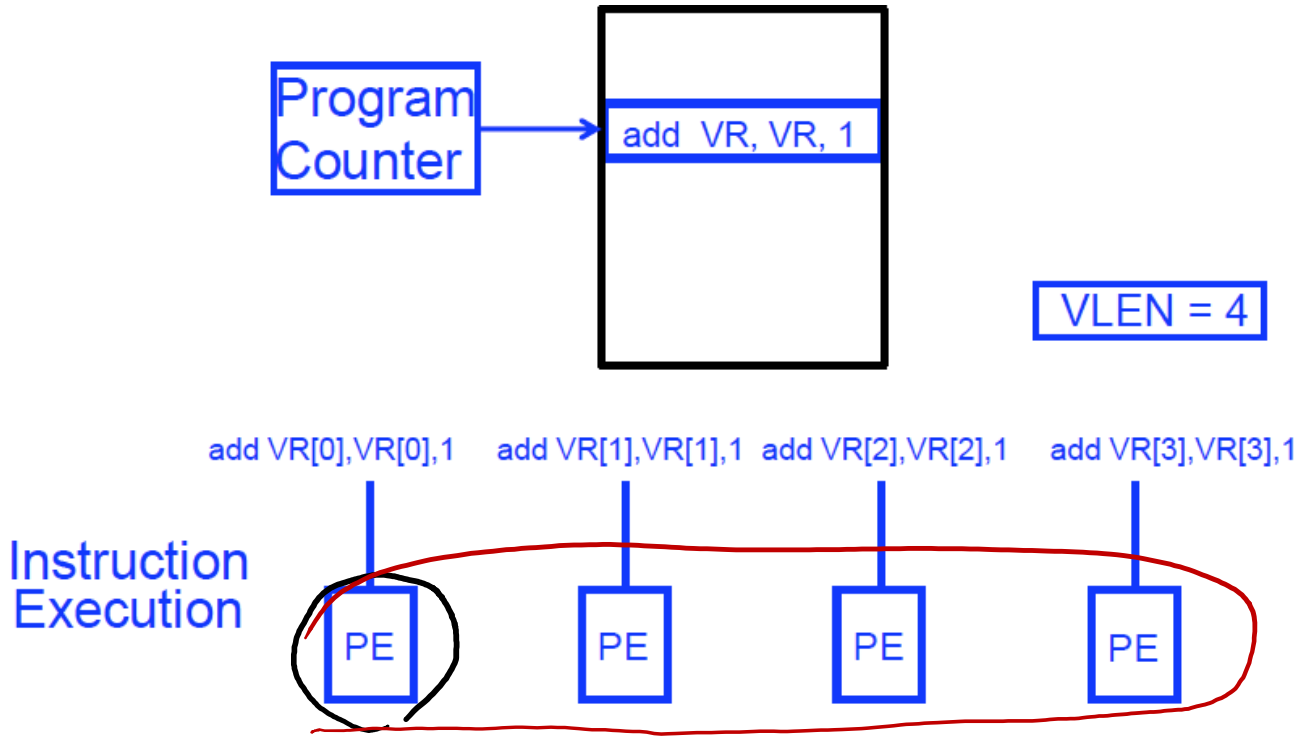
Lecture 20:
Impact of Technology

Review: SIMD Processing

- single instruction controls simultaneous execution of many processing elements
  - lockstep basis



Review: Why SIMD

+more parallelism when parallelism is abundant

+SIMD in addition to ILP

+simple design

+replicated functional units

+energy efficient

+ only needs one instruction for many data operations

+ attractive for personal mobile devices

+small die area

+no heavily ported register files

+die area: +MAX-2(HP): 0.1% +VIS(Sun): 3.0%

-must be explicitly exposed to the hardware

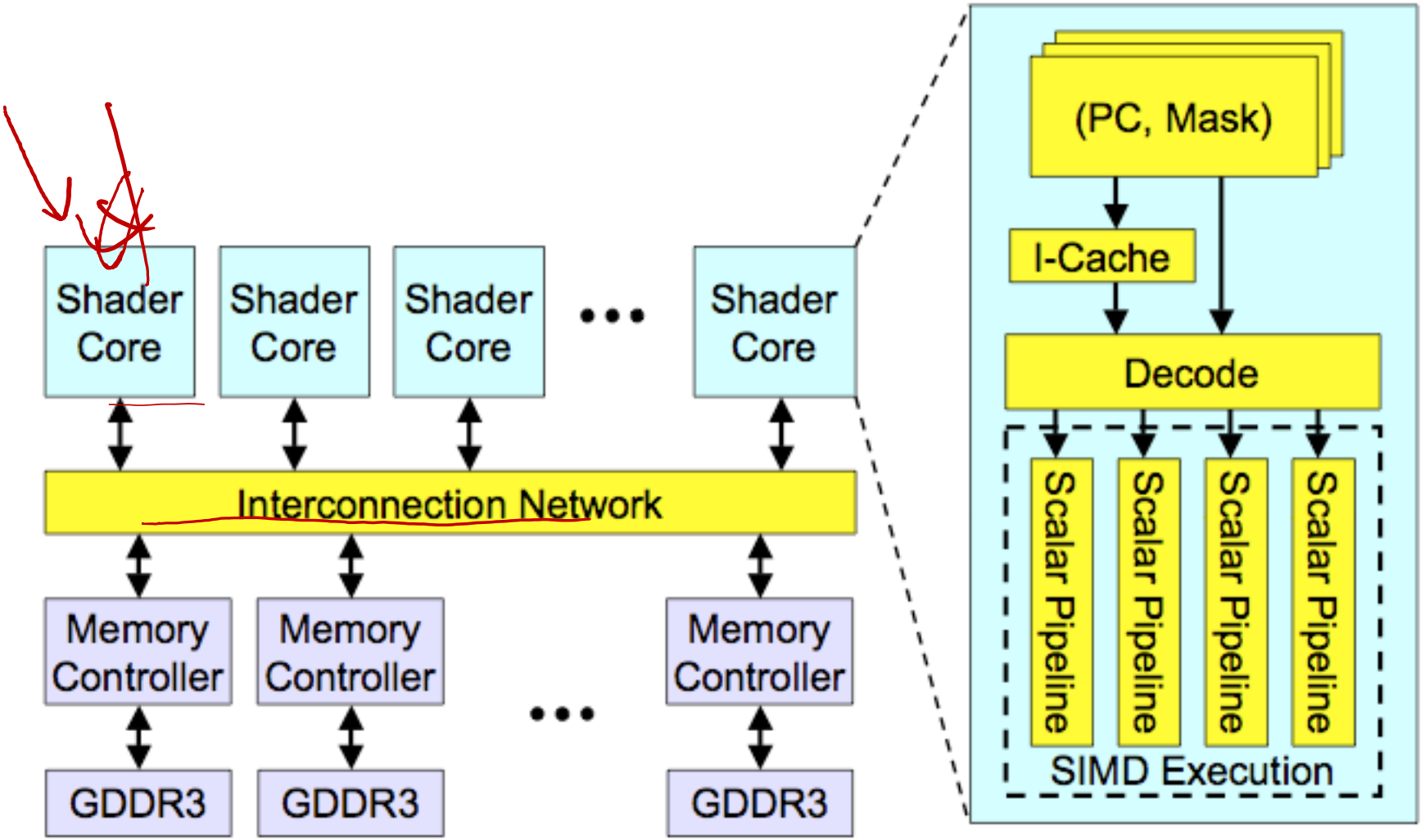
-by the compiler or by the programmer

Review: Exploiting TLP with SIMD

- benefit:
  - multiple ALU ops → one SIMD op
  - multiple ld/st ops → one wide mem op

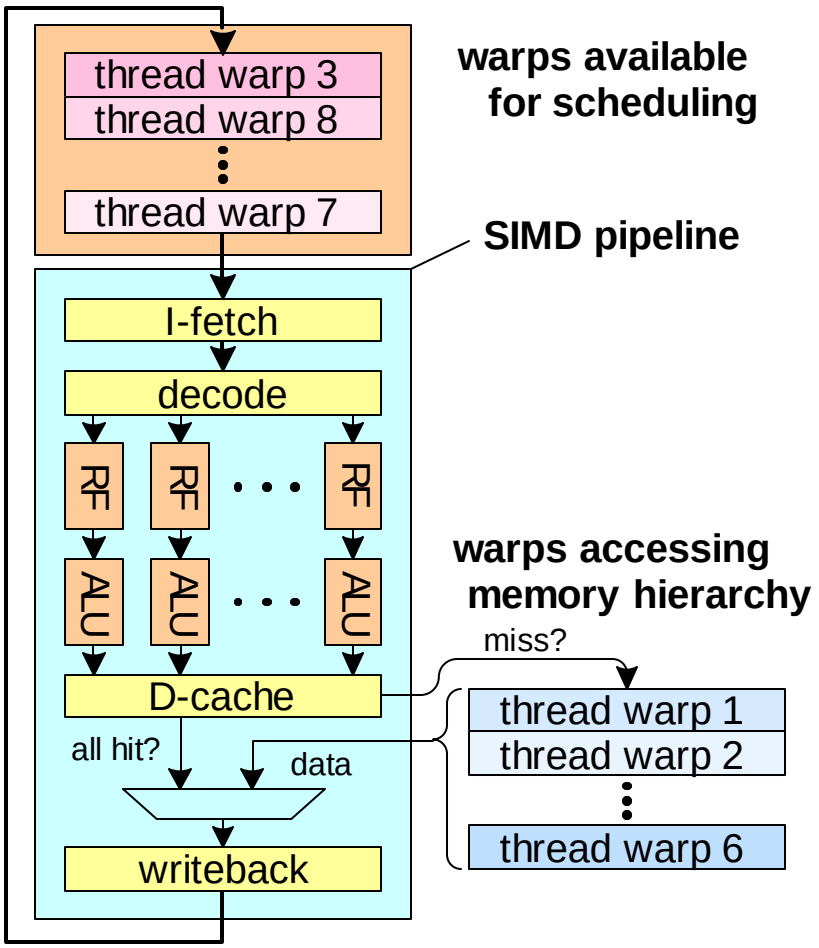
- what are the overheads:
  - packing and unpacking:
    - rearrange data so that it is contiguous
  - alignment overhead
    - accessing data from the memory system so that it is aligned to a "superword" boundary
  - control flow
    - control flow may require executing all paths

Review: High-Level View of a GPU



Review: Latency Hiding

- warp: a set of threads that execute the same instruction (on different data elements)
- fine-grained multithreading
  - One instruction per thread in pipeline at a time (no branch prediction)
  - interleave warp execution to hide latencies
- register values of all threads stay in register file
- no OS context switching
- memory latency hiding
  - graphics has millions of pixels

