Fundamentals of VLSI

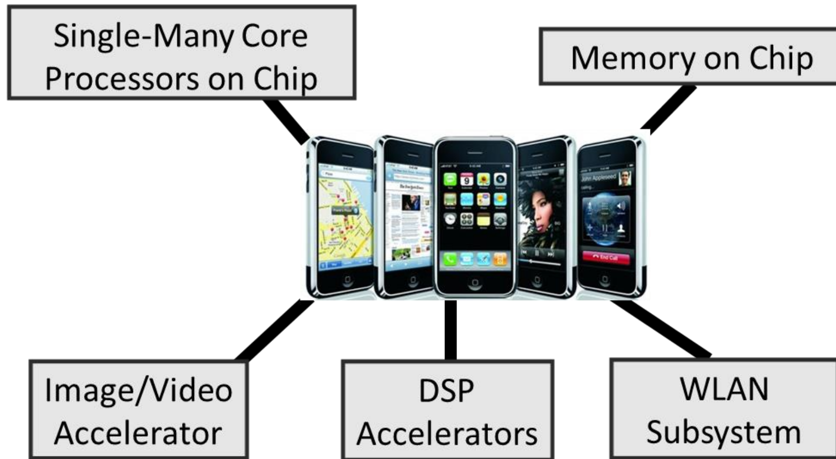
CMOS Power Consumption

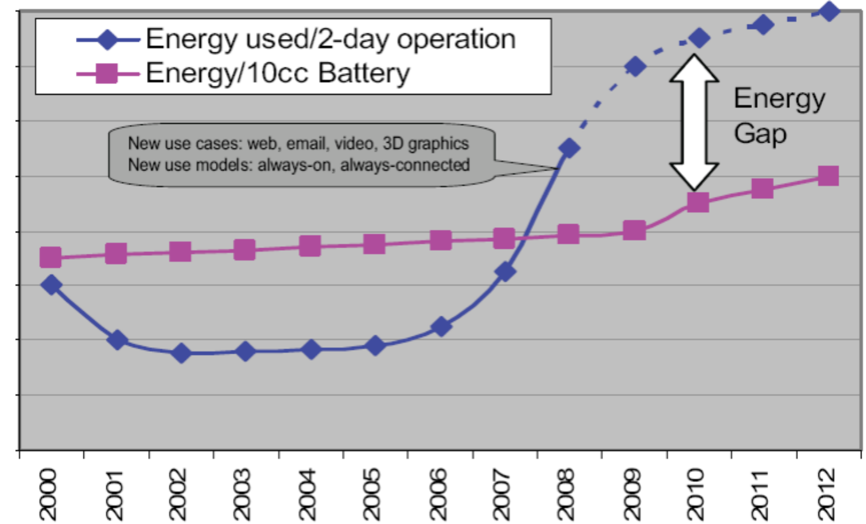
Andreas Burg

Telecommunications Circuits Laboratory

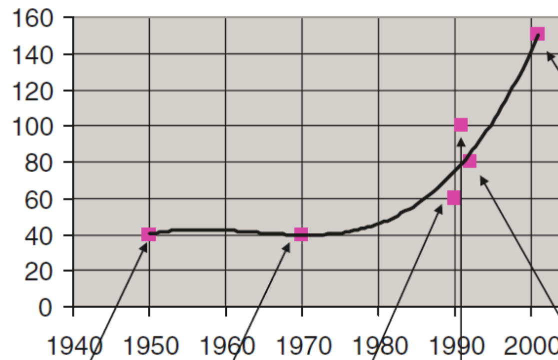
Energy Bottleneck

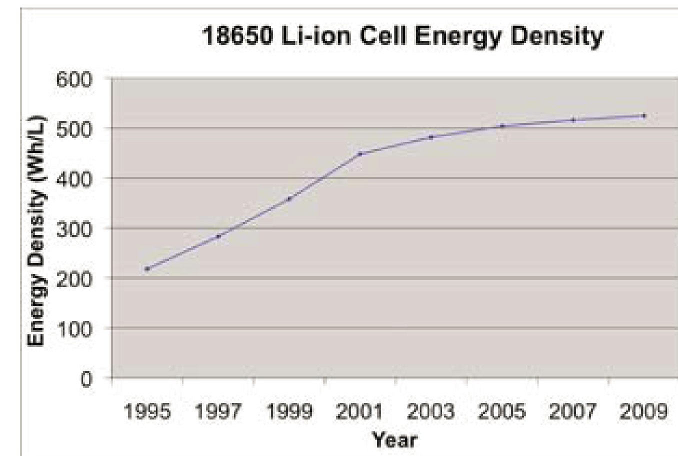
- **Mobile devices: energy-efficiency**



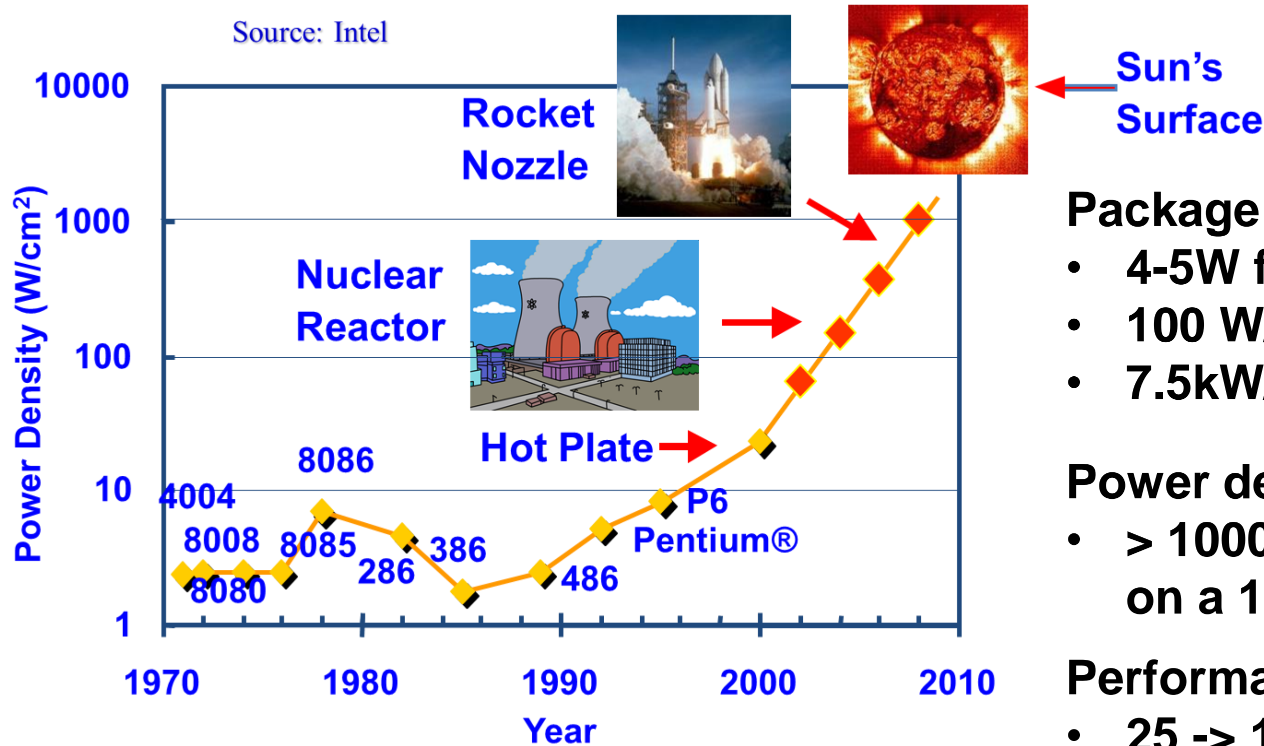


- **Battery capacity grows only very slowly**
  - **Boost in the 1990s due to Mobile Phone introduction**
  - **Capacity growth stalled since 2000 at the limit of Li-ion**
  - **Only 3%-7% annual improvement**





Power Bottleneck



Package cost

- 4-5W for cheap packages
- 100 W/cm^2 for air cooling
- 7.5kW/rack

Power delivery

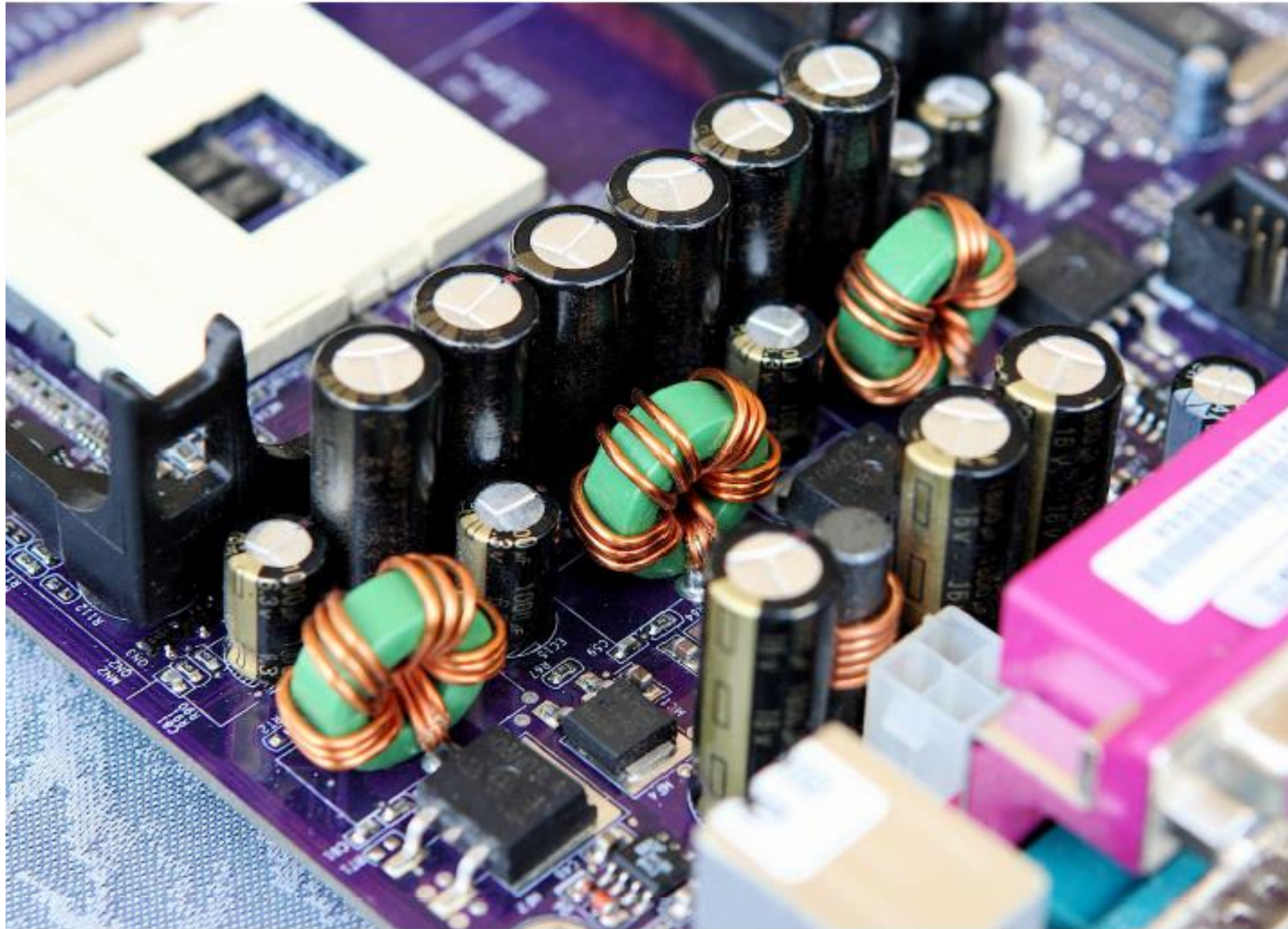
- > 1000 pins for power delivery on a 100W processor

Performance penalty

- 25 -> 100 deg. C : 30%

- **Thermal Design Power:** upper limit on power consumption
  - **Microprocessors for servers: ~30-100 W/cm^2**
  - **Mobile devices: ~3W total (handheld)**

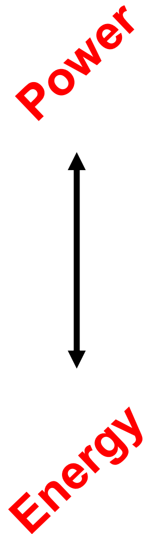
Power Bottleneck



Complex, large and costly power supply circuits: Three-phase step-down converter built from toroidal coils, power MOSFETs, and electrolytic capacitors.

