Published at

IEEE/ACM International Conference Computer Aided Design

Nov 2002

Combined DVS & ABB

- Idea
  - Simultaneous use of ABB and DVS can be used to reduce power in high-performance processors
  - ABB reduces leakage current exponentially, whereas DVS reduces leakage current linearly
  - The difficulty in employing simultaneous DVS and ABB is in determining the optimal trade-off between supply voltage and reverse body bias voltage such that total power consumption at a particular operating frequency in minimized.

Combined DVS & ABB

- Work
  - Analytical models of leakage current, dynamic power, and frequency as functions of supply voltage and body bias.
  - Derived analytical expressions for the optimal supply voltage and body bias for a given frequency and expected duration of operation.
  - The performance of the processor is used as a constraint to reduce a 2-dimensional optimization task to a single dimensional task and is solved through differentiation.

Combined DVS & ABB

- The analytical expression for the optimal Vdd and body bias was verified through SPICE simulations.
- The proposed simultaneous DVS and ABB method was then applied to Crusoe 5600 processor (600MHz, 0.18um) and was compared with using DVS.
- The proposed algorithm is implemented in the Linux Kernel and relies on monitoring system calls and inter-task communication to derive deadlines automatically and without modification of user programs.

Combined DVS & ABB

- Results
  - Vth shows linear dependence on Vdd and Vbs
  - As |Vbs| is increased
    - Current due to junction leakage I_j, increases and counteracts the savings achieved by lowering I_{subn}.
  - The maximum value of |Vbs| before junction leakage overrides subthreshold current reduction is dependent on
    - Process
    - Temperature
  - It varies between -0.6V and -2.5V

Combined DVS & ABB

Microprocessor Results

0.18µm Process	xmms-mp3	mpeg	emacs	os
No scaling	23 J	47 J	13 J	37 J
DVS alone (reduction vs. no scaling)	9.4 J (60%)	21 J (55%)	4.7 J (63%)	18 J (51%)
DVS & ABB (reduction vs. DVS alone)	7.6 J (19%)	19 J (10%)	2.8 J (40%)	14 J (21%)

TABLE 3. Energy consumed and percent reduction in the 0.18µm technology under several workloads with frequency scaling between 50-100% with 16% steps.

- For 0.18um process there isn't much difference in energy reduction for scaling between 50-100% with 16% steps and that between 10-100% with 5% step

0.07µm Process	Freq. scaling between 50-100% in 16% steps				Freq. scaling between 10-100% in 5% steps			
	xmms-mp3	mpeg	emacs	os	xmms-mp3	mpeg	emacs	os
No scaling	65J	111J	50J	119J	65J	111J	50J	119J
DVS alone (reduction vs.no scaling)	26J (60%)	47J (57%)	18J (64%)	53J (55%)	15J (76%)	42J (62%)	11J (78%)	37J (70%)
DVS & ABB (reduction vs. DVS)	16J (38%)	36J (22%)	9.3J (48%)	34J (35%)	8.4J (45%)	32 (22%)	2.1J (80%)	19J (47%)

TABLE 4. Energy consumed and percent reduction for DVS only, and DVS and ABB under two frequency scaling regimes.

A 175-MV multiply-accumulate unit using an adaptive supply voltage and body bias architecture

Authors

Kao, J.T.; Miyazaki, M.; Chandrakasan, A.R.;

Published at

IEEE Journal of Solid-State Circuits, Volume 37, Issue 11

Nov 2002

Adaptive Vdd Scaling & Body Biasing Arch

- Idea
  - Very little research has been devoted to managing leakage currents during the active state
  - Total active power can be minimized by dynamically adjusting Vdd and Vth based on circuit operating conditions such as temperature, workload, circuit architecture
  - For a digital circuit, it is possible to tradeoff dynamic and subthreshold leakage power by balancing between Vdd and Vt to maintain performance.