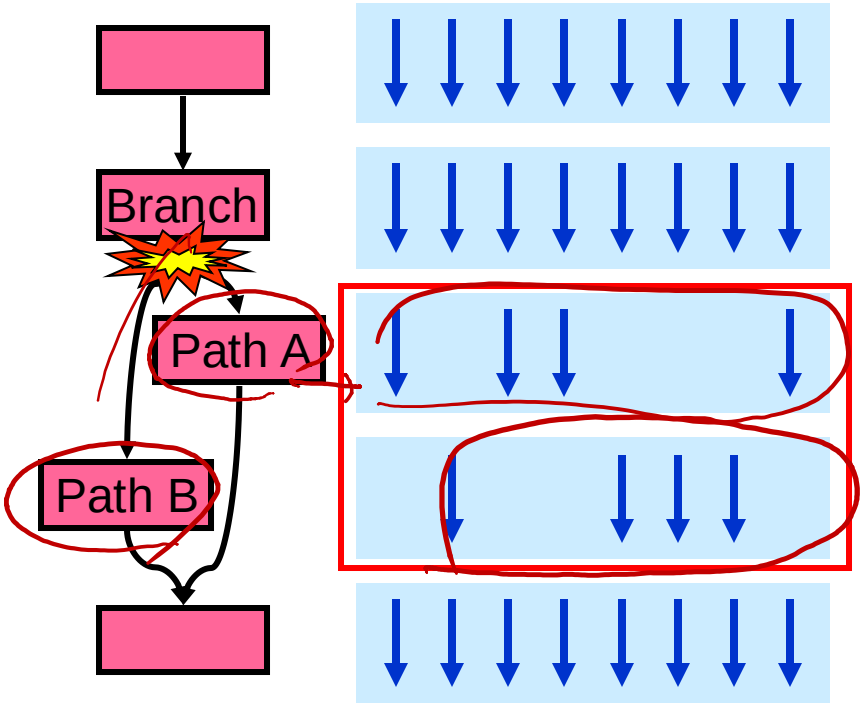


Slide credit: Tor Aamodt

Review: Control Flow Problem

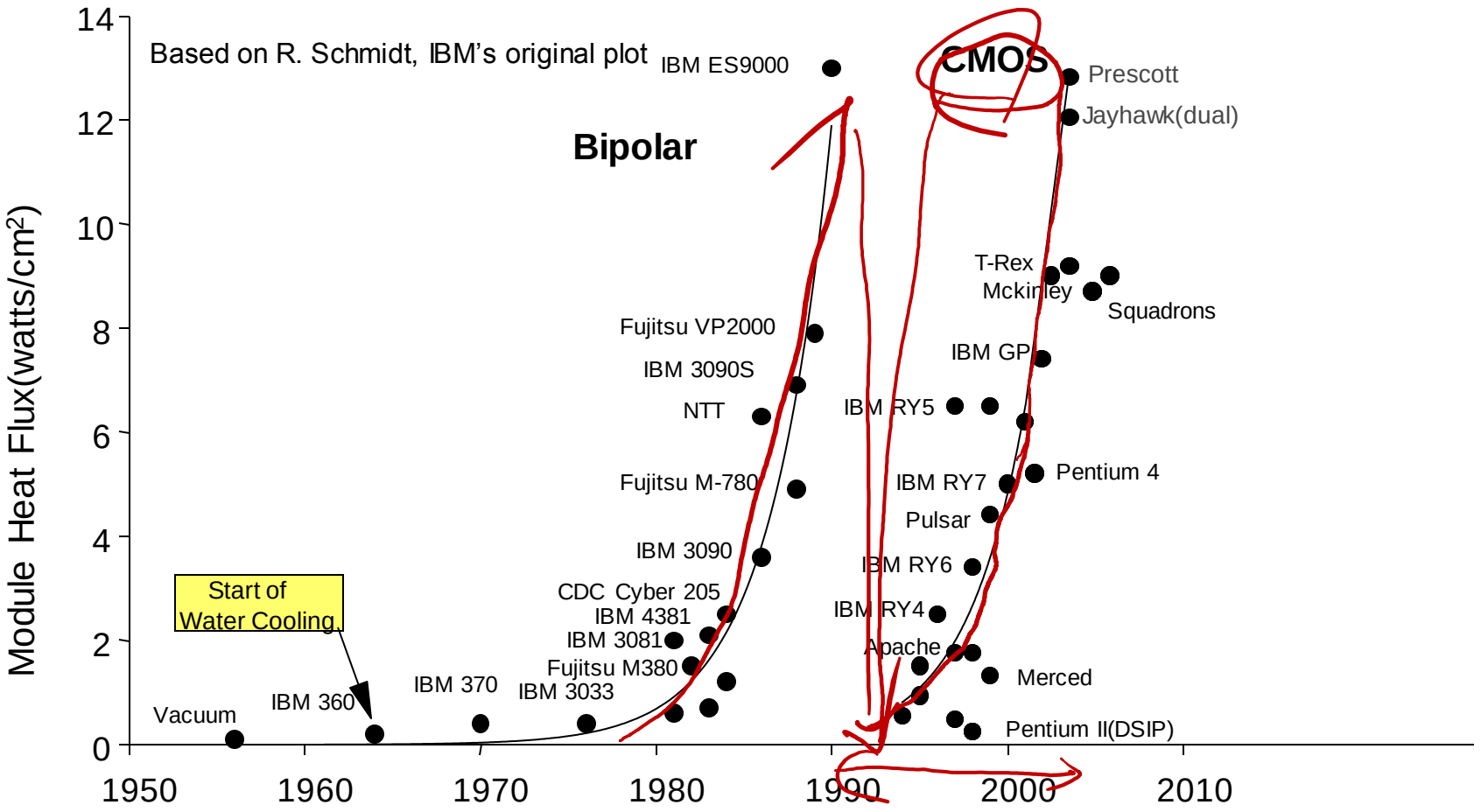
- GPU uses SIMD pipeline to save area on control logic.
  - group scalar threads into warps

- branch divergence occurs when threads inside warps branch to different execution paths.



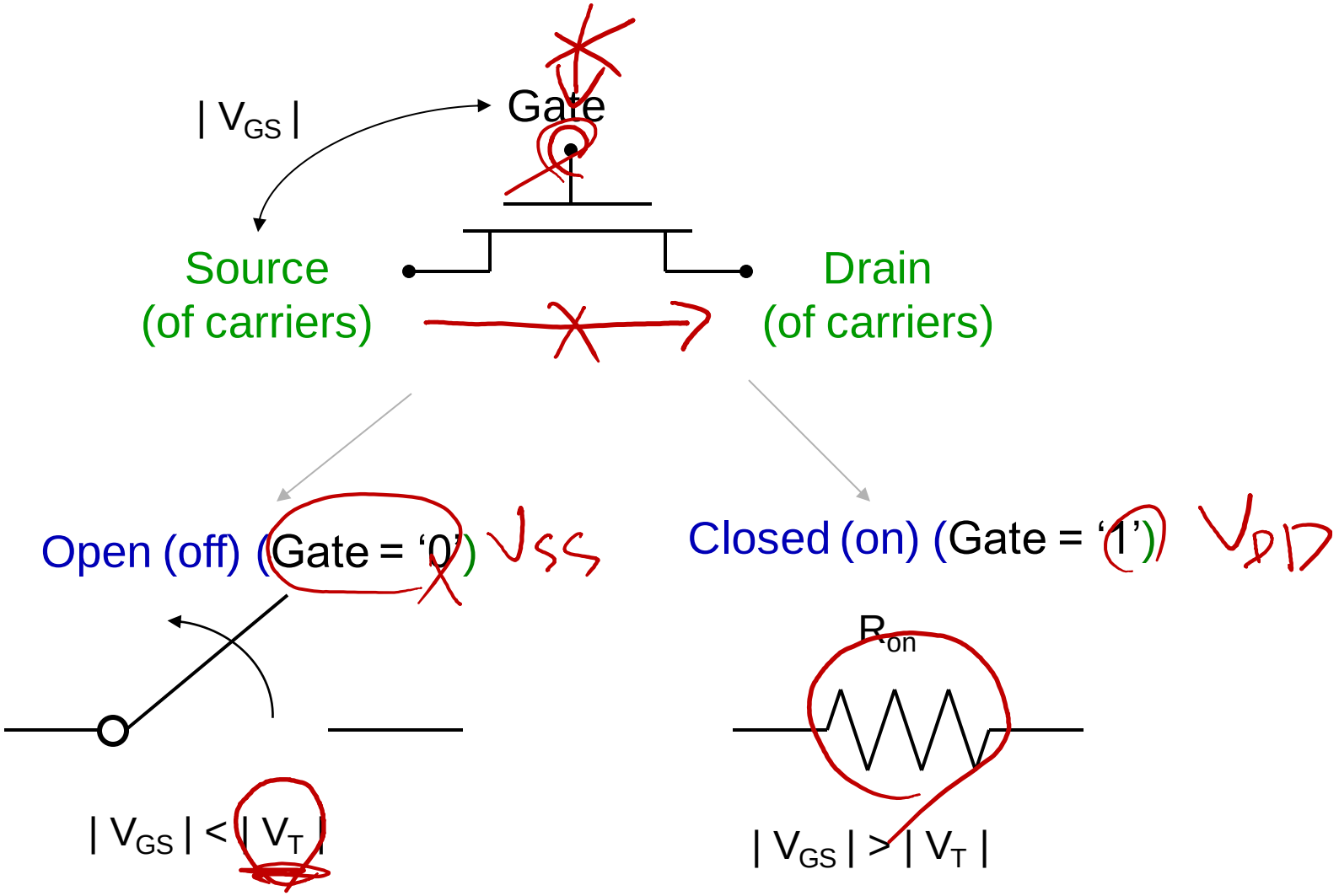
Slide credit: Tor Aamodt

Technology Trend: Bipolar to CMOS

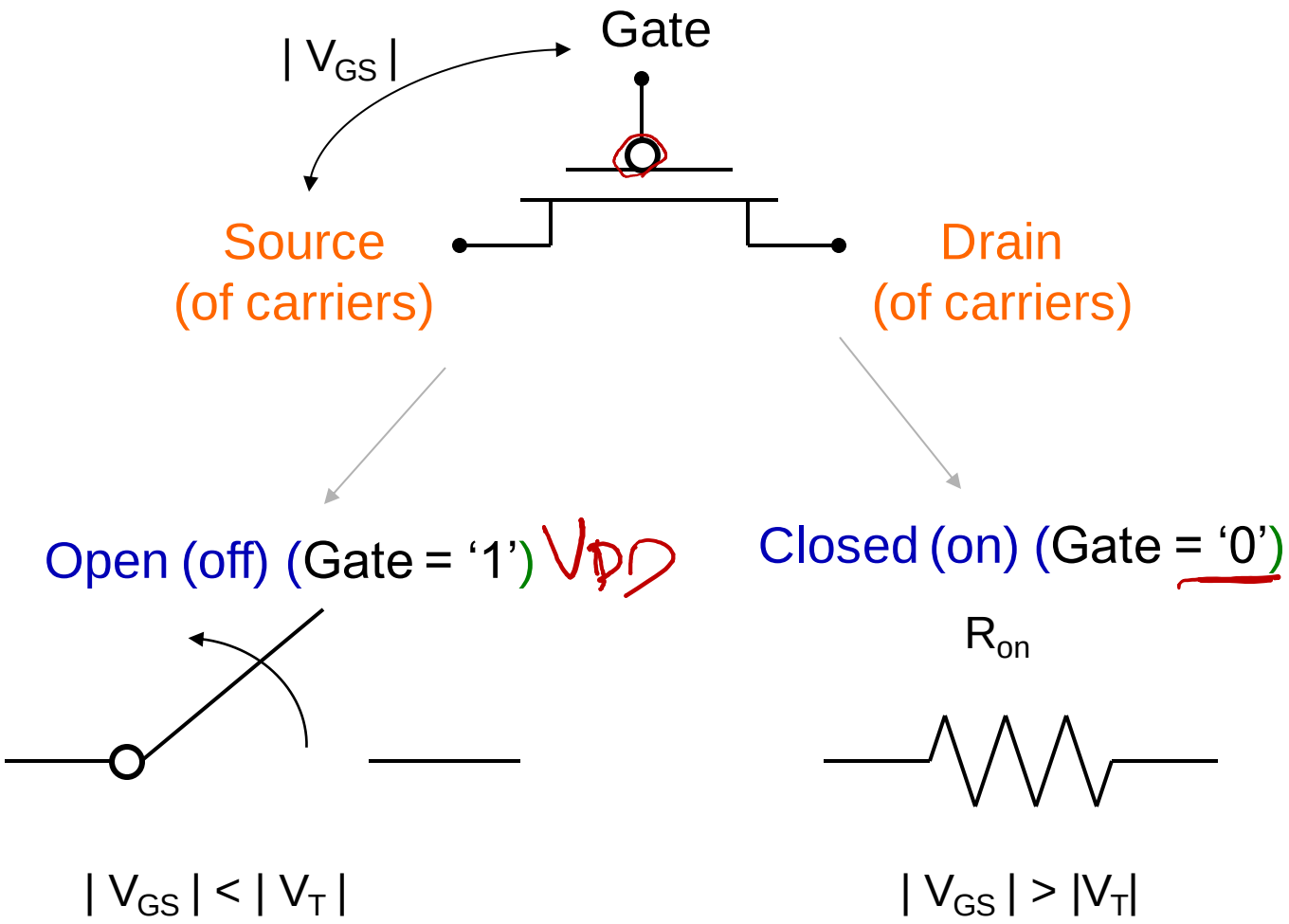


- performance is fundamentally limited by power consumption!
  - ✓ not just a battery issue for mobile computing devices

