inst.eecs.berkeley.edu/~ee241b

EE241B : Advanced Digital Circuits

Lecture 19 – Supply Voltage

Borivoje Nikolić

April 2, AnandTech: Intel Details 10th Gen Comet Lake-H for 45 W Notebooks: Up to 5.3 GHz*

*This CPU can hit this frequency on two cores, when the system is both within its secondary power limits but also Intel's Thermal Velocity Boost is enabled, which means there has to be additional thermal headroom in the system (and it has to be enabled by the OEM). This allows the CPU to go from 5.1 GHz to 5.3 GHz. Every Intel Thermal Velocity Boost enabled CPU requires OEM support in order to get those extra two bins on the single core frequency.

Announcements

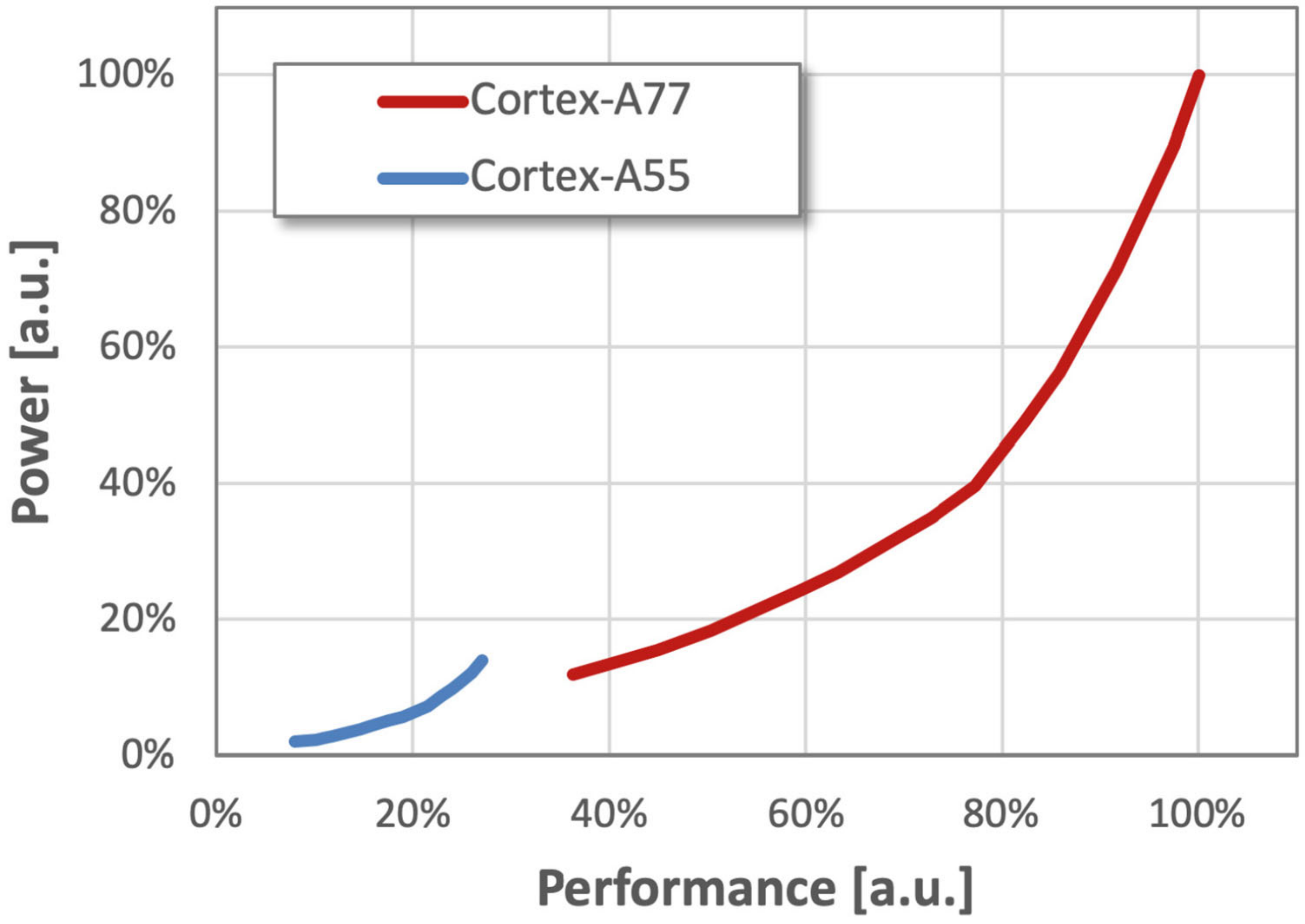
- Assignment 3 due today, April 2.
  - Quiz next Tuesday, end of class

Outline

- Module 5
  - Circuit-level power-performance tradeoffs
  - Reducing supply voltage