Adaptive Vdd Scaling & Body Biasing Arch

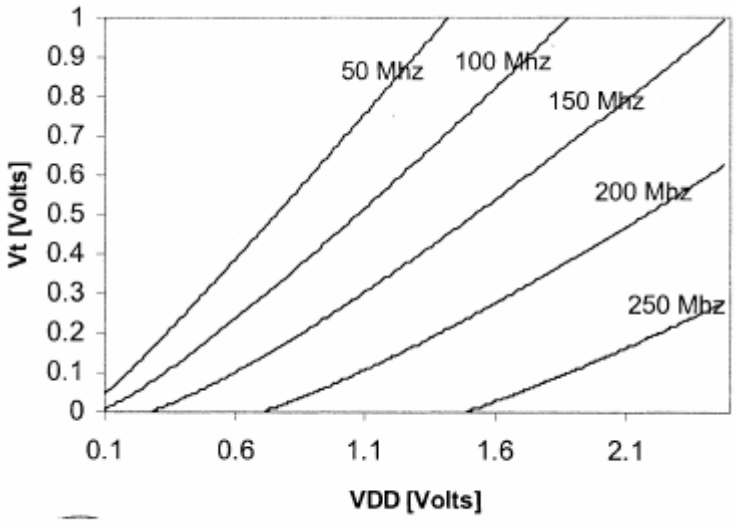


Fig. 2. V_{DD}/V_t combinations to give constant performance.

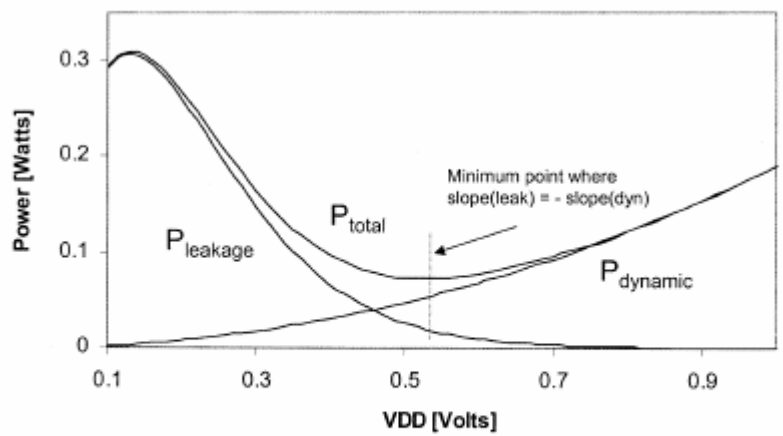
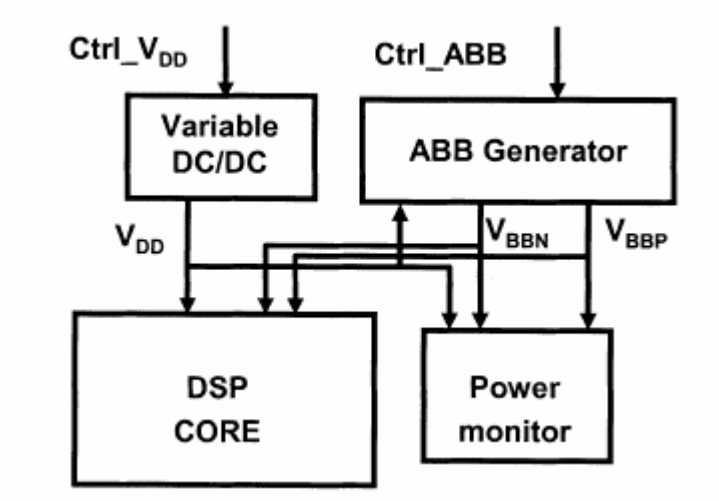


Fig. 3. Dynamic and subthreshold leakage power components for a fixed operating frequency (V_t implicitly set).

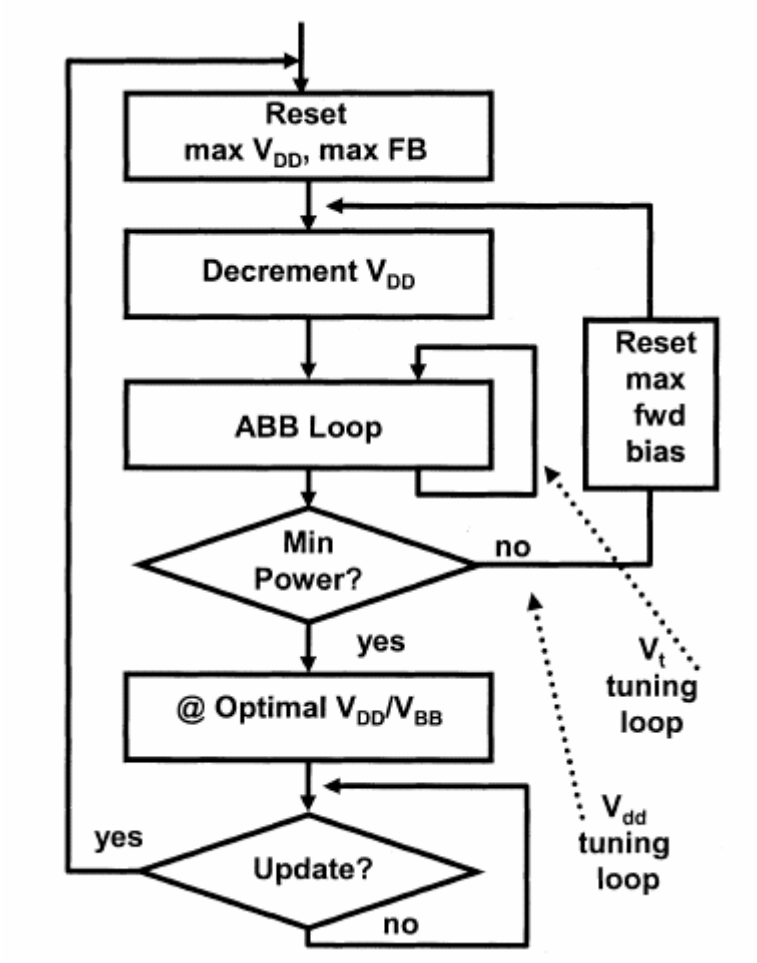
Adaptive Vdd Scaling & Body Biasing Arch