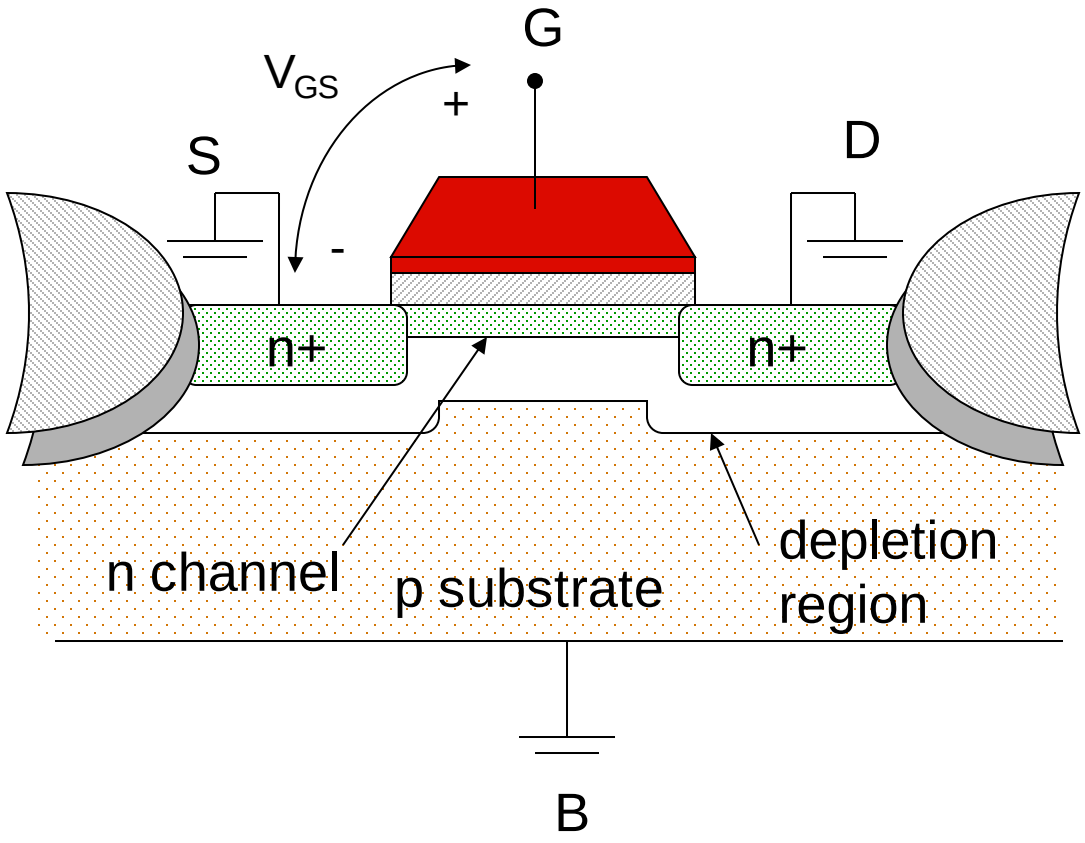
Switch Model of NMOS Transistor



Switch Model of PMOS Transistor



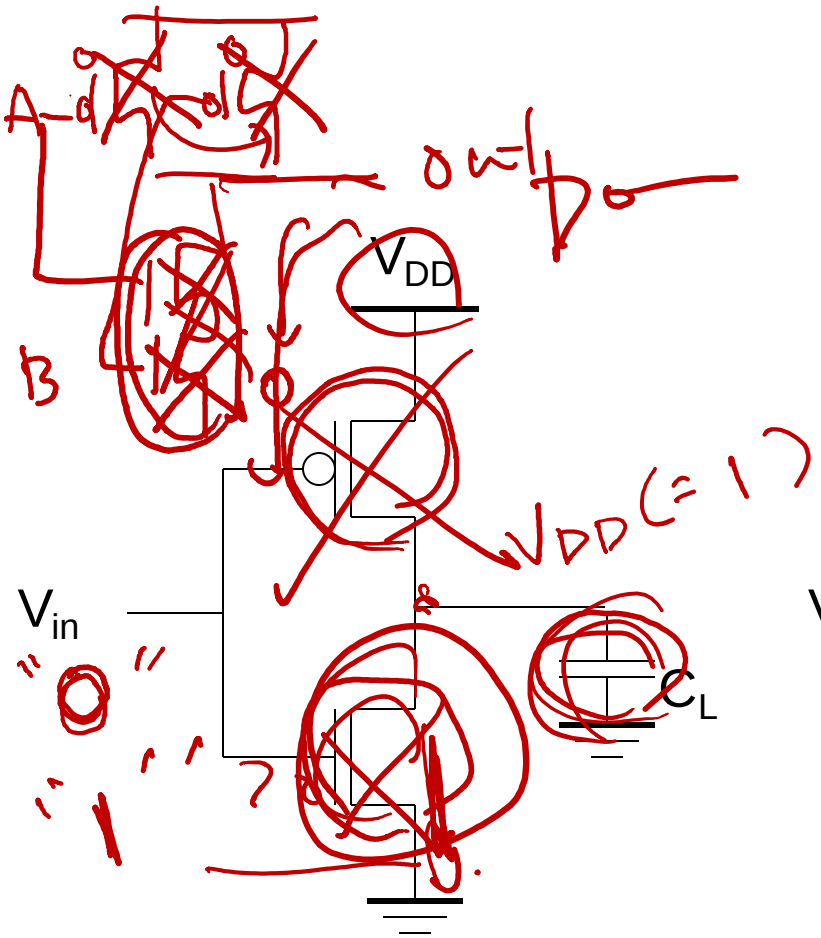
Threshold Voltage Concept

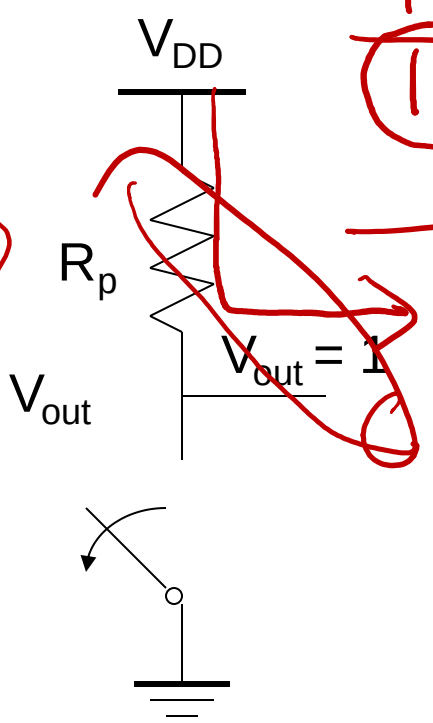


- Threshold voltage (V_T)
  - ✓ The value of V_{GS} where strong inversion occurs