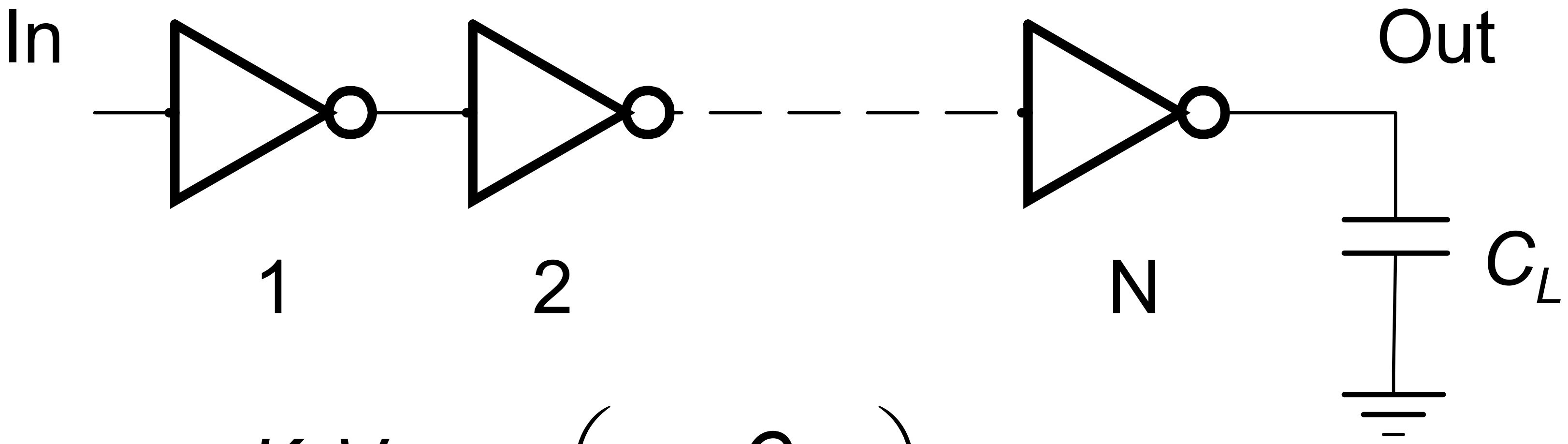
Architectural Tradeoffs

- H, Mair, ISSCC'20



5.D Circuit-Level Tradeoffs

Alpha-Power Based Delay Model



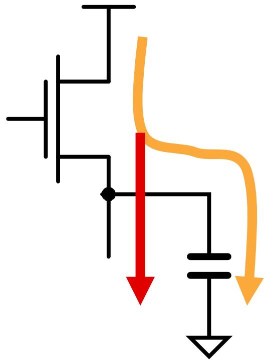
$$t_{pi} = \frac{K_d V_{DD}}{(V_{DD} - V_{Th})^{\alpha}} \left(1 + \frac{C_{L,i}}{C_{in,i}}\right)$$

$$D = \sum t_{pi} = \sum \frac{K_d V_{DD}}{(V_{DD} - V_{Th})^{\alpha}} \left(1 + \frac{W_{L,i}}{W_{in,i}}\right)$$

Energy Models

- **Switching**

$$E_{Sw} = \alpha_{0\to 1}\left(C_{L,i} + C_{\text{int},i}\right)V_{DD}{}^{2}$$



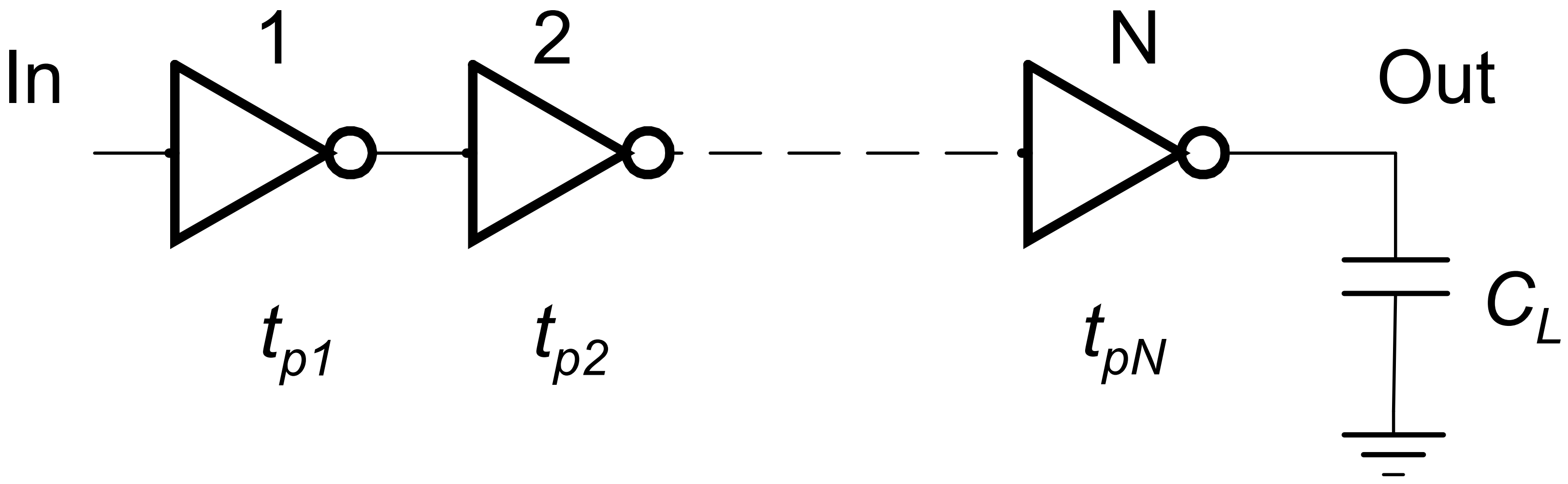
- **Leakage**

$$E_{Lk} = W_{ln} I_0 e^{\frac{-(V_{Th} - \gamma V_{DD})}{nV_t}} V_{DD} D$$

Sizing, Supply, Threshold Optimization

- Transistor sizing can yield large power savings with small delay penalties
  - Gate sizing
  - Beta-ratio adjustments $\beta = Wp/Wn$
  - (Stack resizing)
- Supply voltage affects both active and leakage energy
- Threshold voltage affects primarily the leakage

Apply to Sizing of an Inverter Chain



Unconstrained energy: find min D = Σt_{pi}

$$C_{gin,j} = \sqrt{C_{gin,j-1} C_{gin,j+1}} \qquad W_j = \sqrt{W_{j-1} W_{j+1}}$$

Constrained energy: find min D, under E < E_{max}

Where E = Σe_i

Constrained Optimization

- Find min(D) subject to $E = E_{max}$
  - *Constrained function minimization*

- E.g. Lagrange multipliers

Or dual:

