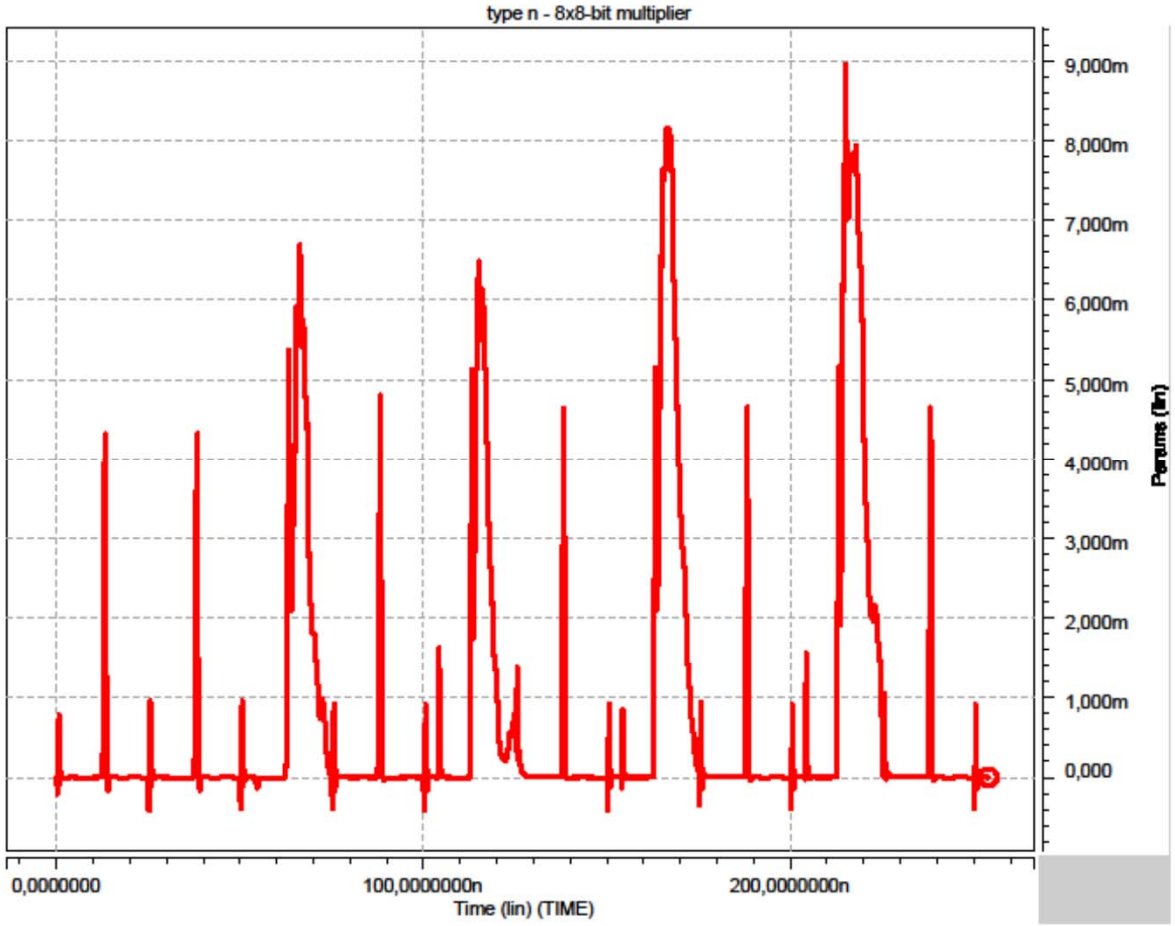
Energy-Aware Computing: Technology and Circuits

Per Larsson-Edefors

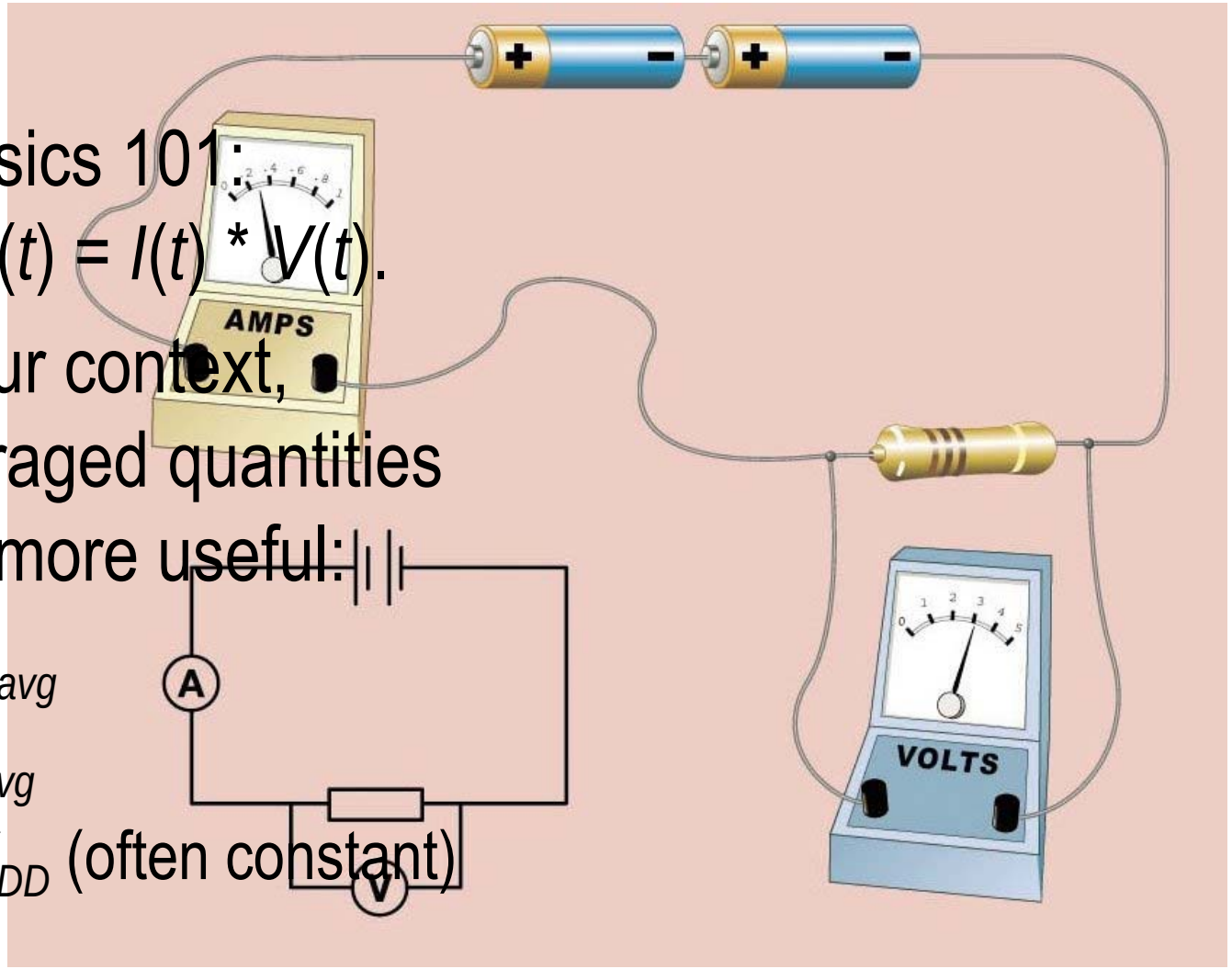
Computer Science and Engineering
Chalmers University of Technology

Computing Circuits Draw Current

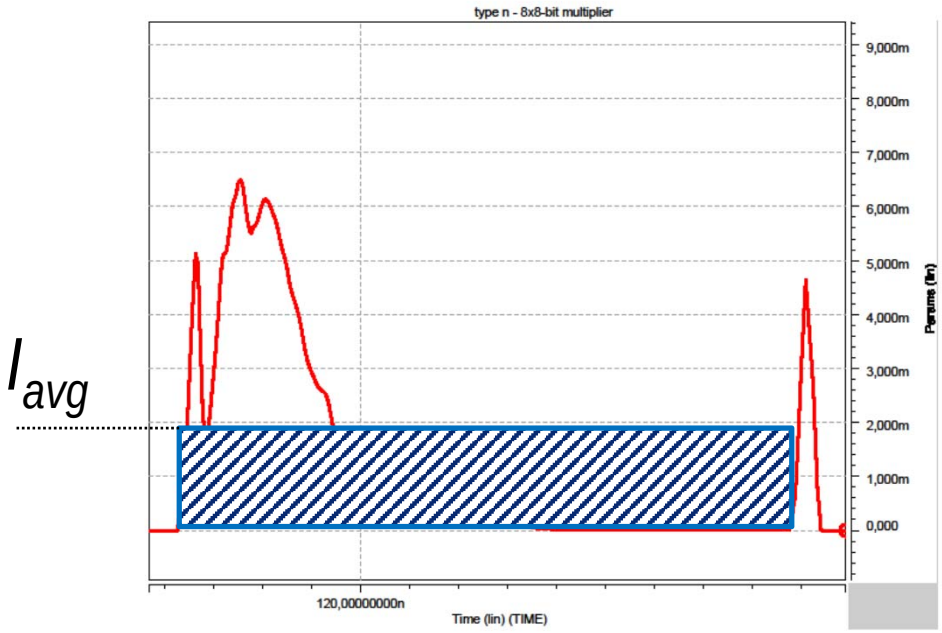


Power = Current * Voltage

- Physics 101:
  $P(t) = I(t) * V(t)$.
- In our context, averaged quantities are more useful:
  - P_{avg}
  - I_{avg}
  - V_{DD} (often constant)

