EECS 427
Lecture 12: Dynamic Power Reduction
Readings: 5.5, 11.7, CBF Ch.4

[Partly adapted from Irwin and Narayanan]

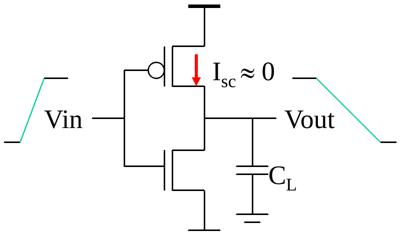
Reminders

- CAD5 is due Wednesday 10/28
  - You can submit it by Thursday 10/29 at noon
- Lecture on 11/2 will be taught by Wei-Hsiang
  - Zhengya’s office hour on 11/2 is moved to 11/4 with extended office hour 3 – 5 pm
- HW4 (detailed proposal) is due 11/16
  - You should be working on your project concurrently
  - Keep a journal of what you did, why you made certain design decisions, intermediate simulation results, etc.
  - HW4 handout is posted online with templates, DAC student design contest information, and past project reports

Last Time

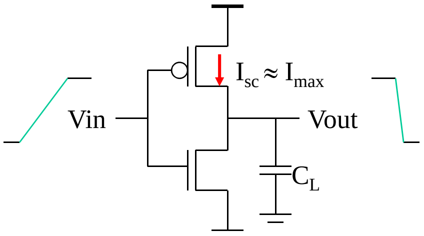
- Power and energy
  - Power (energy consumption per second) matters in package design, cooling, power rail design, and noise immunity etc.
  - Energy matters in battery-powered, portable devices
- Dynamic power
  - Due to switching – reduce VDD, lower capacitance, reduce switching probability
- Leakage power
  - Mainly due to subthreshold currents – increase V_T, sleep
- Short-circuit current
  - Both pull-up and pull-down are on at the same time – control input/output slopes, less important now at lower VDD

Short Circuit Current



Large capacitive load

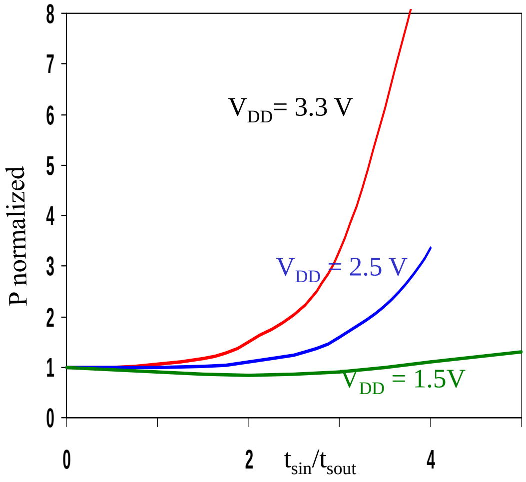
Output fall time significantly larger than input rise time.



Small capacitive load

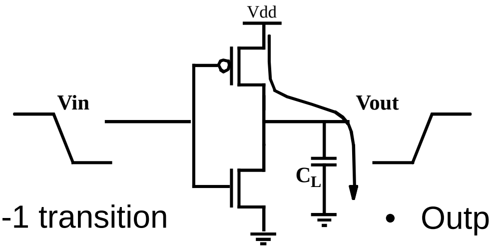
Output fall time substantially smaller than the input rise time.

Supply Voltage Scaling



If $V_{DD} < V_{Tn} + |V_{Tp}|$ then P_{sc} is eliminated since both devices are never on at the same time.