$$\Lambda(x) = D(x) + \lambda\left(E(x) - E_{\max}\right)$$

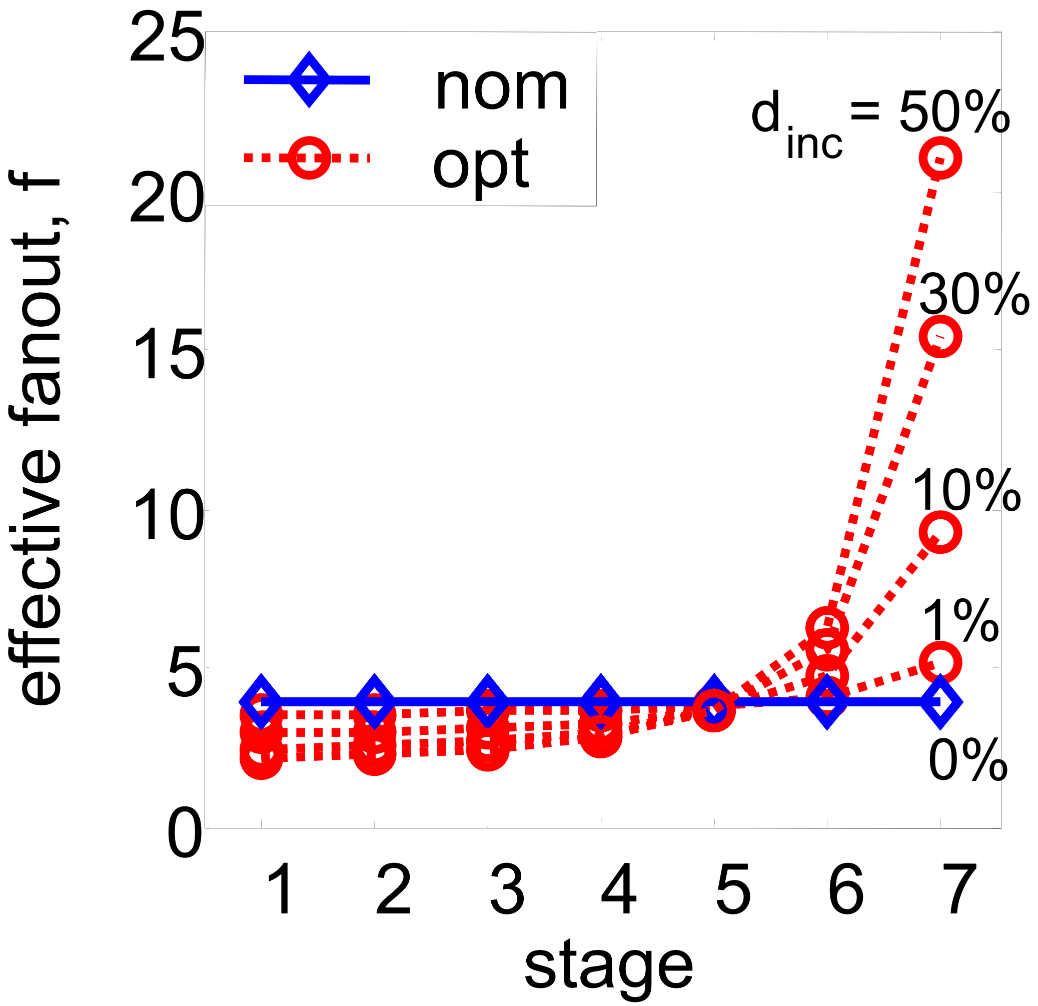
$$\mathrm{K}(x) = E(x) + \lambda\left(D - D_{\max}\right)$$

$$\frac{\partial \Lambda}{\partial x} = 0$$

- Can solve analytically for $x = W_i, V_{DD}, V_{Th}$

Inverter Chain: Sizing Optimization

Inverter Chain: Sizing Optimization



$$W_j = \sqrt{\frac{W_{j-1}W_{j+1}}{1+\lambda W_{j-1}}}$$

[Ma, Franzon, *IEEE JSSC*, 9/94]

$$\lambda = -\frac{2KV_{DD}^{\ 2}}{\tau_{nom}S_W}$$

$$S_W \propto \frac{e_j}{f_j - f_{j-1}}$$

e_i – energy per stage

f_i – fanout per stage

Stojanovic, ICCAD’02

- **Variable taper achieves minimum energy**
- **Reduce number of stages at large d_{inc}**

Sensitivity to Sizing and Supply

- Gate sizing (W_i)

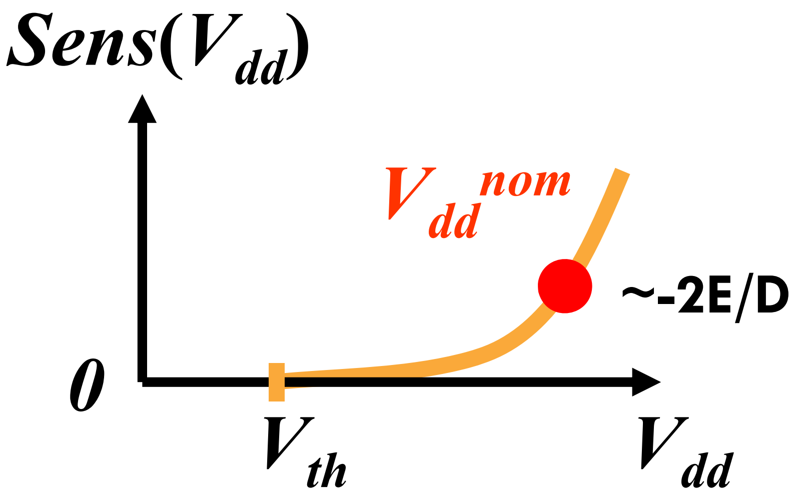
$$-\frac{\partial E_{sw} / \partial W_j}{\partial D / \partial W_j} = \frac{e_j}{\tau_{nom}\left(f_j - f_{j-1}\right)}$$

∞ for equal f_{eff} (D_{min})

- Supply voltage (V_{dd})

$$-\frac{\partial E_{sw} / \partial V_{DD}}{\partial D / \partial V_{DD}} = \frac{E_{sw}}{D} 2 \frac{1 - x_v}{\alpha - 1 + x_v}$$

$x_v = (V_{Th} + \Delta V_{Th}) / V_{dd}$



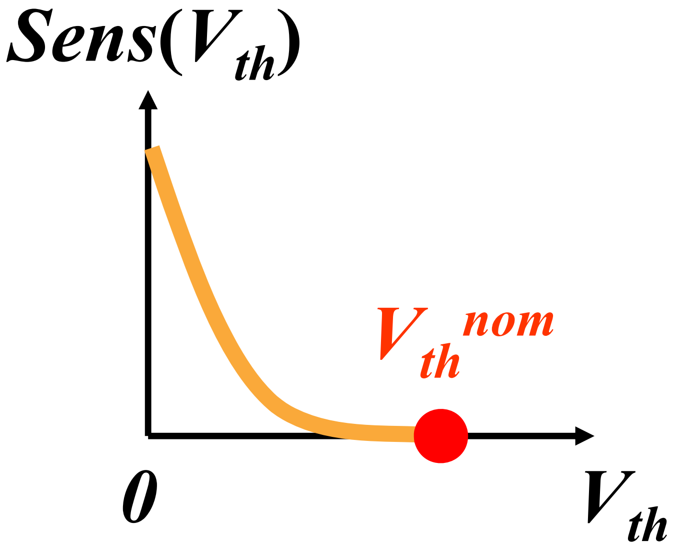
Sensitivity to V_{th}

- Threshold voltage (V_{th})

$$-\frac{\partial E / \partial \Delta V_{Th}}{\partial D / \partial \Delta V_{th}} = P_{Lk}\left(\frac{V_{DD} - V_{Th} - \Delta V_{Th}}{\alpha n V_t} - 1\right)$$

Low initial leakage

⇒ speedup comes for "free"



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}

Energy-Performance Tradeoffs

Enable Time/ Perf. Impact	Design Time	Run Time
Near-zero perf. penalty	Clock gating Architectural switching reduction Multi-V_{Th}	Dynamic V_{DD} Dynamic V_{Th}
True tradeoffs	Fine-granularity clock gating V_{DD}, V_{TH} adjustments Multi-V_{DD} Sizing, logic styles Stack forcing	Power gating