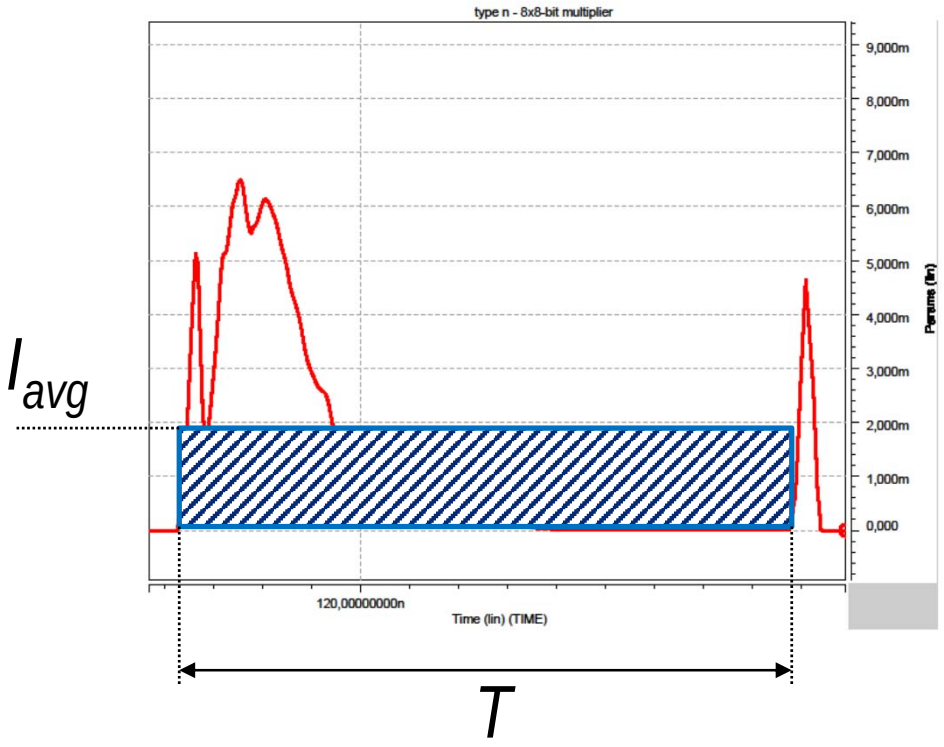


(Average) Current Drawn per Cycle



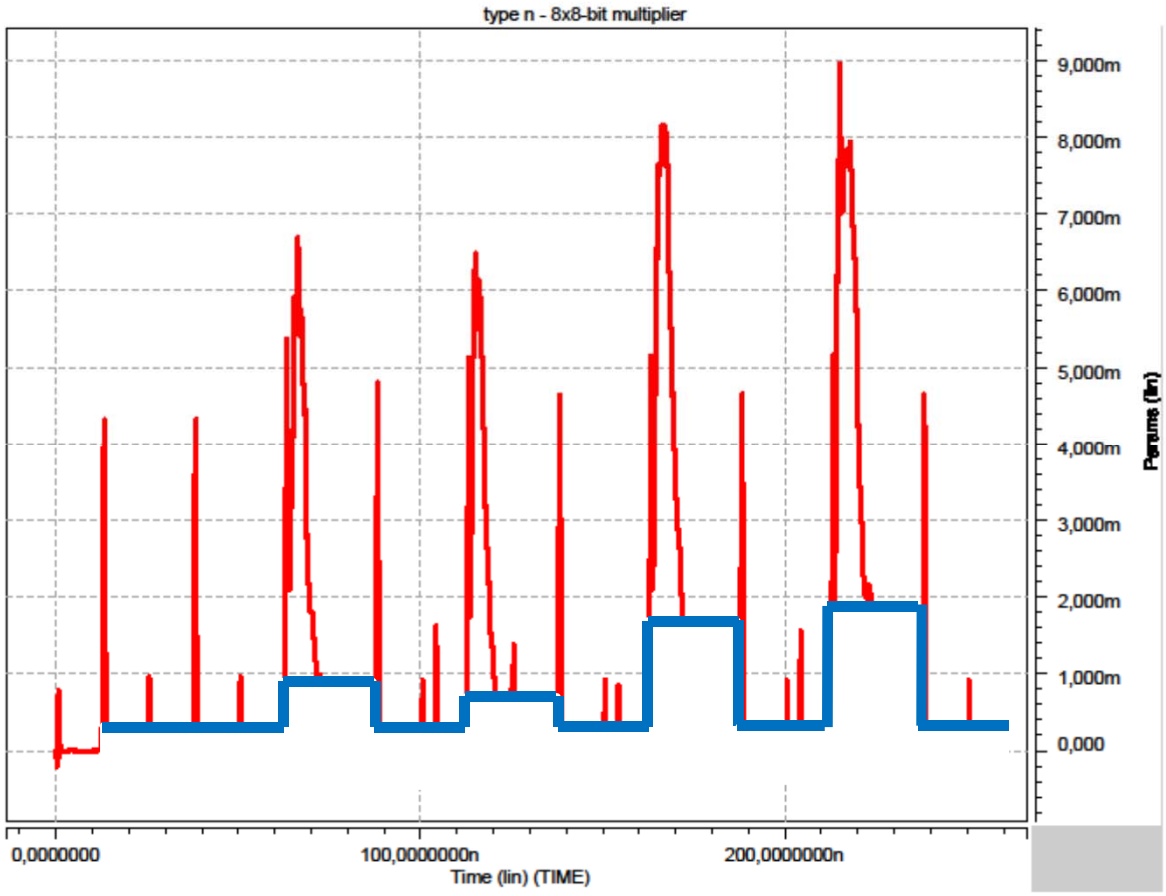
- Average current?
- Consider the instantaneous current throughout one cycle…
- …then take the average.

Power and Energy Dissipated per Cycle



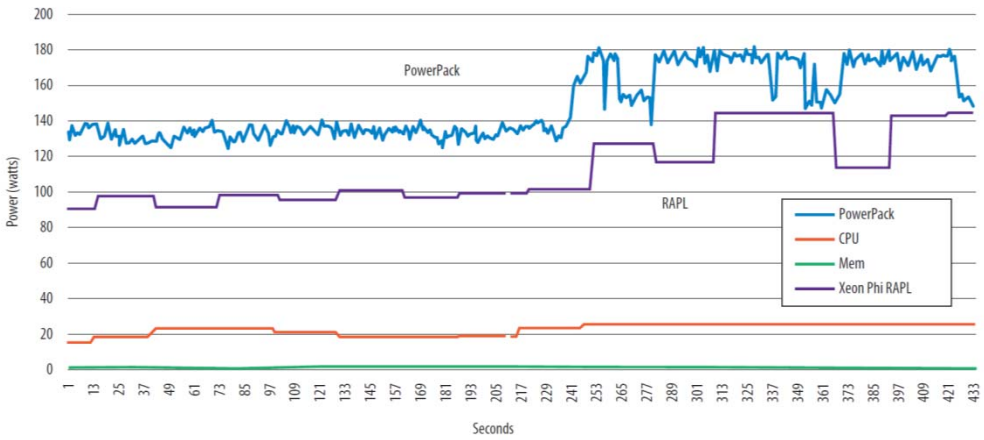
- $P_{avg} = I_{avg} * V_{DD}$, where V_{DD} is the supply voltage of the system.
- Energy/cycle: $E_c = P_{avg} * T$, where T is the clock period.
- P_{avg} and E_c will both be used in this lecture.

Varying Power and Energy Dissipation



- P_{avg} and E_c...
  - depend on what is being computed.
  - vary from cycle to cycle.

Power Measurements vs the TDP Metric



Source: EPM'14

Source: Scythe (Kozuti Cooler)

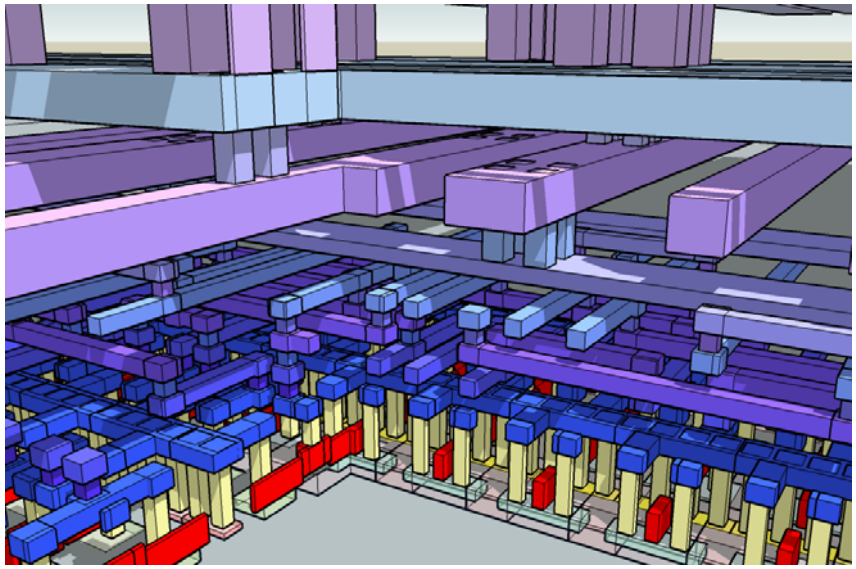
- Power dissipation varies over time. Yet, the TDP (Thermal Design Power) defines certain Watt limit for CPU cooling.
  - Peak power is allowed to exceed TDP, but to what extent?
- Which benchmarks to establish the TDP value?

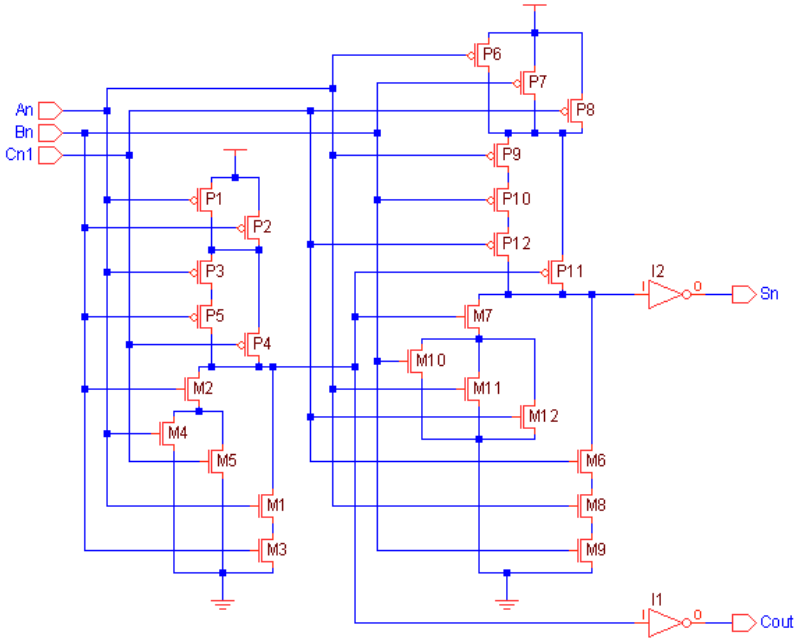
Typical Energy/Cycle

- Find typical energy/cycle:
  - Run system for many cycles.
  - Use statistically relevant benchmarks.
- By associating the typical energy/cycle with different hardware units, energy metrics can be used at software level.

Parameter	Energy (pJ)
Integer Operations	17.93
Floating Point Operations	29.39
Branch	154.22
Local store	47.99
Local load	39.82
Pipeline Stalls	53.65
Shared memory stores	581.72
Shared memory loads	2054.67
NOP	17.07
Idle Cycle	23.59

Technology and Circuits

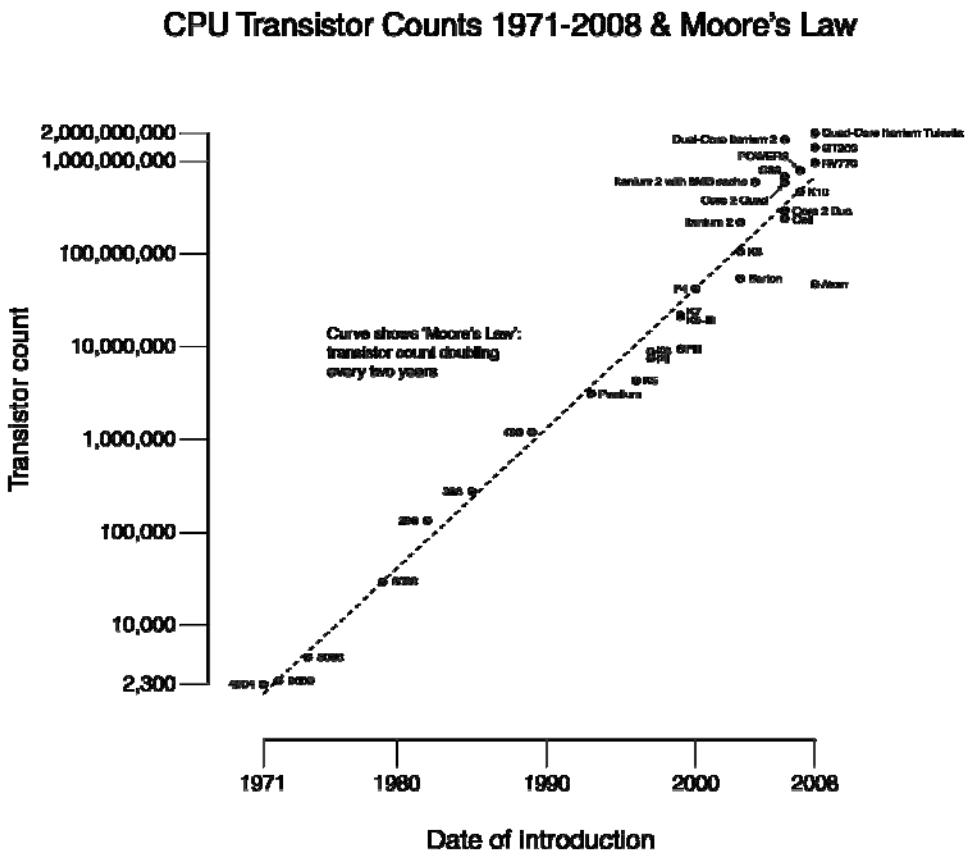




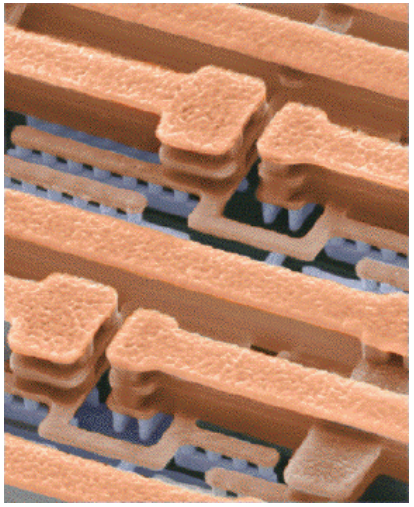
Source: Concept Engineering

- Physical implementation impacts power dissipation.
  - Fabrication process technology (left).
  - Circuit implementation (right).