Motivation for Low Power Design

- **High-performance circuits**
  - How to get the heat out?
  - How to supply massive currents at very low voltages?
  - How to avoid critical voltage drops on supply rails?



- **Battery-operated circuits**
  - How long can we operate a device on a battery charge?

Basics in CMOS Power Consumption and Low Power Design

- Active Power Consumption in CMOS
- Leakage Power Consumption
- Voltage Scaling and Sub-VT Design

- Four phenomena dissipate energy in digital CMOS circuits
  - Charging and discharging of capacitive loads
  - Crossover currents
  - Leakage currents
  - (Driving of resistive loads)

Power (charging and discharging of capacitive loads; crossover currents)

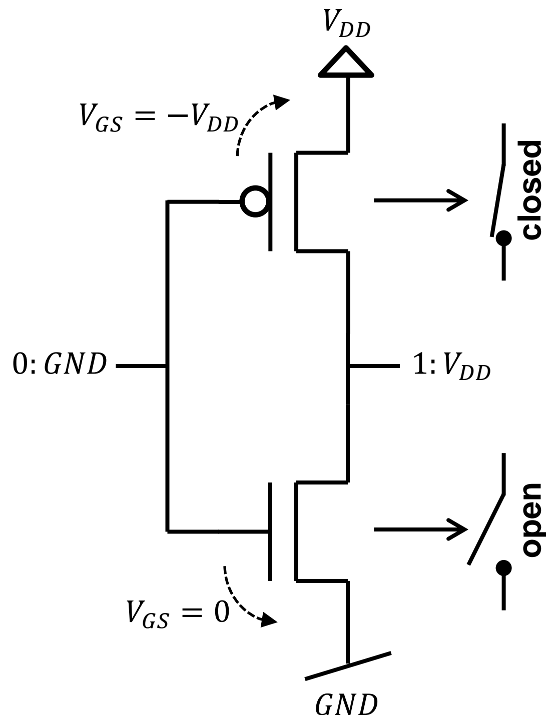
Energy (leakage currents; driving of resistive loads)

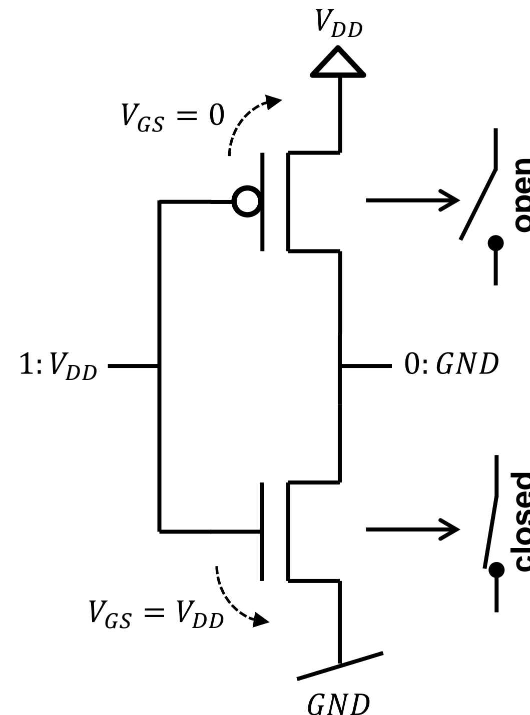
Active Power Consumption in CMOS

Basic Inverter in Steady State (ideal)

- PMOS performs pull-up to V_{DD}
- NMOS performs pull-down to GND

- **Complementary gate:** output connected to either V_{DD} or GND
  - Static (steady state)





  - Ideally, no current path from V_{DD} to GND
  - Ideally, no static power consumption

CMOS Transistors: Parameters

The gain factor β is a function of **process parameters** and **layout geomerty**

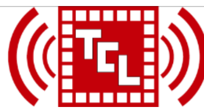
$$\beta = \frac{\mu\varepsilon_{OX}}{t_{OX}}\frac{W}{L}$$

where

t_{OX}	gate dielectric thikness
ε_{OX}	gate dielectric permitivity
μ	effective carrier mobility in inversion layer
W	channel (gate) width
L	channel (gate) length

W, L: **Design parameters**

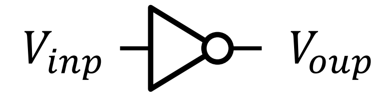
- **Designer sets the drive-strength by controlling width and length of the transistor**



Inverter as non-linear amplifier with a large, but finite gain in the transition region

range	applies when	n-channel ▼	p-channel ▲
A	$0 \leq U_{inp} \leq U_{th\,n}$	subthreshold	linear
B	$U_{th\,n} < U_{inp} < U_{inv}$	saturation	linear
C	$U_{inp} \approx U_{inv}$	saturation	saturation
D	$U_{inv} < U_{inp} < U_{dd} + U_{th\,p}$	linear	saturation
E	$U_{dd} + U_{th\,p} \leq U_{inp} \leq U_{dd}$	linear	subthreshold

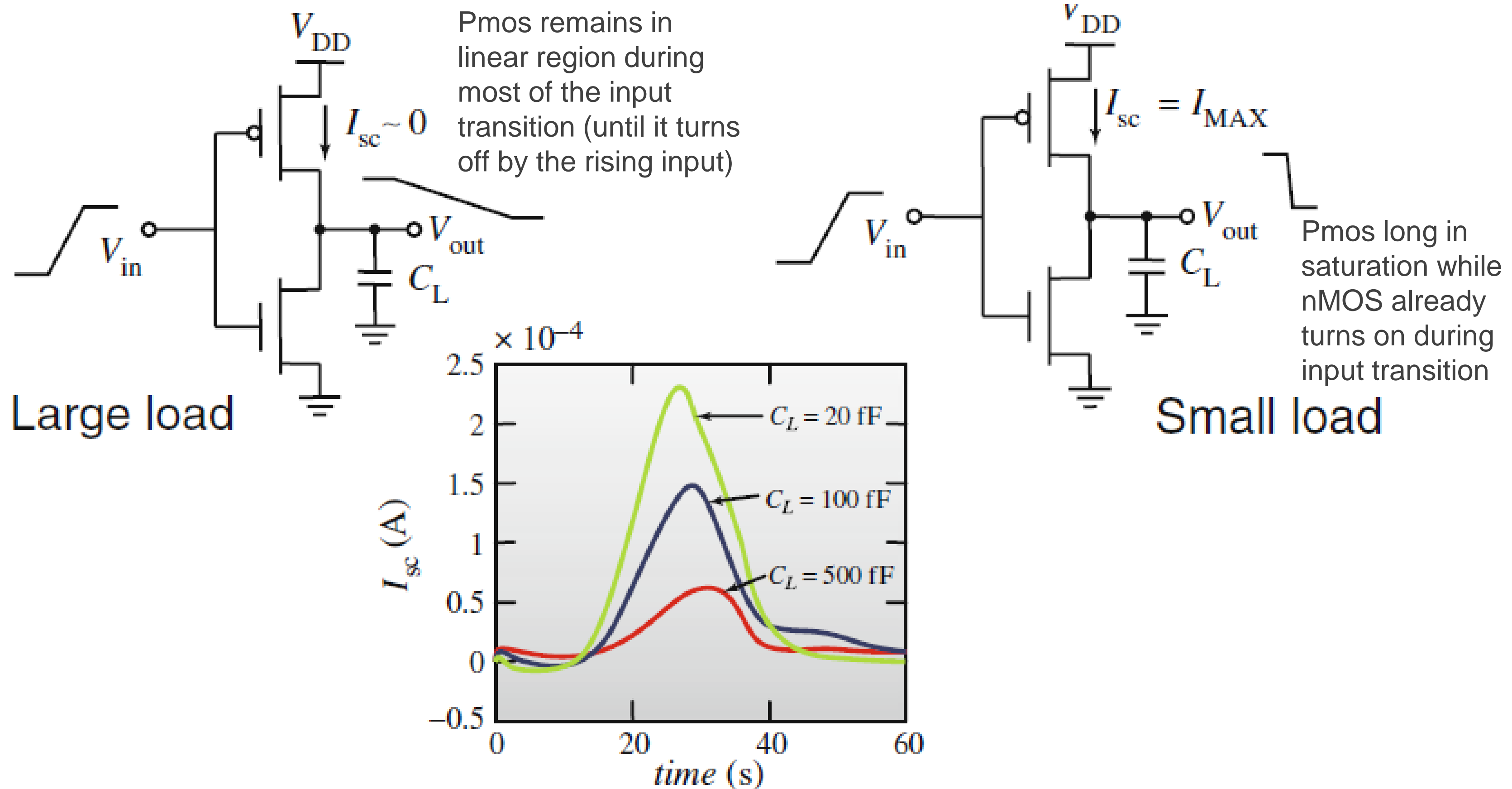
Dominant during transition region (rows B and D)



(a) Transfer characteristic (b) Crossover current (c) Logic states

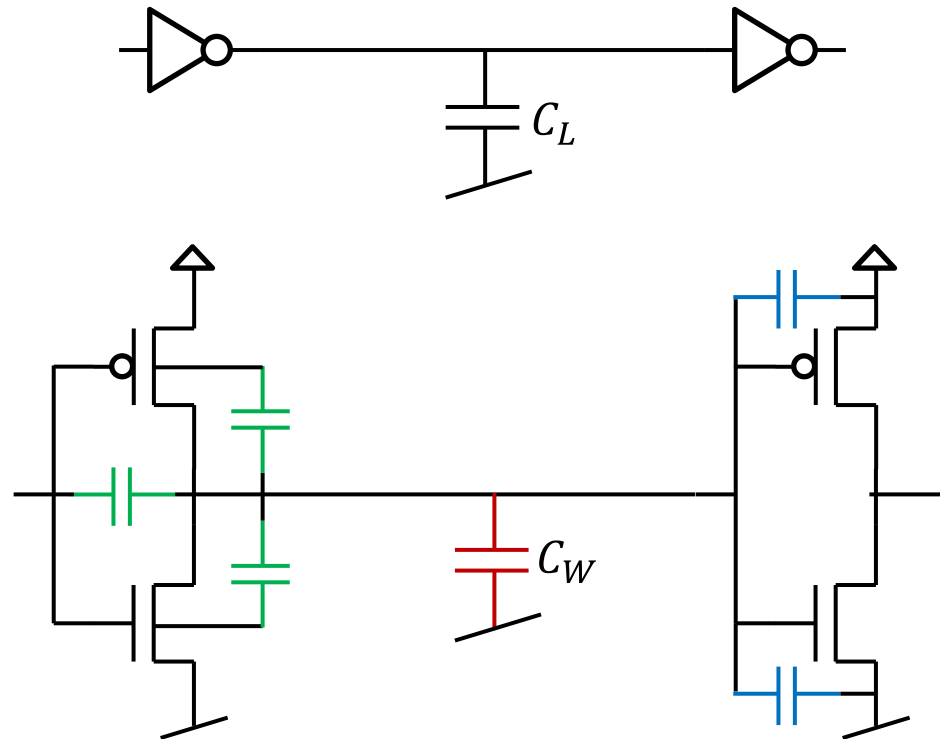
- **Cross-over currents lead to power consumption during transients**