5.E Scaling Supplies

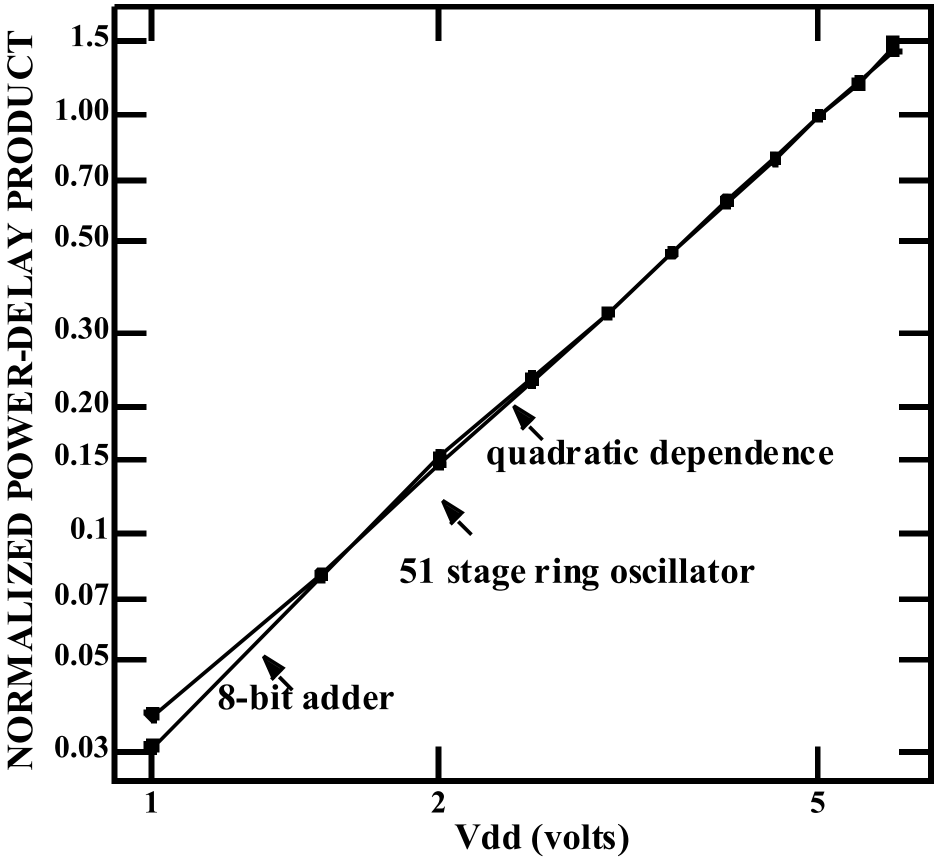
Power /Energy Optimization Space

	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}

Supply Voltage Adjustment

- How to maintain throughput under reduced supply?
- Introducing more parallelism/pipelining
  - Area increase
  - Cost/power tradeoff
- Multiple voltage domains
  - Separate supply voltages for different blocks
  - Lower VDD for slower blocks
  - Cost of DC-DC converters
- Dynamic voltage scaling – with variable throughput
- Reducing V_{TH} to improve speed
  - Leakage issues

Reducing V_{dd}



$$\boxed{P \times t_d = E_t = C_L * V_{dd}^2}$$

$$\frac{E_{(Vdd=2)}}{E_{(Vdd=5)}} = \frac{(C_L) * (2)^2}{(C_L) * (5)^2}$$

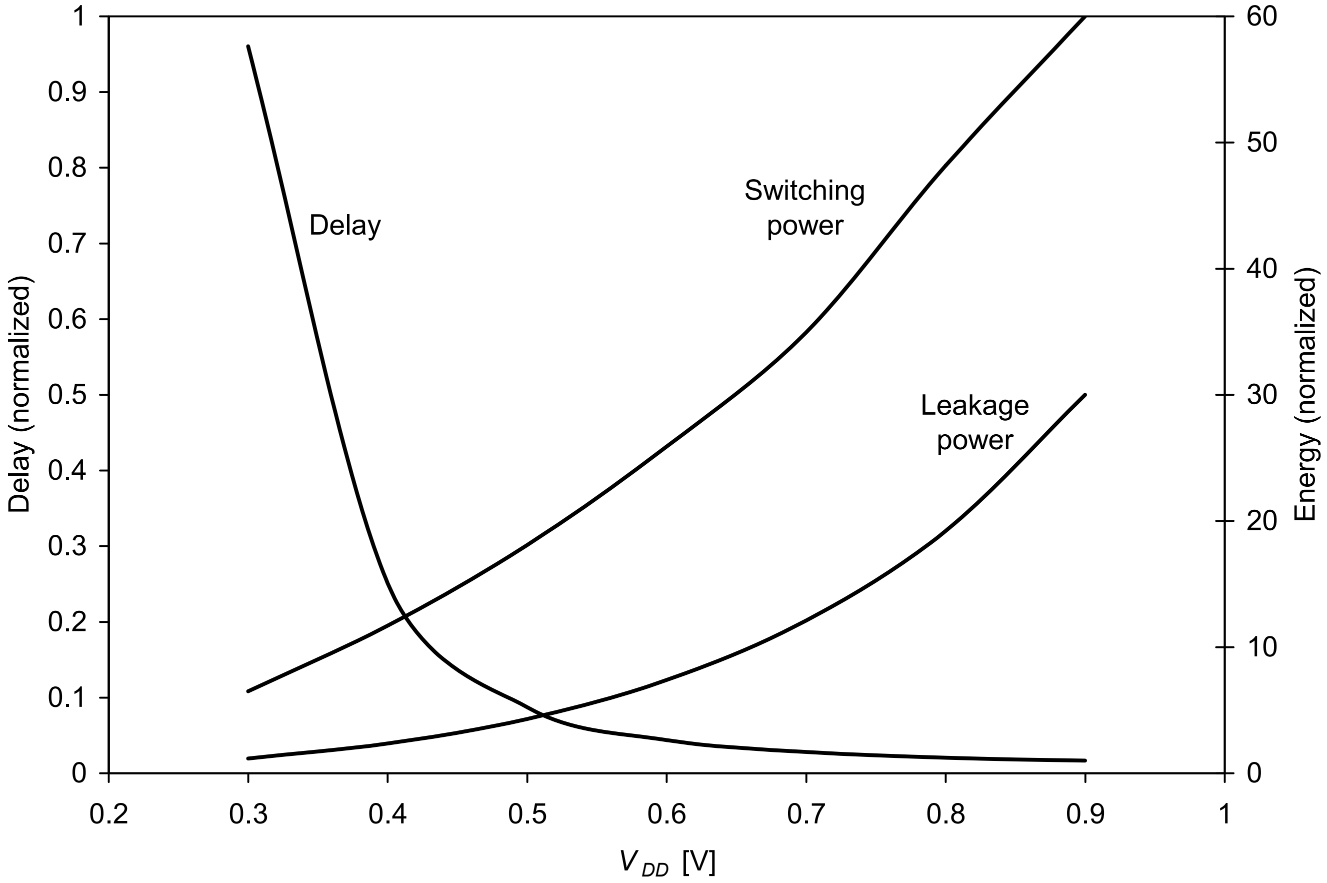
$$E_{(Vdd=2)} \approx 0.16\ E_{(Vdd=5)}$$

- Strong function of voltage (V^2 dependence).
- Relatively independent of logic function and style.
- Power Delay Product Improves with lowering V_{DD}.