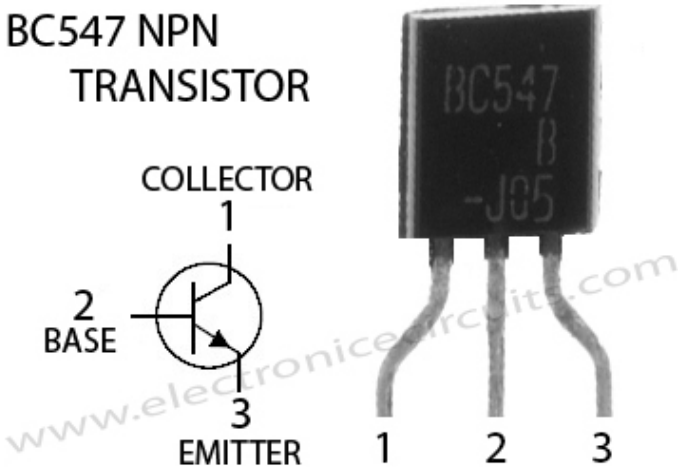
Technology: Wires and Transistors



Source: wikipedia



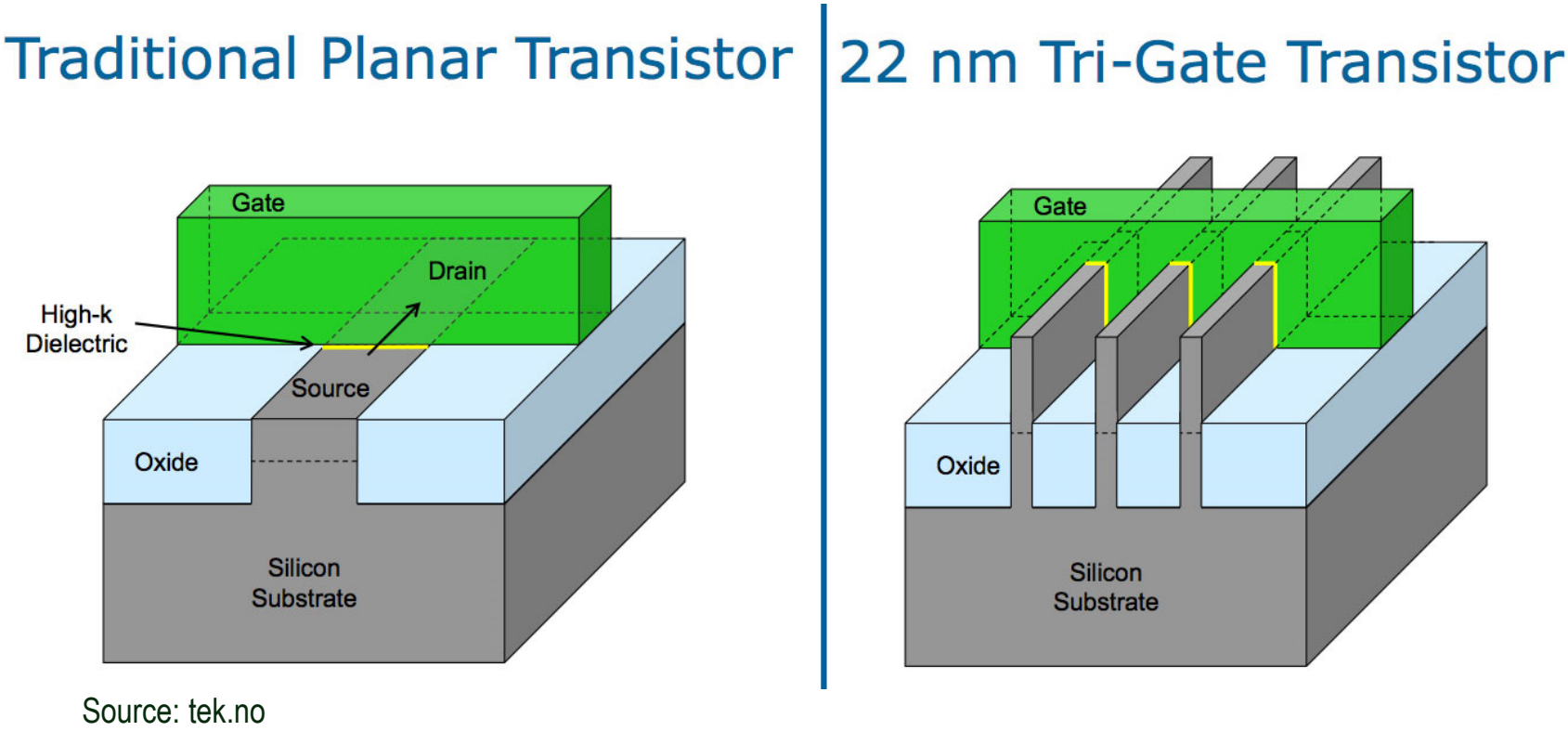
Source:
Univ. of Florida



Source: electronicecircuits.com

Field-Effect Transistors (FETs)

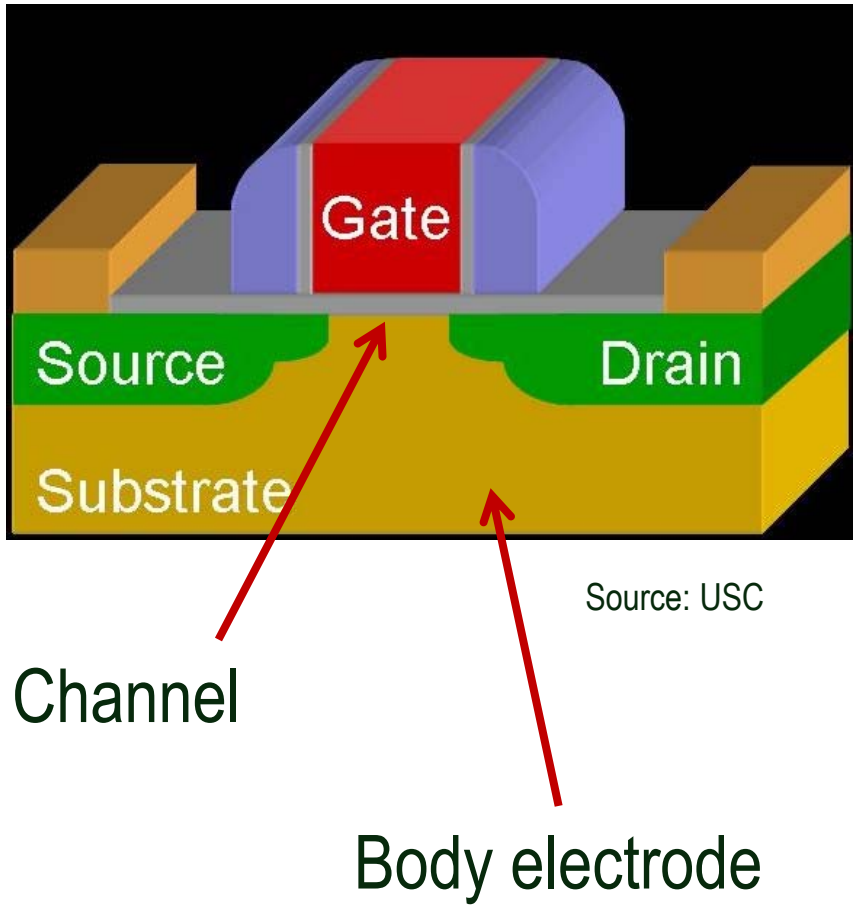
The FET; the work horse of all digital systems



Source: tek.no

Field-Effect Transistor Basics

- Voltage is applied on gate.
- Electric field regulates channel properties.
- Threshold voltage, V_T (or V_{TH}) is the gate voltage required to create a conducting channel.

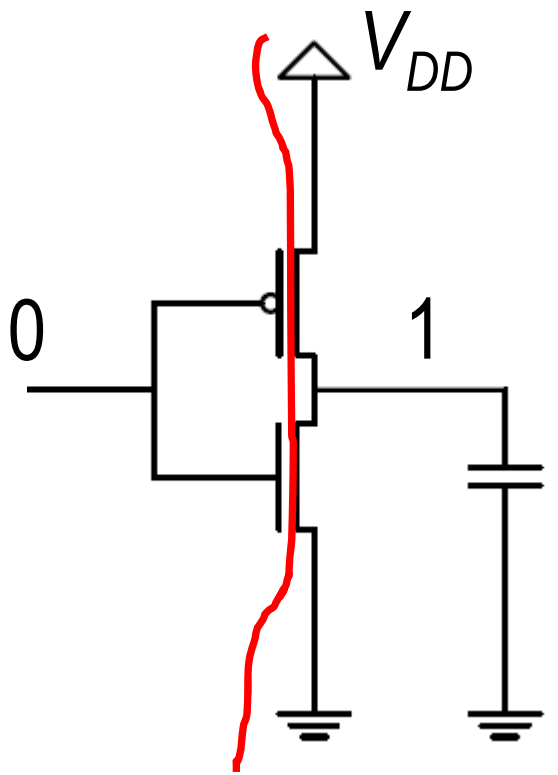


Source: USC

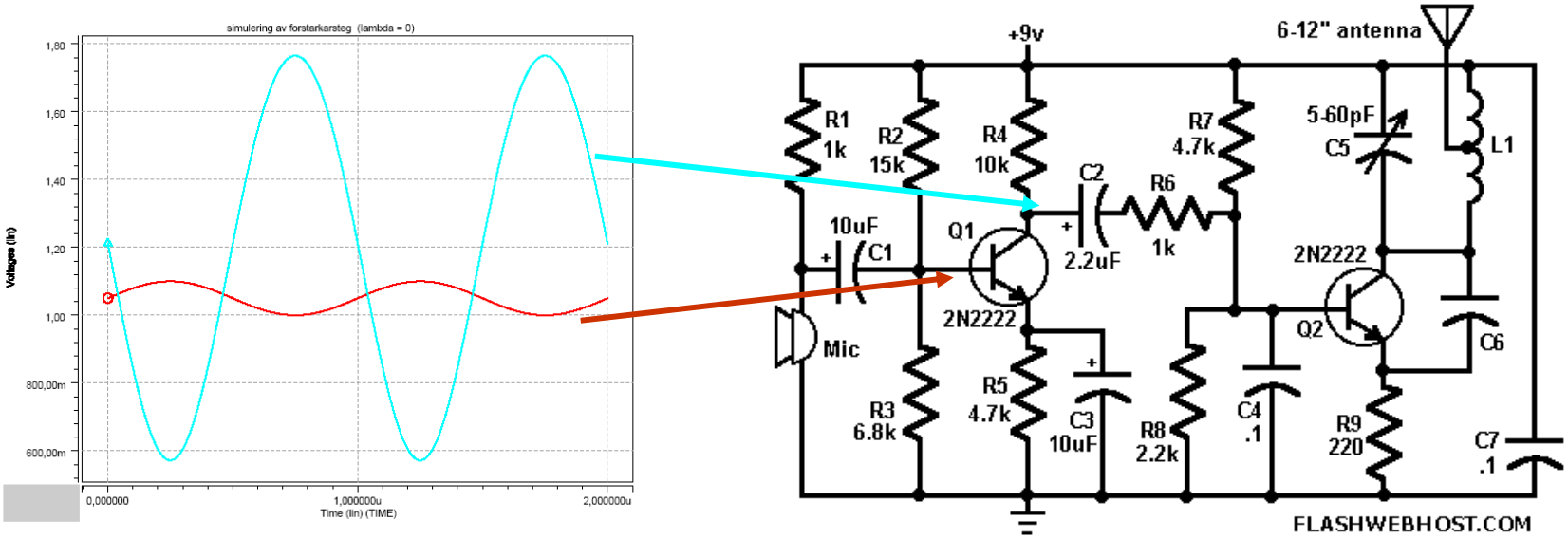
The CMOS Technique

- Terminology:
  - MOSFET = Metal Oxide Semiconductor FET.
  - CMOS = Complementary MOS.
- CMOS is the foundation for all digital circuits.
  - Key property: Gate isolated from channel $\Rightarrow$ “no” current flows through gate insulator.

The CMOS inverter



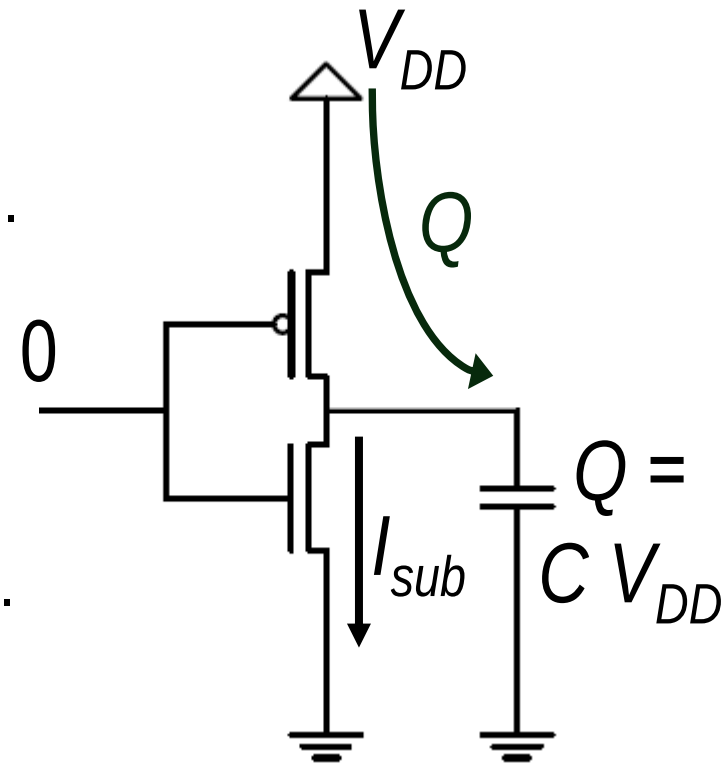
Contrast to BJTs and Analog Circuits



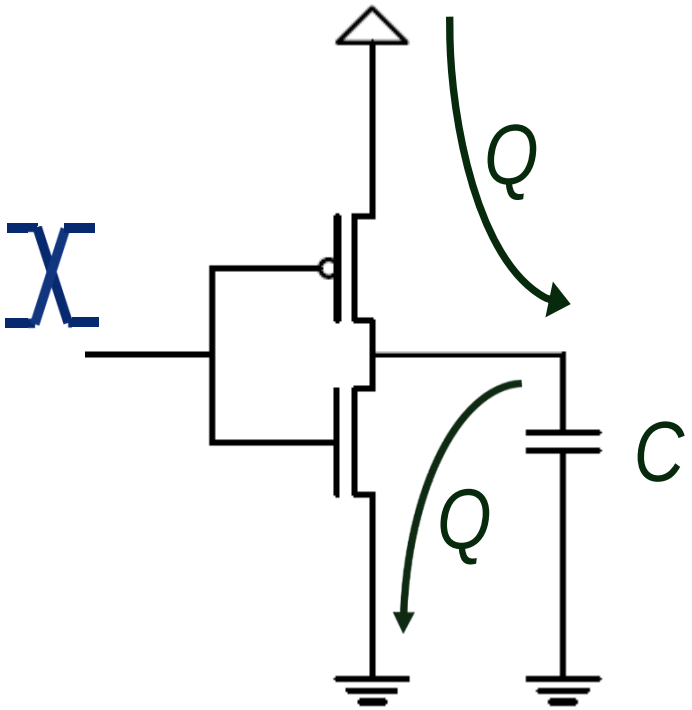
- Analog circuits: biasing required ⇒ transistors are always on.
- Bipolar transistors: current flows into base; no isolation.