- Work
  - Main goal is to show how frequency and operating conditions affect optimal Vdd-Vt point
  - A theoretical model is first developed to predict how the minimum active power point depends on circuit parameters.
  - A 175-mV MAC test chip using a triple-well technology with tunable supply and body bias values is measured to experimentally verify the tradeoffs between dynamic and leakage currents as workload changes
  - A preliminary automatic supply and body biasing architecture (ASB) developed that automatically configures a circuit to operate with the lowest possible active power consumption.

Adaptive Vdd Scaling & Body Biasing Arch



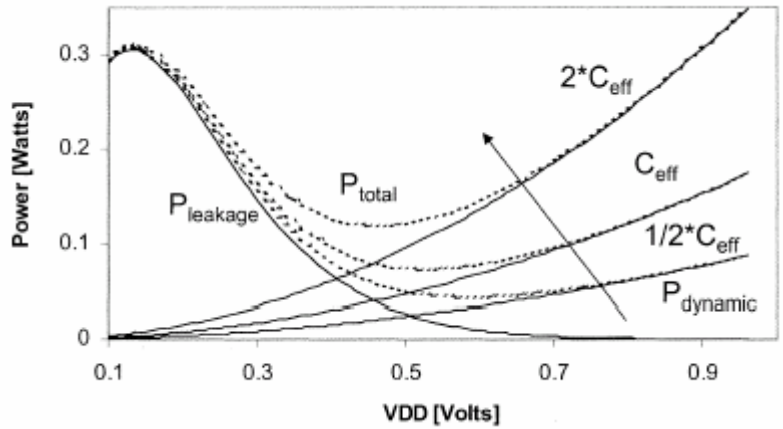
ASB Architecture



Flowchart for Vdd / Vt Optimization

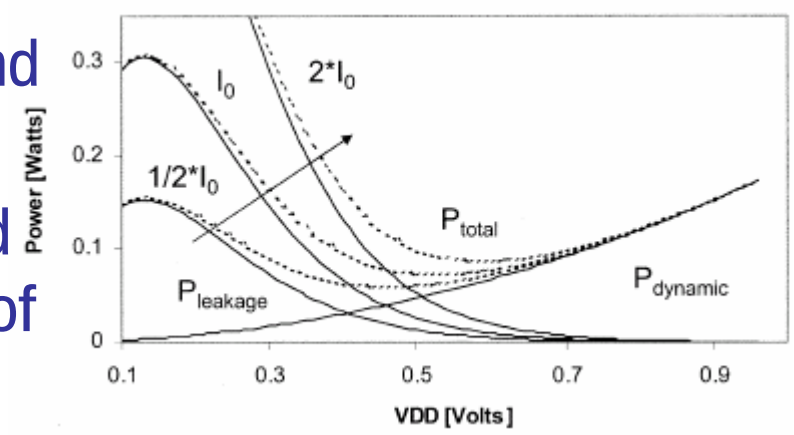
Adaptive Vdd Scaling & Body Biasing Arch

- Circuit Parameters that affect the optimal Vdd-Vtpoint
  - Architecture - determines Ceff and logic depth
  - Technology and temperature - determines Vt and Io
  - Workload requirements - operating frequency
- Effect of Ceff
  - As Ceff increases, optimal Vdd-Vt point shifts towards lower Vdd
  - Circuits that are heavily skewed to having larger dynamic currents, can be operated at lower Vdd, at the expense of increased leakage power to reduce total active power