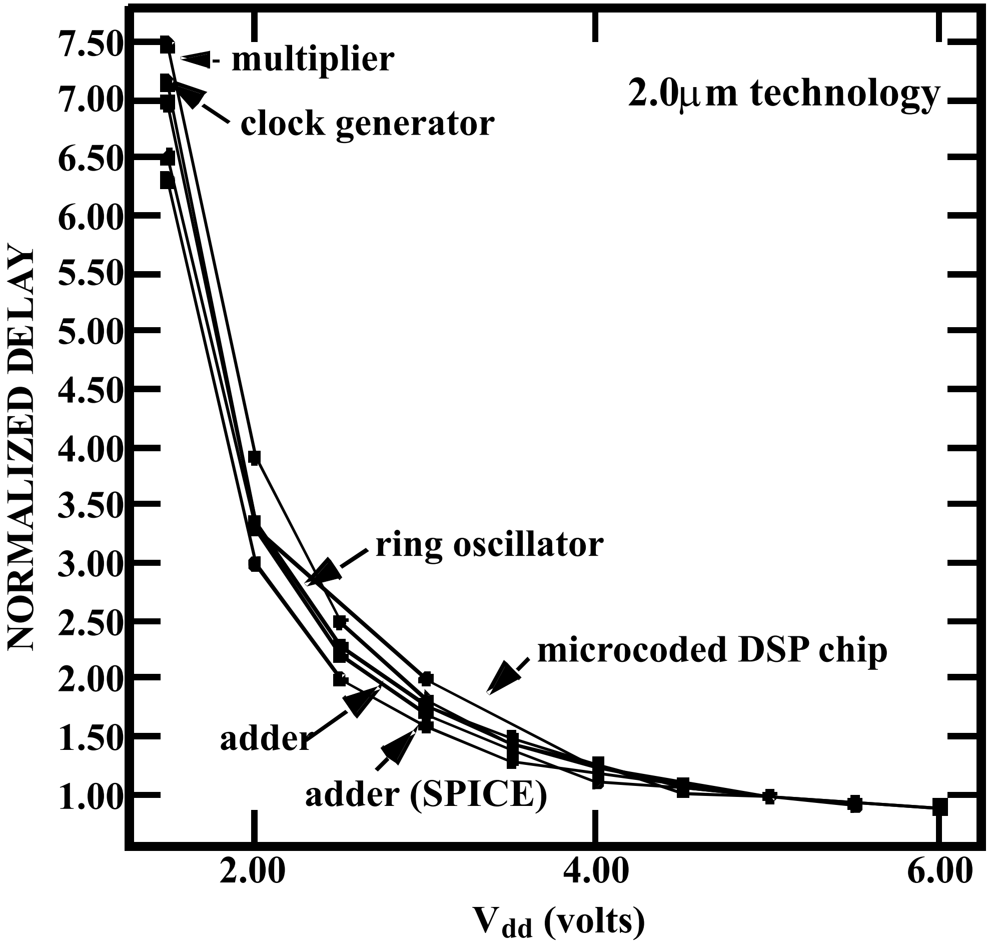
Chandrakasan, JSSC'92

Reducing V_{DD}

32nm process



Lower V_{DD} Increases Delay



$$\boxed{T_d = \frac{C_L * V_{dd}}{I}}$$

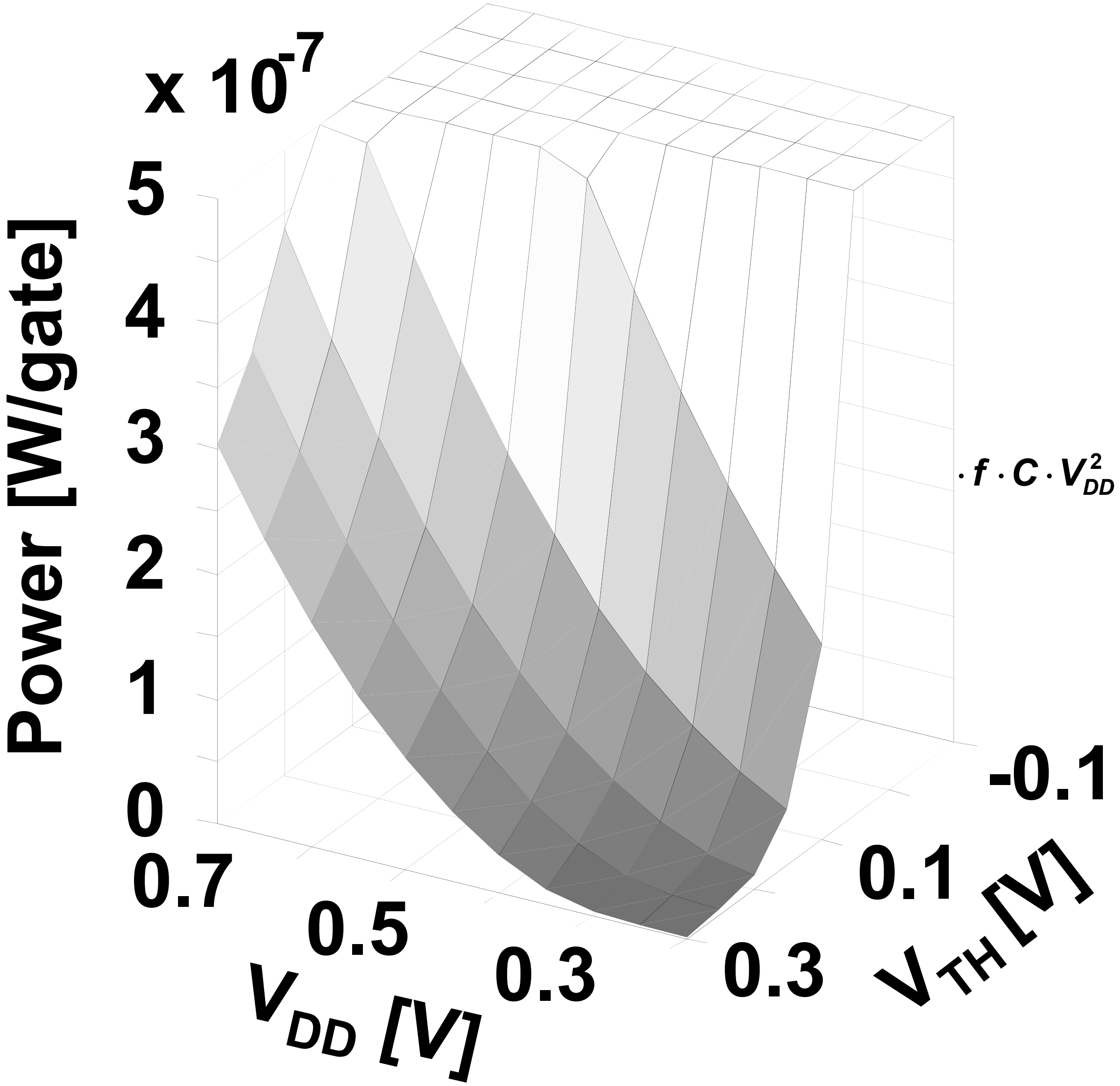