Minimizing transient currents



- – **Fast input slow output: driving device quickly shuts of completely**
- – **Slow input fast output: driving device remains long in linear region**
- **Input of one device is output of the other device: balance input-output delay for optimum power consumption**

CMOS Gates With Capacitive Load

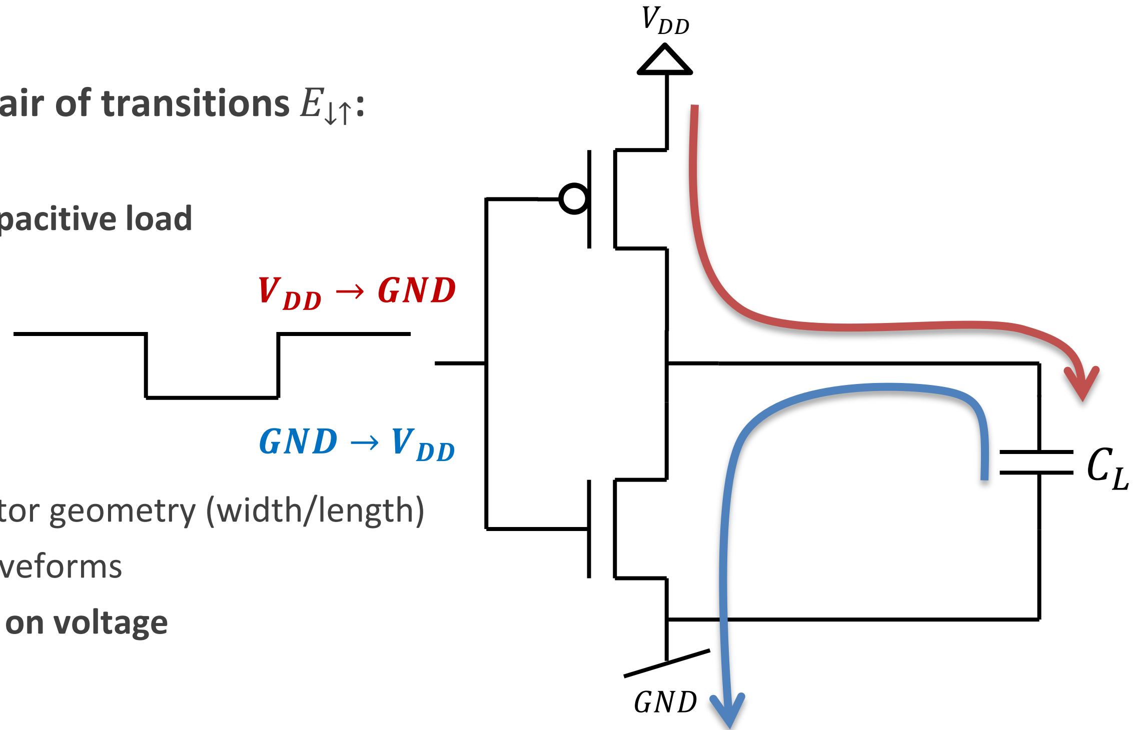


Wider transistors increase the gain factor (drive) but also increase the load (capacitance)

- **Various capacitances are merged into a single load capacitor C_L**
  - **Intrainsic MOS transistor capacitors (driver)**
  - Extrinsig **(fanout)** MOS transistor capacitances
  - **Interconnect capacitance**

Reminder: Power/Energy of an Inverter with Capacitive Load

- **Energy consumed during one pair of transitions $E_{\downarrow\uparrow}$:**
  - **Cross-over currents**
  - **Charge pumped onto the capacitive load (dominant):**



- $E_{\downarrow\uparrow} = (C_L V_{dd})V_{dd} = C_L V_{dd}^2$
  - independent of transistor geometry (width/length)
  - Independent of the waveforms
  - **quadratic dependency on voltage**

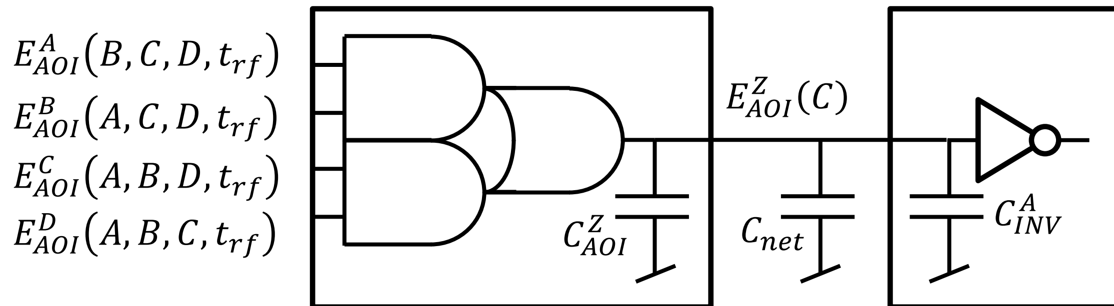
- Energy/transition
- $E_t = C_L V_{dd}^2/2$

- Power consumption = Energy/transition * transition/cycle (α) * frequency (f_{clk})
- $P = \frac{\alpha}{2} C_L V_{dd}^2 f_{clk}$

Gate-Level Power Modeling

- Power consumption is divided into
  - Net switching power
  - Internal power:
    - Internal power depends on actual input values
    - Power is consumed even if output does not change

- Library files: internal energy characterization for each cell at given supply voltage
  - Internal energy (cross-current, switching) per change in each input and output (as functions of input slope t_{rf} and output load C)
  - Contribution to capacitance of the connected net (input/output load)



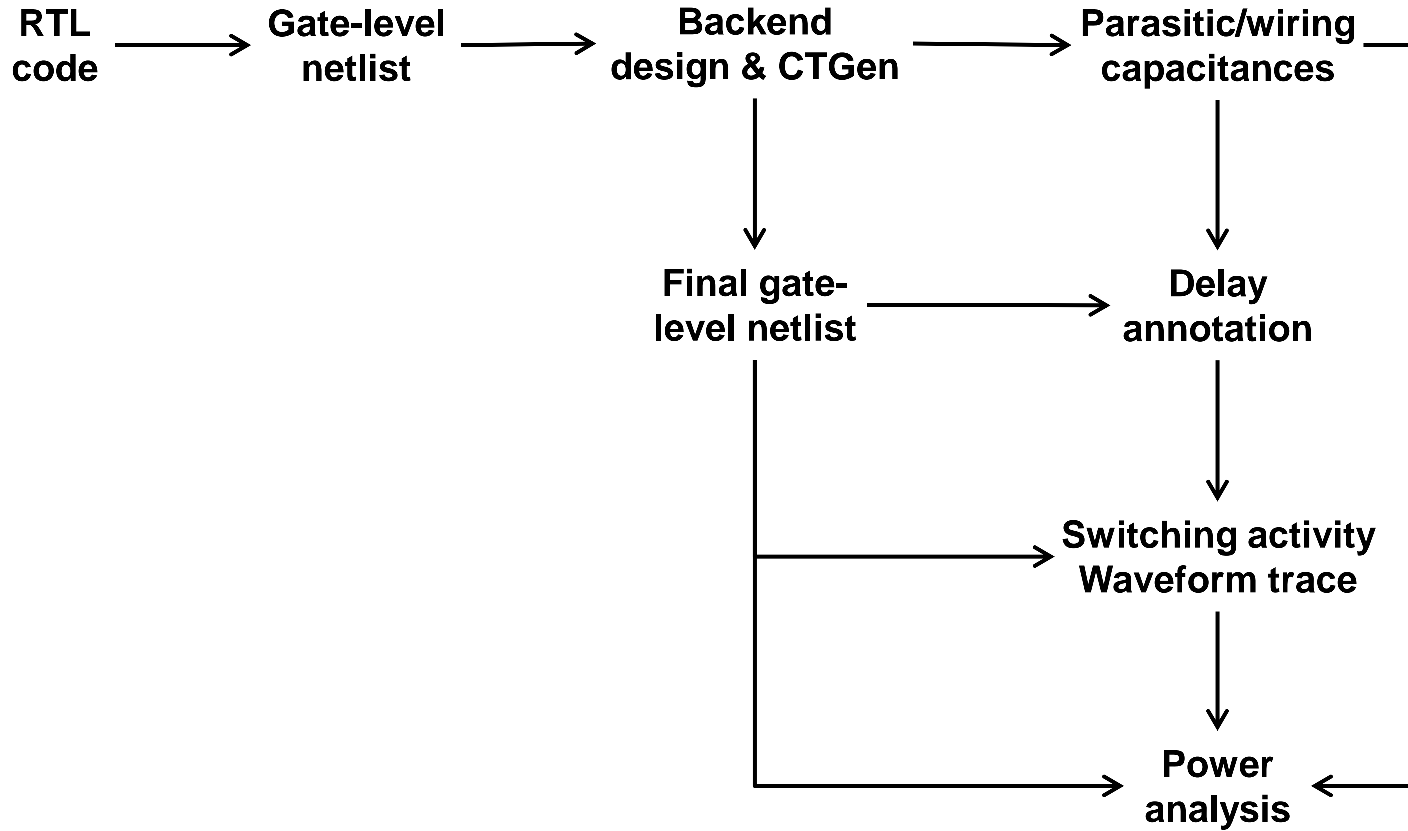
$$C = C_{AOI}^{Z} + C_{net} + C_{INV}^{A}$$

Methods for Power Analysis

What about the activity factor(s)?

- Fixed activity:
  Assume a constant activity factor for all nodes in the circuit
  - Very rough estimate and highly inaccurate
- Statistical power analysis:
  Assumes a given toggle activity at the input and propagates the activity throughout the circuit using statistical models of the gates
  - Does not account for correlation between signal values
  - No accounting for glitching activity
- Simulation based:
  Obtains toggle statistics from gate level simulations
  - Most accurate method
  - Slow

Gate-Level Power Analysis Flow

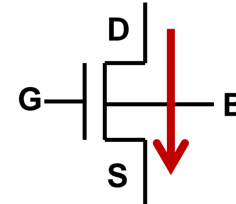


Leakage Power Consumption

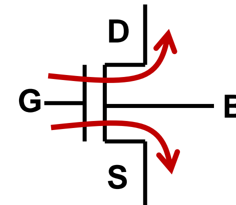
Leakage Power

- Transistors leak currents even when in off-state

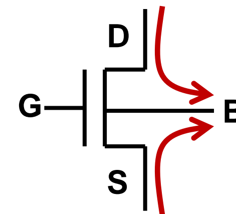
- Sources for leakage
  - Sub-threshold leakage
    - Dominant component in most circuits



  - Gate tunneling
    - Generally low, even in modern technologies due to high-k gate dielectrics
    - Decreases very rapidly with decreasing V_{dd}

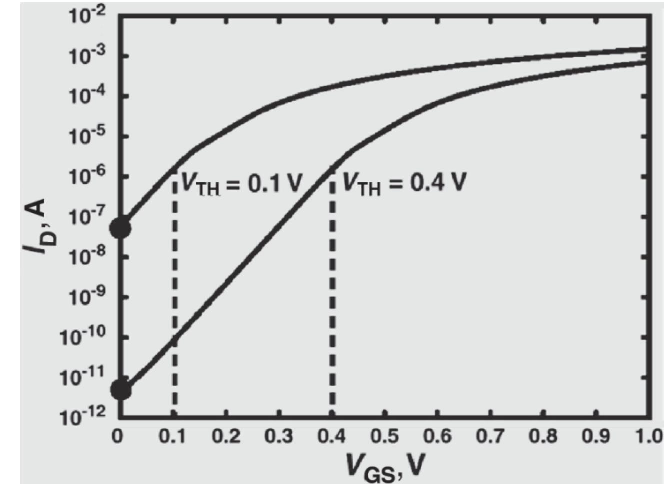


  - Junction current
    - Generally low
    - Decreases very rapidly with decreasing V_{dd}



