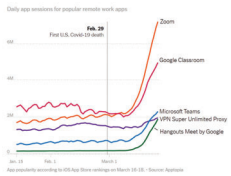
EE241B : Advanced Digital Circuits

Lecture 20 – Dynamic Voltage Scaling

Borivoje Nikolić

April 7, NY Times: The Virus Changed the Way We Internet

By Ella Koeze and Nathaniel Popper



Stuck at home during the coronavirus pandemic, with movie theaters closed and no restaurants to dine in, Americans have been spending more of their lives online.

But a New York Times analysis of internet usage in the United States from SimilarWeb and Apptopia, two online data providers, reveals that our behaviors shifted, sometimes starkly, as the virus spread and pushed us to our devices for work, play and connecting.

Announcements

- Assignment 4 due in two weeks.
- Reading
  - T. Burd, et al, JSSC, Nov 2000.

Outline

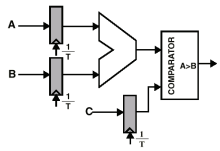
- Module 5
  - Reducing supply voltage
  - Multiple supply voltages
  - Dynamic voltage scaling

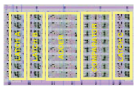
5.E Scaling Supplies

Power /Energy Optimization Space

	Constant Throughput/Latency		Variable Throughput/Latency
Energy	**Design Time**	**Sleep Mode**	**Run Time**
Active	Logic design Scaled V_{DD} Trans. sizing Multi-V_{DD}	Clock gating	DFS, DVS
Leakage	Stack effects Trans sizing Scaling V_{DD} + Multi-V_{Th}	Sleep T's Multi-V_{DD} Variable V_{Th} + Input control	DVS Variable V_{Th}

Architecture Trade-off for Fixed-rate Processing Reference Datapath



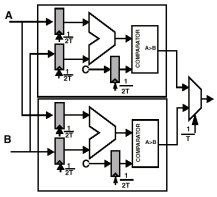


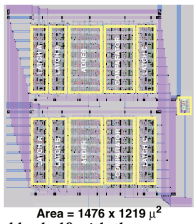
Area = 636 x 833 μ^2

- Critical path delay $\Rightarrow T_{adder} + T_{comparator}$ (= 25ns)
  $\Rightarrow f_{ref}$ = 40Mhz
- Total capacitance being switched = C_{ref}
- $V_{dd} = V_{ref}$ = 5V
- Power for reference datapath = $P_{ref} = C_{ref} V_{ref}^2 f_{ref}$

from [Chandrakasan92] (*IEEE JSSC*)

Parallel Datapath

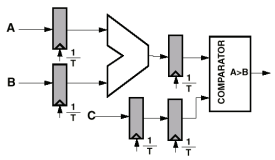


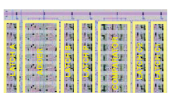


Area = 1476 x 1219 μ^2

- The clock rate can be reduced by half with the same throughput $\Rightarrow f_{par} = f_{ref} / 2$
- $V_{par} = V_{ref} / 1.7$, $C_{par} = 2.15C_{ref}$
- $P_{par} = (2.15C_{ref}) (V_{ref}/1.7)^2 (f_{ref}/2) \approx 0.36\ P_{ref}$

Pipelined Datapath





Area = 640 x 1081 μ^2

- Critical path delay is less $\Rightarrow$ max [T_{adder} , $T_{comparator}$]
- Keeping clock rate constant: $f_{pipe} = f_{ref}$
  Voltage can be dropped $\Rightarrow V_{pipe} = V_{ref} / 1.7$
- Capacitance slightly higher: $C_{pipe} = 1.15C_{ref}$
- $P_{pipe} = (1.15C_{ref}) (V_{ref}/1.7)^2 f_{ref} \approx 0.39\ P_{ref}$

A Simple Datapath: Summary

Architecture type	Voltage	Area	Power
Simple datapath (no pipelining or parallelism)	5V	1	1
Pipelined datapath	2.9V	1.3	0.39
Parallel datapath	2.9V	3.4	0.36
Pipeline-Parallel	2.0V	3.7	0.2

5.F Multiple Supplies

Power /Energy Optimization Space

	Constant Throughput/Latency		Variable Throughput/Latency
Energy	**Design Time**	**Sleep Mode**	**Run Time**
Active	Logic design Scaled V_{DD} Trans. sizing Multi-V_{DD}	Clock gating	DFS, DVS
Leakage	Stack effects Trans sizing Scaling V_{DD} + Multi-V_{Th}	Sleep T's Multi-V_{DD} Variable V_{Th} + Input control	DVS Variable V_{Th}