Dynamic Power



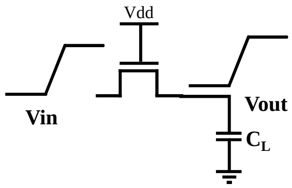
- Output 0-1 transition

Energy drawn from supply =
Energy stored in C_L =
Energy dissipated =

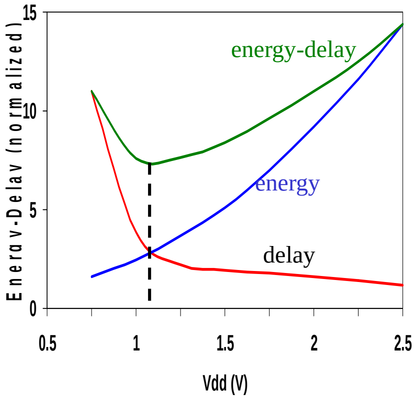
- Output 1-0 transition

Energy dissipated =

Total switching power =



Energy and Delay



Power and Energy Design Space

	Constant Throughput/Latency		Variable Throughput/Latency
Energy	Design Time	Non-active Modules	Run Time
Active	Logic Design Reduced V_{dd} Sizing Multi-V_{dd}	Clock Gating	DFS, DVS (Dynamic Freq, Voltage Scaling)
Leakage	+ Multi-V_T	Sleep Transistors Multi-V_{dd} Variable V_T	+ Variable V_T

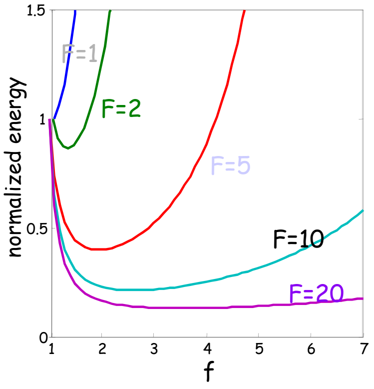
Power and Energy Design Space

	Constant Throughput/Latency		Variable Throughput/Latency
Energy	Design Time	Non-active Modules	Run Time
Active	Logic Design Reduced V_{dd} Sizing Multi-V_{dd}	Clock Gating	DFS, DVS (Dynamic Freq, Voltage Scaling)
Leakage	+ Multi-V_T	Sleep Transistors Multi-V_{dd} Variable V_T	+ Variable V_T

Device Sizing

- Device sizing affects dynamic energy consumption
  - gain is largest for networks with large overall effective fan-outs ($F = C_L/C_{g,1}$)

- The optimal gate sizing factor (f) for dynamic energy is smaller than the one for performance, especially for large F's
  - e.g., for F=20, f_{opt}(energy) = 3.53 while f_{opt}(performance) = 4.47
- If energy is a concern avoid oversizing beyond the optimal



From Nikolic, UCB

Power and Energy Design Space

	Constant Throughput/Latency		Variable Throughput/Latency
Energy	Design Time	Non-active Modules	Run Time
Active	Logic Design Reduced V_{dd} Sizing Multi-V_{dd}	Clock Gating	DFS, DVS (Dynamic Freq, Voltage Scaling)
Leakage	+ Multi-V_T	Sleep Transistors Multi-V_{dd} Variable V_T	+ Variable V_T

Data Dependency

- Switching activity, $P_{0\rightarrow1}$, has two components
  - A static component – function of the logic topology
  - A dynamic component – function of the timing behavior (glitching)

Static transition probability

$$P_{0\rightarrow1} = P_{out=0} \text{ x } P_{out=1}$$
$$= P_0 \text{ x } (1-P_0)$$