$$V_T = V_{T0} + \gamma(\sqrt{|-2\phi_F + V_{SB}|} - \sqrt{|-2\phi_F|})$$

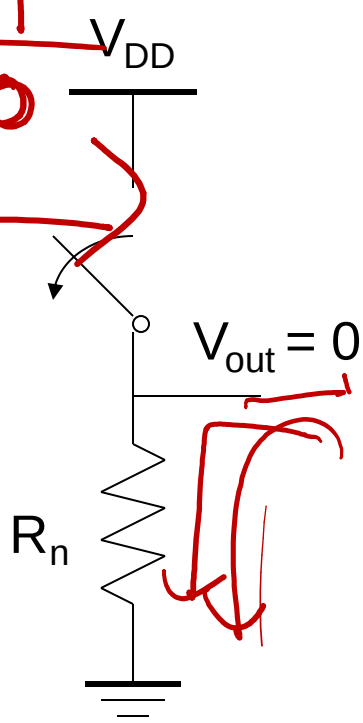
CMOS Inverter: A First Look





$V_{OH} = V_{DD}$

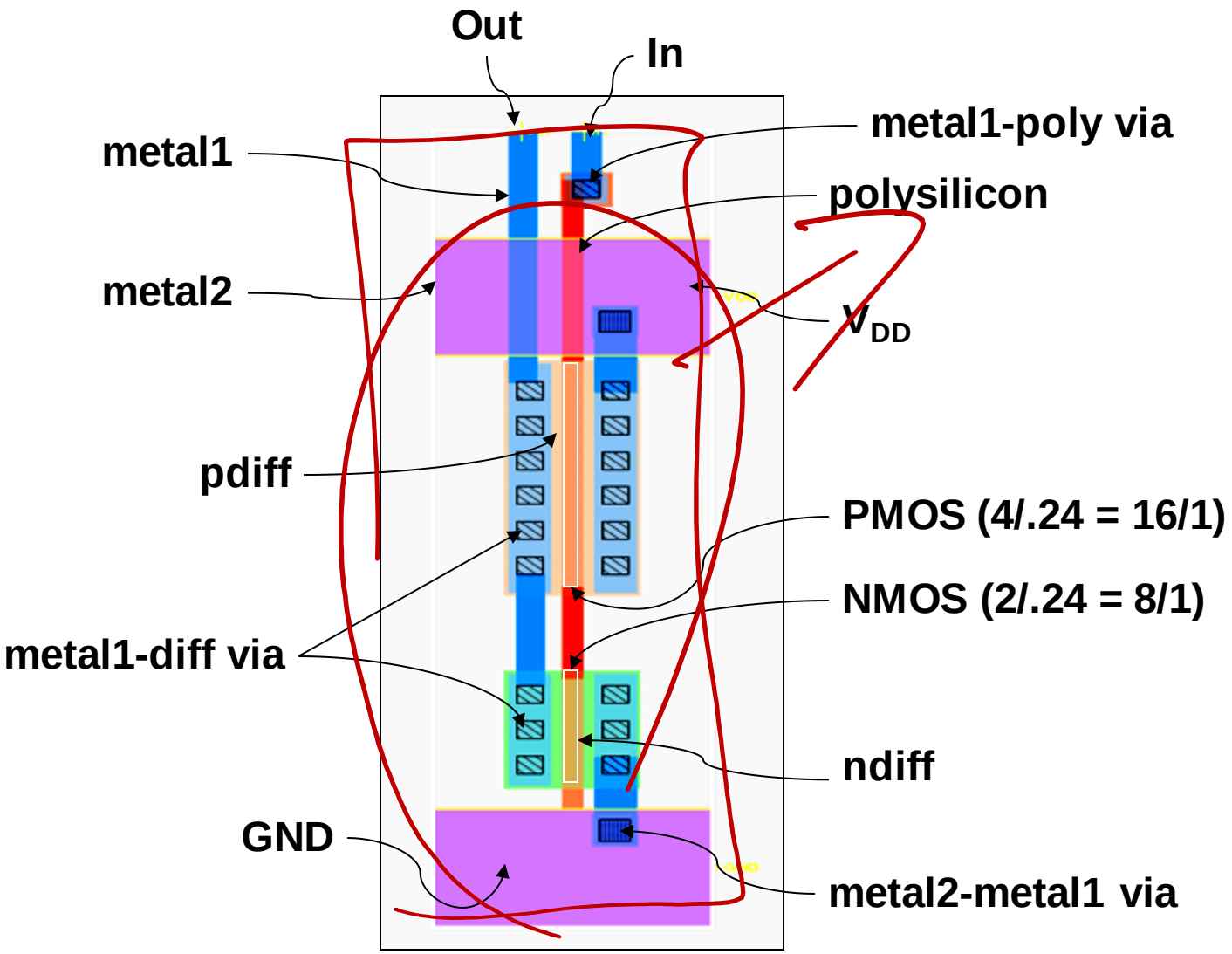
$V_{in} = 0$



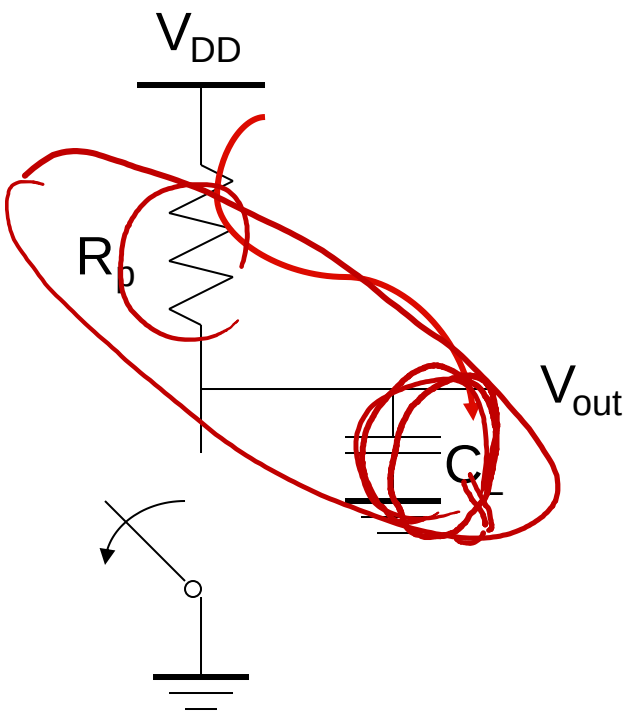
$V_{OL} = 0,$

$V_{in} = V_{DD}$

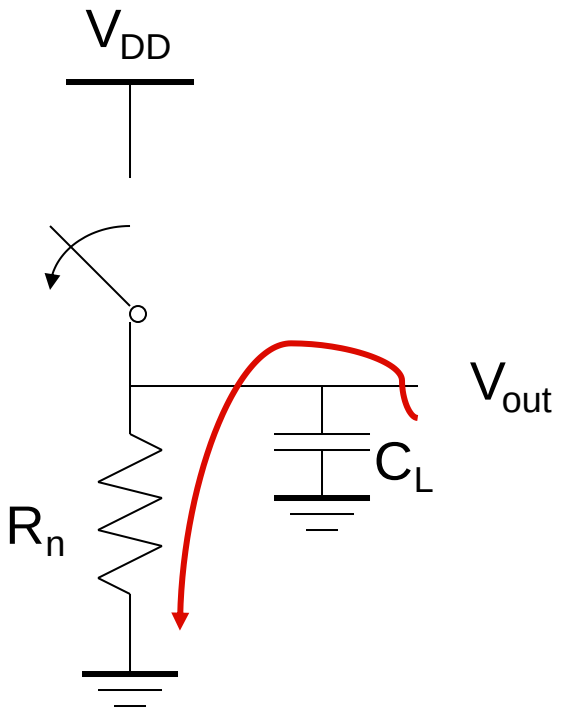
CMOS Inverter Layout



Switch Model of Dynamic Behavior



$V_{in} = 0$



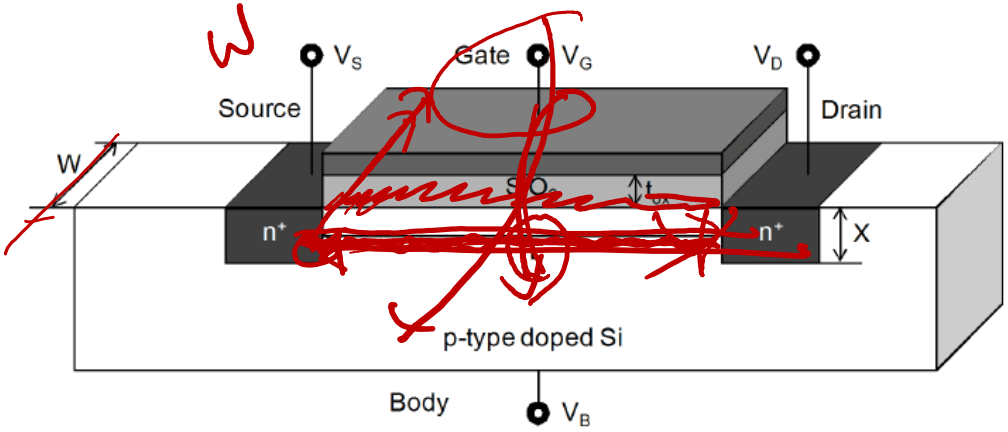
$V_{in} = V_{DD}$

Gate response time is determined by the time to charge C_L through R_p (discharge C_L through R_n)

Technology Scaling

- Moore's Law
  - ✓ all these features scale by 1/S

$\frac{1}{S} = 0.7$ $S = 1.4$



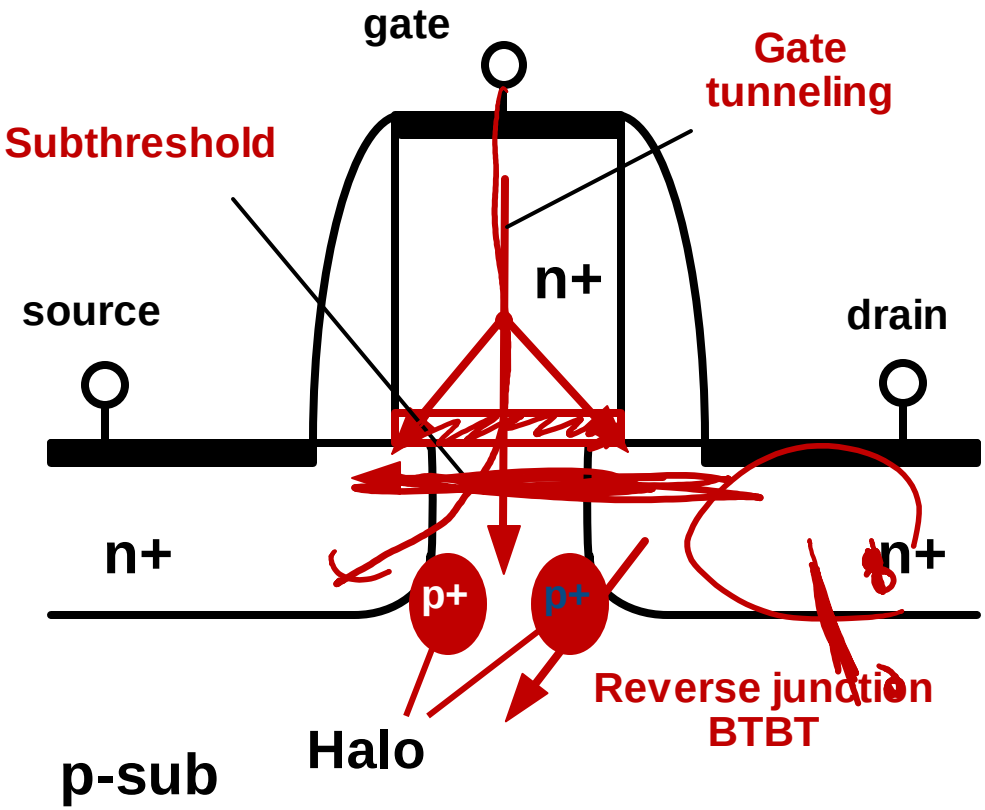
Feature/Voltage	Variable
Channel Length	L
Channel Width	W
Oxide Thickness	t_{ox}
Junction Depth	X
Supply Voltage	V_{dd}
Threshold Voltage	V_{th}
Wire width, space, height	w,s,h

Impact of Scaling

Device Characteristics	Feature Dependence	Scaling
Transistor Gain (β)	$W/(L.\ t_{ox})$	S
Current (I_{ds})	$\beta(V_{dd}-V_{th})^2$	S
Resistance	V_{dd}/I_{ds}	1
Gate Capacitance	$(W.L)/t_{ox}$	1/S
Gate delay	R.C	$1/S^2$
Clock Frequency	1/(R.C)	S
Circuit Area	W.L	$1/S^2$
Wire Resistance	l/(w.h)	S^2
Wire Capacitance	h/s	1

CMOS Scaling & Leakage Power

Long Channel (L > 1μm)	Short Channel (L > 180nm, t_{ox} > 20Aº)	Very Short Channel (L > 90nm, t_{ox} > 15Aº)	Nano-Scaled (L < 90nm, t_{ox} < 15Aº)
Negligible leakage	Subthreshold leakage	Subthreshold + gate leakage	Subthreshold + gate + BTBT leakage



- Power is consumed when tr's are switching, but
- 30~40% of total power in 45nm microprocessors is leakage.
- Leakage is becoming a major limiting factor for a large scale CMOS integration.