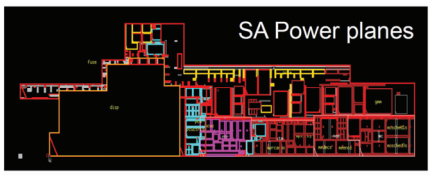
Multiple Supply Voltages

- Block-level supply assignment
  - Higher throughput/lower latency functions are implemented in higher V_{DD}
  - Slower functions are implemented with lower V_{DD}
  - Often called "Voltage islands"
  - Separate supply grids, level conversion performed at block boundaries
- Multiple supplies inside a block ("power domains" or "voltage islands")
  - Non-critical paths moved to lower supply voltage
  - Level conversion within the block
  - Physical design challenging

Power Domains

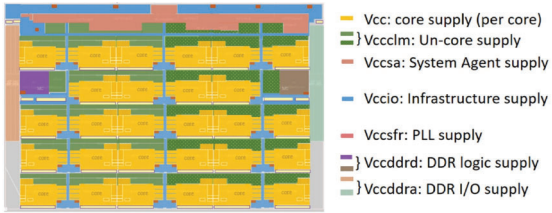
Practical Examples

- Intel Skylake (ISSCC'16)
  - Four power planes indicated by colors



Practical Examples

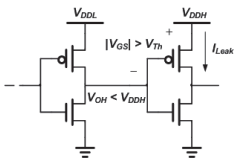
- Intel 28-core Skylake-SP (ISSCC'18)

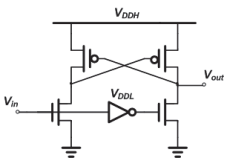


- 9 primary VCC domains are partitioned into 35 VCC planes

Leakage Issue

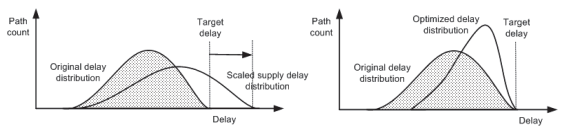
- Driving from V_{DDL} to V_{DDH} › Level converter



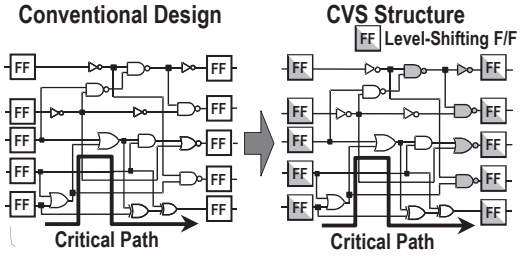


Multiple Supplies Within A Block

- Downsizing, lowering the supply on the critical path will lower the operating frequency
- Downsize (lowering supply) non-critical paths
  - Narrows down the path delay distribution
  - Increases impact of variations



Multiple Supplies in a Block



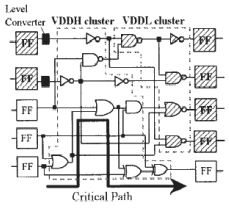
Lower V_{DD} portion is shaded

"Clustered voltage scaling"

M.Takahashi, ISSCC'98.

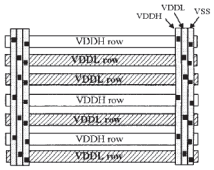
Multiple Supplies in a Block

CVS

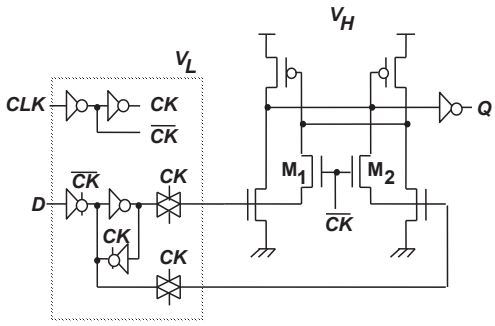


Usami'98

Layout:



Level-Converting Flip-Flop

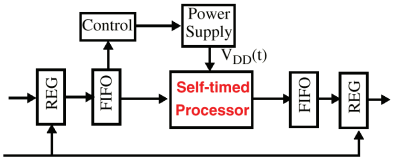


5.F Dynamic Voltage Scaling

Power /Energy Optimization Space

	Constant Throughput/Latency		Variable Throughput/Latency
Energy	**Design Time**	**Sleep Mode**	**Run Time**
Active	Logic design Scaled V_{DD} Trans. sizing Multi-V_{DD}	Clock gating	DFS, DVS
Leakage	Stack effects Trans sizing Scaling V_{DD} + Multi-V_{Th}	Sleep T's Multi-V_{DD} Variable V_{Th} + Input control	DVS Variable V_{Th}