2-input NOR Gate

A	B	Out
0	0	1
0	1	0
1	0	0
1	1	0

With input signal probabilities

$P_{A=1} = 1/2$

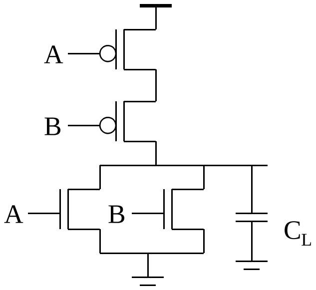
$P_{B=1} = 1/2$

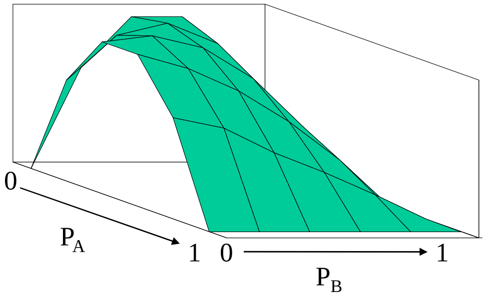
NOR static transition probability

$= 3/4 \text{ x } 1/4 = 3/16$

Transition Probabilities

- Switching activity is a strong function of the input signal statistics
  - P_A and P_B are the probabilities that inputs A and B are one

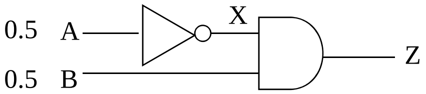




$$P_{0\to1} = P_0 \text{ x } P_1 = (1-(1-P_A)(1-P_B))\,(1-P_A)(1-P_B)$$

Transition Probabilities

	$P_{0\to1} = P_{out=0} \text{ x } P_{out=1}$
NOR	$(1 - (1 - P_A)(1 - P_B)) \text{ x } (1 - P_A)(1 - P_B)$
OR	$(1 - P_A)(1 - P_B) \text{ x } (1 - (1 - P_A)(1 - P_B))$
NAND	$P_AP_B \text{ x } (1 - P_AP_B)$
AND	$(1 - P_AP_B) \text{ x } P_AP_B$
XOR	$(1 - (P_A + P_B - 2P_AP_B)) \text{ x } (P_A + P_B - 2P_AP_B)$

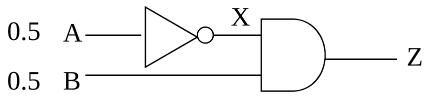


For X: $P_{0\to1} =$

For Z: $P_{0\to1} =$

Transition Probabilities

	$P_{0\rightarrow1} = P_{out=0}$ x $P_{out=1}$
NOR	$(1 - (1 - P_A)(1 - P_B))$ x $(1 - P_A)(1 - P_B)$
OR	$(1 - P_A)(1 - P_B)$ x $(1 - (1 - P_A)(1 - P_B))$
NAND	$P_A P_B$ x $(1 - P_A P_B)$
AND	$(1 - P_A P_B)$ x $P_A P_B$
XOR	$(1 - (P_A + P_B - 2P_A P_B))$ x $(P_A + P_B - 2P_A P_B)$



For X: $P_{0\rightarrow1} = P_0$ x $P_1 = (1-P_A)\,P_A$

$= 0.5$ x $0.5 = 0.25$

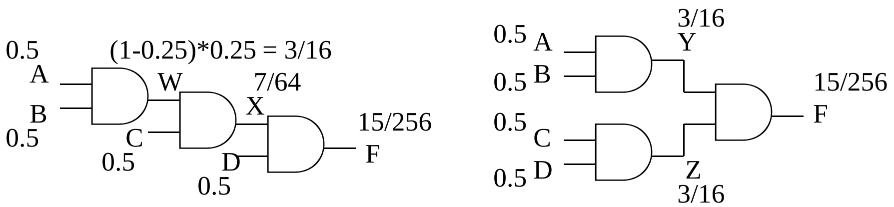
For Z: $P_{0\rightarrow1} = P_0$ x $P_1 = (1-P_X P_B)\,P_X P_B$

$= (1 – (0.5$ x $0.5))$ x $(0.5$ x $0.5) = 3/16$

Logic Restructuring

- Logic restructuring: changing the topology of a logic network to reduce transitions

AND: $P_{0\rightarrow1} = P_0$ x $P_1 = (1 - P_A P_B)$ x $P_A P_B$

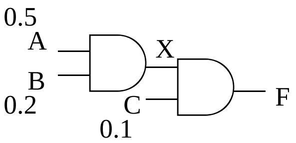


Chain implementation has a lower overall switching activity than the tree implementation for random inputs