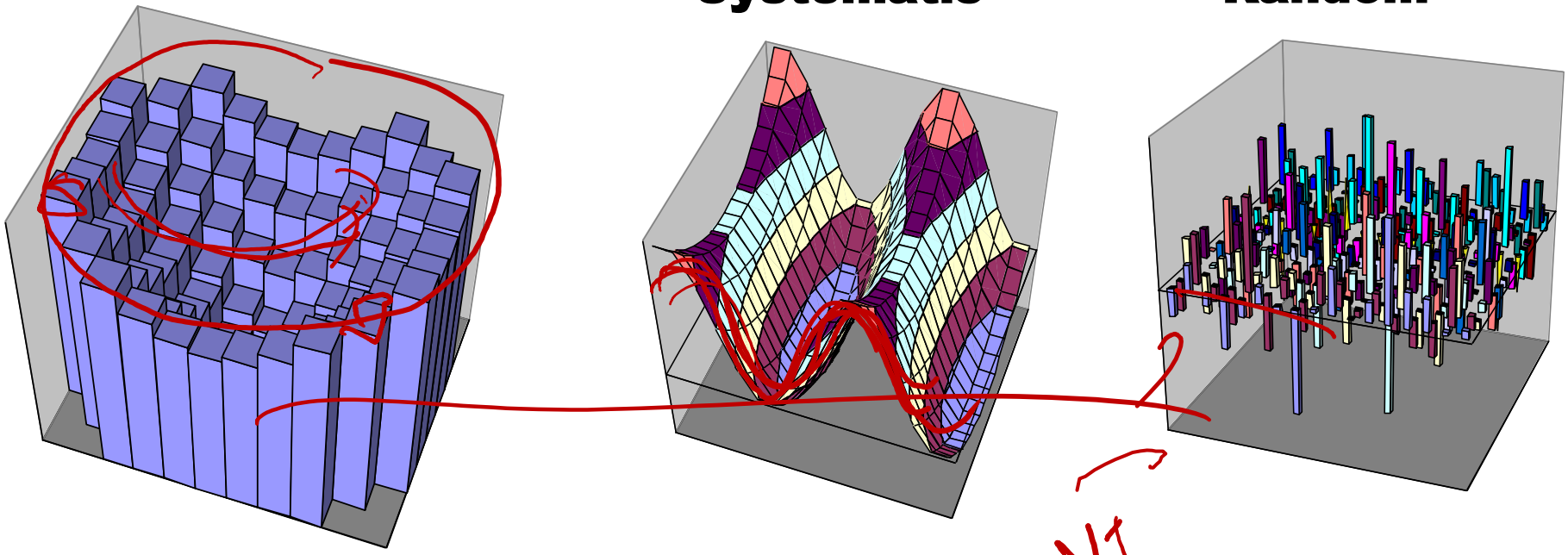
Scale of Process Variations

Die-to-Die (D2D) Variations

Within-Die (WID) Variations

Systematic

Random



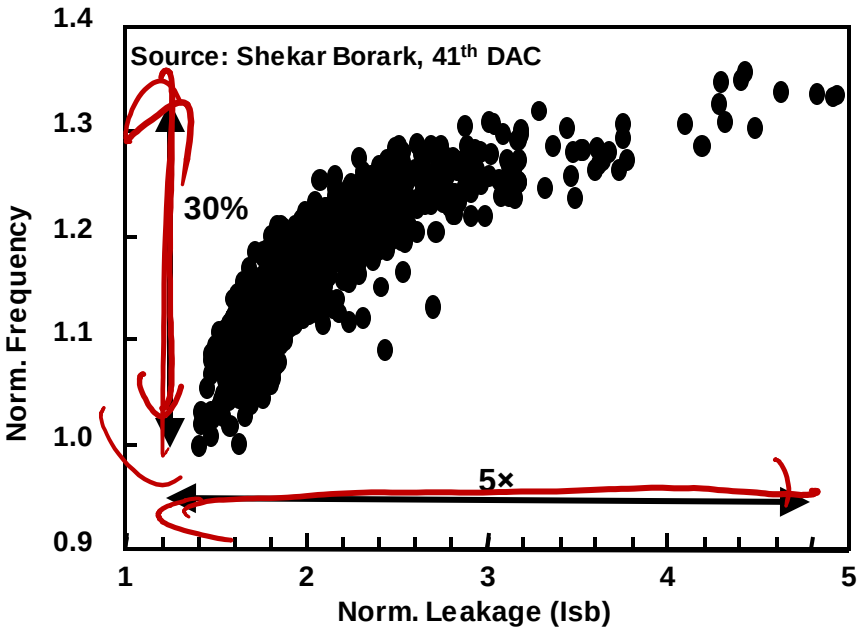
Wafer Scale

Die Scale

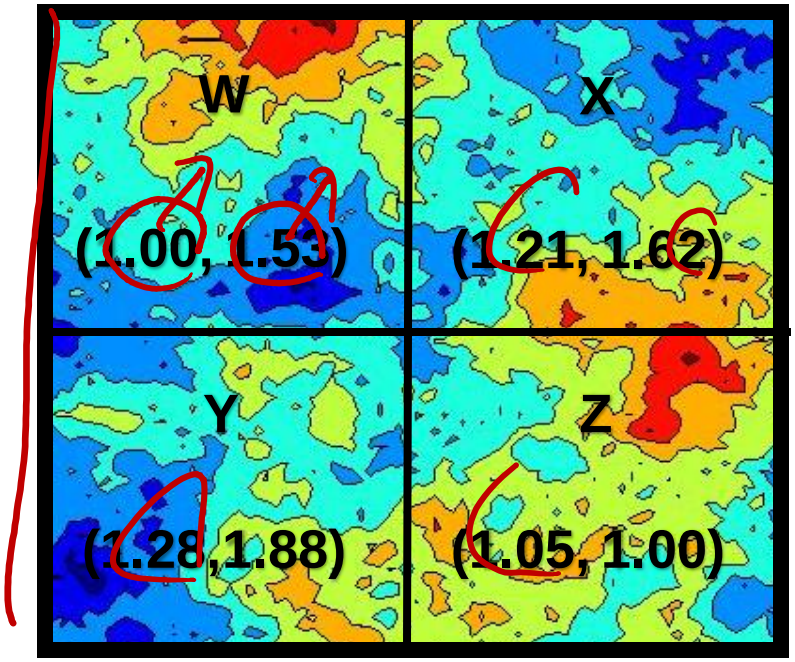
Feature Scale

Courtesy: K. Bowman from Intel

Negative Impact of Tech Scaling



130nm ~1000 die samples

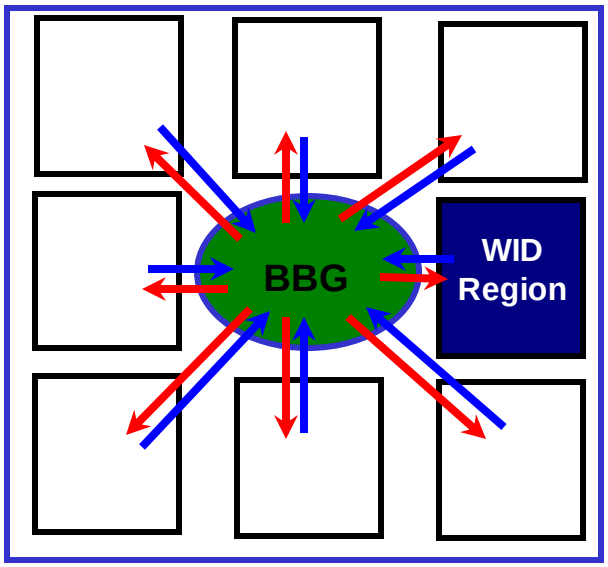


32nm WID variation map quad-core processor

- D2D variations
  - ✓ 30-50% frequency and 5-10x leakage variations
- WID systematic spatial variations
  - ✓ ~88% more leakage power and ~28% frequency variations

Mitigation of Process Variations

[1]



BBG: Body Bias Generator

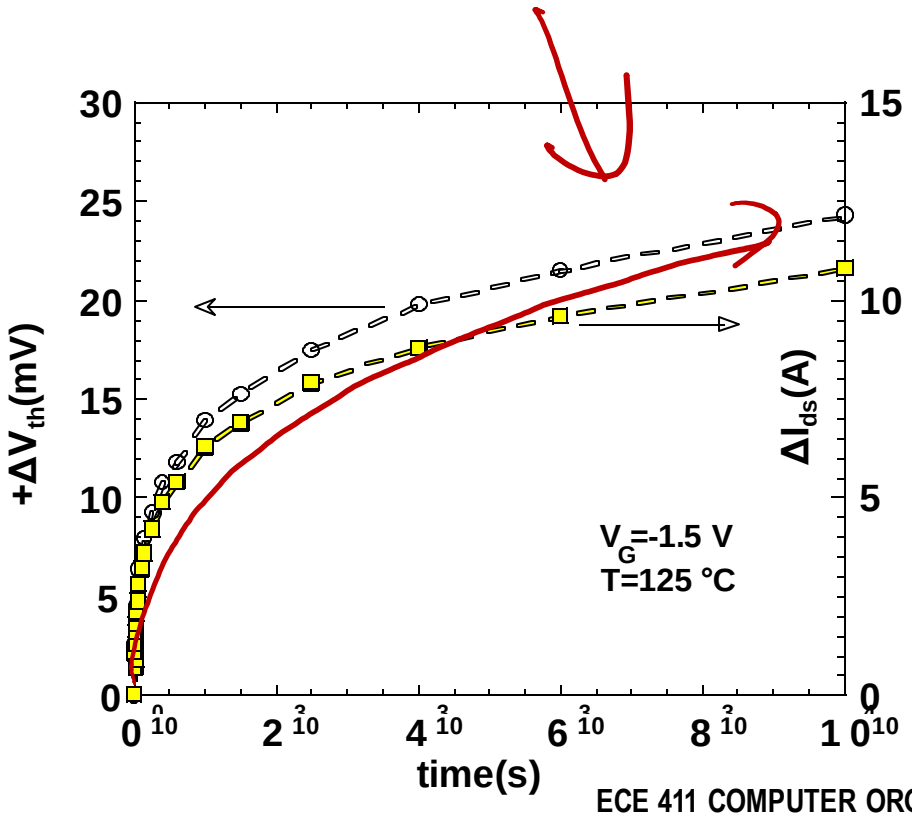
- adaptive body biasing (ABB)
  - requiring triple-well process
    - Apply ABB for both NMOS/PMOS
  - dual-V_{th} design: difficult to dealing w/ two different body coefficients
  - effectiveness diminishes w/ technology scaling

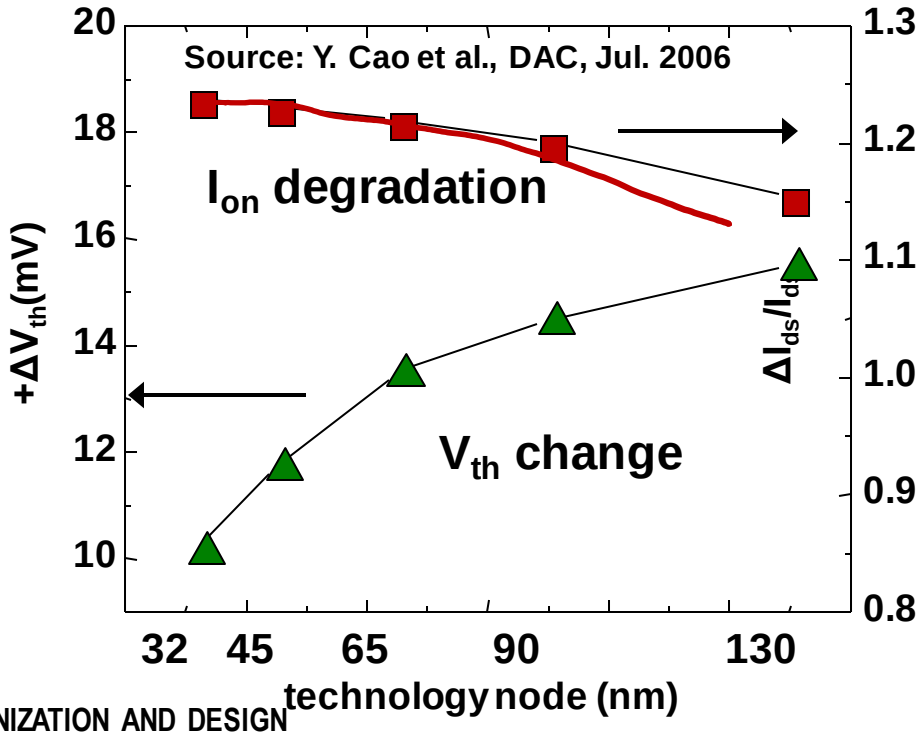
- adaptive voltage scaling (AVS)
  - ✓ difficult to apply fine-grain tuning for individual WID regions unless designs have multiple voltage domains
  - ✓ multiple voltage domains require
    - more voltage regulators and on-die decap issues
    - more design, verification, and testing time due to domain crossing issues

[1] S. Kumar et al., Mathematically Assisted Adaptive Body Bias (ABB) for Temperature Compensation in Gigascale LSI Systems, ASP-DAC06.

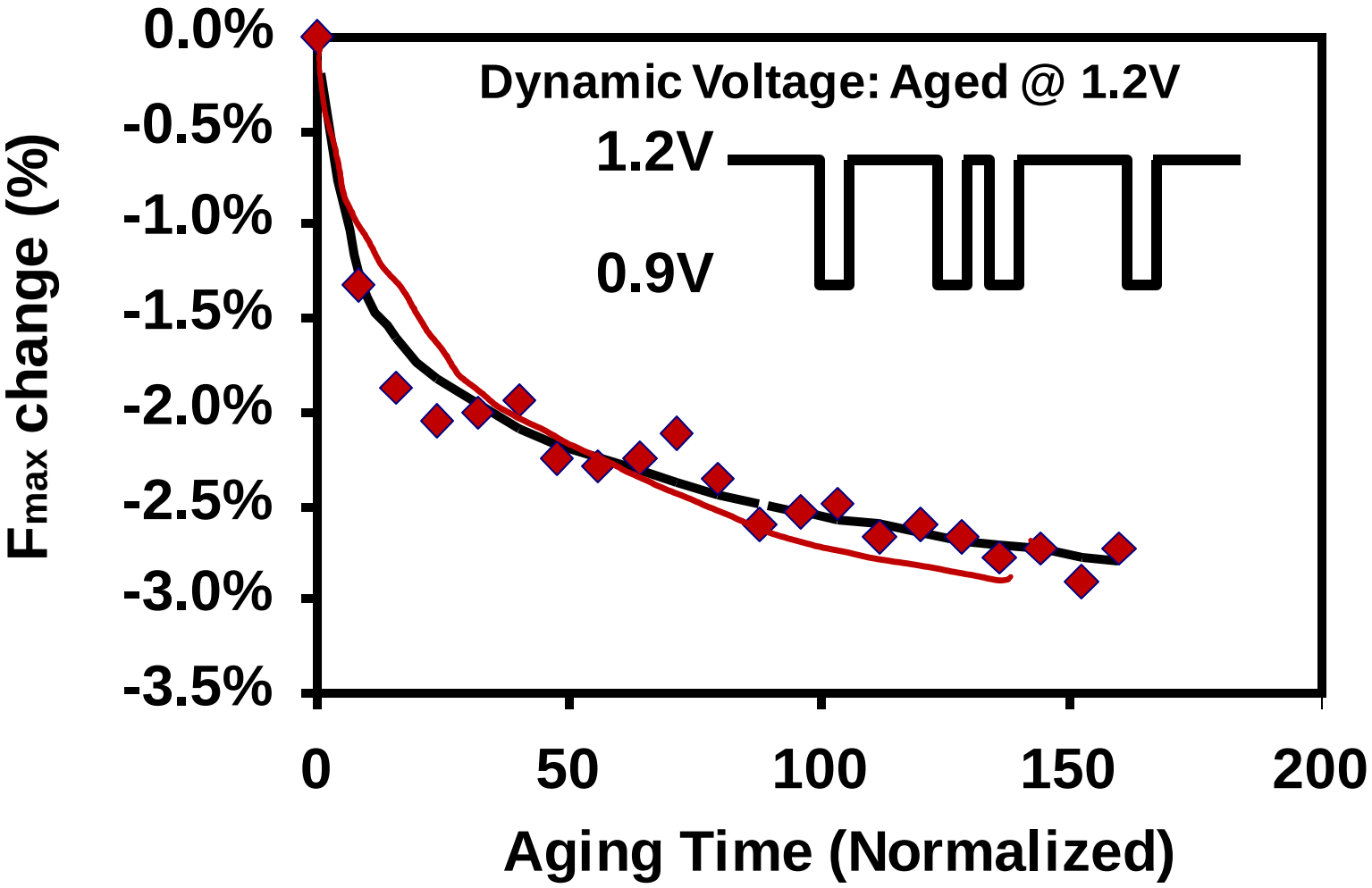
Reliability – Device Aging

- voltage stress over time generates interface traps resulting in device V_t shift and I_{ds} reduction (NBTI).
- technology scaling increasing aging susceptibility showing more $\Delta I_{ds}/I_{ds}$ degradation.