Adaptive Vdd Scaling & Body Biasing Arch

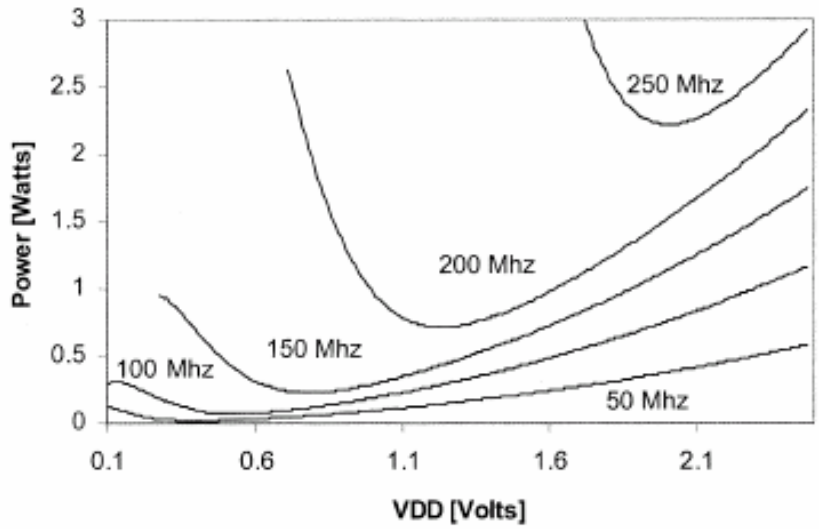
- Effect of leakage constant, I_0
  - As I_0 increases, optimal Vdd-Vt point shifts towards higher Vdd and Vt combination
  - It effectively reduces subthreshold leakage currents at the expenses of higher dynamic power



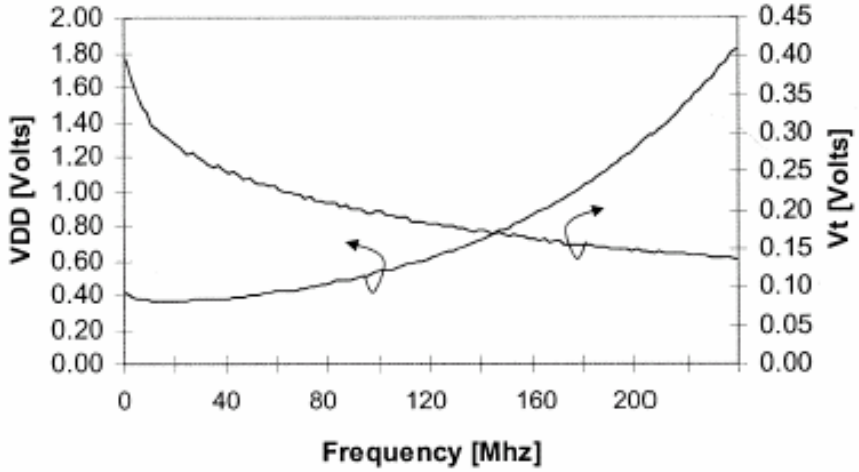
- Effect of temperature
  - As T increases, optimal Vdd-Vt point shifts towards higher Vdd to minimize leakage power at the expense of increased dynamic power

Adaptive Vdd Scaling & Body Biasing Arch

- Effect of workload
  - Operating frequency should be varied during the runtime based on changing workload requirements
  - At higher frequency, Vdd must increase and Vt must decrease
  - At lower frequency, Vdd must decrease and Vt must increase