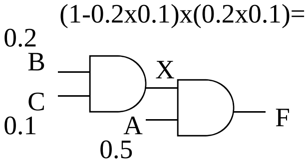
Ignores glitching effects

Input Ordering

(1-0.5x0.2)x(0.5x0.2)=0.09

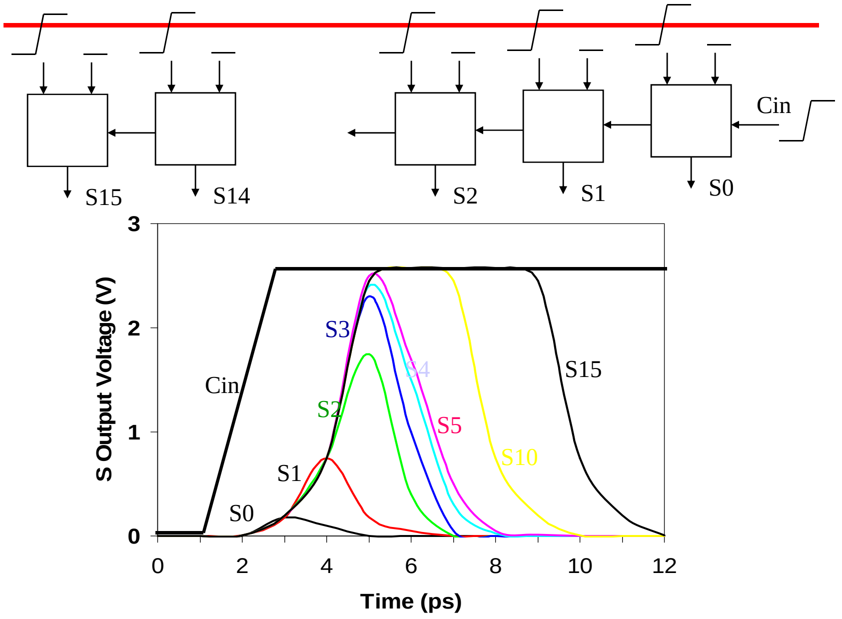


(1-0.2x0.1)x(0.2x0.1)=0.0196



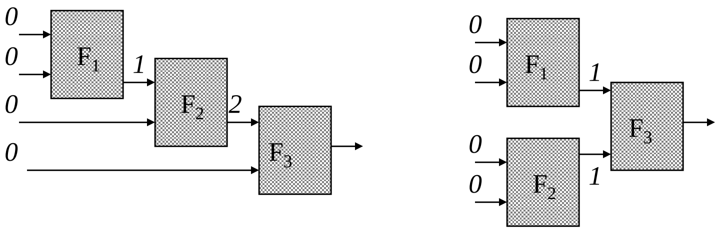
Beneficial to postpone the introduction of signals with a high transition rate (signals with signal probability close to 0.5)

Glitching in an RCA



Balanced Delay Paths

- Glitching is due to a mismatch in the path lengths in the logic network; if all input signals of a gate change simultaneously, no glitching occurs

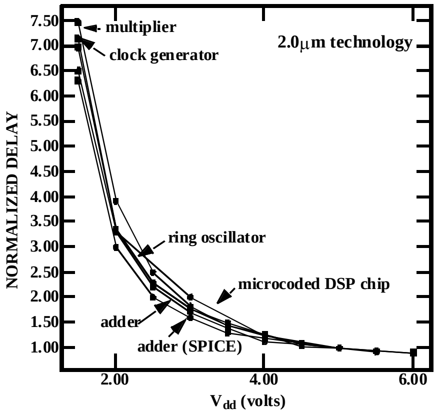


So equalize the lengths of timing paths through logic

Power and Energy Design Space

	Constant Throughput/Latency		Variable Throughput/Latency
Energy	Design Time	Non-active Modules	Run Time
Active	Logic Design Reduced V_{dd} Sizing Multi-V_{dd}	Clock Gating	DFS, DVS (Dynamic Freq, Voltage Scaling)
Leakage	+ Multi-V_T	Sleep Transistors Multi-V_{dd} Variable V_T	+ Variable V_T