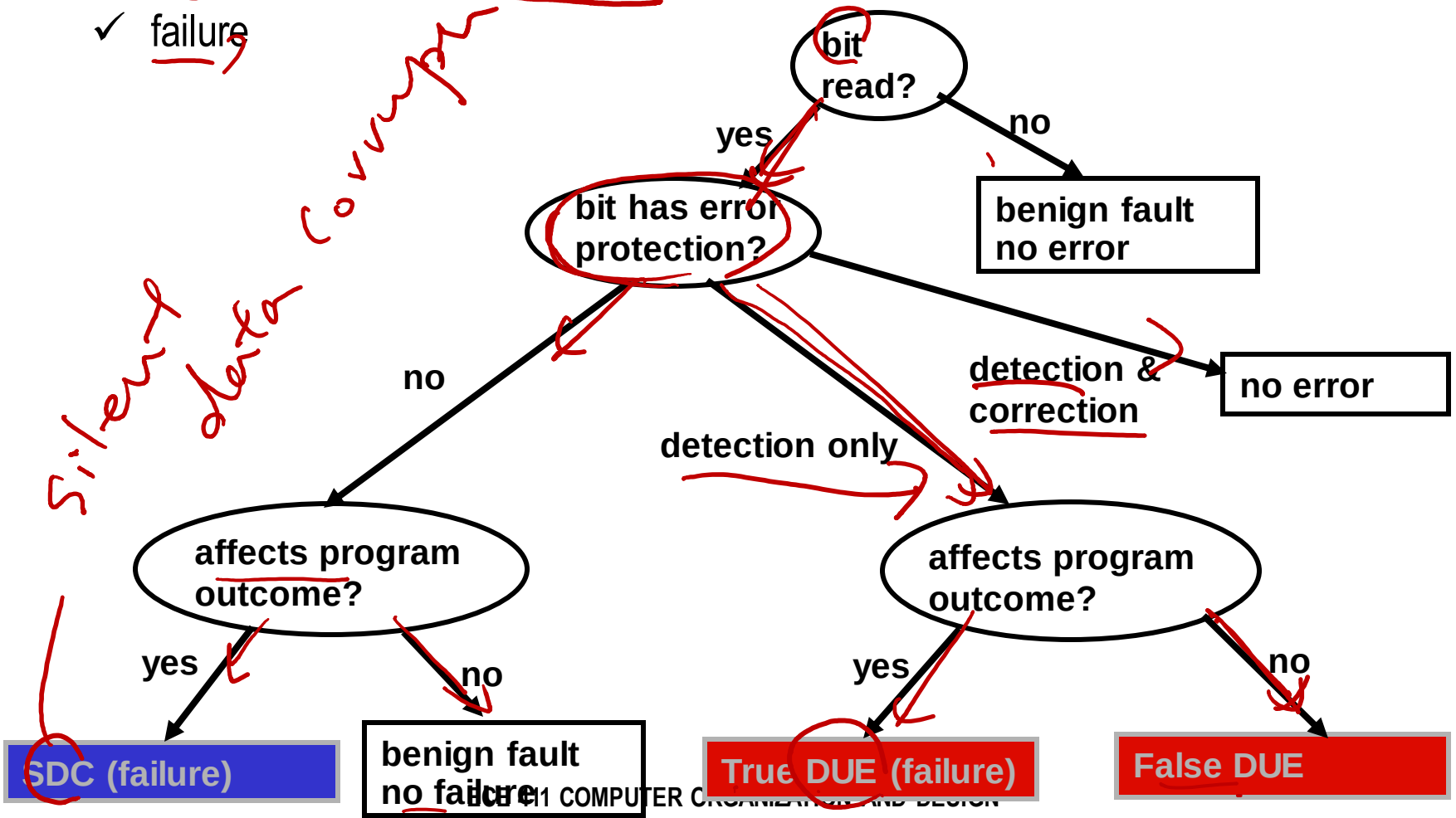
Reliability – Device Aging



- ✓ losing ~10% potential performance at the EOL
- ✓ no way to detect actual Fmax loss b/w BOL and EOL

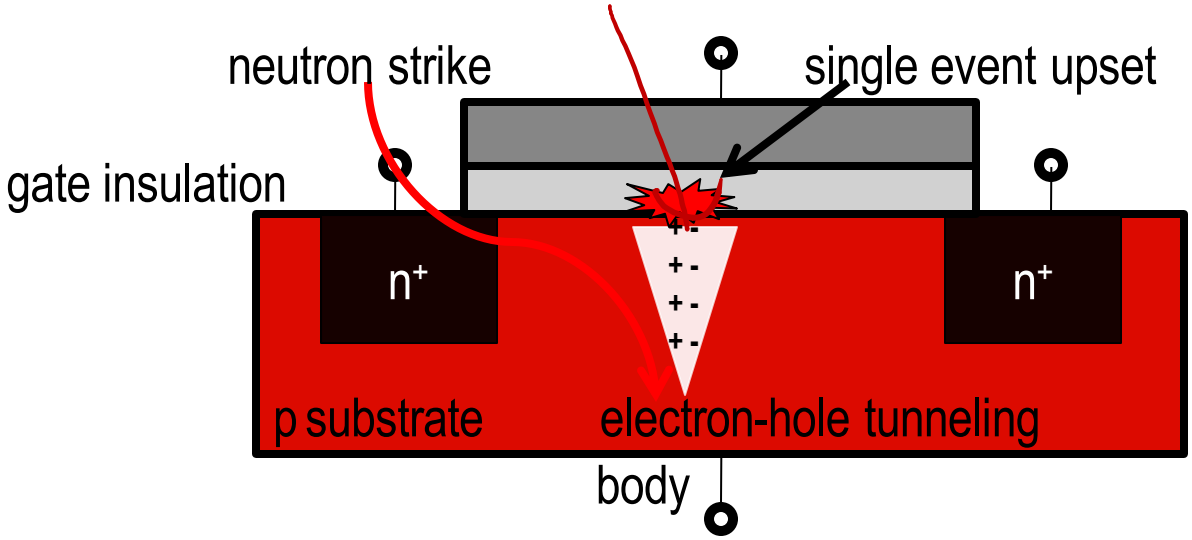
Reliability: Fault vs Error

- when a fault propagates outside the context (scope)
  - ✓ an error
- when an error affects correct execution
  - ✓ failure



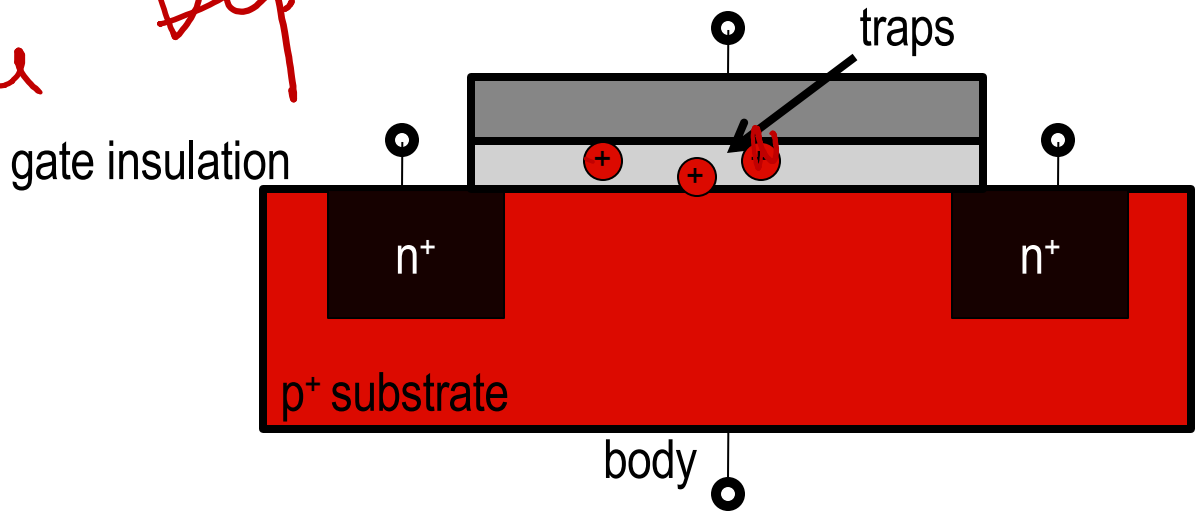
Transient Faults: SEUs

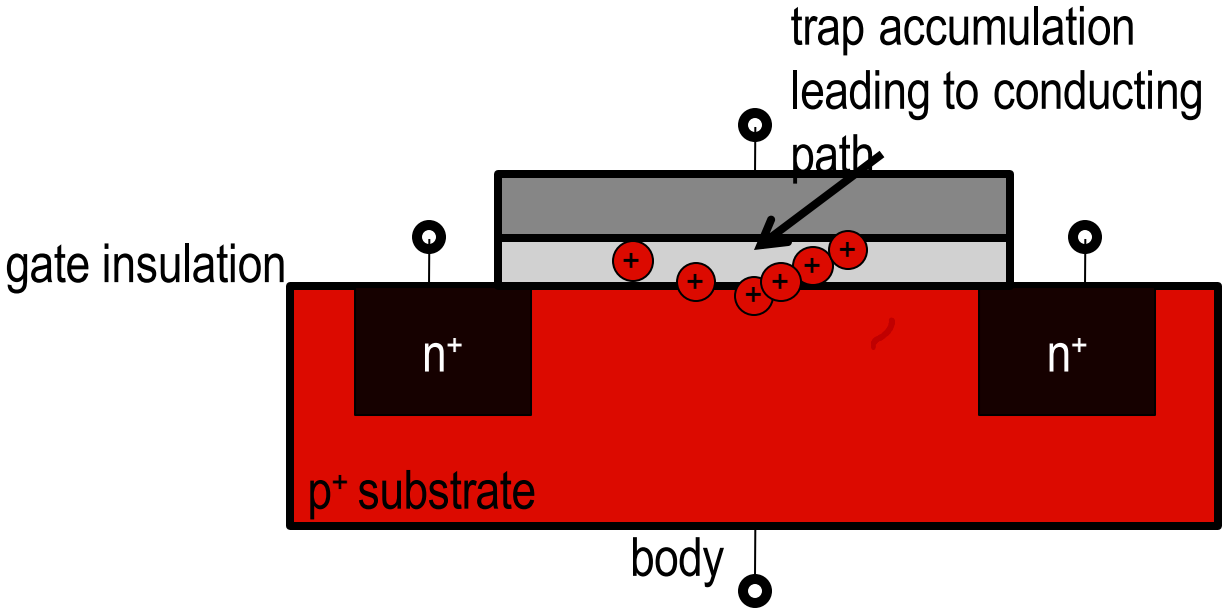
- high energy neutron strike
  - ✓ creates electron-hole pairs by splitting silicon nucleus
    - ○ charges from the pairs travels toward gate diffusion region
  - ✓ causes the transistor charge to flip
  - ✓ causes a bit to flip
  - ✓ both values 0 or 1 can be flipped (depending on holes or electron interactions)