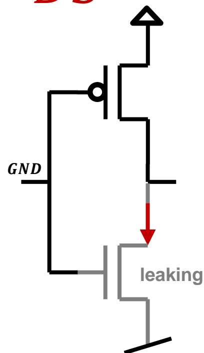
Leakage Power

- Long channel deices (>130nm): $I_{DS} = I_0 e^{\frac{V_{GS}-V_{th}}{v_t n}}$
  - I_{DS} mostly independent from Drain-Source Voltage
  - Leakage current depends strongly on $V_{GS} - V_{th}$
    - Decreasing threshold voltage increases leakage



- Impact of technology scaling on sub-threshold leakage (<130nm)
  - Drain-Induced Barrier Lowering (DIBL): V_{DS} modulates threshold voltage
  - I_{DS} becomes a function of V_{DS}

- $$I_{DS} = I_0 e^{\frac{V_{GS}-V_{th}+\lambda_{DS}V_{DS}}{v_t n}}$$



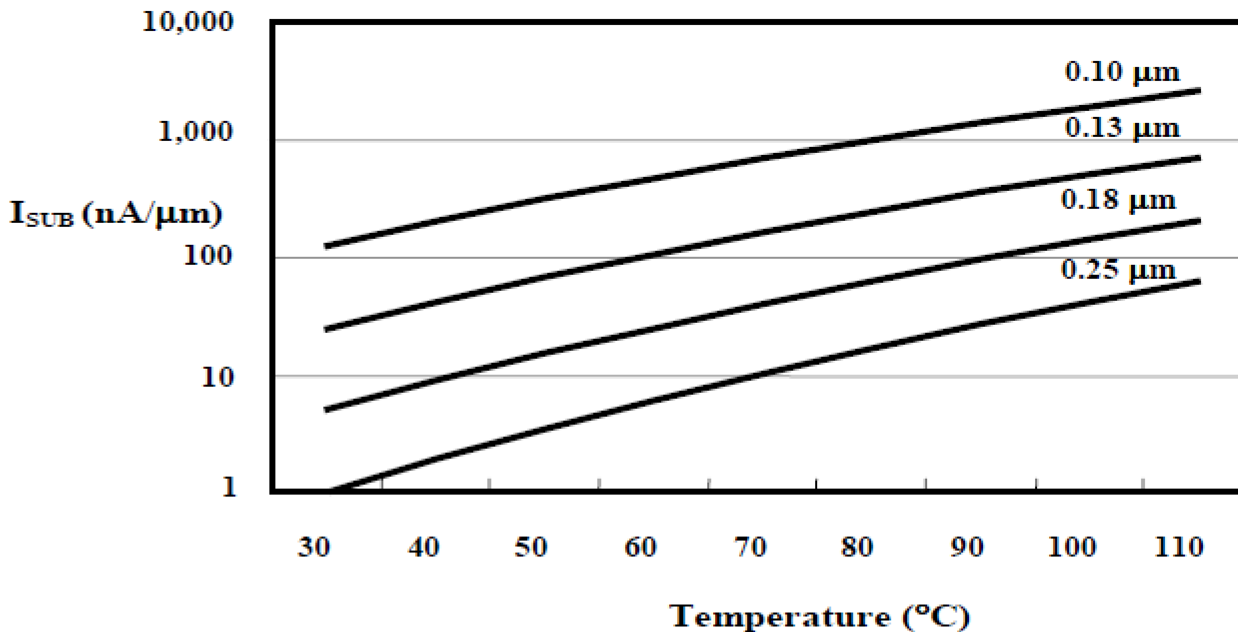
$$I_{leak} = I_0 e^{\frac{-V_{th}+\lambda_{DS}V_{DD}}{v_t n}}$$ ← **Voltage scaling reduces leakage**

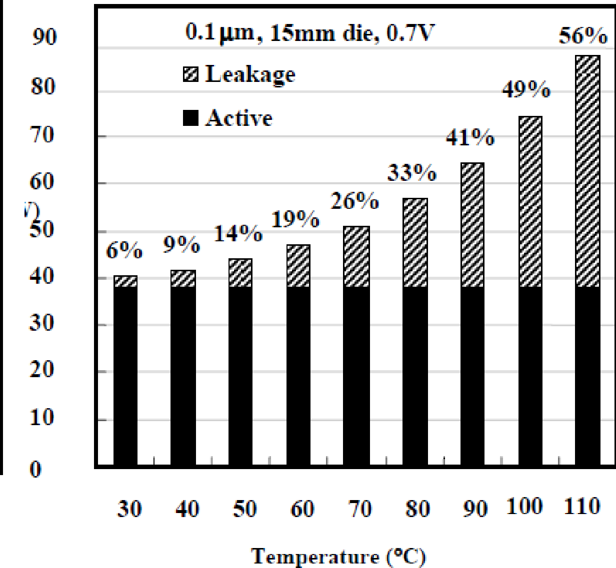
Drain current depends exponentially on thermal voltage $v_t = kT/q$

$$I_{DS} = I_0 e^{\frac{V_{GS}-V_{th}}{v_t n}}$$

$n<0$

- Exponential I_{DS} increase with temperature





Example: 0.7V, 100nm process, 15mm2 die

Vivek De, Intel

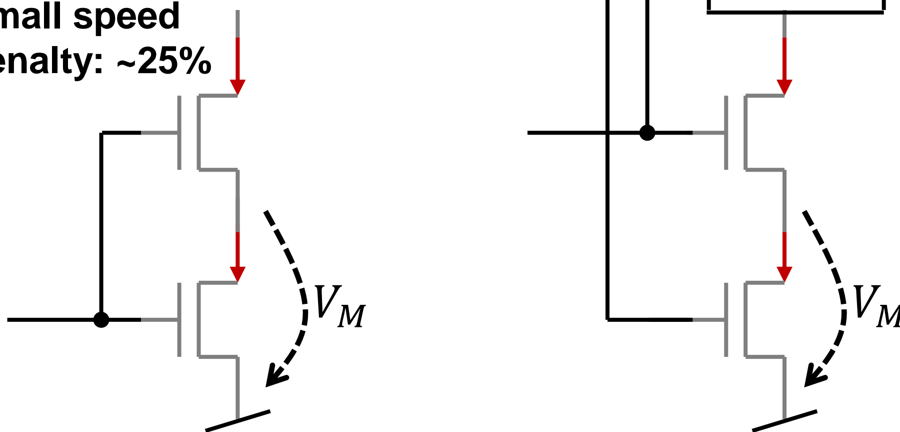
Leakage in Transistor Stacks

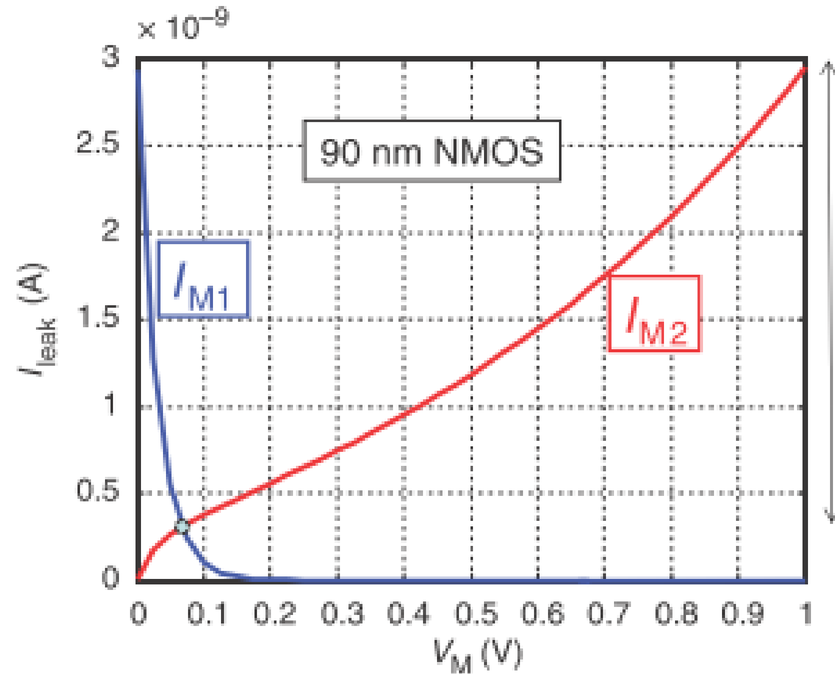
- **Stacking occurs**
  - In many logic gates (> 1 input)
  - When introduced intentionally for leakage reduction

$$I_{leak,M1} = I_0 e^{\frac{-V_M - V_{th} + \lambda_{DS}(V_{dd} - V_M)}{v_t n}}$$

$$I_{leak,M2} = I_0 e^{\frac{-V_{th} + \lambda_{DS} V_M}{v_t n}}$$

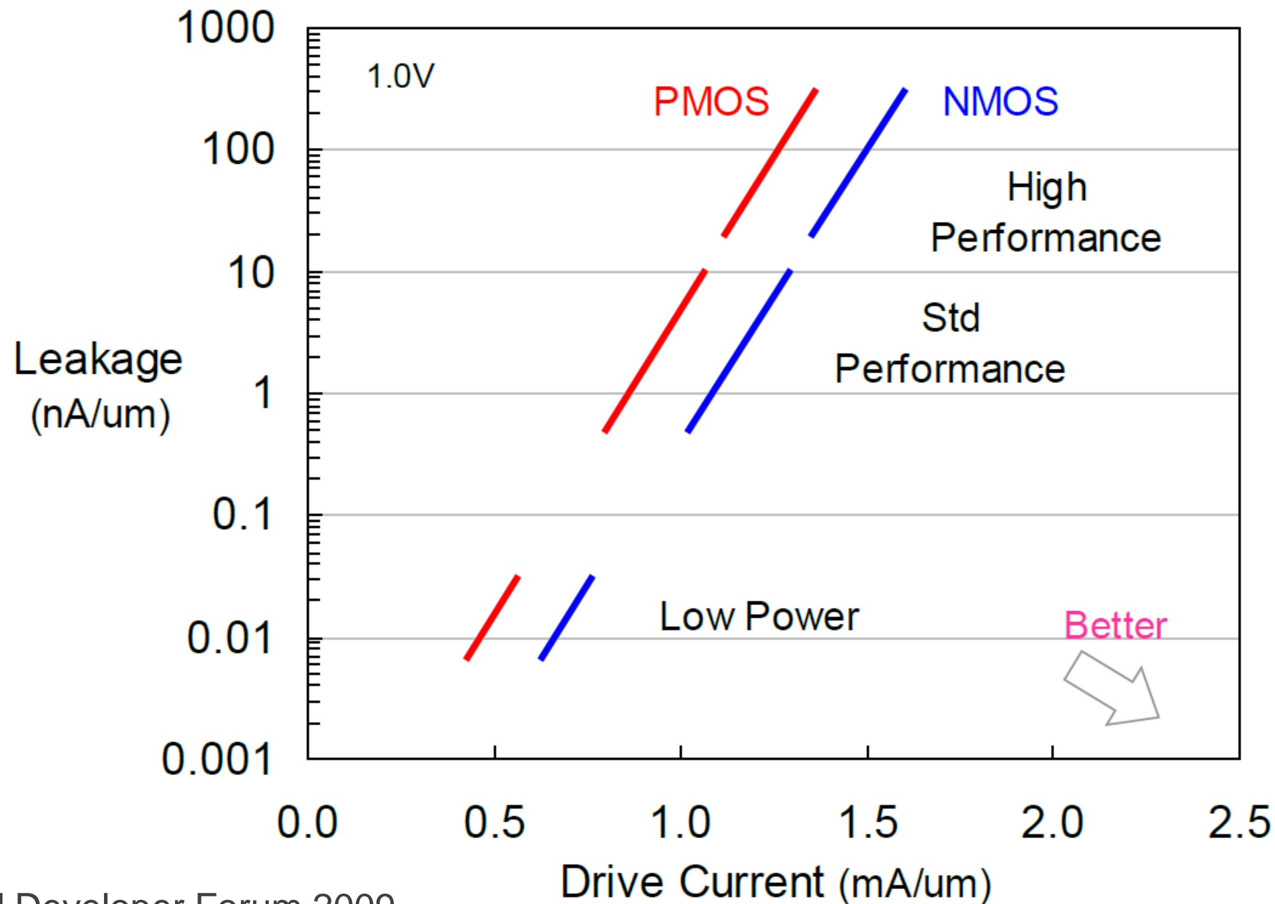
Leakage Reduction	
2 NMOS	9
3 NMOS	17
4 NMOS	24
2 PMOS	8
3 PMOS	12
4 PMOS	16