Lower V_{dd} Increases Delay



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

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

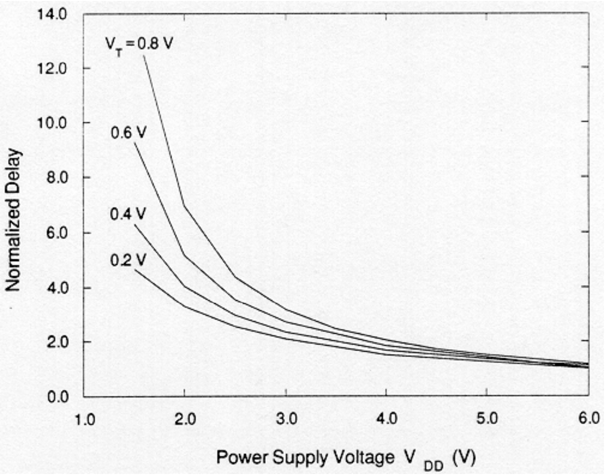
$$\frac{T_{d(Vdd=2)}}{T_{d(Vdd=5)}} = \frac{(2) * (5 - 0.7)^2}{(5) * (2 - 0.7)^2}$$

$$\approx 4$$

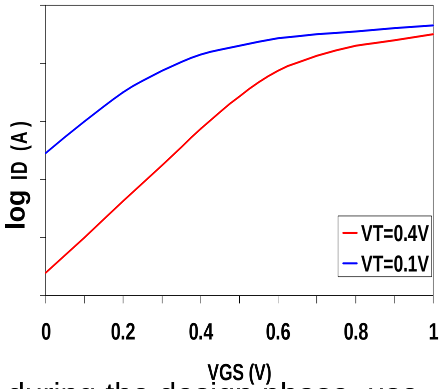
The exponent will change next time we see this…

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

Reducing V_{th} to Offset Delay Penalty

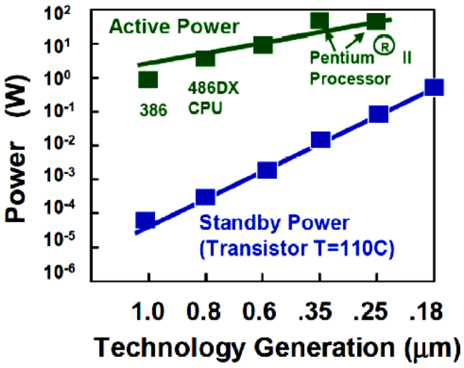


Leakage as a Function of V_T



- Reducing the V_T **increases** the sub-threshold leakage current (<u>exponentially</u>)
  - ~90mV reduction in V_T increases leakage by 10X
- But, reducing V_T **decreases** gate delay (increases performance)
- Determine the critical paths during the design phase, use low V_T devices on those paths for speed (requires dual-Vt)
- Use a high V_T in rest of logic to control leakage
  - Can provide total leakage reduction of up to 80%

Leakage Power



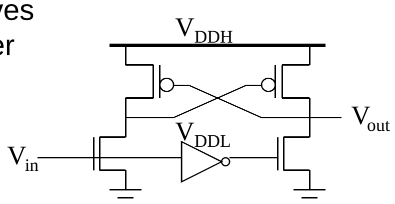
Power and Energy Design Space

	Constant Throughput/Latency		Variable Throughput/Latency
Energy	Design Time	Non-active Modules	Run Time
Active	Logic Design Reduced V_{dd} Sizing Multi-V_{dd}	Clock Gating	DFS, DVS (Dynamic Freq, Voltage Scaling)
Leakage	+ Multi-V_T	Sleep Transistors Multi-V_{dd} Variable V_T	+ Variable V_T