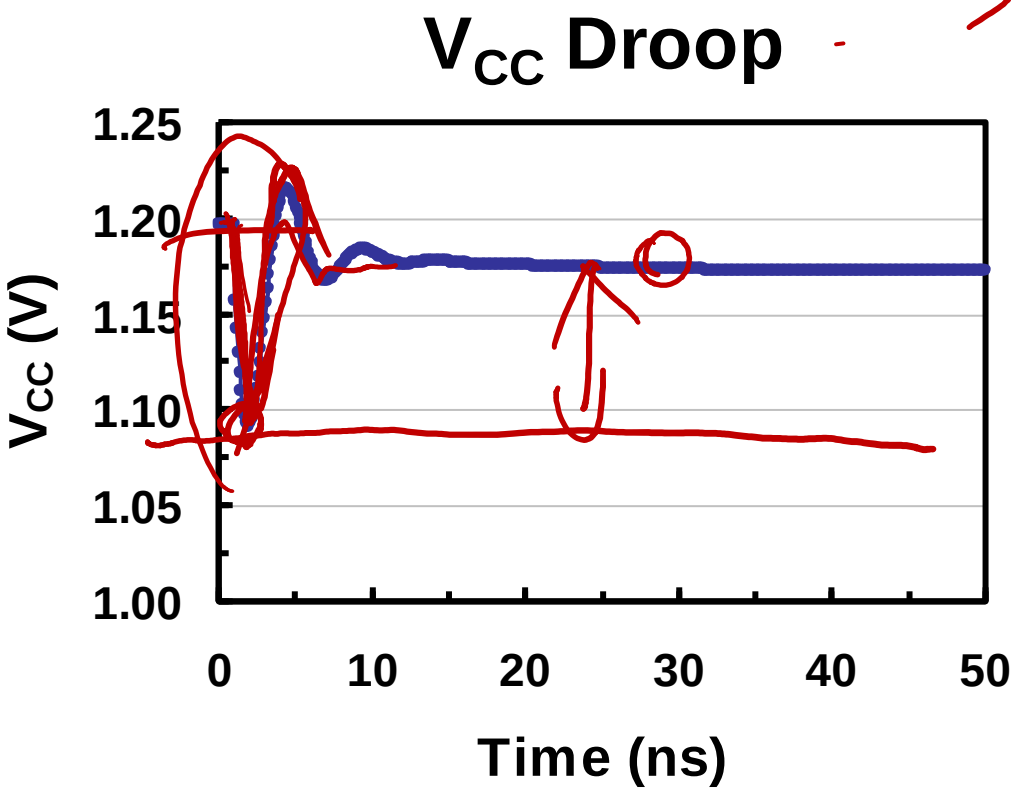
TDDB

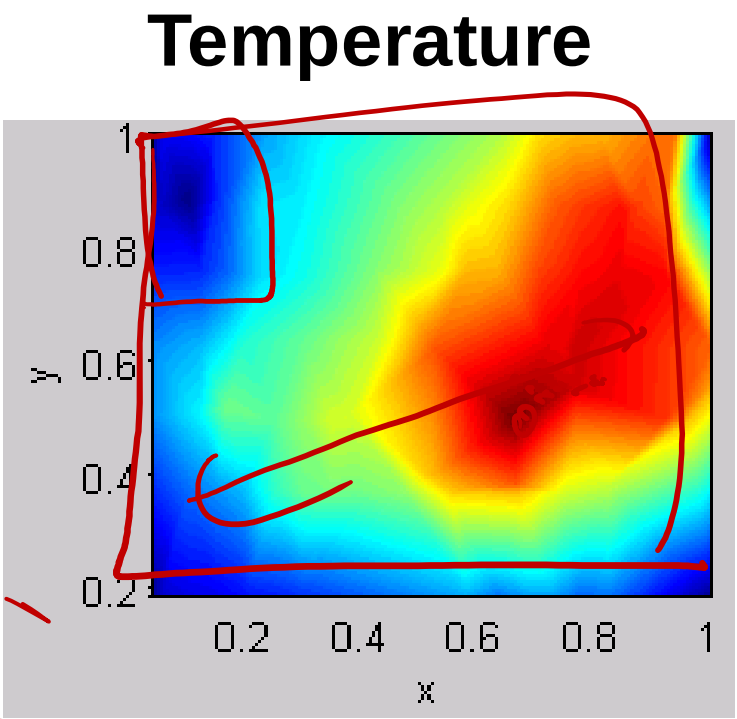
Time Dependent Dielectric Breakdown





Impact of Dynamic Variations

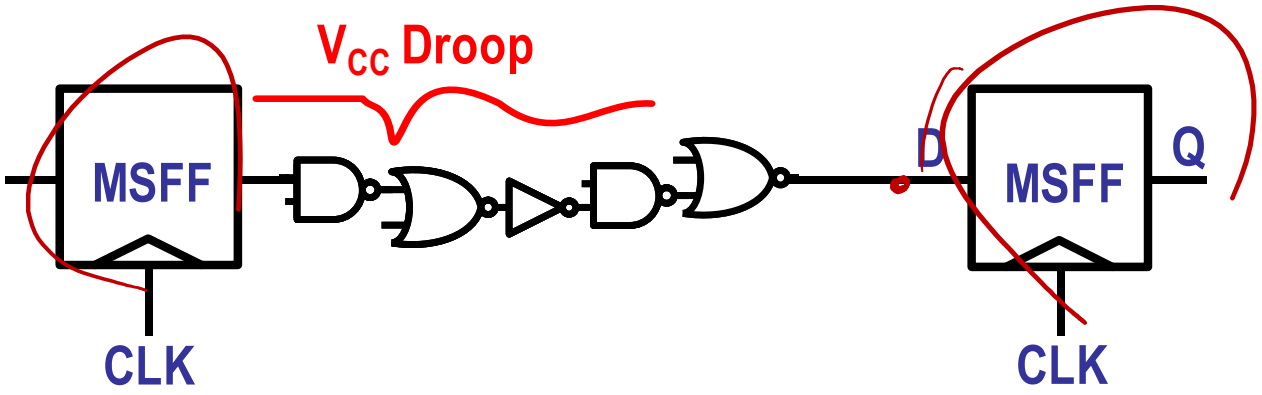


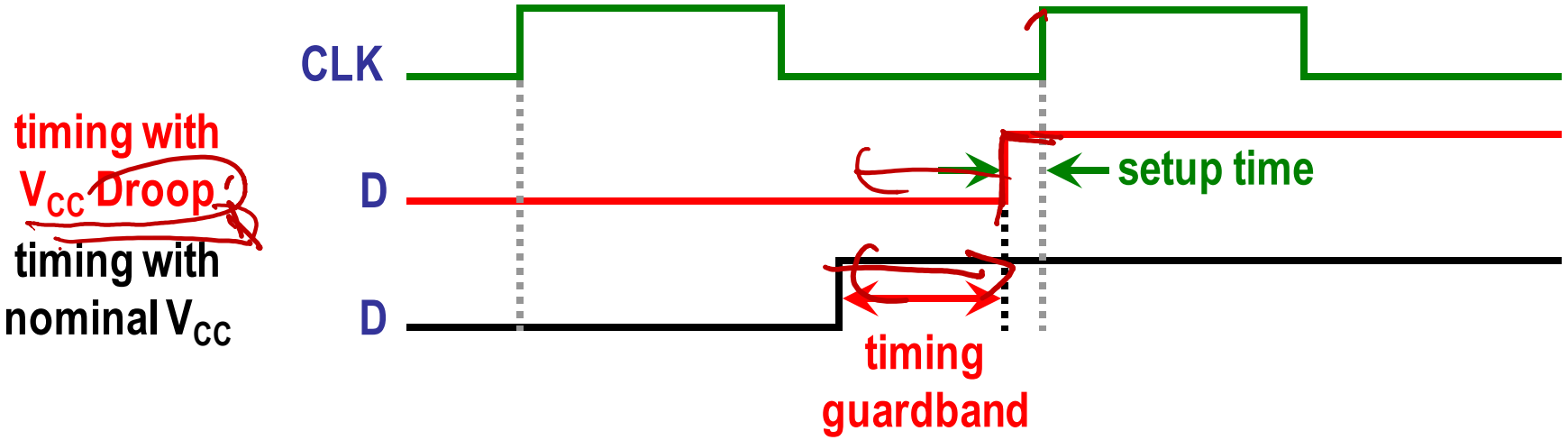


- microprocessor clock frequency (F_{CLK}) based on maximum V_{CC} droop & temperature specifications
- infrequent dynamic variations severely limit F_{CLK}