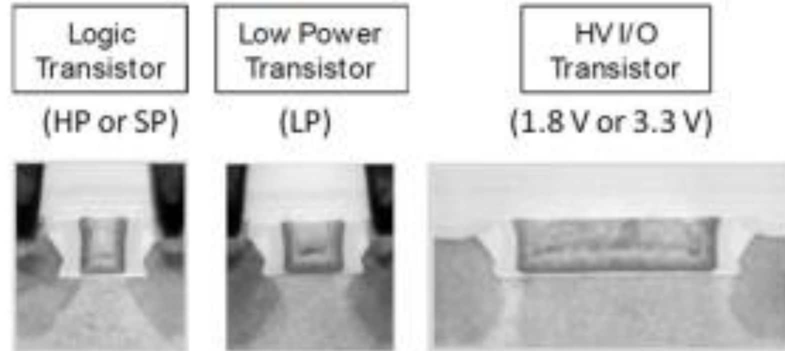
Modern technologies offer different device flavors

- **Devices with different threshold voltages => can often be combined on same die/wafer**
- **Different process flavors (can typically not be mixed on same wafer)**



M. Bohr, Intel Developer Forum 2009

Modern technologies offer different device flavors



Transistor Type	Logic (option for HP or SP)		Low Power	HV I/O (option for 1.8 or 3.3 V)	
	HP	SP	LP	1.8V	3.3V
EOT(nm)	0.95	0.95	0.95	~ 4	~ 7
Vdd (V)	.75/ 1	.75/ 1	0.75/1.2	1.5 /1.8	1.5 /3.3
Pitch(nm)	112.5	112.5	126	min. 338	min. 675
Lgate (nm)	30	34	46	>140	>320
NMOS Idsat (mA/um)	1.53 @ 1 V	1.12 @ 1 V	0.71 @ 1 V	0.68 1.8 V	0.7 3.3 V
PMOS Idsat (mA/um)	1.23 @ 1V	0.87 @ 1 V	0.55 @ 1 V	0.59 1.8 V	.34 @3.3 V
Ioff (nA/um)	100	1	0.03	0.1	<0.01

- **Sometimes IO transistors are an interesting option: low-leakage, high-VT but large distance to core transistors in the layout required**

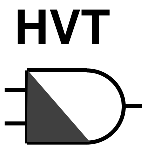
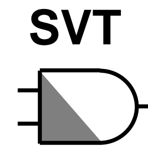
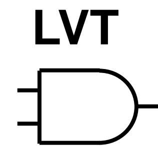
Threshold Voltage Selection

- **Modern process technologies support devices with different threshold voltages**
  - **Typically three flavors: low-VT, standard-VT, high-VT**
  - **Often all three flavors can be mixed in the same design**

- **VT-selection:** tradeoff between speed and leakage

$$t_{pd} = \frac{t_{OX}}{\mu \varepsilon_{OX}} \frac{L}{W} C_L \frac{V_{DD}}{(V_{DD} - V_{th})^{\alpha}} \qquad I_{leak} = I_0 e^{\frac{-V_{th} + \lambda_{DS} V_{DS}}{v_t n}}$$

- **Example**: 55nm process

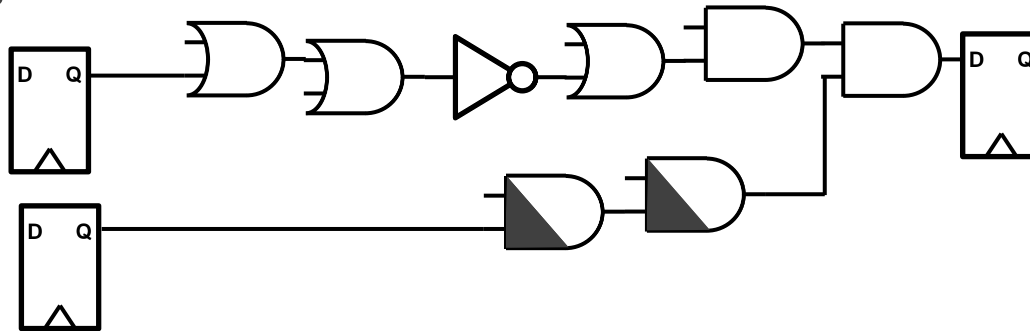
	HVT	SVT	LVT
			
Delay	**20ps**	**16ps**	**14ps**
Leakage	**30nW**	**60nW**	**200nW**

  - **Small increase in speed comes with a significant leakage penalty**

Multi-VT Design

- **Design tradeoff when choosing a VT flavor:**
  - Less leakage (high-VT) increases delay and vice versa
  - **Threshold voltage types can often be mixed**

- **Multi-VT design**



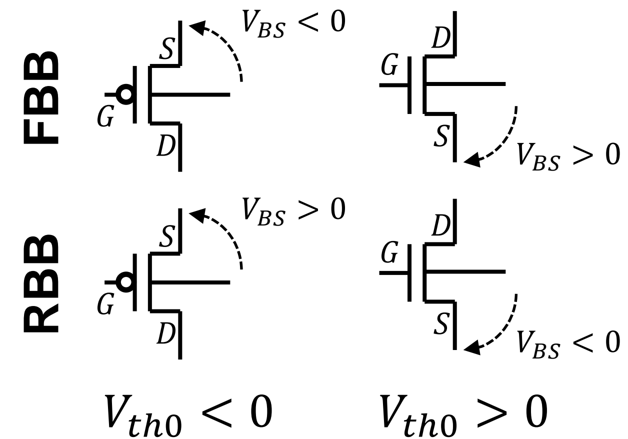
  - **Use low-VT cells only on critical paths**
  - **High-VT cells are used in all other paths**

Caveat: can be very problematic for near-VT or sub-VT design: path delays scale very differently

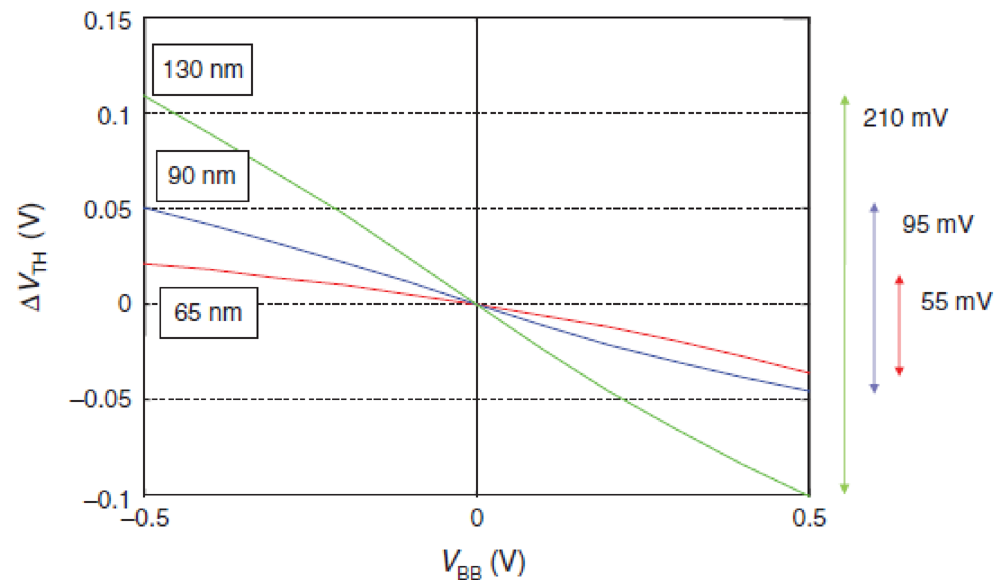
- **Methodology:**
  - **Either done by replacing non-critical cells in the backend OR already during synthesis by providing multiple libraries (HVT/SVT *and* LVT)**

Body Bias Modulates Threshold Voltage

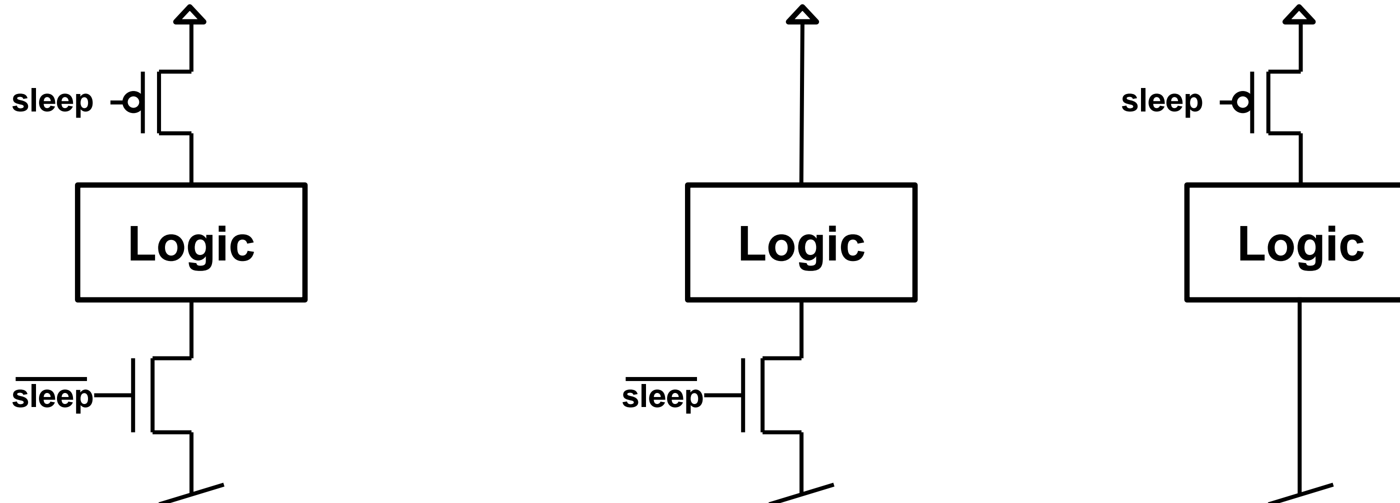
- Body of the transistor is often connected to the source (no body bias)
- Introducing a body bias modulates threshold voltage
  - Forward Body Bias (FBB): increases threshold voltage
  - Reverse Body Bias (RBB): reduces threshold voltage



- $V_{th} = V_{th0} - \lambda_{BS} V_{BS}$
- BULK CMOS:
  - Effect of body bias decreases for technologies below 100nm
  - FBB is limited to ~300mV to avoid operating junction diodes in forward direction