Multiple V_{DD} Considerations

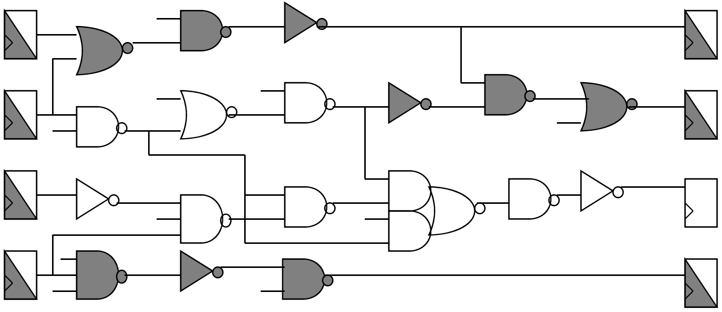
- How many V_{DD}? – Two is common
  - Two supplies: one for core and one for I/O

- When combining multiple supplies, level converters are required whenever a module at the lower supply drives a gate at the higher supply (step-up)
  - If a gate supplied with V_{DDL} drives a gate at V_{DDH}, the PMOS never turns off
  - Level converters are not needed for step-down
  - Overhead of level converters can be mitigated by doing conversions at register boundaries and embedding the level conversion inside the flipflop



Dual-Supply Inside a Logic Block

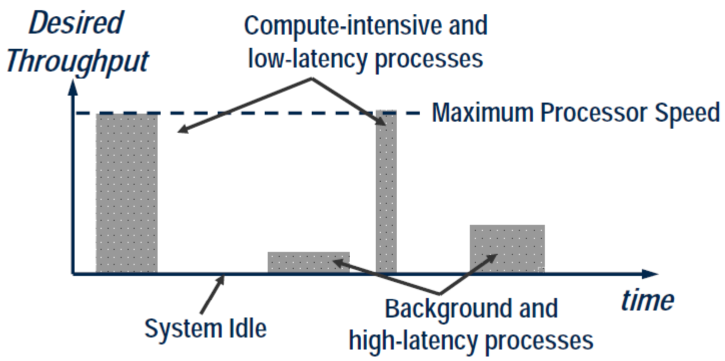
- Minimum energy consumption is achieved if all logic paths are critical (have the same delay)
- Clustered voltage-scaling
  - Each path starts with V_{DDH} and switches to V_{DDL} (gray logic gates) when delay slack is available
  - Level conversion is done in the flipflops at the end of the paths



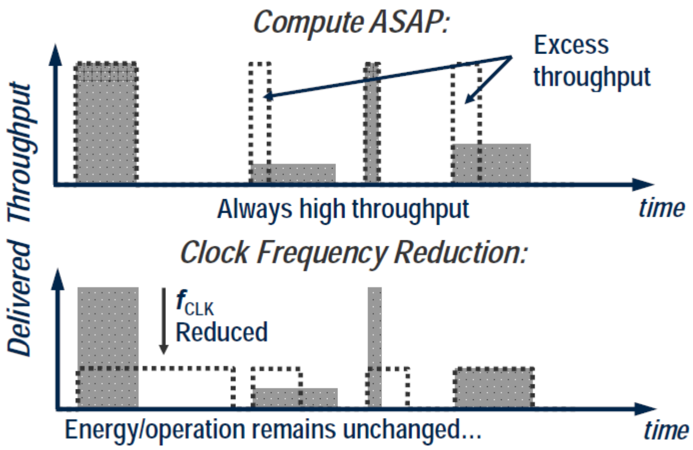
Power and Energy Design Space

	Constant Throughput/Latency		Variable Throughput/Latency
Energy	Design Time	Non-active Modules	Run Time
Active	Logic Design Reduced V_{dd} Sizing Multi-V_{dd}	Clock Gating	DFS, DVS (Dynamic Freq, Voltage Scaling)
Leakage	+ Multi-V_T	Sleep Transistors Multi-V_{dd} Variable V_T	+ Variable V_T