Dynamic and Static Power Dissipation

- Dynamic power:
  - Switching power, P_{sw}.
  - Switching logic levels, i.e., computation.
  - Charge and discharge, Q.
- Static power:
  - Leakage power, P_{leak}.
  - Mainly due to subthreshold current, I_{sub}.
  - Caused by small-size effects, i.e., advanced FETs are never fully off.

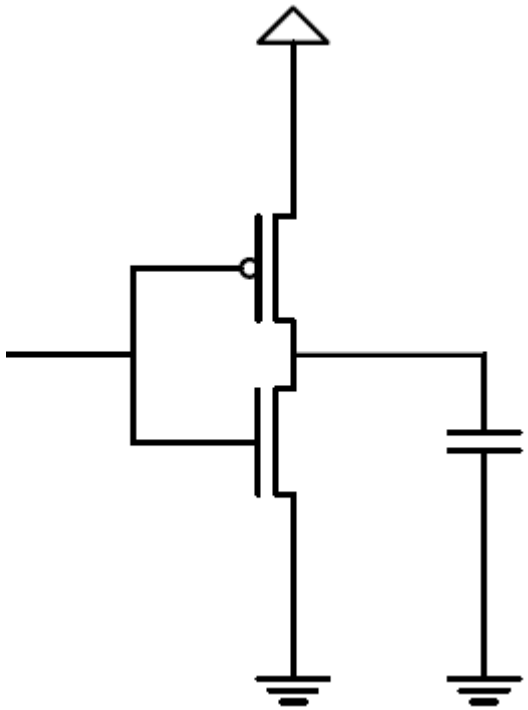


Switching Power Dissipation, P_{sw}



- Input transition 1→0 ⇒ output node 0→1, requiring charge $Q = C\,V_{DD}$ from the power supply.
- Later, input 0→1 ⇒ output node falls, draining the charge to the ground.

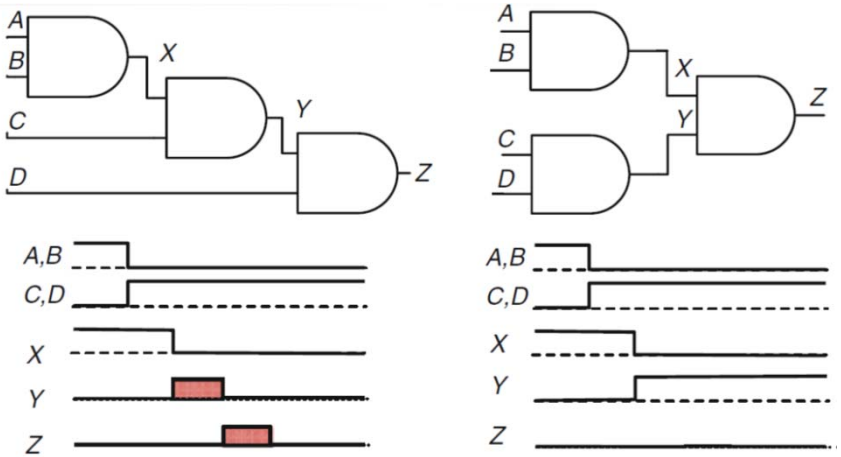
While P_{sw} Mostly Depends on V_{DD} …



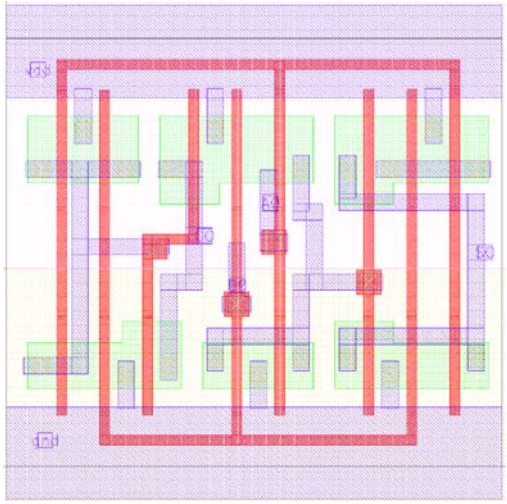
- After one full transition, the energy of the charge has been converted into heat:
  $P_{sw} = E/T = (Q\ V_{DD})/T =$
  $= (CV_{DD}\ V_{DD})\ f.$
- This gives us this famous expression
  $P_{sw} = f\ \alpha C\ V_{DD}^2$
  where α represents switching/cycle.
- <u>To reduce switching power, focus on the supply voltage.</u>

… All Tricks Are Necessary

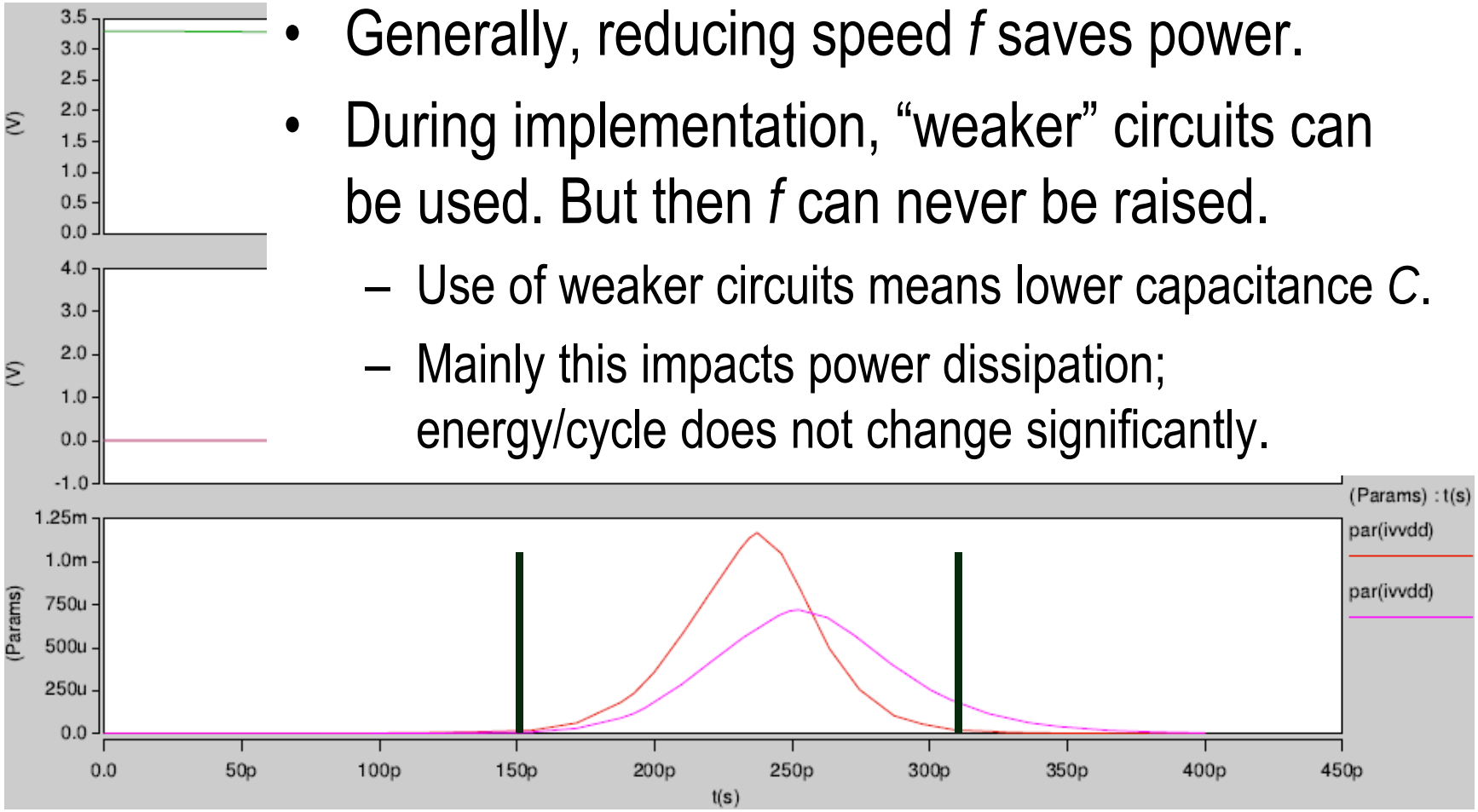
- We want to reduce $P_{SW} = f\,\alpha C\,V_{DD}^2$
- Aside from V_{DD}, there are several implementation best design practices:
  - Reduce signal activity (α), e.g., by eliminating glitches.
  - Reduce nodal capacitance (C), by optimizing layout of transistors and wires.



Source: LPDE'09/Ch 3



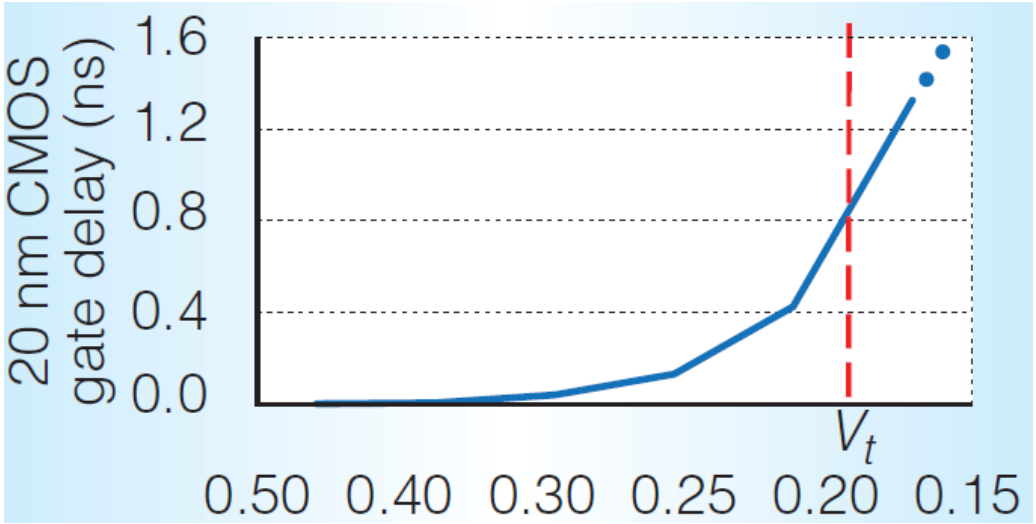
Impact of Speed (*f*) on Power and Energy



- Generally, reducing speed *f* saves power.
- During implementation, “weaker” circuits can be used. But then *f* can never be raised.
  - Use of weaker circuits means lower capacitance *C*.
  - Mainly this impacts power dissipation; energy/cycle does not change significantly.