Adaptive Supply Voltages

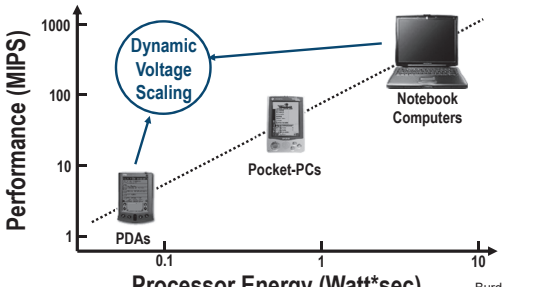


Exploit Data Dependent Computation Times To Vary the Supply

from [Nielsen94]

(IEEE Transactions on VLSI Systems)

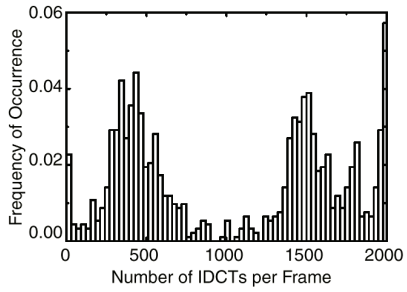
Processors for Portable Devices



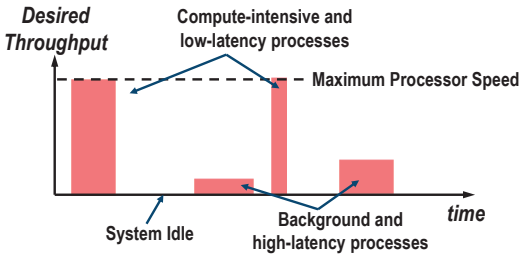
Burd ISSCC'00

- Eliminate performance ↔ energy trade-off

Typical MPEG IDCT Histogram



Processor Usage Model

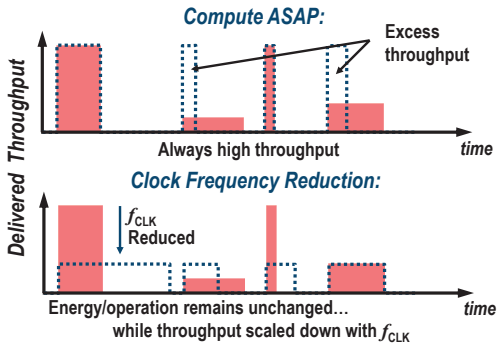


System Optimizations:

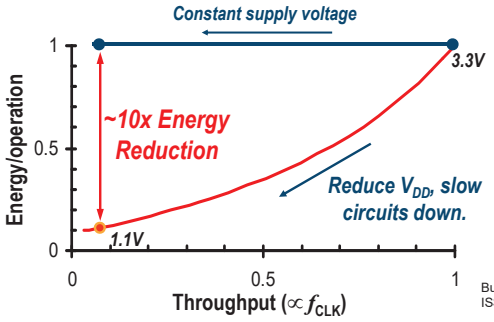
- Maximize Peak Throughput
- Minimize Average Energy/operation

Burd ISSCC'00

Common Design Approaches (Fixed VDD)

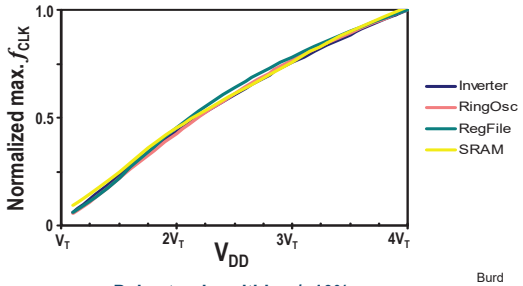


Scale V_{DD} with Clock Frequency



Burd ISSCC'00

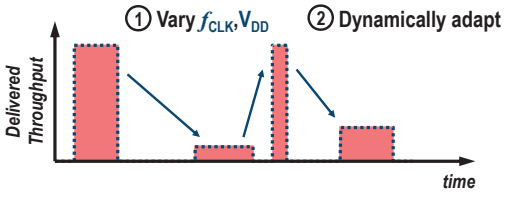
CMOS Circuits Track Over V_{DD}



← Delay tracks within +/- 10% →

Burd ISSCC'00

Dynamic Voltage Scaling (DVS)



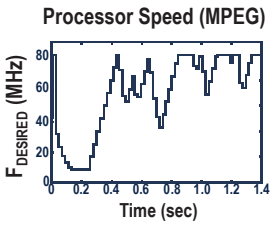
Burd ISSCC'00

- Dynamically scale energy/operation with throughput.
- Always minimize speed → minimize average energy/operation.
- Extend battery life up to 10x with the exact same hardware!