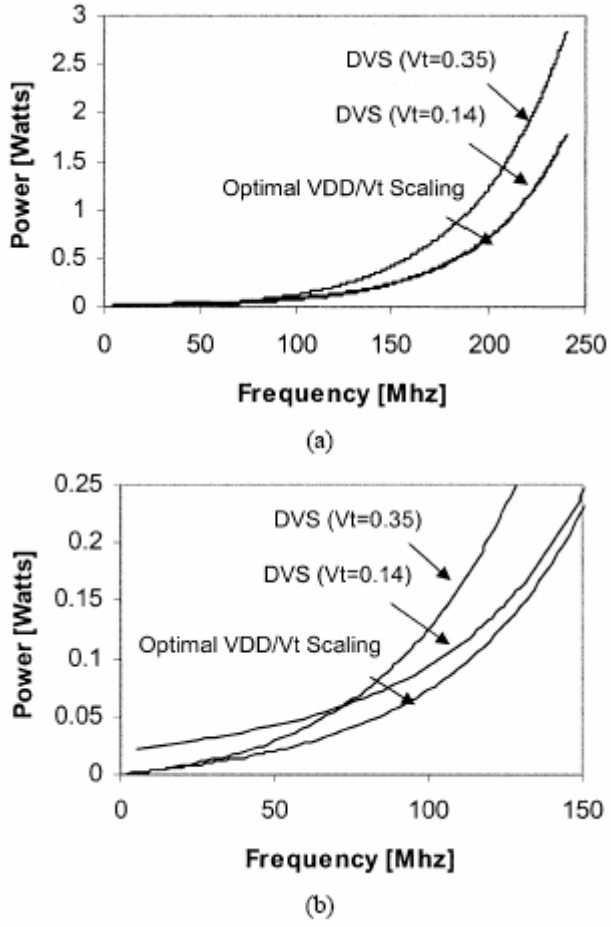


(a)



Adaptive Vdd Scaling & Body Biasing Arch

- Results
  - Comparison with fixed low and high Vt DVS and optimal Vdd/Vth scaling
  - Optimum Vdd-Vt operating point is higher than theoretical optimal point
  - With technology scaling, nominal Vt will become closer to the theoretical optimum Vt
    - Only small amount of fwd (or rev) bias would be needed to dynamically adjust Vt for all operating conditions.



Adaptive Vdd Scaling & Body Biasing Arch

- Forward bias benefits become limited at very low supply voltages
- Complication in the ASB arch is that this is a 3-D control problem because the supply voltage, PMOS bias, NMOS bias can all be independently tuned.
  - By tuning PMOS and NMOS devices by the same amounts, the problem can be reduced to a 2-D control problem.

Note: Here PMOS and NMOS bias are tuned by the same amount. Independent tuning can be tried.

A dynamic voltage scaled microprocessor system

Authors

Burd, T.D.; Pering, T.A.; Stratakos, A.J.; Brodersen, R.W.;

Published at

IEEE Journal of Solid-State Circuits, Volume 35, Issue 11

Nov 2000

A DVS Microprocessor System

- Idea
  - Reducing clock frequency during non-compute-intensive activity reduces power but doesn't affect the total energy consumed per task
  - Reducing the voltage of the processor improves its energy efficiency, but compromises its peak throughput.
  - If both clock frequency and supply voltage are dynamically varied in response to computational load demands, then the energy consumed per task can be reduced for the low computational periods, while retaining peak throughput when required.
  - 3 key components for implementing DVS in a general purpose uP system

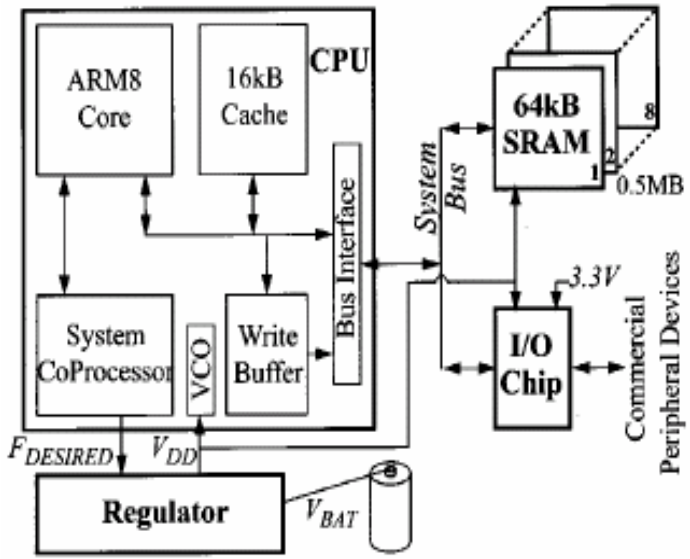
A DVS Microprocessor System

- An OS that can intelligently vary the processor speed
- A regulation loop that can generate the minimum voltage required for the desired speed
- A uP that can operate over a wide voltage range

- Control of the processor speed must be under the OS control. The h/w alone cannot distinguish whether the currently executing instruction is a part of a compute-intensive task or a non-speed-critical task.
- S/w is not aware of the minimum required supply voltage for a desired clock frequency since it is a function of the underlying h/w implementation, process variation, and operating temperature.
- DVS introduces two new performance parameters
  - Transition time
  - Transition energy