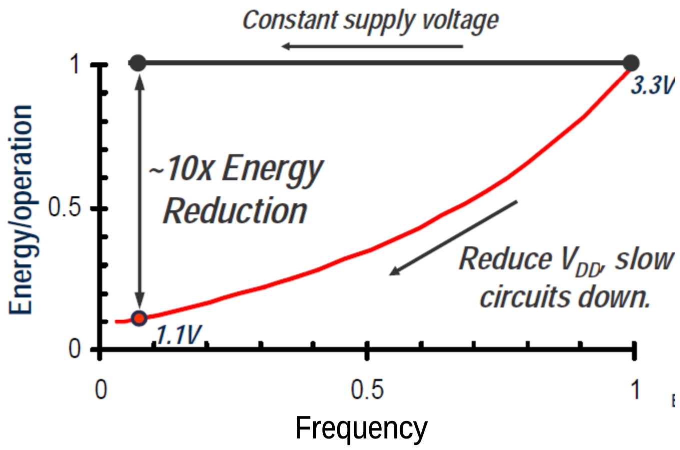
Processor Usage Model



Fixed Clock Frequency

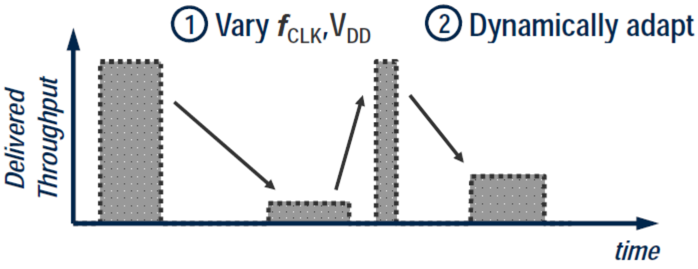


Voltage and Frequency Scaling



Burd, ISSCC'00

Dynamic Voltage and Frequency Scaling



Burd, ISSCC'00

Recent Examples

- Intel Xscale: 180nm 1.8V bulk-CMOS (2000)
  - 0.7-1.75V, 200-1000MHz, 55-1500mW (typ)
  - Max energy efficiency: 23 MIPS/mW

- IBM PowerPC: 180nm 1.8V bulk-CMOS (2002)
  - 0.9-1.95V, 11-380MHz, 53-500mW (typ)
  - Max energy efficiency: 11 MIPS/mW

- Transmeta Crusoe: 130nm 1.5V bulk-CMOS (2003)
  - 0.8-1.3V, 300-1000MHz, 0.85-7.5W (peak)

- Intel Pentium M: 130nm 1.5V bulk-CMOS (2003)
  - 0.95-1.5V, 600-1600MHz, 4.2-31W (peak)

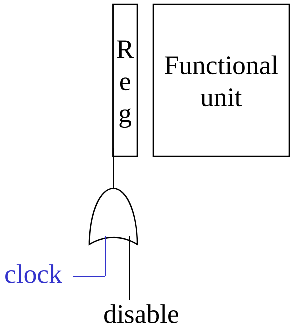
Power and Energy Design Space

	Constant Throughput/Latency		Variable Throughput/Latency
Energy	Design Time	Non-active Modules	Run Time
Active	Logic Design Reduced V_{dd} Sizing Multi-V_{dd}	Clock Gating	DFS, DVS (Dynamic Freq, Voltage Scaling)
Leakage	+ Multi-V_T	Sleep Transistors Multi-V_{dd} Variable V_T	+ Variable V_T