$$I \sim (V_{dd} - V_t)^2$$

$$\frac{T_{d(Vdd=2)}}{T_{d(Vdd=5)}} = \frac{(2) * (5 - 0.7)^2}{(5) * (2 - 0.7)^2} \approx 4$$

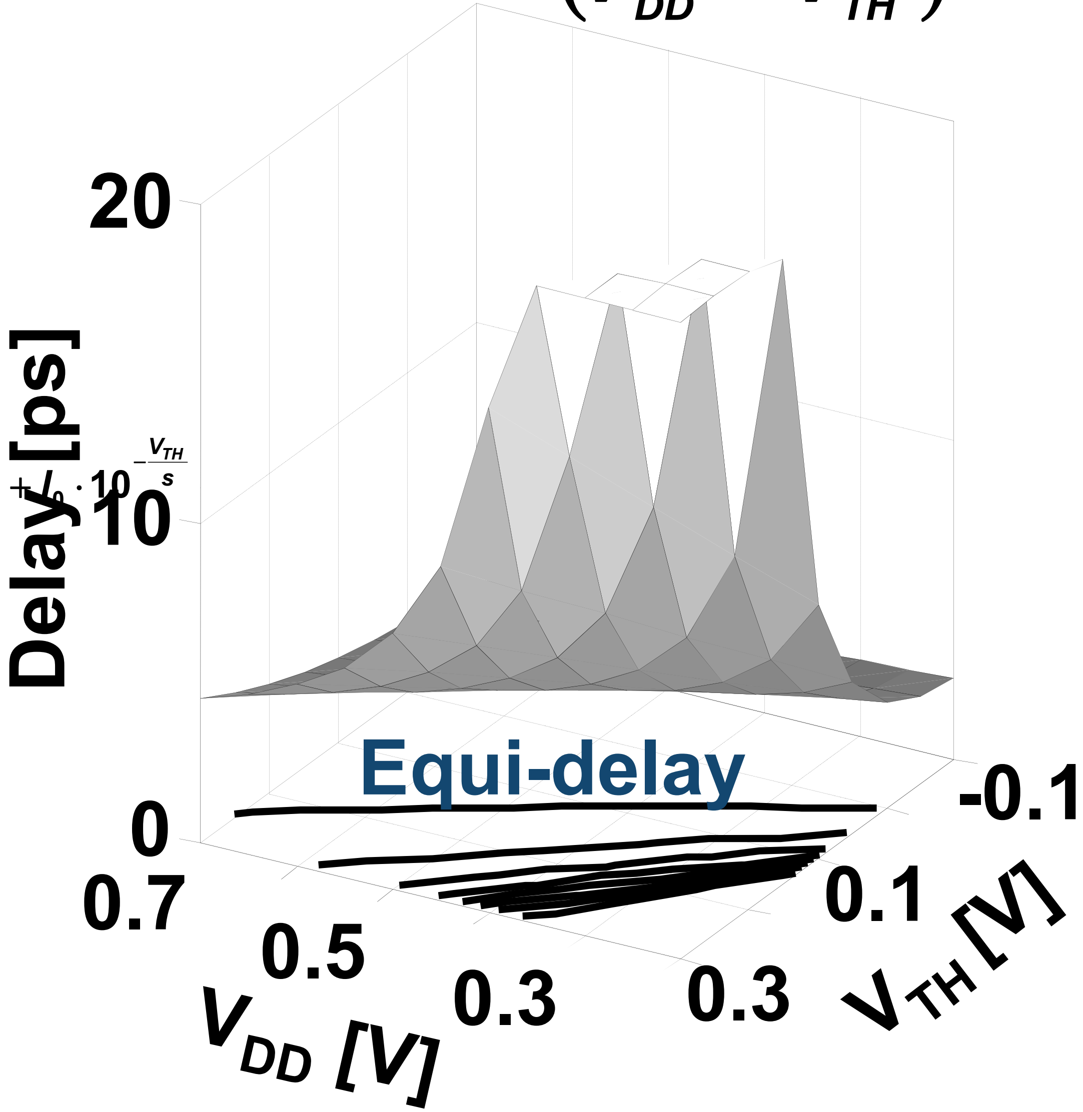
- **Relatively independent of logic function and style.**