A DVS Microprocessor System

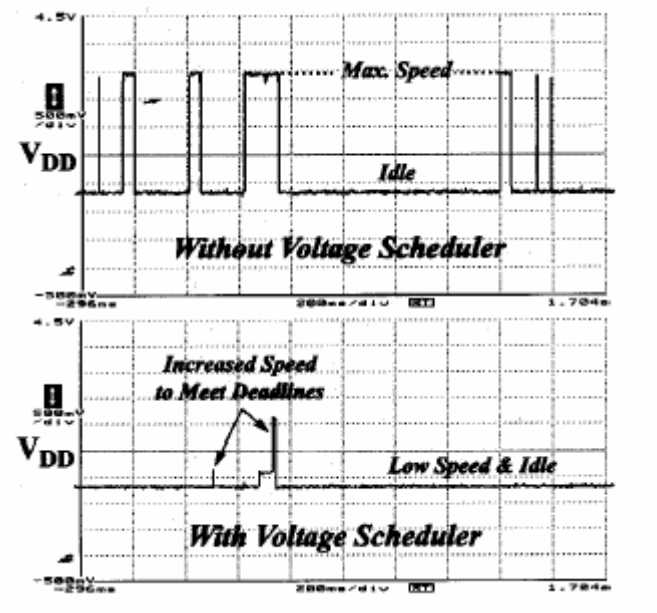
- Work
  - DVS of a general purpose uP, under direct OS control, and over a complete chip-set
  - The complete uP system is comprised of 4 custom chips all of which were designed for DVS to maximize system energy efficiency
  - Voltage scheduler is a new OS component for use in a DVS system.
    - It controls the processor speed by writing the desired clock freq to a system control register
    - Voltage scheduler can be separate from the temporal scheduler



System architecture—four custom chips.

A DVS Microprocessor System

- Results



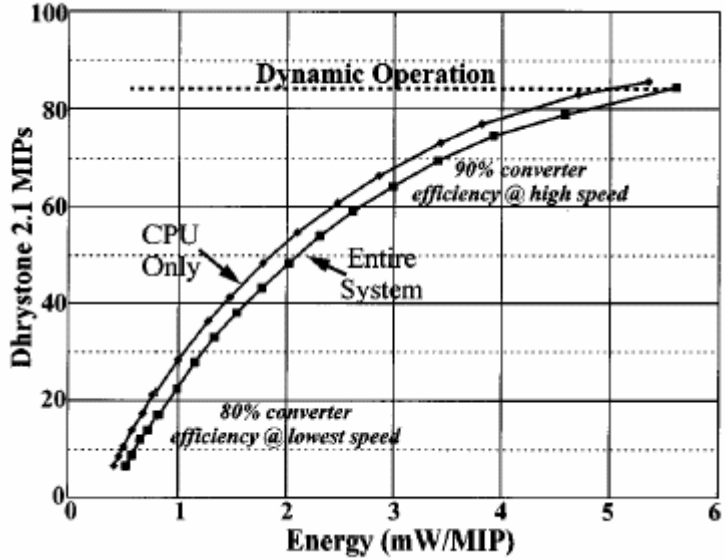


TABLE I
MEASURED BENCHMARK ENERGY CONSUMPTION (NORMALIZED)

	Benchmark Programs		
Algorithm	MPEG	UI	AUDIO
Maximum Performance	100%	100%	100%
Optimal	67%	25%	16%
Voltage Scheduler	89%	30%	22%

A practical transistor-level dual threshold voltage assignment methodology

Authors

Gupta, P.; Kahng, A.B.; Sharma, P.

Published at

Sixth International Symposium Quality of Electronic Design

March 2005

Tx Level Dual Vt Assignment Methodology

- Idea
  - TLVA can yield better reduction in leakage & total power as compared to CLVA.
  - Dual Vth processes are now standard and used together with other power reduction techniques.
  - Sensitivity based downsizing is better than upsizing.
    - Downsizing begins with low Vth assigned to all Tx and assigns nominal Vth to non-critical Tx.
    - Upsizing begins with nominal Vth assigned to all Tx and assigns low Vth iteratively to timing critical transistors.
    - Sensitivity is the ratio of change in leakage to the change in slack due to upsizing or downsizing

Tx Level Dual Vt Assignment Methodology

- To reduce the runtime of the algorithm, first CLVA and then TLVA is carried out
  - CLVA first assigns nominal Vth to the cells. TVLA does optimizations on the remaining low Vth cells
- Timing arc
  - An intra cell path from an input transition to a resulting rise (or fall) output transition.
  - For an n-input gate there are 2n timing arcs.
  - Due to different parasitics as well as PMOS/NMOS asymmetries, these timing arcs can have different delay values associated with them.
  - In this work, these asymmetries are exploited using TVLA to "recover" leakage from non-critical timing arcs within a cell.