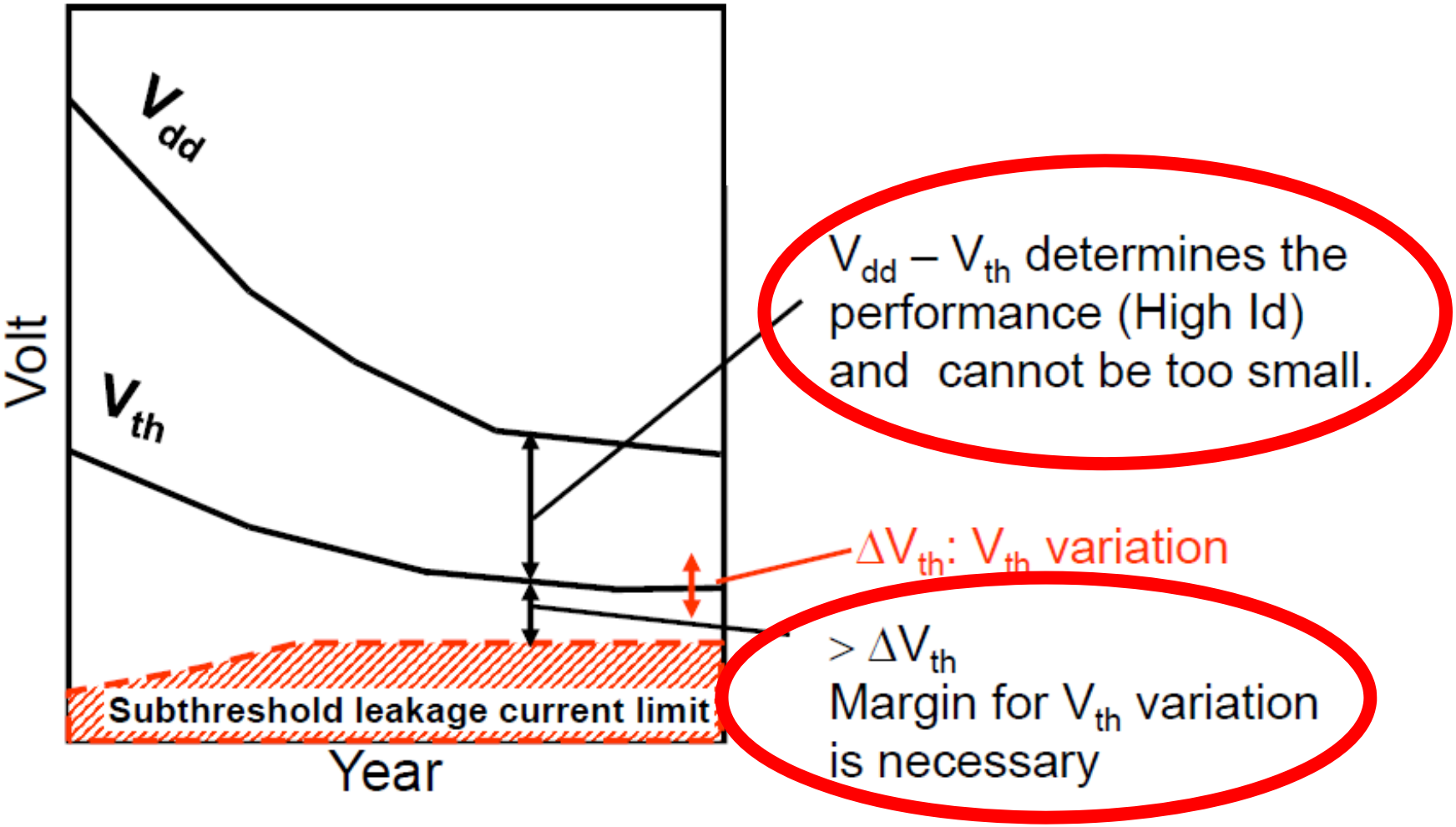
Scaling Supply Voltage for Reduced P_{sw}

- So V_{DD} is decreased to save switching power.
- Since performance deteriorates rapidly as V_{DD} approaches V_T, V_T has to be decreased as well.



Source: SSD'13

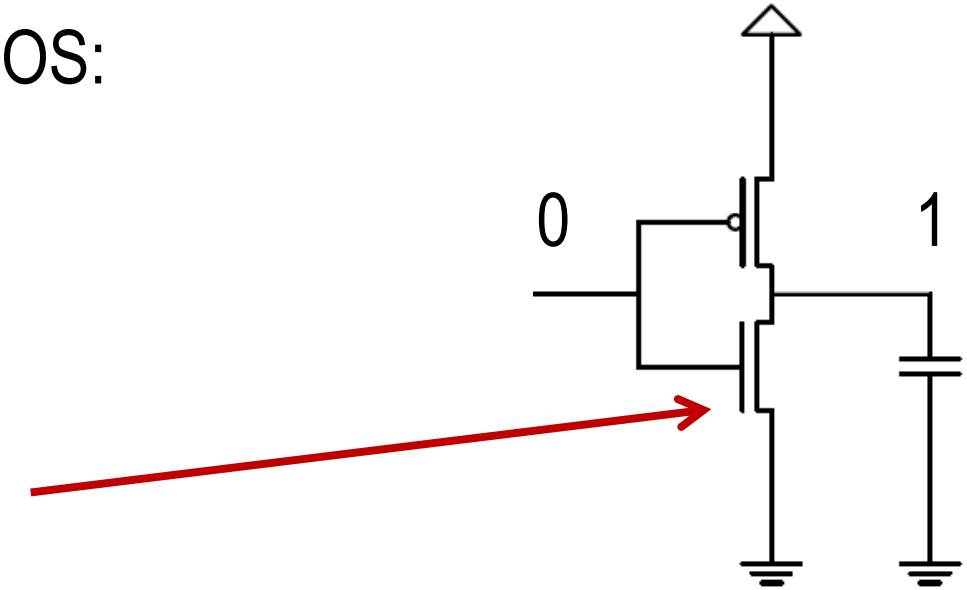
Voltage Scaling Issues



Source: H. Iwai, Technology Scaling and Roadmap, IEDM'08/Short Course

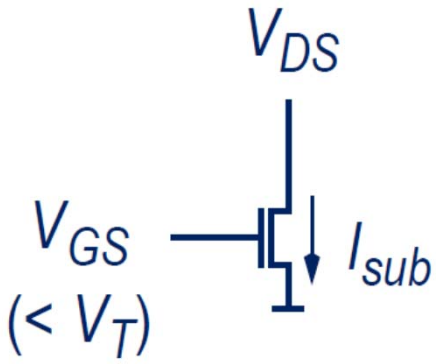
CMOS Stability Reduces with V_{DD}

- Classic advantage of CMOS: High I_{ON}/I_{OFF} ensures stable digital operation.
- Because of scaling, leakage increases $\Rightarrow$ I_{OFF} increases.
- Degrading I_{ON}/I_{OFF} ratio limits how far V_{DD} can be scaled; especially serious for SRAMs.



Subthreshold Current

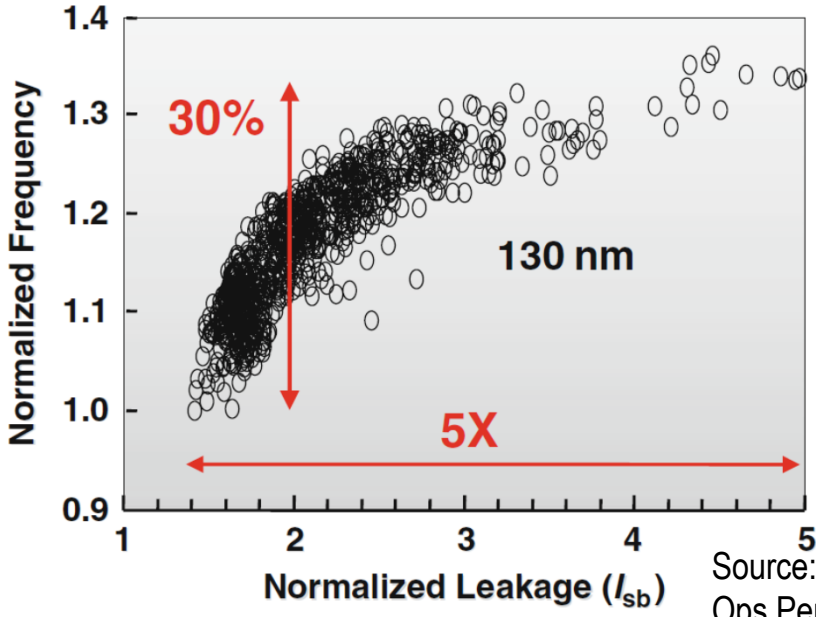
- Smaller FETs and decreasing V_{DD} (and V_T) $\Rightarrow$ increasing I_{sub}.
- I_{sub} function of semiconductor energy states (quantum mech.).
- Three important features:
  - Exponential dependence on V_{GS}.
  - Exponential dependence on V_T.
    - Since V_T depends on V_{DS}, static power strongly depends on supply voltage!
  - Temperature matters.



$$I_{sub} \propto e^{\frac{q(V_{GS}-V_T)}{kT}}\left(1-e^{-\frac{qV_{DS}}{kT}}\right)$$

$$V_{thermal} = \frac{kT}{q} = 26\text{ mV (room temp)}$$

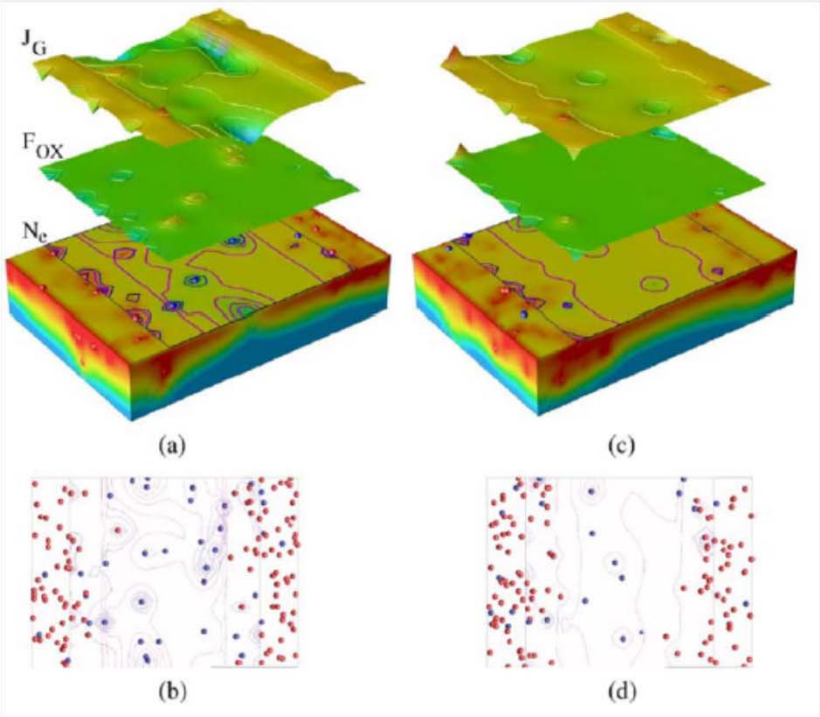
CMOS Stability - Variability



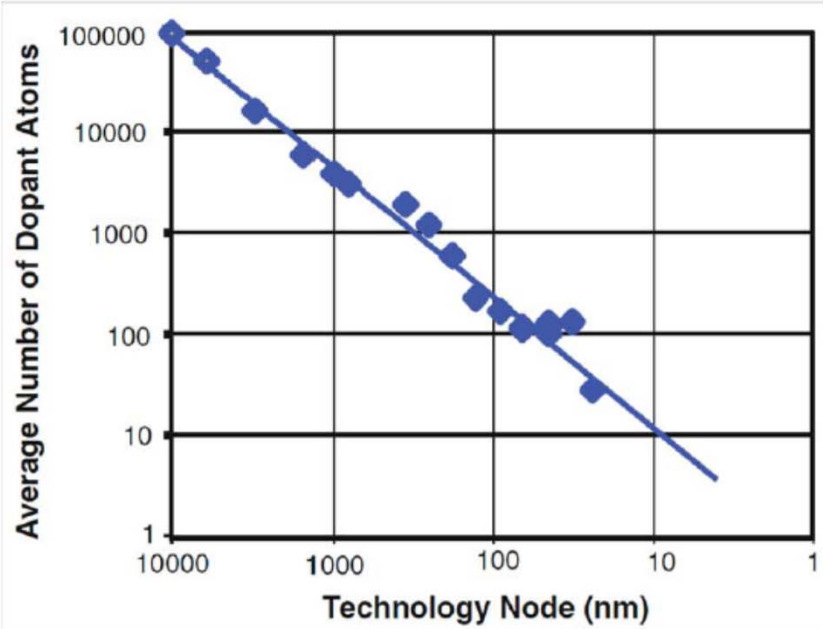
Source: "Giga-scale Integration for Tera Ops Performance", P. Gelsinger, DAC'04

- V_T variations impact leakage exponentially.
  - Variability increases with scaling.
- Generally, technology variations have a stronger impact on designs where V_{DD} and V_T are reduced.

Some Variations Are Random in Nature

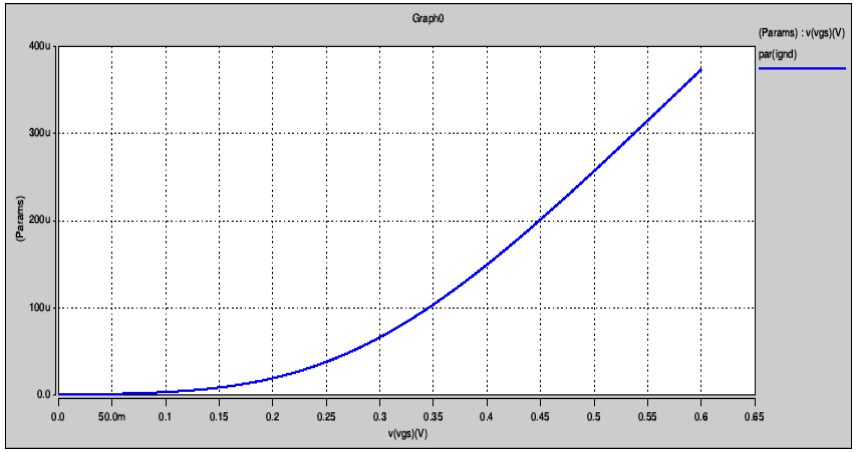


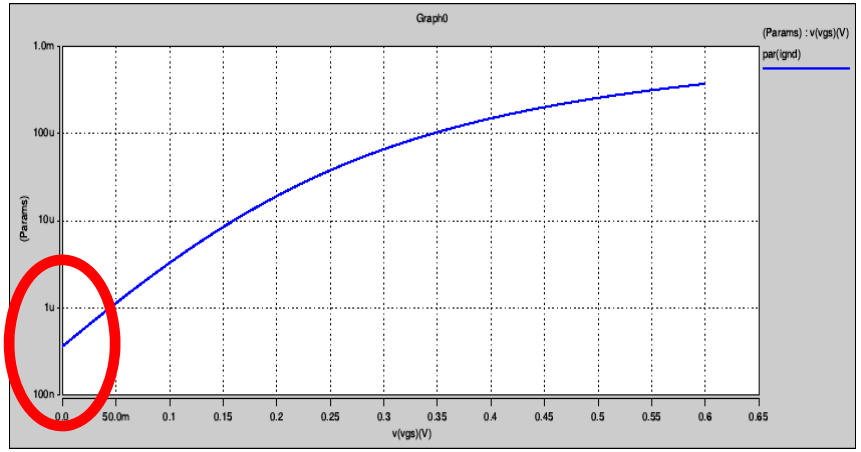
Source: Direct Tunnelling Gate Leakage Variability in Nano-CMOS Transistors, IEEE TED, 2010.