Trade-off Between Power and Delay

$$Delay \propto \frac{C \cdot V_{DD}}{(V_{DD} - V_{TH})^{1.3}}$$



$\cdot f \cdot C \cdot V_{DD}^2 + I_0 \cdot 10^{-\frac{V_{TH}}{s}}$

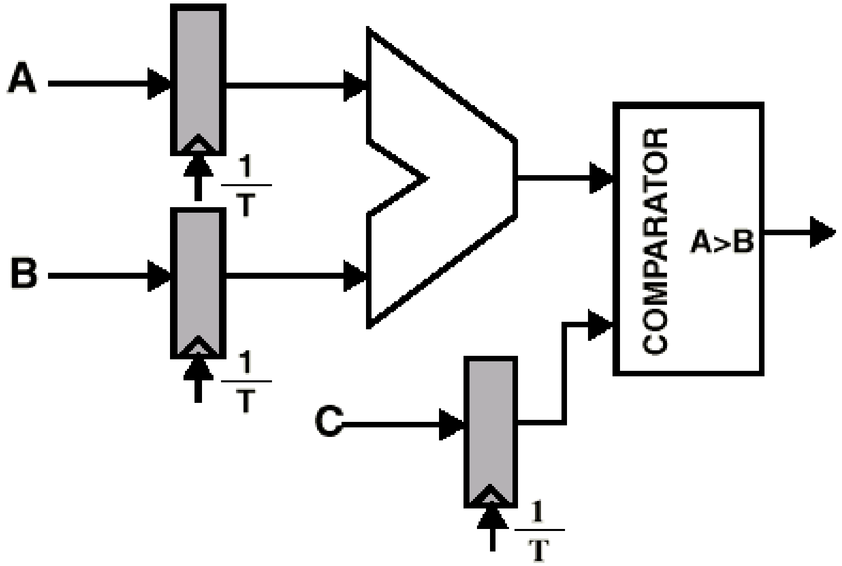


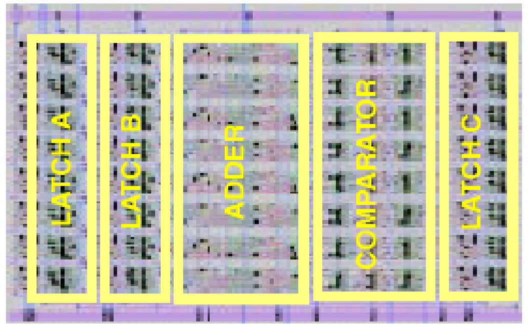
50nm node, FO3 INV

Two Types of Processing

- Fixed-rate processing (e.g. signal processing for multimedia or communications)
  - Stream-based computation
  - No advantage in obtaining throughput in excess of the real-time constraint
- Variable-rate or burst-mode computation (e.g. general purpose computation)
  - Mostly idle (or low-load) with bursts of computation
  - Faster is better

Architecture Trade-off for Fixed-rate Processing Reference Datapath





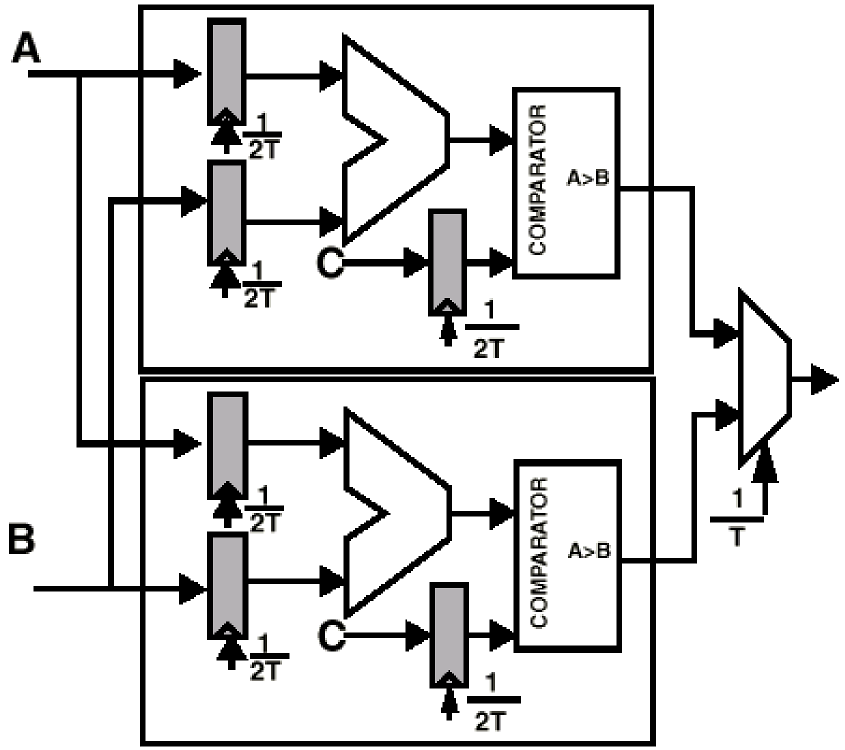
Area = 636 x 833 μ^2

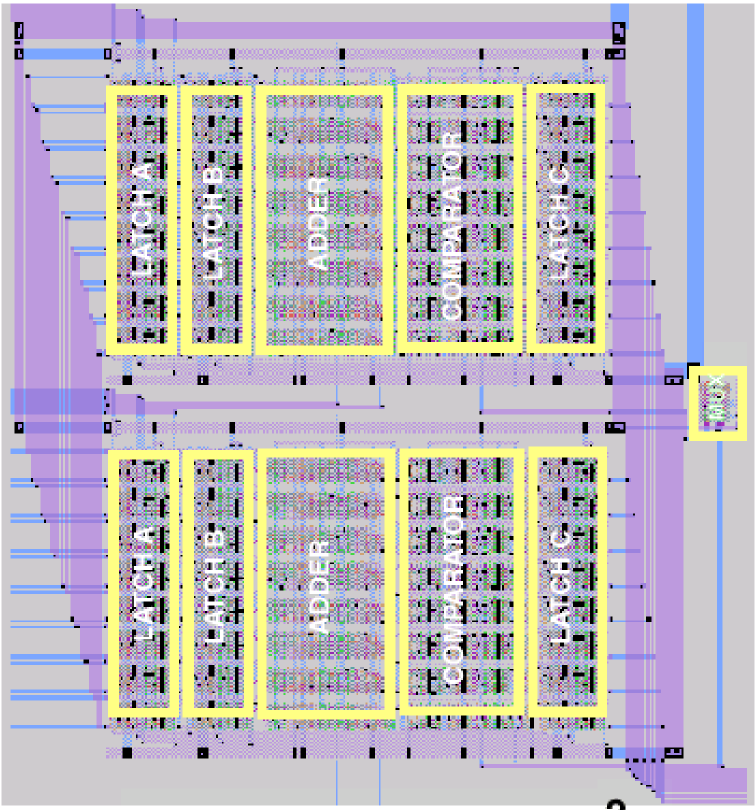
- Critical path delay $\Rightarrow T_{adder} + T_{comparator}$ (= 25ns)

  $\Rightarrow f_{ref} = 40$Mhz
- 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