Tx Level Dual Vt Assignment Methodology

Timing Arc	Propagation Delay (ps)	Transition Delay (ps)
A → Y ↑	99.05	104.31
A → Y ↓	73.07	79.12
B → Y ↑	107.20	112.98
B → Y ↓	70.65	76.37

Table 1: Asymmetry in delays of various timing arcs within a NAND2X2 cell.

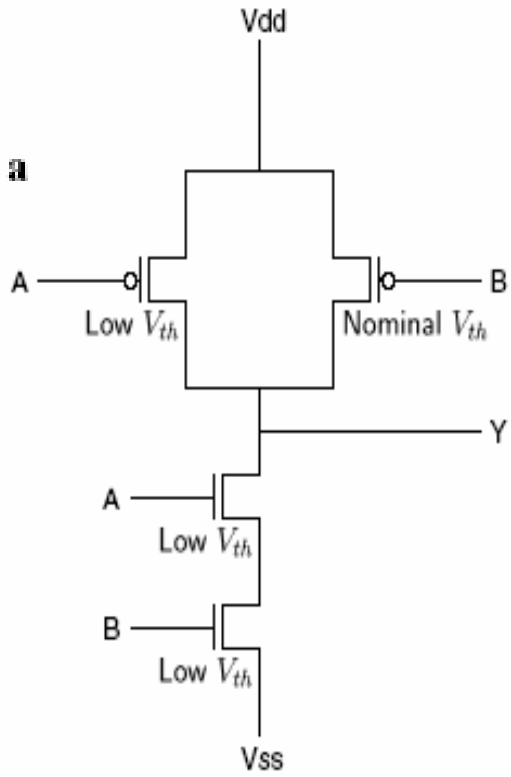


Figure 1: V_{th} assignment in NAND2X1 when only the rise and fall timing arcs from input A to the output are critical.

Tx Level Dual Vt Assignment Methodology

- Work
  - Cell variant creation
  - Library generation
  - Optimization for leakage
- Cell Variant Creation
  - A gate with n inputs has 2n timing arcs and therefore 2^{2n} variants (including the original cell).
  - Given a set of critical arcs, our goal is to assign nominal Vth to some transistors in the cell and low Vth to the remaining transistors to meet two criteria:
    - Cell leakage power is minimized
    - Critical timing arcs have a delay penalty of under 1% with respect to the cell in which all Tx are assigned low Vth

Tx Level Dual Vt Assignment Methodology

- Process of Vth assignment to Tx in a cell
  - For each cell, enumerate all configurations in which low Vth is assigned to some Tx and nominal Vth to others.
  - Find delay and power for each configuration under a canonical load of an inverter using SPICE
  - For each possible subset of timing arcs that can be simultaneously critical, one Vth assignment configuration is chosen based on the two criteria above.

- Library Generation
  - A restricted library is generated comprising of variants of the 25 most frequently used cells in our test cases.
  - The optimal nominal and low Vth voltages are not assumed but the voltages specified with the IBM 130nm SPICE device models are used.

Tx Level Dual Vt Assignment Methodology

- Optimization of Leakage
  - Synopsys Design Compiler is used for Tx sizing prior to Vth assignment
  - Timing Analyzer is used to compute delay sensitivity to Vth assignment
    - Standard Static Timing Analysis (SSTA) - slack is calculated
    - Exact Incremental STA (EISTA) - it adjusts slack after a node has been modified starting from the fan-in nodes of the node
    - Constrained Incremental STA (CISTA)

Tx Level Dual Vt Assignment Methodology

- Results
  - 62-89% reduction in leakage and 23-63% reduction in total power is achieved in comparison to when all transistors are assigned low Vth.
  - Larger leakage reduction in sequential circuits is observed when no delay penalty is allowed. But the reduction is insignificant if large delay penalties is allowed
  - Reduction in power using TLVA becomes comparable to CVLA for large sequential circuits
  - Run time of Sensitivity based Downsizing increases with circuit size and when there is less slack on nodes. This is because CLVA assigns low Vth to a smaller number of cells, and TLVA must then optimize a larger number of cells.
  - The no of cell variants required is significantly larger than those required for CLVA.

Standby power reduction using dynamic voltage scaling and canary flip-flop structures

Authors

Calhoun, B.H.; Chandrakasan, A.P.