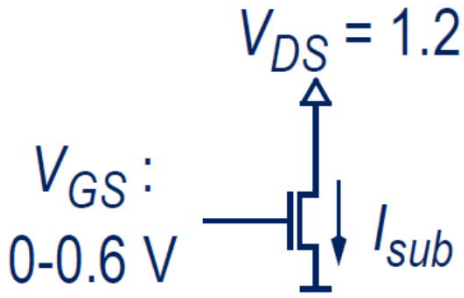


Source: Analog IC Reliability in Nanometer CMOS, Springer, 2013.

I_{OFF} Isn't 0 (Due to Leakage)



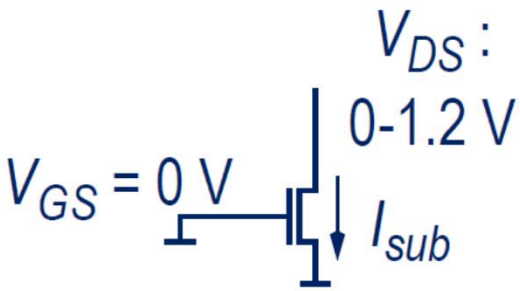


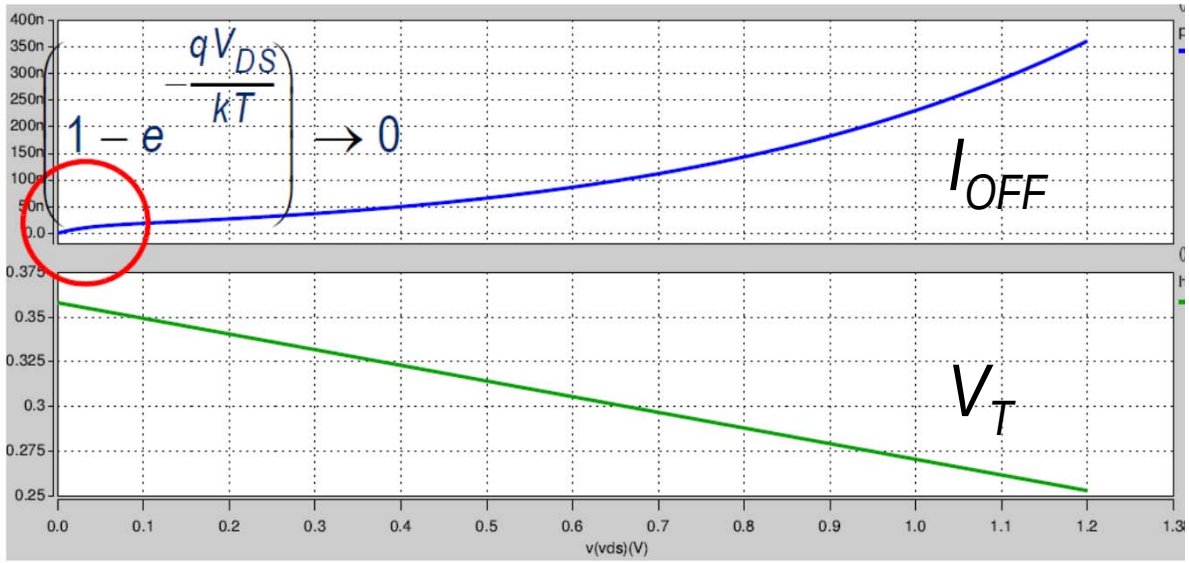


$$I_{sub} \propto e^{\frac{q(V_{GS} - V_T)}{kT}}$$

High $I_{OFF} \Rightarrow$ high static power + stability issues!

Note That I_{OFF} Strongly Depends on V_{DD}

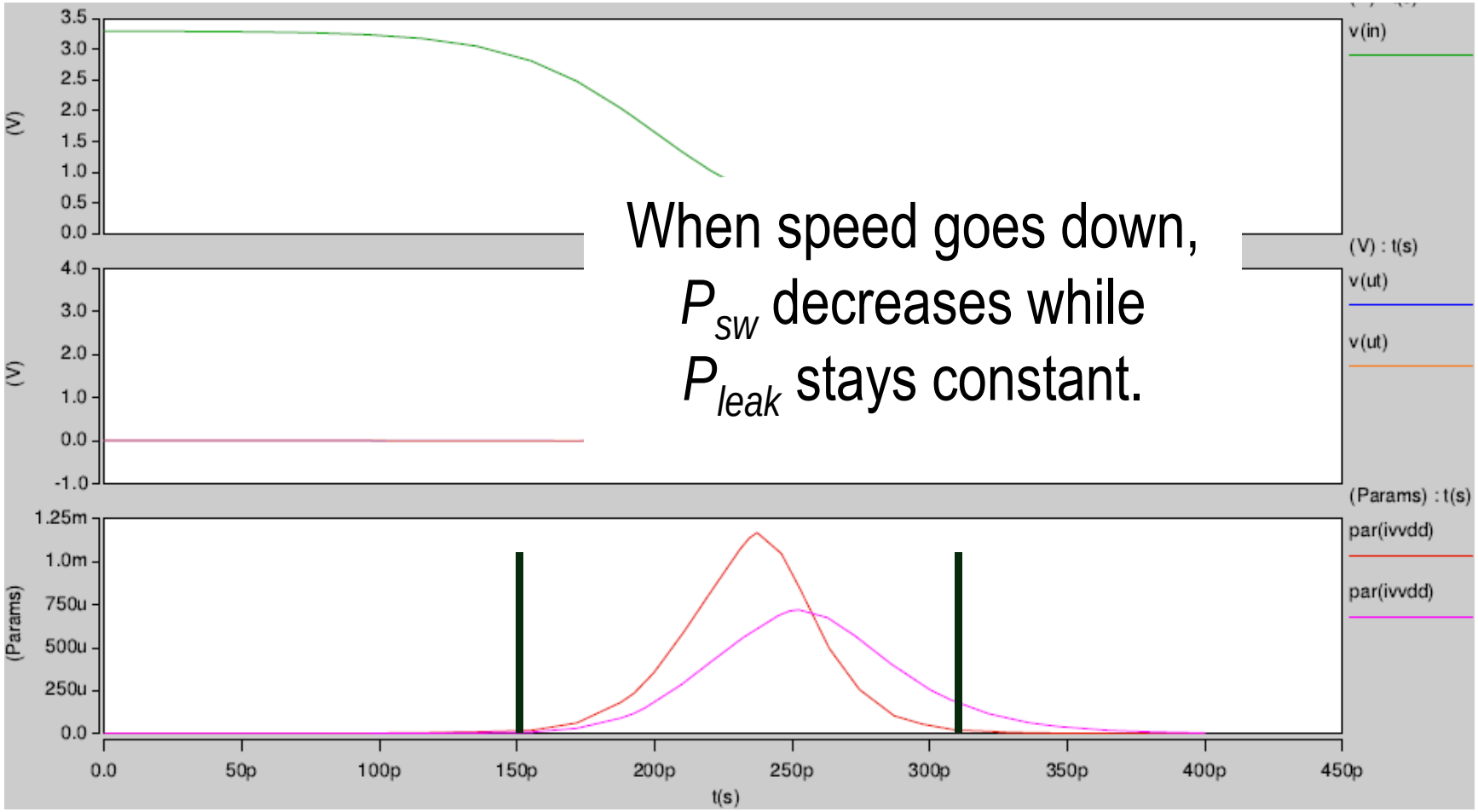




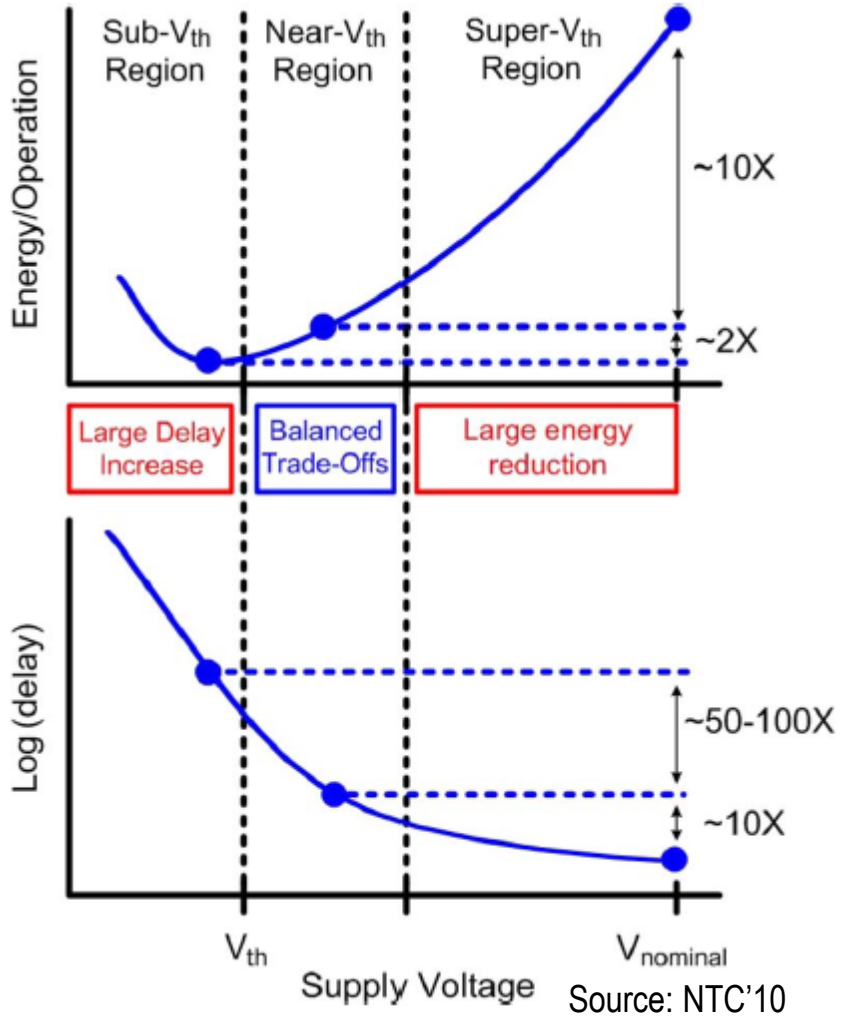
$$I_{sub} \propto e^{\frac{q(-V_T)}{kT}} \left(1 - e^{-\frac{qV_{DS}}{kT}}\right)$$

- For short channels, V_T increases with decreasing V_{DD}.
  - Cause: Drain-induced barrier lowering (DIBL).
  - DIBL lower in e.g. FinFETs.

Reduce Speed to Save Power?



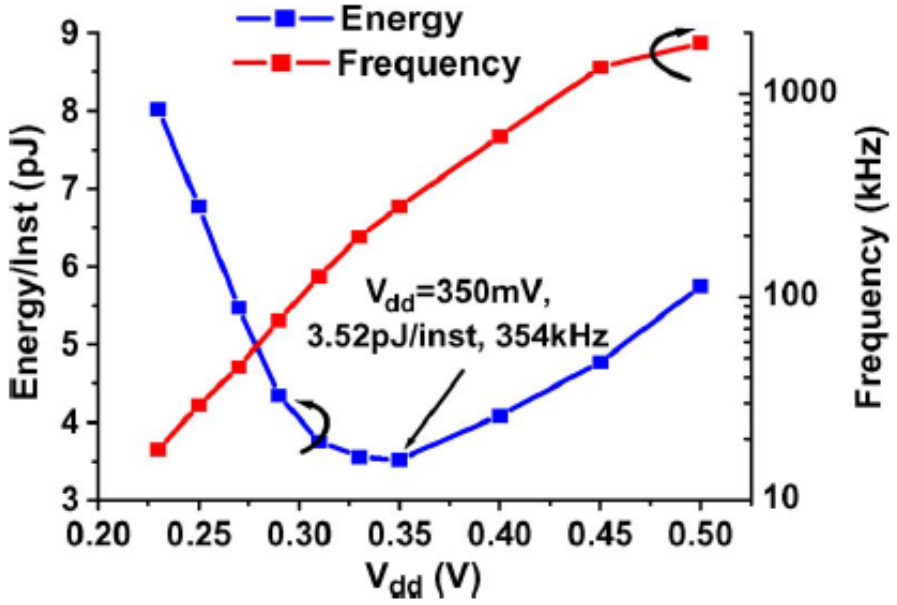
Subthreshold/Nearthreshold Computing



Source: NTC'10

- Lower speed $f \Rightarrow$ lower P_{sw}:
  - $P_{sw} = f\, \alpha C\, V_{DD}^2$
- and …reduced speed $\Rightarrow$ increased delay slack $\Rightarrow$ V_{DD} can be reduced $\Rightarrow$ P_{sw} dramatically reduced.
- Also P_{leak} is reduced due to reduced V_{DD}, but the exponentially deteriorating performance makes E_{leak} very significant.