Clock Gating

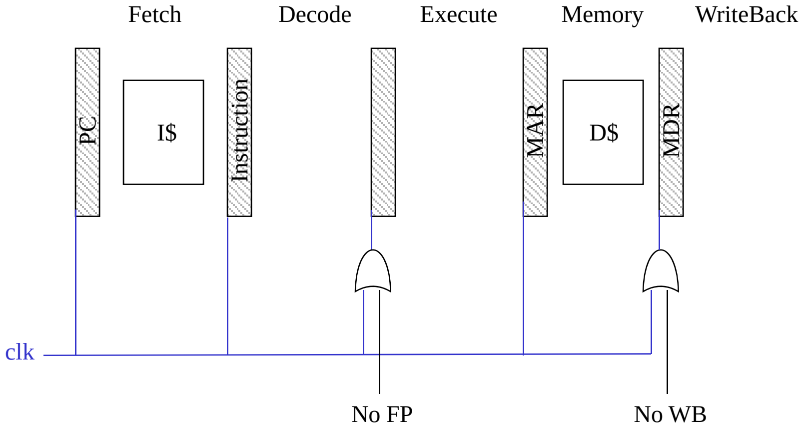
- Most popular method for power reduction of clock signals and functional units

- Gate off clock to idle functional units
  - e.g., floating point units
  - need logic to generate disable signal
    - increases complexity of control logic
    - consumes power
    - timing critical to avoid clock glitches at OR gate output
  - additional gate delay on clock signal
    - gating OR gate can replace a buffer in the clock distribution tree



Pipelined Datapath

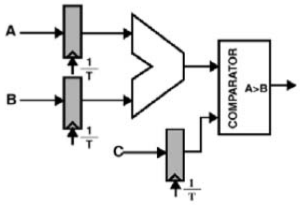
- For idle units (e.g., floating point units in Exec stage, WB stage for instructions with no write back operation)

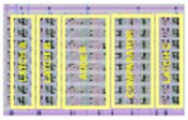


Power and Energy Design Space

	Constant Throughput/Latency		Variable Throughput/Latency
Energy	Design Time	Non-active Modules	Run Time
Active	Logic Design Reduced V_{dd} Sizing Multi-V_{dd}	**Architectural considerations** Clock Gating	DFS, DVS (Dynamic Freq, Voltage Scaling)
Leakage	+ Multi-V_T	Sleep Transistors Multi-V_{dd} Variable V_T	+ Variable V_T

Architectural Consideration



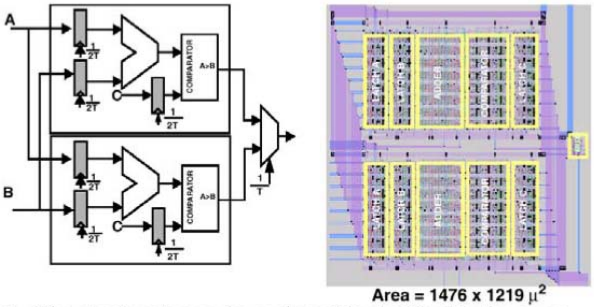


Area = 636 x 833 μ^2

- 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*)

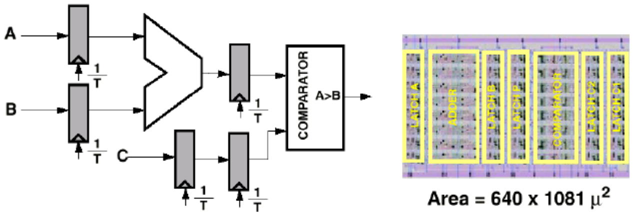
Parallel Datapath



Area = 1476 x 1219 μ^2

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

Pipelined Datapath



Area = 640 x 1081 μ^2

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

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**

Power and Energy Design Space

	Constant Throughput/Latency		Variable Throughput/Latency
Energy	Design Time	Non-active Modules	Run Time
Active	Logic Design Reduced V_{dd} Sizing Multi-V_{dd}	Clock Gating	DFS, DVS (Dynamic Freq, Voltage Scaling)
Leakage	+ Multi-V_T	Sleep Transistors Multi-V_{dd} Variable V_T	+ Variable V_T