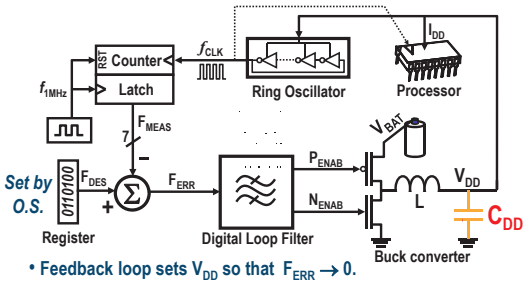
Operating System Sets Processor Speed

- DVS requires a *voltage scheduler* (VS).
- VS predicts workload to estimate CPU cycles.
- Applications supply completion deadlines.

$$\frac{CPU\ cycles}{\Delta\ time} = F_{DESIRED}$$



Converter Loop Sets V_{DD}, f_{CLK}



- Feedback loop sets V_{DD} so that $F_{ERR} \rightarrow 0$.
- Ring oscillator delay-matched to CPU critical paths.
- Custom loop implementation → Can optimize C_{DD}.

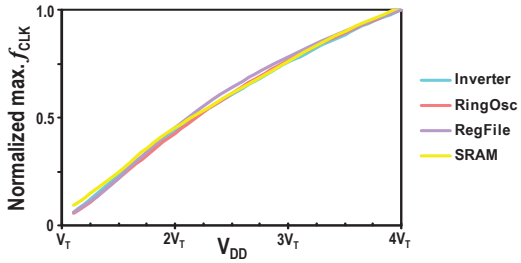
Burd ISSCC'00

Design Over Wide Range of Voltages

- Circuit design constraints. (Functional verification)
- Circuit delay variation. (Timing verification)
- Noise margin reduction. (Power grid, coupling)
- Delay sensitivity. (Local power distribution)

Design verification complexity similar to high-performance processor design @ fixed V_{DD}

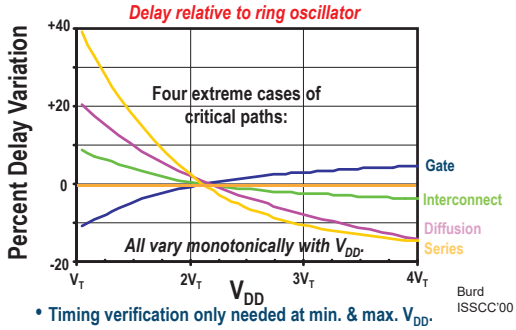
Delay Variation & Circuit Constraints



- Cannot use NMOS pass gates – fails for $V_{DD} < 2V_T$.
- Functional verification only needed at one V_{DD} value.

Burd ISSCC'00

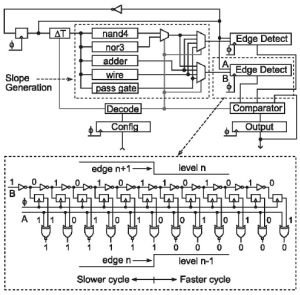
Relative Delay Variation



Burd ISSCC'00

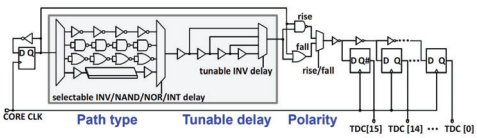
- Timing verification only needed at min. & max. V_{DD}.

Multiple Path Tracking



A. Drake, ISSCC'07

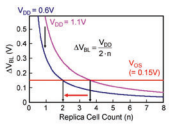
Multiple Path Tracking



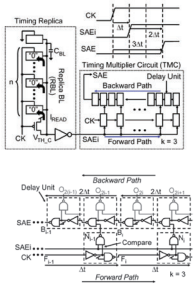
Cho, ISSCC'16

Tracking with SRAM in Critical Path

- **Mismatch between logic and SRAM**

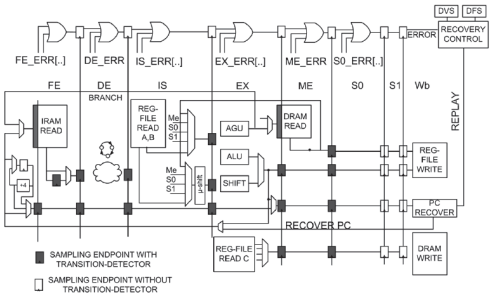


- **SRAM multiplicative replica**



Niki, JSSC'11

Alternative: Error Detection



Bull, ISSCC'2010

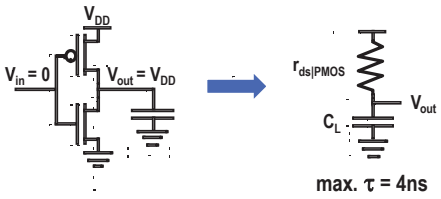
Design for Dynamically Varying VDD

- Static CMOS logic.
- Ring oscillator.
- Dynamic logic (& tri-state busses).
- Sense amp (& memory cell).