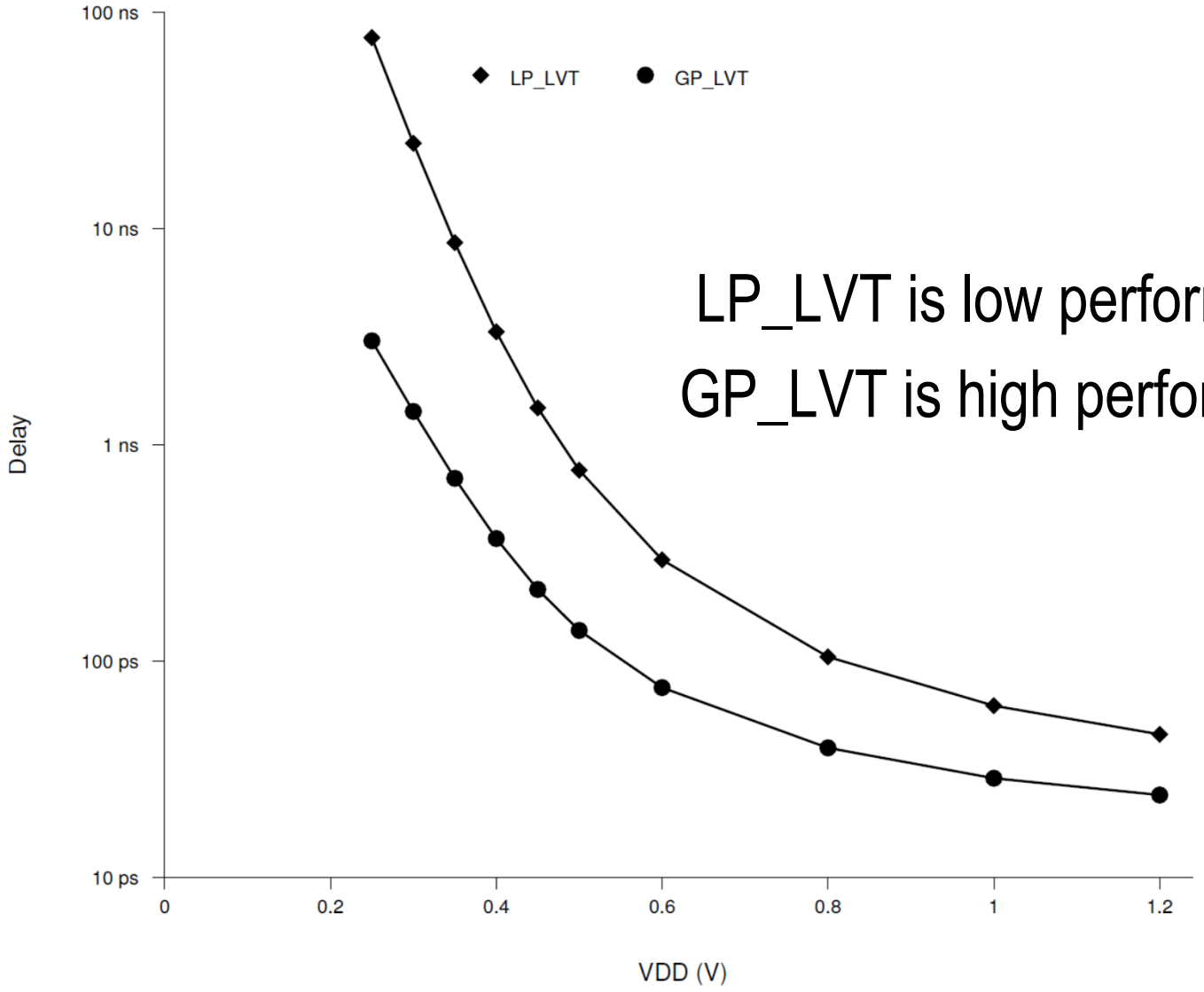
Subthreshold/Nearthreshold Computing



Source: NTC'10

- Subthreshold/nearthreshold operation is interesting because it leads to minimal energy/operation!
- Note though that performance is very poor in these regions.
- Some simulations follow for 65nm GP_LVT and LP_LVT.

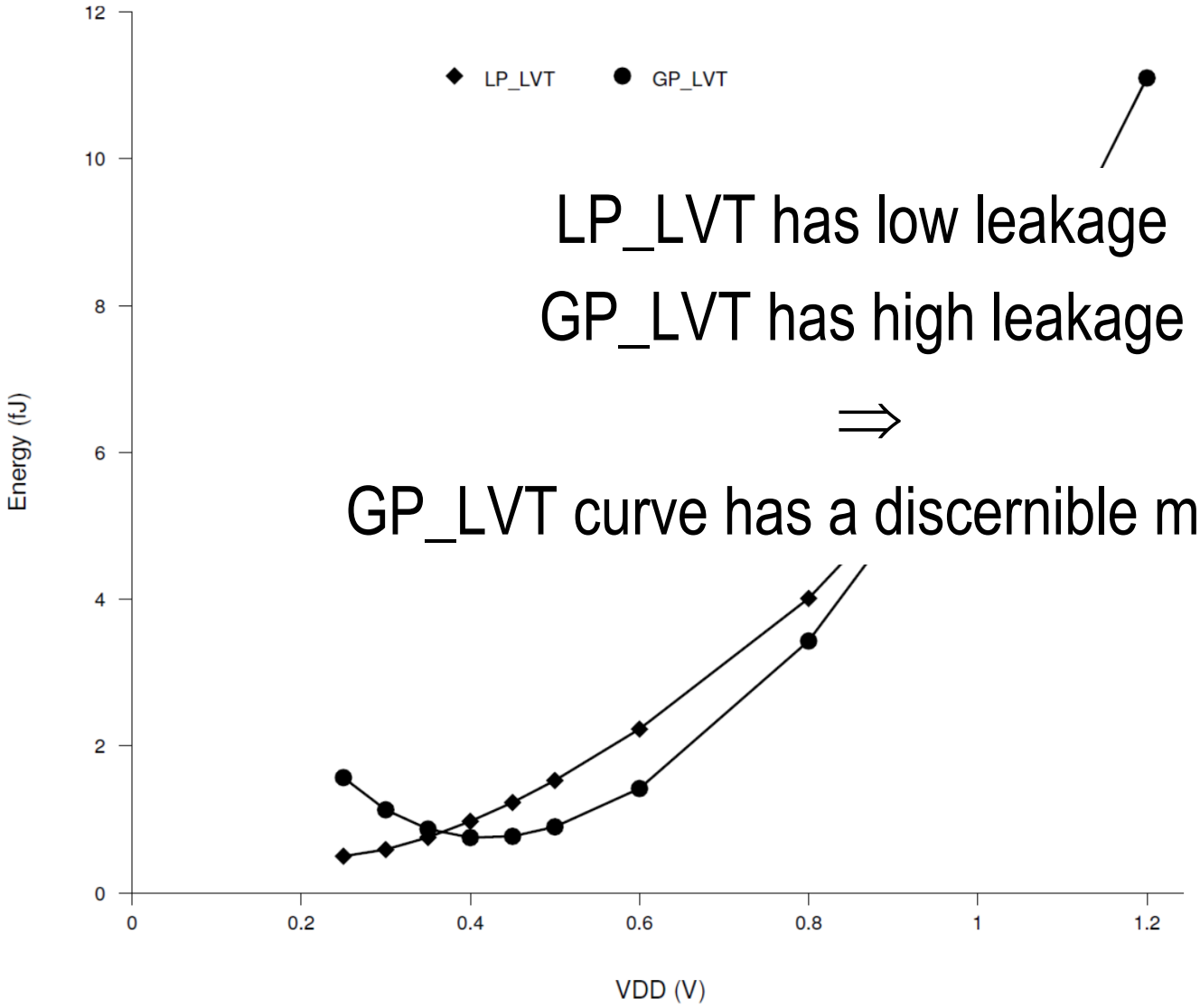
Delay vs V_{DD}



LP_LVT is low performance

GP_LVT is high performance

Energy/Operation vs V_{DD}



LP_LVT has low leakage

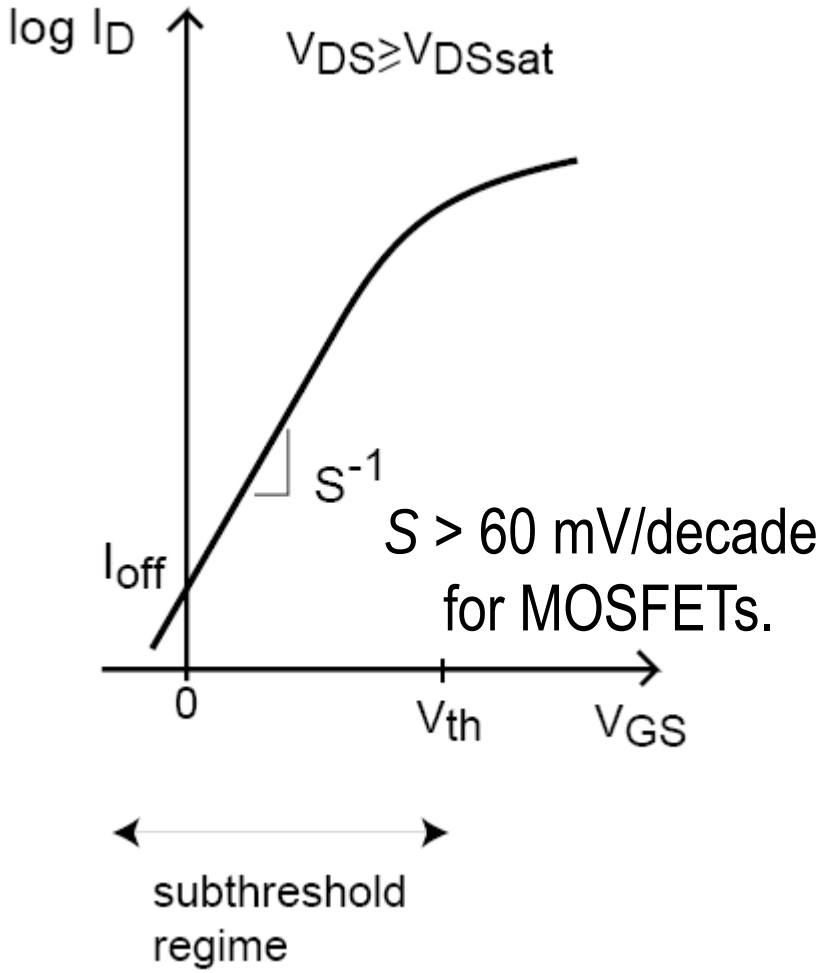
GP_LVT has high leakage

$\Rightarrow$

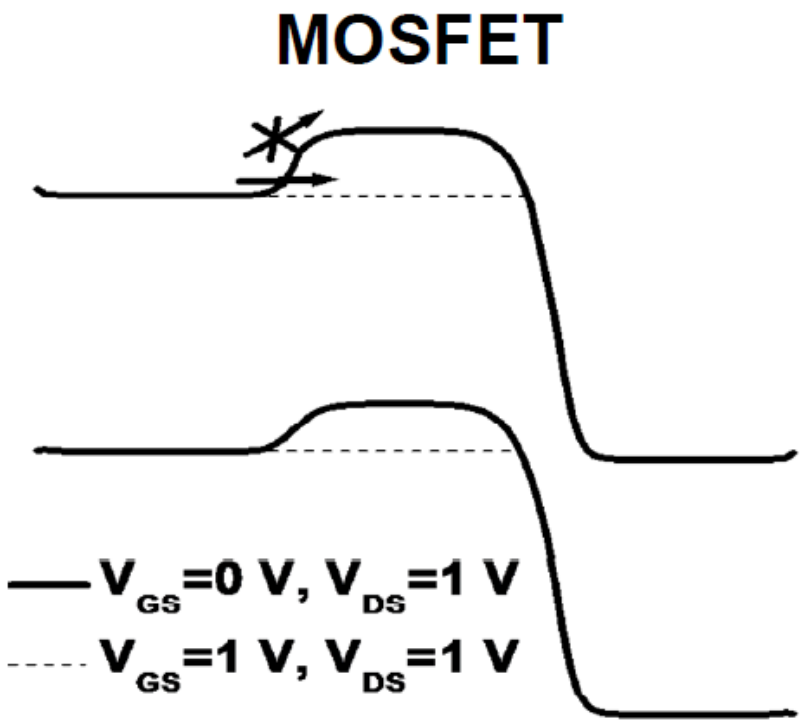
GP_LVT curve has a discernible minimum

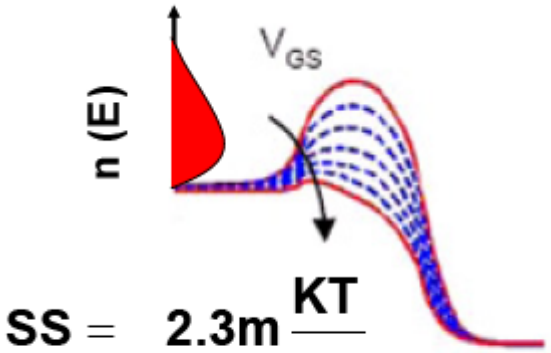
Subthreshold Swing/Slope Factor (*S*)

- To efficiently reduce I_{OFF}, a smaller subthreshold swing (*S*) than that of MOSFETs is desirable.
  - Priority: Good *S* for low V_{DD}s.
- This, however, requires radical changes to the fundamental transistor operation.