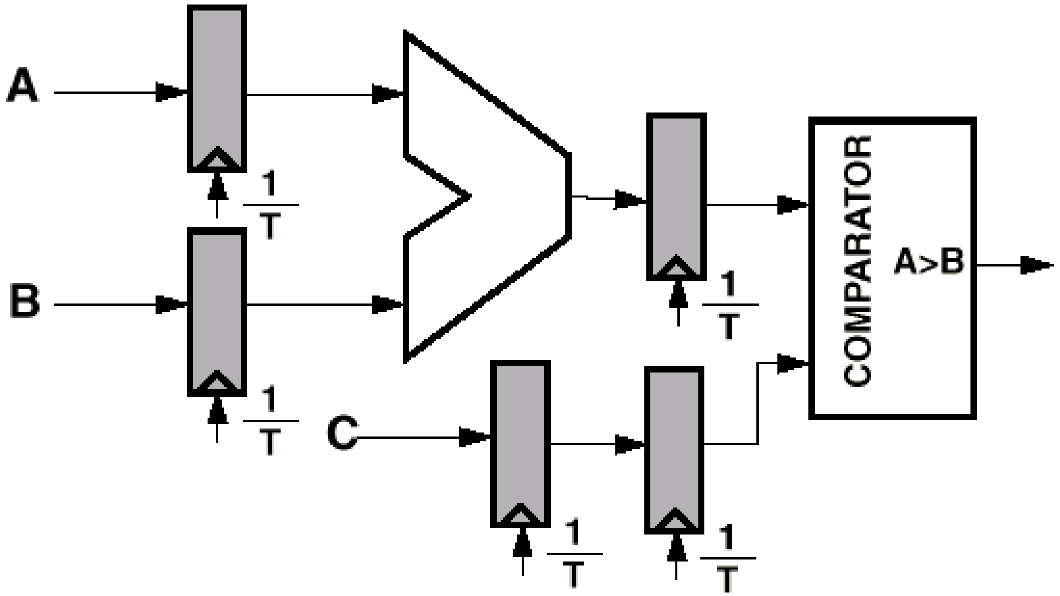


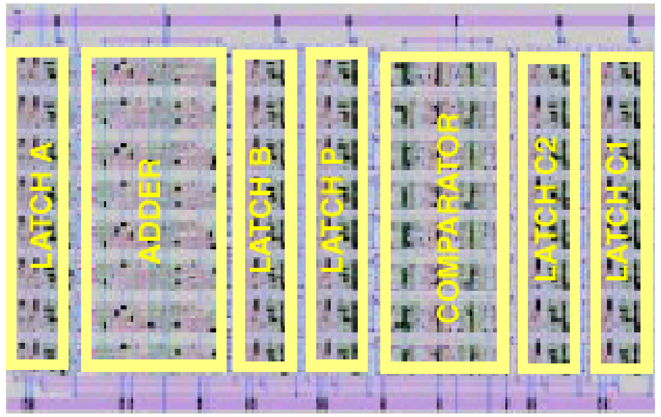


Area = 1476 x 1219 μ^2

- 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

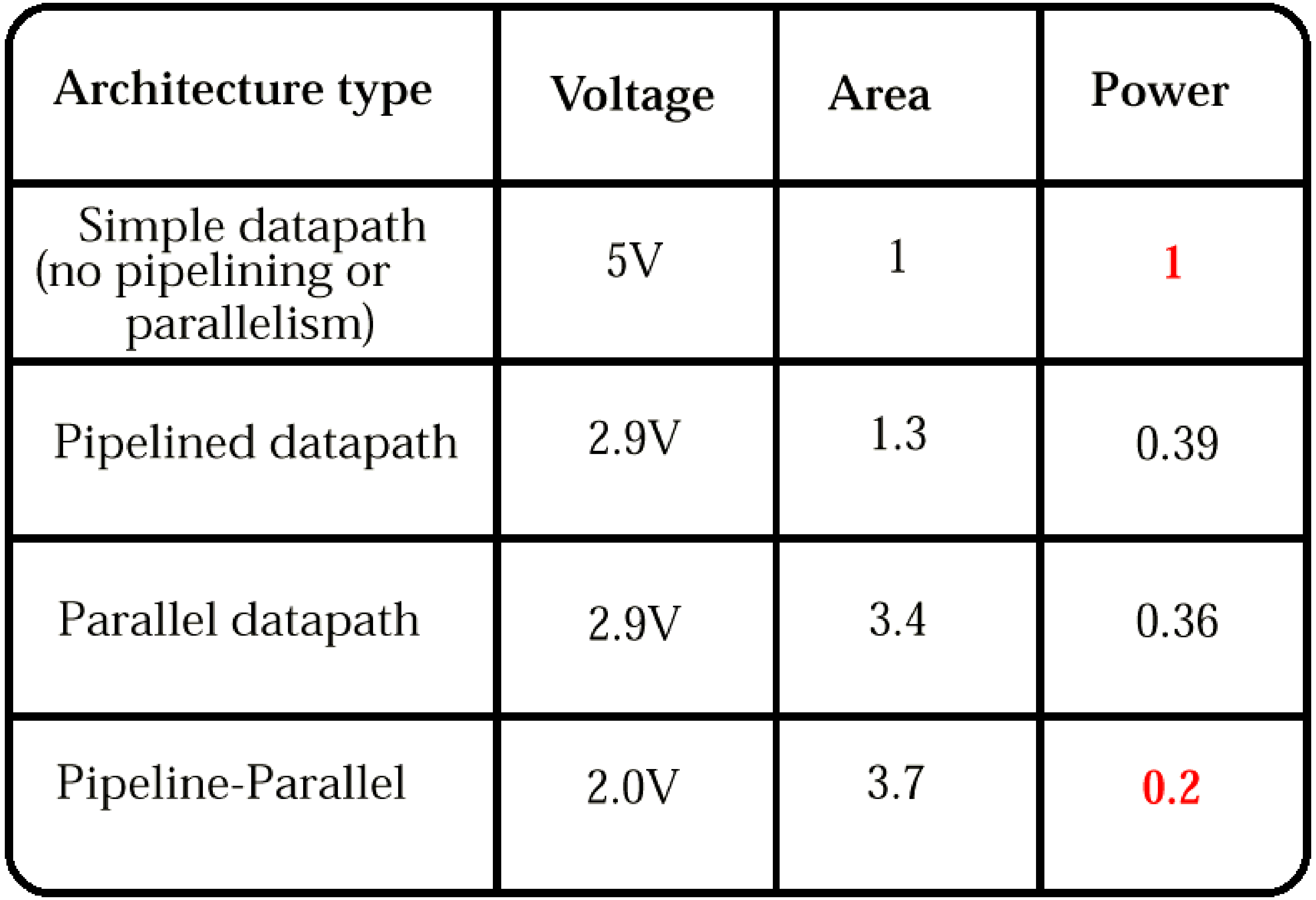




Area = 640 x 1081 μ^2

- 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

Next Lecture

- Low-power design
  - Multiple supplies
  - Dynamic voltage scaling