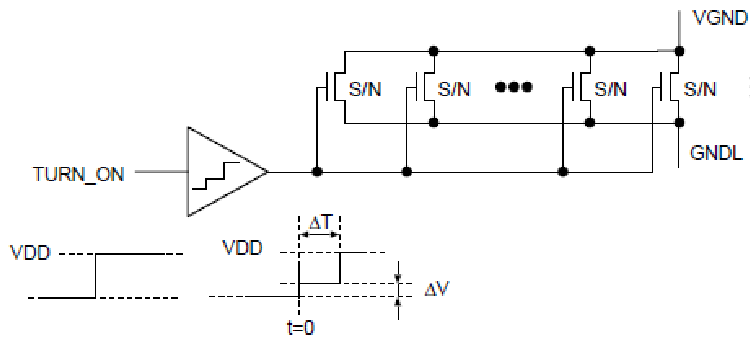
- Avoid leakage almost completely when individual design units are not used:
  - Disconnect entire modules from the supply with headers and/or footers)

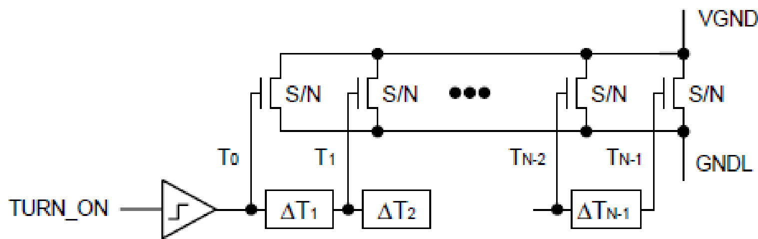


- Objectives with conflicting requirements
  - Sleep mode: large off-resistance to avoid leakage (stacking)
    - PMOS preferred over NMOS and HVT over LVT, header+footer
  - Active mode: minimize on-resistance to reduce negative impact on timing
    - Sleep transistors require large area
    - NMOS preferred over PMOS, LVT over HVT, footer-only

Power Mode Transition

- **Rapid re-activation of a power gated block can cause large spikes on the supply network of the entire circuit**
- **Popular solutions:**

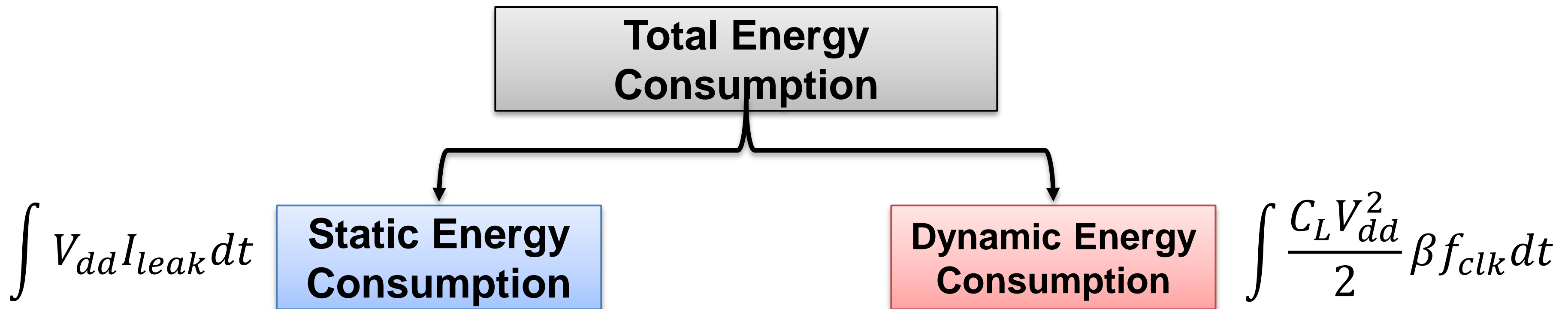




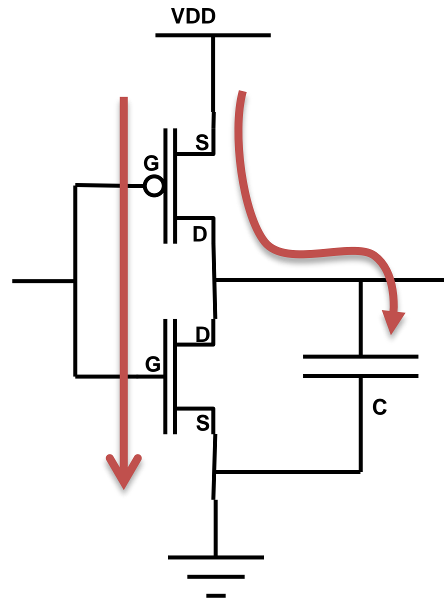
Suhwan Kim, Stephen V. Kosonocky, and Daniel R. Knebel. 2003. Understanding and minimizing ground bounce during mode transition of power gating structures. In *Proceedings of the 2003 international symposium on Low power electronics and design* (ISLPED '03). ACM, New York, NY, USA, 22-25. DOI=10.1145/871506.871515 http://doi.acm.org/10.1145/871506.871515

Voltage Scaling and Sub-VT Design

Voltage Scaling

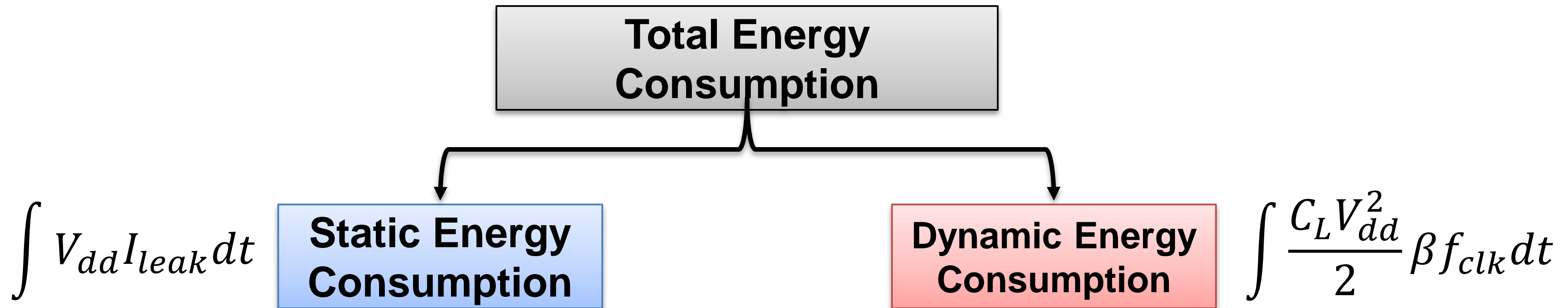


Minimize leakage energy by:
- Reducing voltage
- Reducing Vdd to GND paths



Minimize active energy by:
- Reducing voltage
- Switching activity
- Capacitance

Voltage Scaling



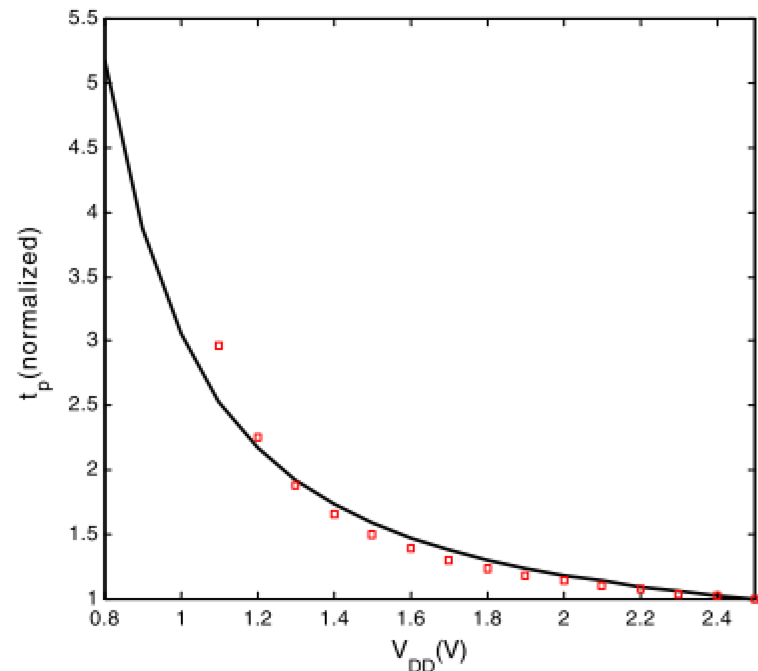
- Reducing supply voltage below nominal
  - Most popular and most effective low-power strategy
  - Voltage-scaling
    - Reduces active power
    - Reduces leakage power (but not necessarily energy/Op)
    - Reduces speed : need to compensate with architectural changes (e.g., parallel processing)

- **Inverter Delay**

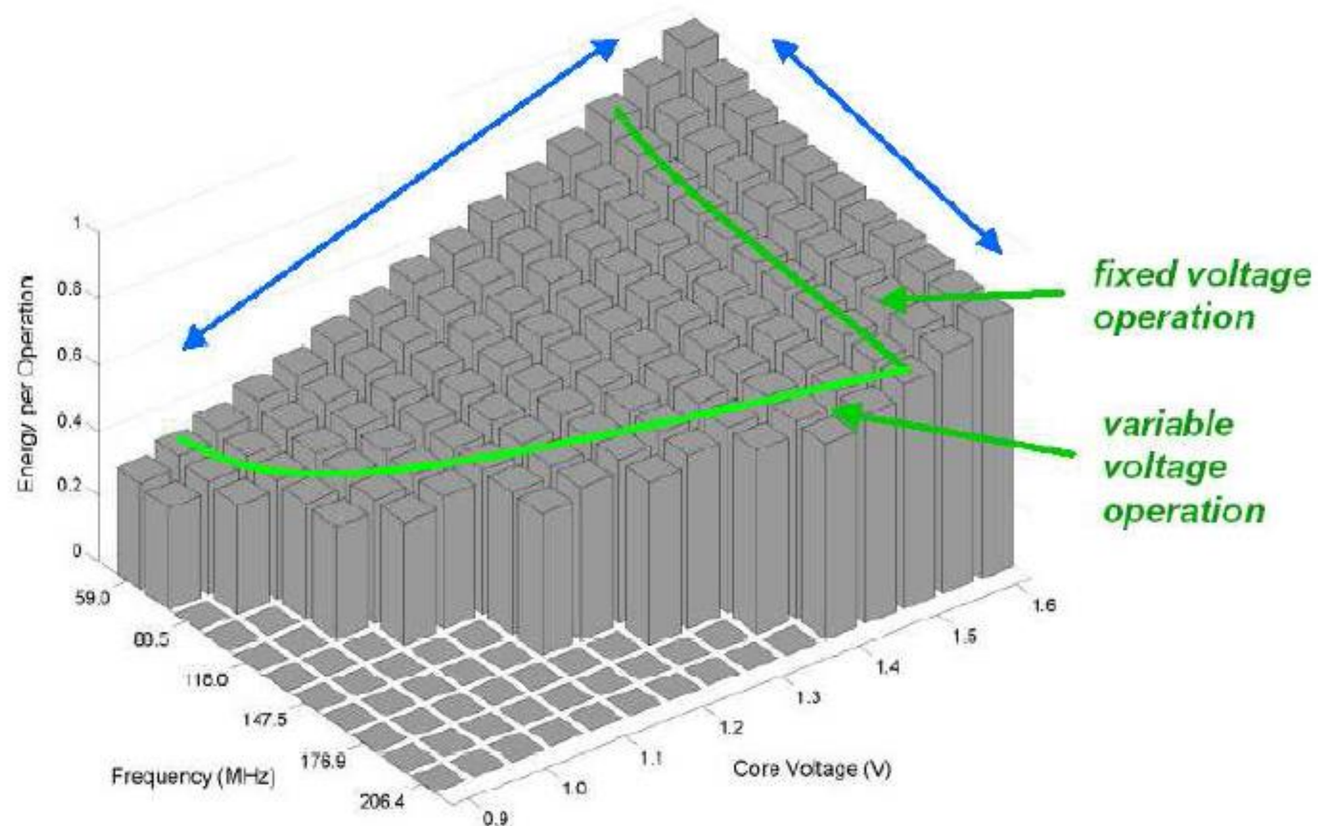
$$t_{pHL} = 0.69\,\frac{3}{4}\frac{V_{DD}C_{load}}{I_{DSATn}} = 0.52\,\frac{V_{DD}C_{load}}{k_n V_{DSATn}\left(V_{DD} - V_{THn} - {V_{DSATn}}/{2}\right)}$$