Published at

IEEE Journal of Solid-State Circuits, Volume 39, Issue 9

Sept 2004

DVS & Canary Flip Flops

- Idea
  - Reducing Vdd to near the point where state is lost gives the best power savings.
  - “Canary” flip-flops provide a mechanism for observing the proximity to failure for the flip-flops.
    - They enable closed-loop standby voltage scaling for achieving savings near the optimum.
    - Open-loop design
      - value for the scaled voltage supply is fixed at design time.
      - This predetermined value doesn’t accounts for variations in process corner, temperature, threshold voltage, etc.
    - A closed loop design
      - By monitoring how close the critical path flip-flops are to failure, the supply voltage can be dynamically adjusted for maximum savings under different environmental conditions such as varying temperatures

DVS & Canary Flip Flops

- A closed-loop control of the standby voltage supply based on feedback from the canary flip-flops can lower the supply voltage very close to the minimum value without causing the critical path flip-flops to fail.
- The closed-loop approach tracks changes in the environment. Any change that affects the core flip-flops has a similar impact on the canary flip-flops. Since the canary flip-flops fail before the core devices, then the system continues to work as the environment fluctuates.

DVS & Canary Flip Flops

- Results
  - Canary flip-flops consistently fail at higher supply voltages than the core flip-flops at all process corners and operating temperatures.
  - The success of the canary flip-flops makes a closed-loop approach to standby voltage scaling feasible.
  - Power savings of over 40x in a 0.13- m, dual- test chip.
    - This is over 2x better than an optimal open-loop approach.

Theoretical and practical limits of dynamic voltage scaling

Authors

Bo Zhai; Blaauw, D.; Sylvester, D.; Flautner, K.

Published at

Proceedings of 41st Design Automation Conference

2004

Theoretical & Practical Limits to DVS

- Idea
  - It is possible to construct designs that operate over a much larger voltage range: from full Vdd to subthreshold voltages
- Work
  - Theoretically shown that for subthreshold supply voltages leakage energy becomes dominant
  - An analytical model derived for the minimum energy optimal voltage and study its trends with technology scaling.
  - This model is used to analyze the energy efficiency of an actual processor as a function of the lower limit of voltage scaling.
  - The optimal voltage limit depends on two factors:
    - Power/delay trade-offs at low operating voltages
    - Workload characteristics of the specific processors

Theoretical & Practical Limits to DVS

- Results
  - Extending the voltage range below 1/2 Vdd will improve the energy efficiency for most processor designs, while extending this range to subthreshold operation is beneficial only for very specific applications.
  - Applications that spend extensive time in near idle mode will benefit significantly from a voltage scaling ability from full to subthreshold voltages
  - Operation deep in the subthreshold voltage range is never energy-efficient

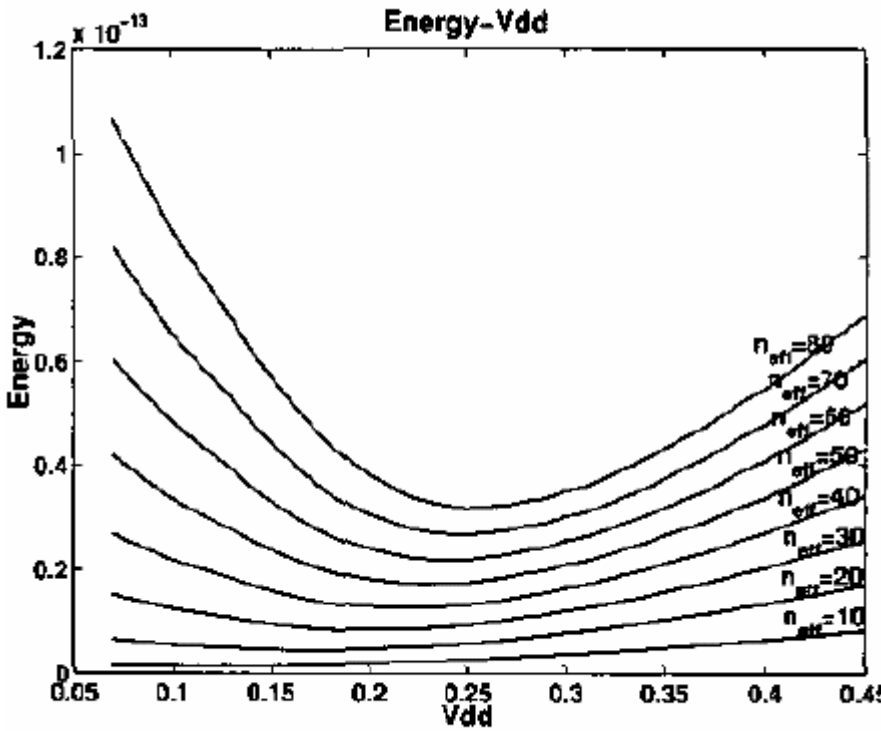


Figure 8. NAND2 chain Energy-Vdd (SPICE)

Thank You