Limited Subthreshold Swing of FETs



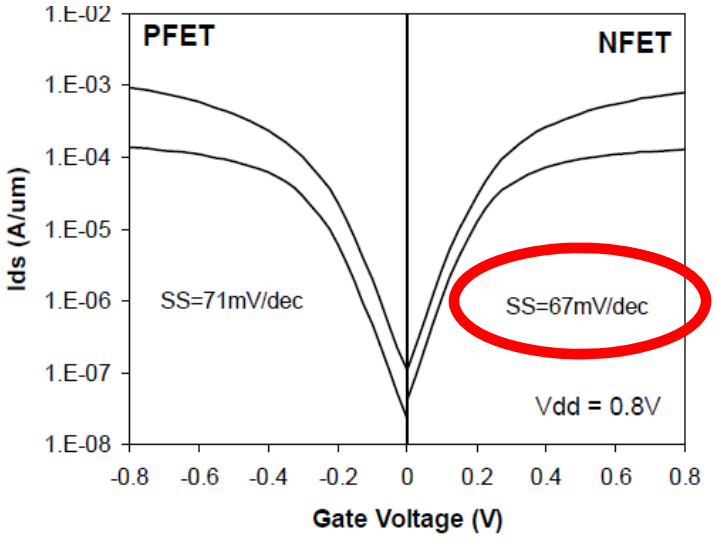


$$SS = \underbrace{2.3m\frac{KT}{q}}_{\geq 60mv/dec}$$

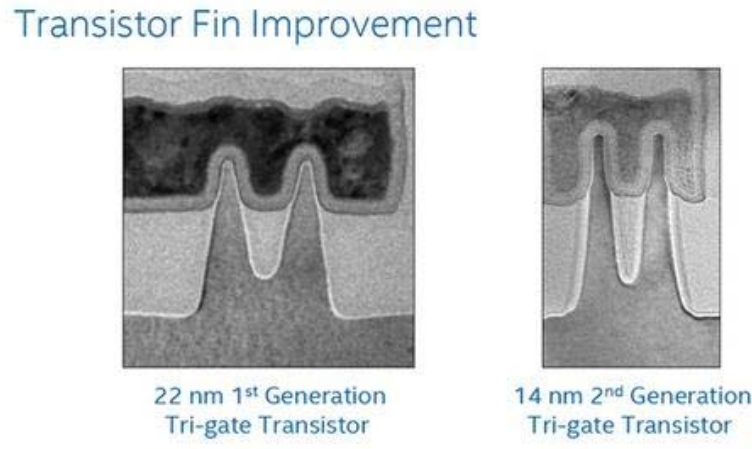
Source: BTBTFET'10

- *S* (called *SS* above) >= 2.3 * 0.026 = 59.8 mV/decade.
- $m = 1 + C_{depletion\text{-}layer}/C_{gate\text{-}oxide}$

The FinFET - A “3D Gate” FET

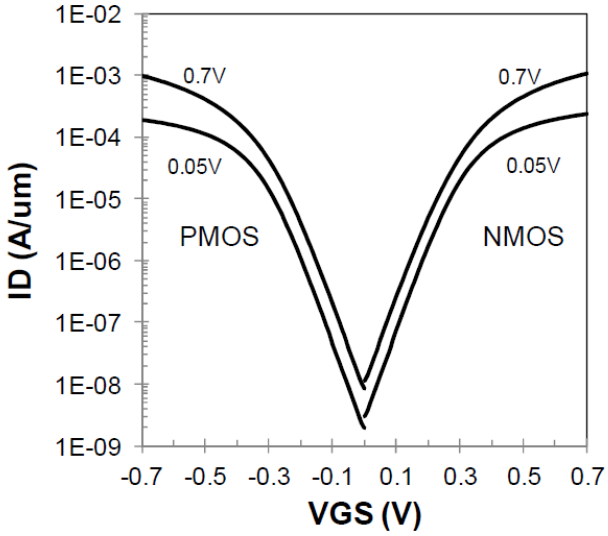


Source: IBM14nm

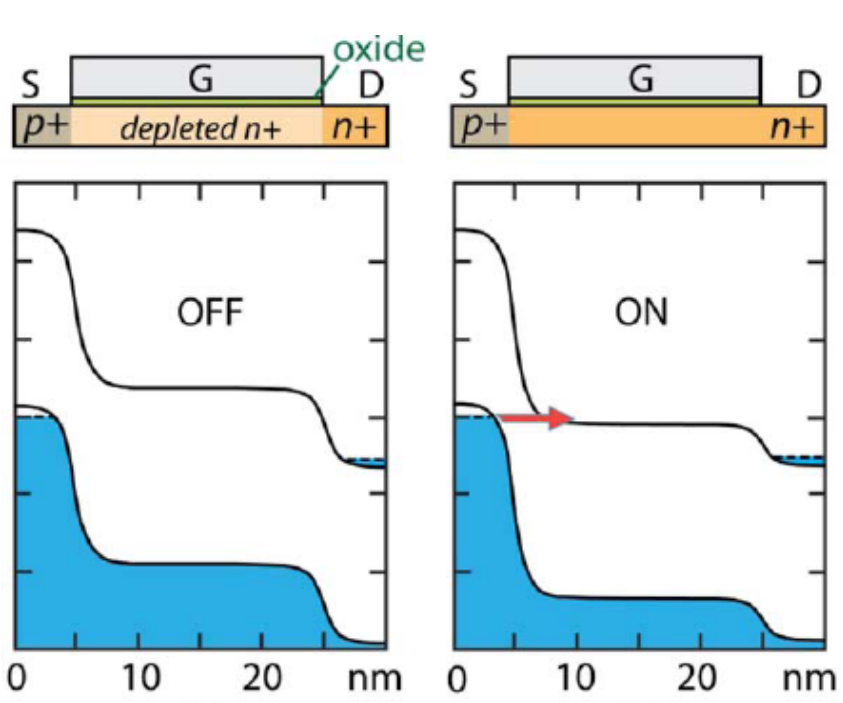


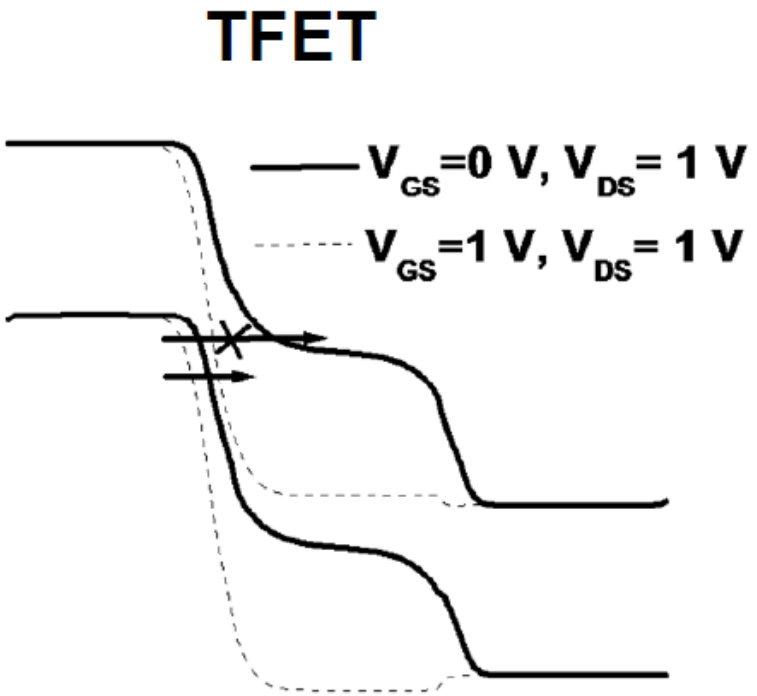
Source: Intel14nm

- FinFETs:
  - Decent *S* at low V_{DD}.
  - But still MOSFETs !
  - IBM – SOI-based FinFETs
  - Intel – bulk FinFETs



Post/Beyond-CMOS - Tunnel-FETs ?





Source: BTBTFET'10

- A lower subthreshold swing gives acceptable I_{ON}/I_{OFF} ratios at low supply voltages.

TFET vs CMOS

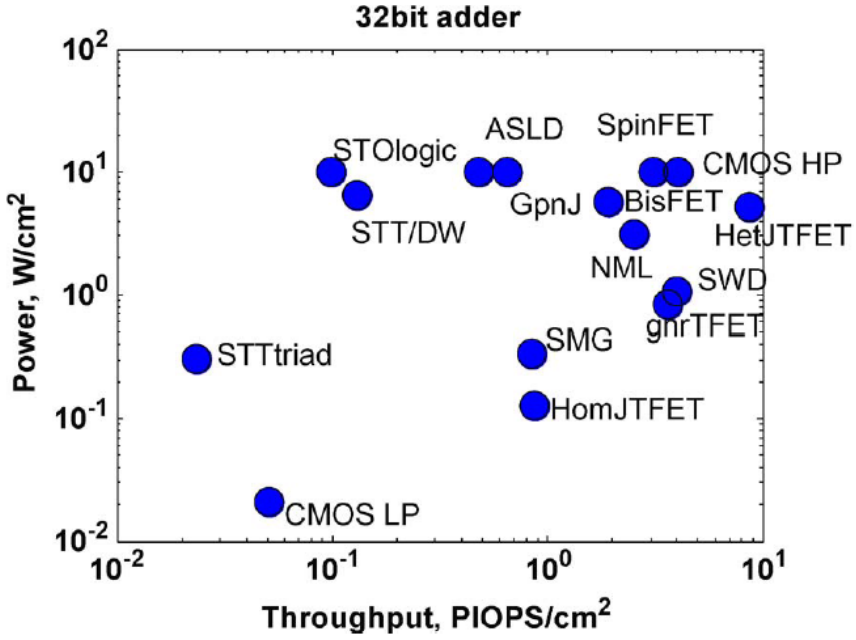


Fig. 52. ***Throughput versus dissipated power density of devices. The preferred corner is bottom right.***

Source: OBCD'13

- TFETs (e.g. HetJFET) are promising but CMOS is not doing that bad…
- The challenge is to make TFETs that can both have
  - subthreshold swings << 60mV/decade
  - high I_{ON} currents

Conclusion

- Implementation aspects, technology and circuits, strongly impact power and energy dissipation.
- Several power dissipating mechanisms $\Rightarrow$ need different low-power techniques (next lecture).
- While reducing V_{DD} is the most effective way to reduce power, this also has disadvantages:
  - Lower speed, which hurts performance, or
  - lower V_T to maintain speed; this in turn increases leakage.
  - Larger impact of variability in any case.

References

- BTBTFET'10: "Band-to-Band Tunneling Field Effect Transistor for Low Power Logic and Memory Applications", S. A. Mookerjea, PhD Thesis, Penn State Univ, 2010.
- CATPE'08: "Computer Architecture Techniques for Power-Efficiency", S. Kaxiras and M. Martonosi, Morgan & Claypool, 2008.
- EPM'14: "The Evolution of Power Measurement", K. Cameron, Computer, IEEE, Mar. 2014.
- IBM14nm: "High Performance 14nm SOI FinFET CMOS Technology with 0.0174μm2 embedded DRAM and 15 Levels of Cu Metallization", C.-H. Lin et al., IEDM 2014.
- Intel14nm: "A 14nm logic technology featuring 2nd-generation FinFET, air-gapped interconnects, self-aligned double patterning and a 0.0588 μm2 SRAM cell size", S. Natarajan et al., IEEE Int. Electron Devices Meeting (IEDM), 2014.
- LPDE'09: "Low Power Design Essentials", J. Rabaey, Springer, 2009.
- NTC'10: "Near-Threshold Computing: Reclaiming Moore's Law Through Energy Efficient Integrated Circuits", R. G. Dreslinski, et al., Proc. of IEEE, Feb. 2010.
- OBCD'13: "Overview of Beyond-CMOS Devices and a Uniform Methodology for Their Benchmarking", D. E. Nikonov, Proc. of IEEE, Dec. 2013.
- SSD'13: "Steep-Slope Devices: From Dark to Dim Silicon", K. Swaminathan et al., IEEE Micro, 2013.