- **Delay is a function of the supply voltage above VTh**
  - *Depends strongly on the overdrive*
  - *Decreases as overdrive decreases*
- **Delay in the sub-VTh regime**
  - *Exponential dependency on overdrive*

$$t_{pd} \propto \frac{V_{DD}C_{load}}{I_0 e^{\frac{V_{DD}-V_{th}}{v_t n}}}$$



Example: StrongARM SA-1100 processor



Energy per operation as a function of voltage and clock rate

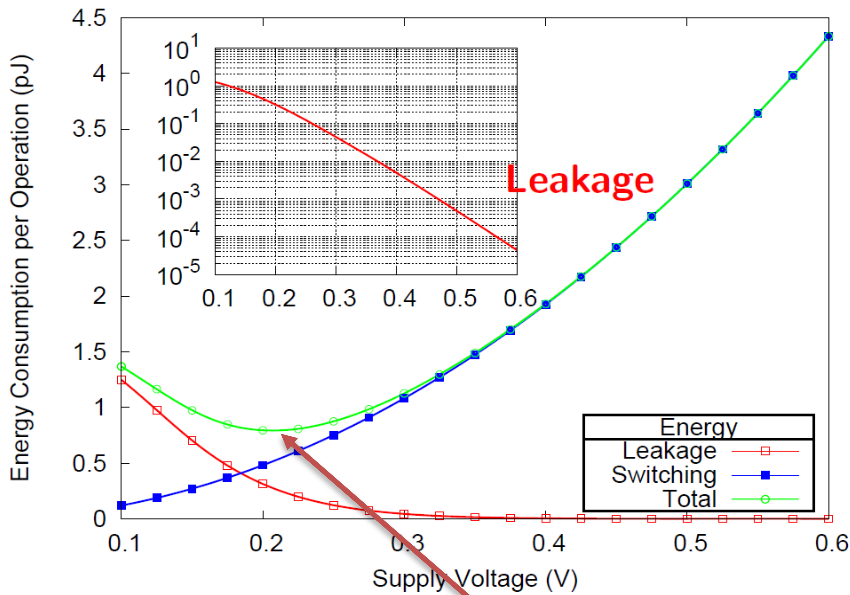
Observation: better energy efficiency for higher frequency at constant voltage.

- **Reduced overhead per operation due to constant (leakage) currents**

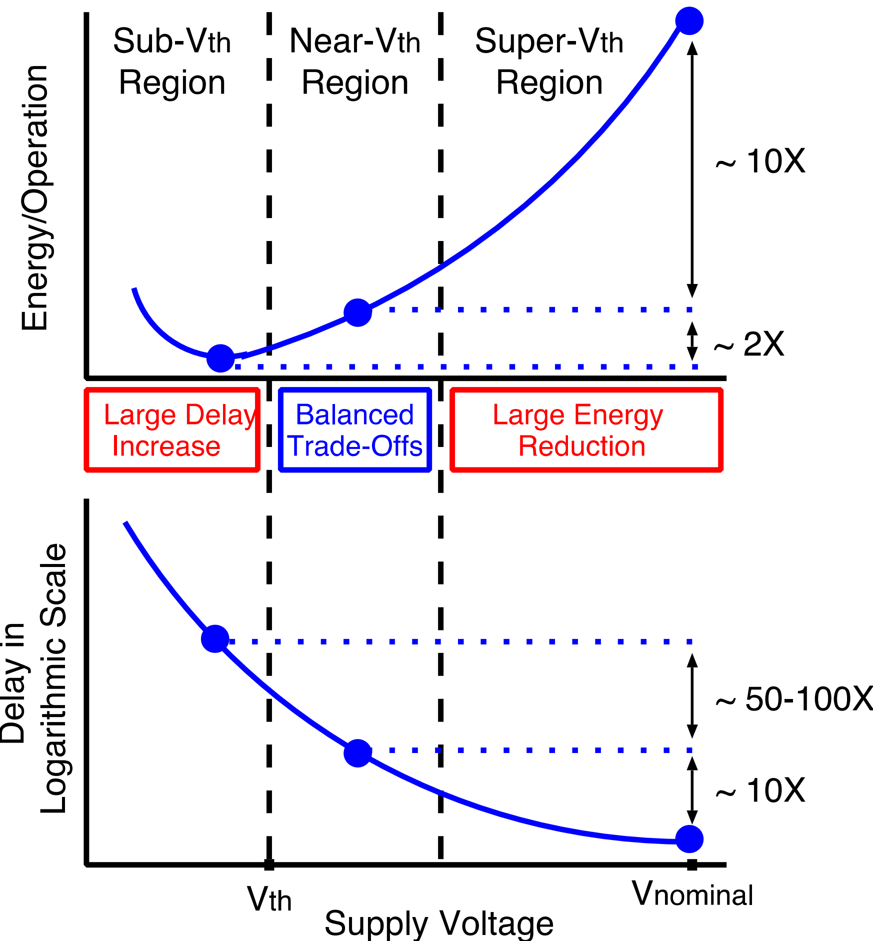
Ultra-Low-Power Design: Sub-Threshold Operation

- **Near/below VT operation:**
  - Exponential delay/leakage increase
  - Minimum energy voltage: balance between leakage and active power consumption

J. Rodrigues, PATMOS 2011, Keynote



Relatively flat around EMV



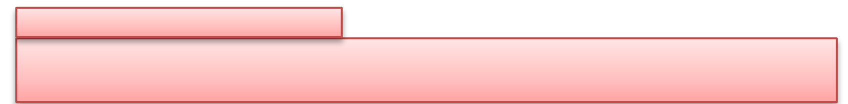
Ultra-Low-Power Design: Sub-Threshold Operation

- Real-time embedded system requirements
  - Handle a given workload with lowest power consumption
- Optimum solution
  - Operation at the energy minimum voltage with power gating during idle periods to avoid leakage
  - But, power gating is only effective when idle periods are long and memories can often not be power gated and are the major source of leakage

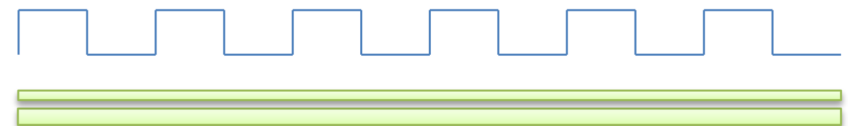
Operation @ EMV

with power gating

without power gating



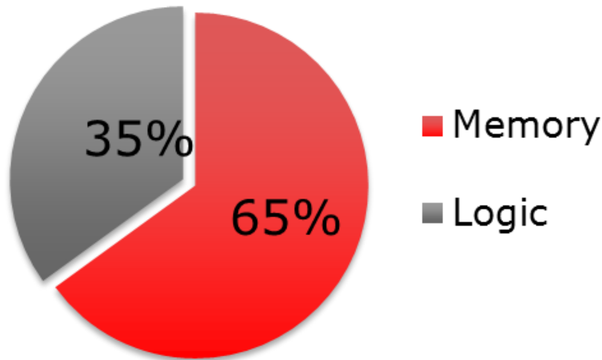
Operation below EMV without power gating

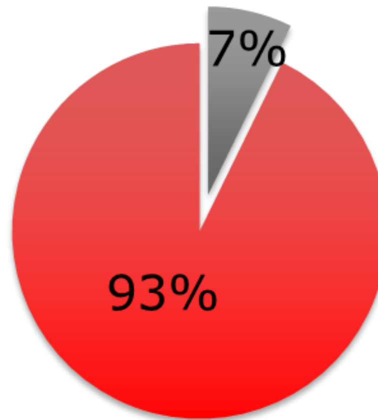


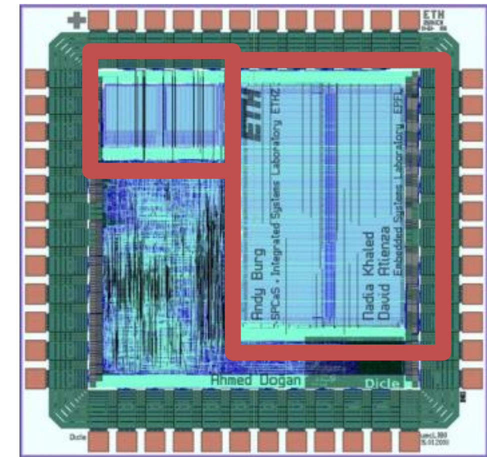
Example:

Power Consumption and Energy Efficiency in Memories

Example: Memories Consume a Large Amount of Power



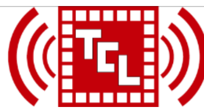




typical embedded processor

- For embedded processors, memories occupy a large percentage of the silicon area
- Active mode:
  - Data and program memory can consume 2/3 of total power
  - Low-frequency: SRAM leakage becomes visible
- Sleep modes:
  - Generally, no power gating to retain SRAM content
  - SRAM leakage becomes dominates system power consumption (3-4 pJ/bit in 180nm): 32kByte -> 400nW @ 1.8V

Example: Standard Cell Based Low Power Memory



Write Logic

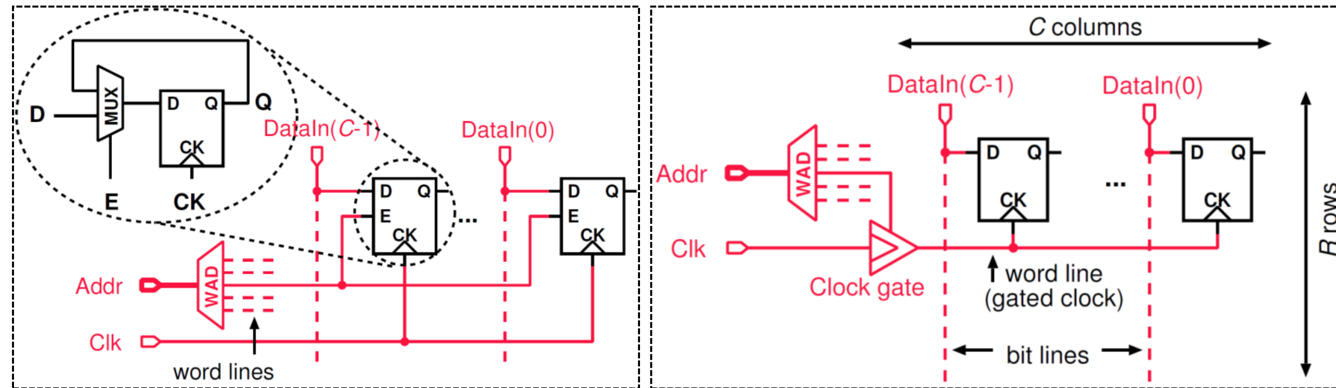
- *Clock-gates (b):* smaller and less power than *enable flip-flops (a)*