Max. allowed $|dV_{DD}/dt| \rightarrow$ Min. C_{DD} = 100nF (0.6µm)

Circuits continue to properly operate as V_{DD} changes

Static CMOS Logic

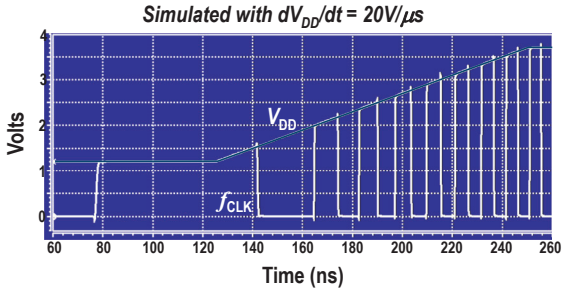


max. τ = 4ns

0.6µm CMOS: $|dV_{DD}/dt| < 200V/\mu s$

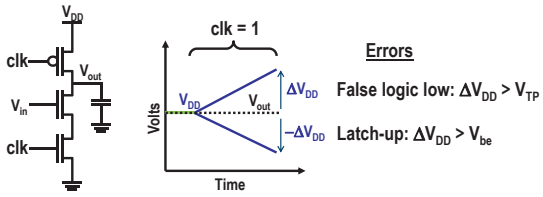
- **Static CMOS robustly operates with varying V_{DD}.**

Ring Oscillator



- **Output f_{CLK} instantaneously adapts to new V_{DD}.**

Dynamic Logic



Errors

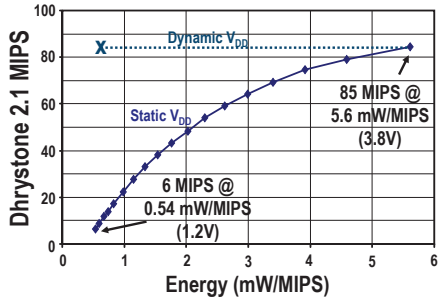
False logic low: $\Delta V_{DD} > V_{TP}$

Latch-up: $\Delta V_{DD} > V_{be}$

0.6µm CMOS: $|dV_{DD}/dt| < 20V/\mu s$

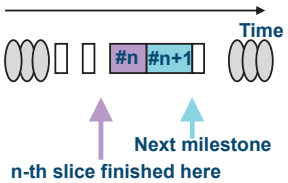
- **Cannot gate clock in evaluation state.**
- **Tri-state busses fail similarly → Use hold circuit.**

Measured System Performance & Energy

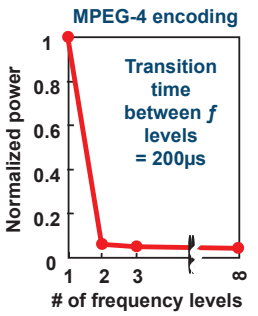


- **Dynamic operation can increase energy efficiency > 10x.** Burd ISSCC'00

V_{DD}-Hopping

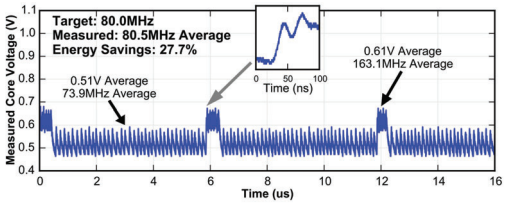


Application slicing and software feedback guarantee real-time operation.



Two hopping levels are sufficient.

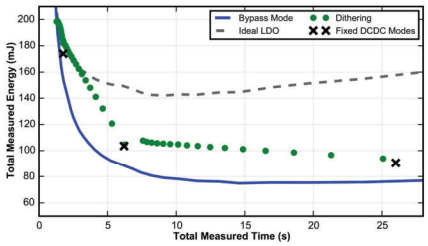
Dithering Between Supply Levels



Keller et al, ESSCIRC'16

Dithering Between Supply Levels

- Dithering fills in between fixed DC-DC modes



Next Lecture

- Low-power design
  - Clock gating
  - Power gating