Read Logic

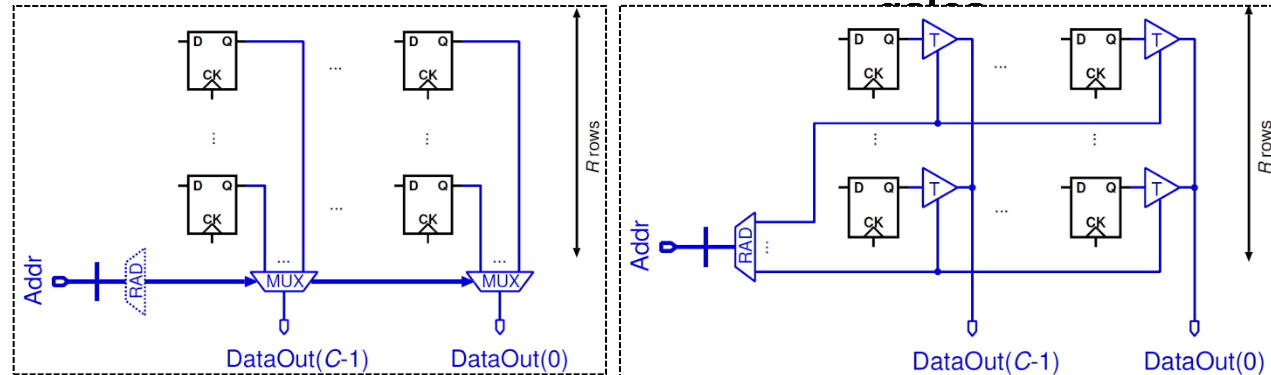
- Above-VT
  - ✓ ***Multiplexers (c)*: smaller, faster, and less power than tri-state buffers**
- Sub-VT
  - ✓ ***Tri-state buffers (d)*: less leakage (energy) than multiplexers**

Array of Storage Cells

- *Latch* arrays smaller than *flip-flop* arrays, but longer write-address setup time



(a) Enable flip-flops **(b) Clock gates**



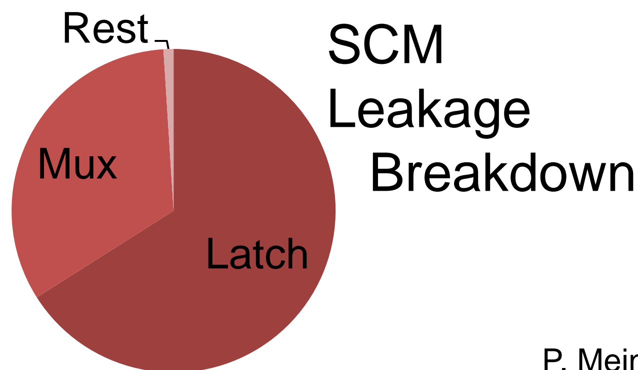
(c) Multiplexers **(d) Tri-state buffers**

Sub-V_T SCM: Insights and Leakage Breakdown

Large memory arrays: little switching activity

- **Total energy is dominated by leakage**
- **Active energy negligible, except for smallest SCMs**
- **Only smallest SCMs reach EMV in sub-V_T domain**

→ Minimize leakage!



Energy [fJ]
10^3
10^2
10^1
10^0
R=128
R=32
R=8
EMV
EMV
C=128
C=32
C=8
0.1 0.15 0.2 0.25 0.3 0.35 0.4
V_{DD} [V]

P. Meinerzhagen *et al.*, JETCAS'11

Custom Cell: Low-Leakage Latch with Tri-State Output

Best practice for low leakage

1. **Lowest number of V_{DD}-ground paths**
2. **Highest resistance on each such path**

- **Tri-state buffers**
- **Stacking & stretching for inverters**

Stacking factor: max 2

Channel length stretching: $2L_{min}$

Tri-state-enabled output

32 columns

Custom low-leakage standard-cell

128 rows

W=Wmin
L=1.5Lmin

3-state read logic (1 column)

Stacking & stretching

Convert output buffer to tri-state buffer to avoid static CMOS muxes

4kb SCM Test Chip in LP-HVT 65nm CMOS

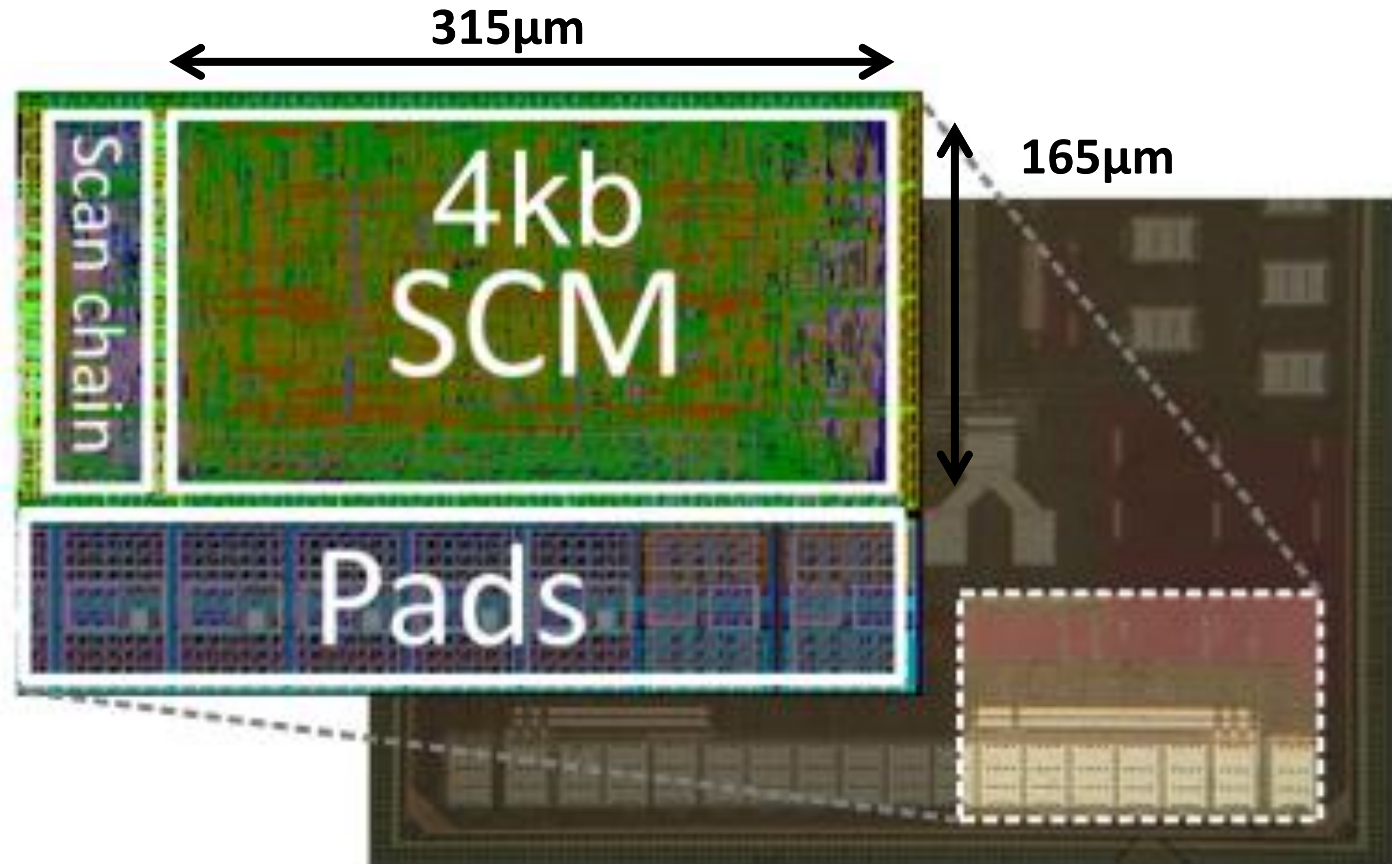
Chip microphotograph and zoomed-in layout picture

Area cost of 12.7 μm^2 per bit (including peripherals)

Scan-chain test interface

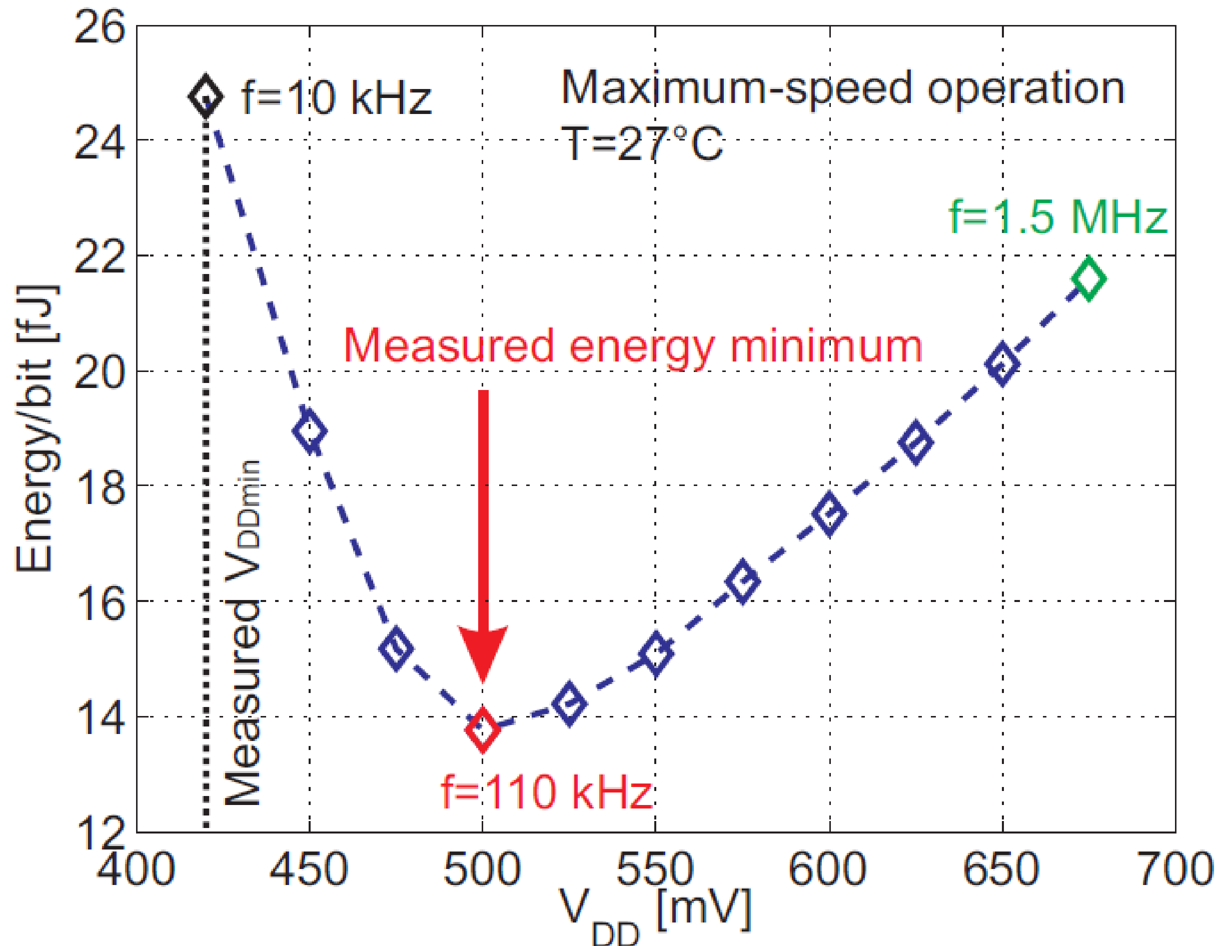
Functionality verification: W/R random and checker-board patterns

Oven to control temperature: 27 or 37°C



Measured energy per bit-access performed at maximum speed

Measured energy minimum is 14fJ/bit at 500mV, 110kHz



Silicon Measurements: Leakage Power is 500fW/bit

At VDDhold=220mV, data is correctly held with a leakage power of 425-500fW per bit (best and worst out of 4 measured dies)

At 37°C (typical for biomedical implants)

- VDDmin=400mV (instead of 420mV at 27°C)
- Maximum operating frequency doubles
- But: higher leakage power
- Low retention voltage is key for low power