Variation Tolerant Techniques

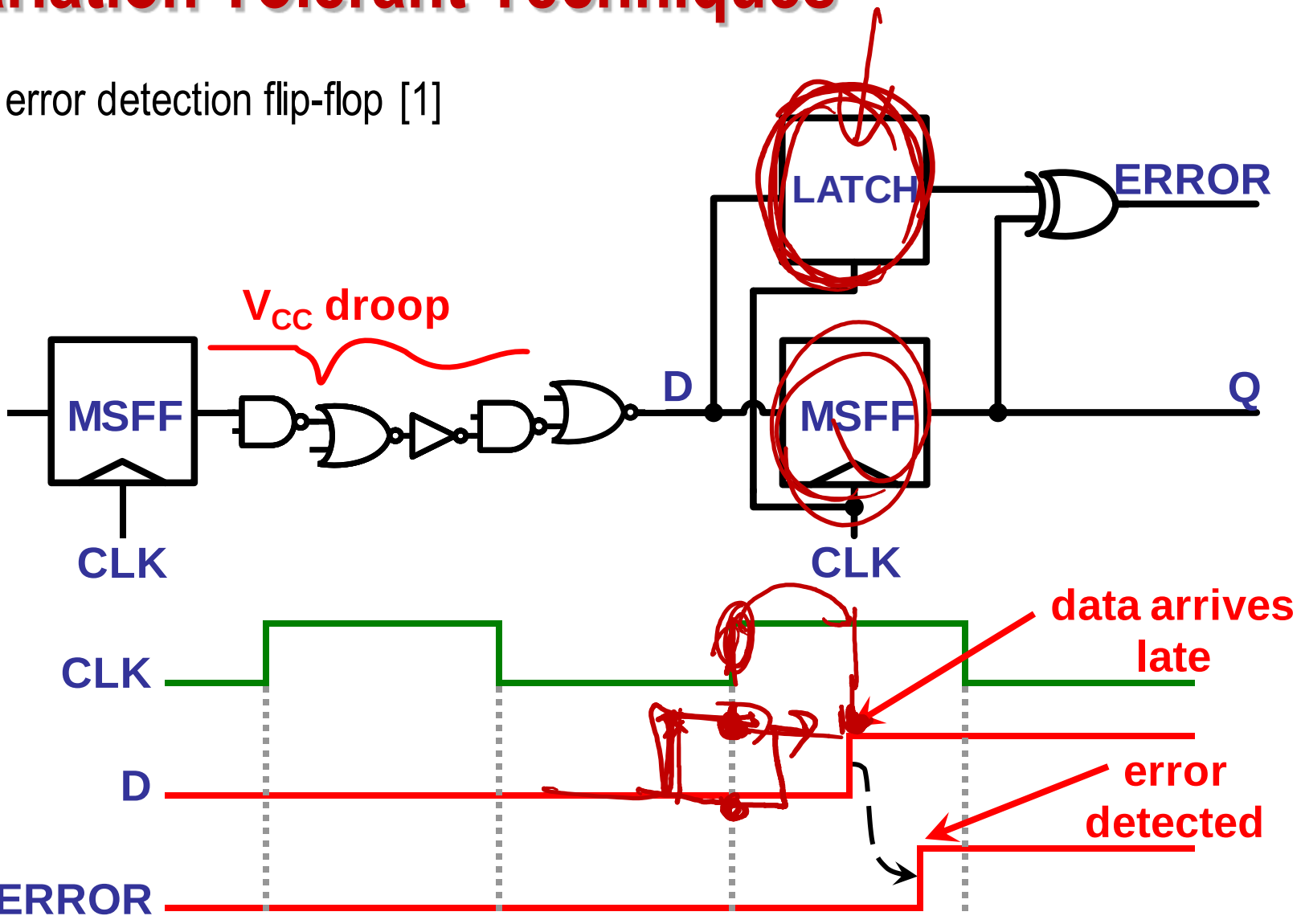
- master-slave flip-flop





Variation Tolerant Techniques

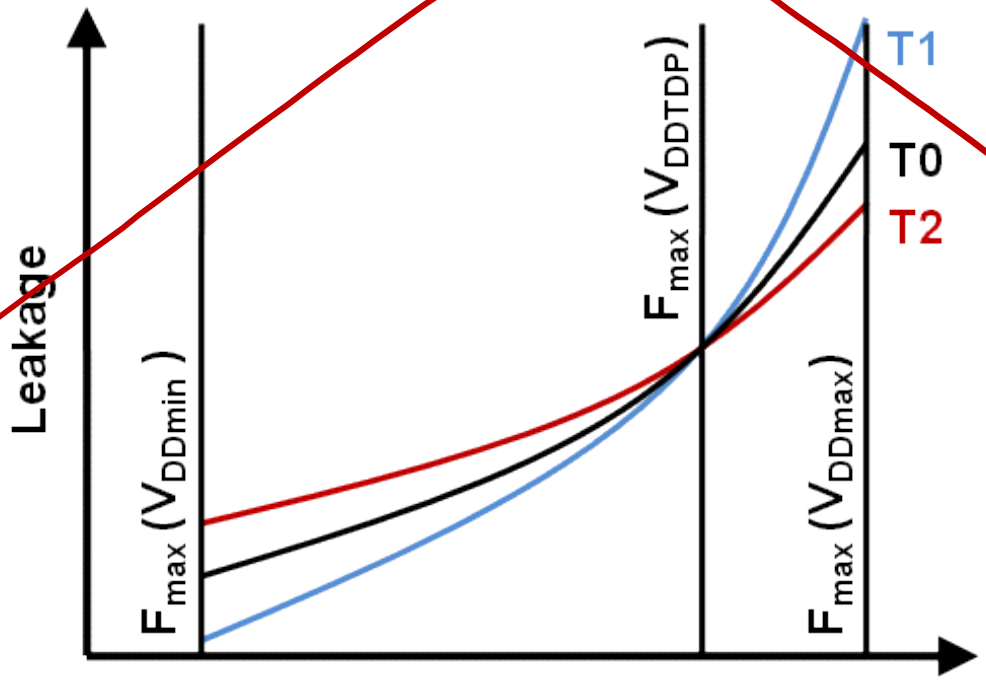
- error detection flip-flop [1]



[1] D. Ernst, N. Kim, et al., "Razor: A Low-Power Pipeline Based on Circuit-Level Timing Speculation", *MICRO*, 2003.

Tech. Tuning: Choices

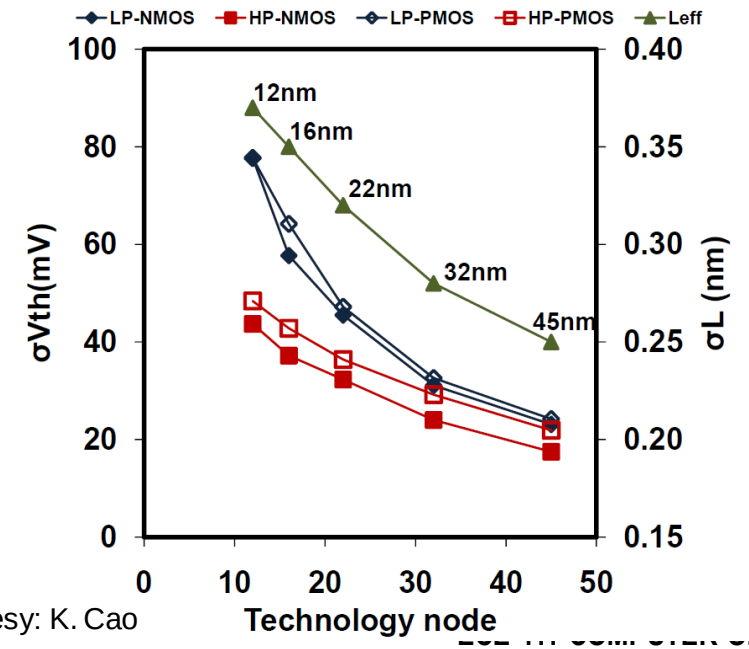
- key objectives:
  - ✓ maintaining the same max performance (F_{max}) and leakage power (P_{leak}) at V_{DDTDP}
  - ✓ minimizing average leakage energy (thus total energy) at various performance and power states



SRAM V_{DDMIN} Challenges(1)

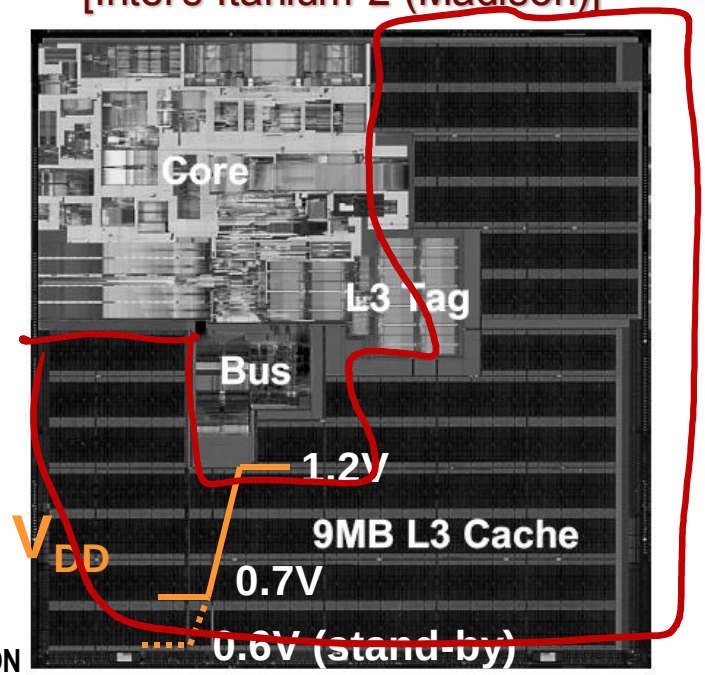
- dynamic V/F scaling (DVFS) for power efficient computing
  - ✓ Require SRAMs to operate at low V_{DD}
- increasing SRAM cell failures w/ tech. scaling due to growing
  - ✓ manufacturing process variations (ΔV_{TH} & ΔL)
  - ✓ demand for larger on-chip cache capacity (> 50% of die area)
    - ○ more cells, higher on-chip cache failure probability

[ITRS Projection for V_{th} and L_{eff} Vairations]



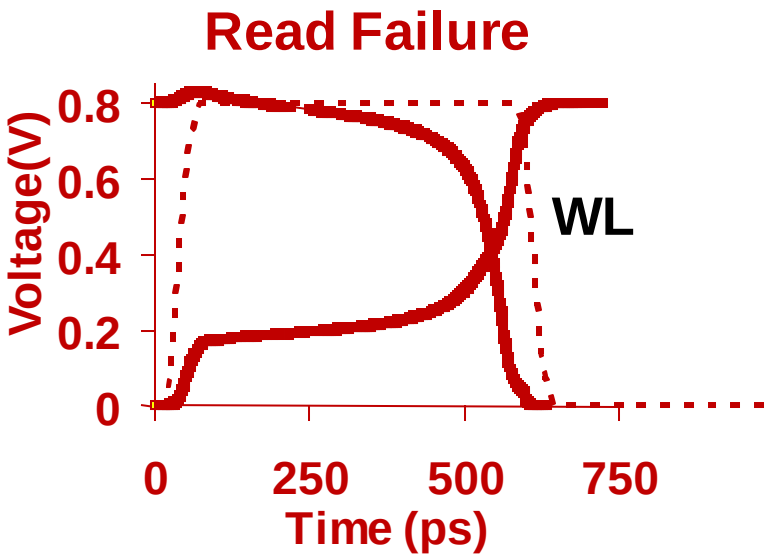
Courtesy: K. Cao

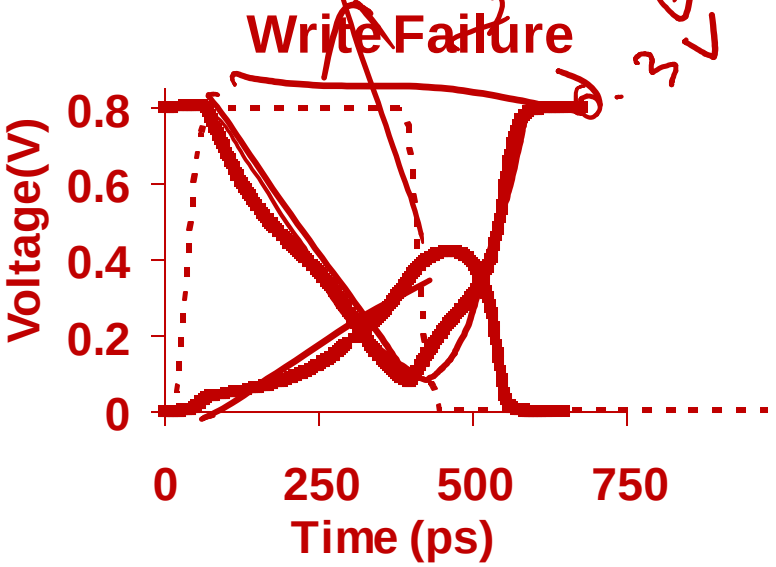
[Intel's Itanium 2 (Madison)]



V_{DDMIN} Challenges(2)

- V_{DDMIN}: lowest voltage satisfying a target yield



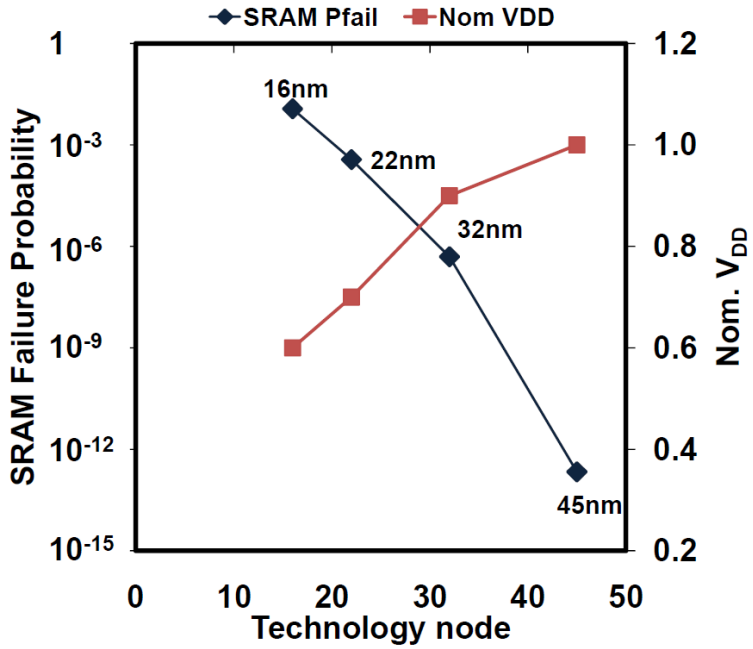


- V_{DDMIN} in conventional 6T cell heavily impacted by:
  - ✓ local (WID) random variations (ΔV_{th})
- SRAM w/ FinFET devices
  - ✓ less ΔV_{th} but harder to tune transistor sizes for V_{DDMIN} due to quantization effect

SRAM V_{DDMIN} Challenges(2)

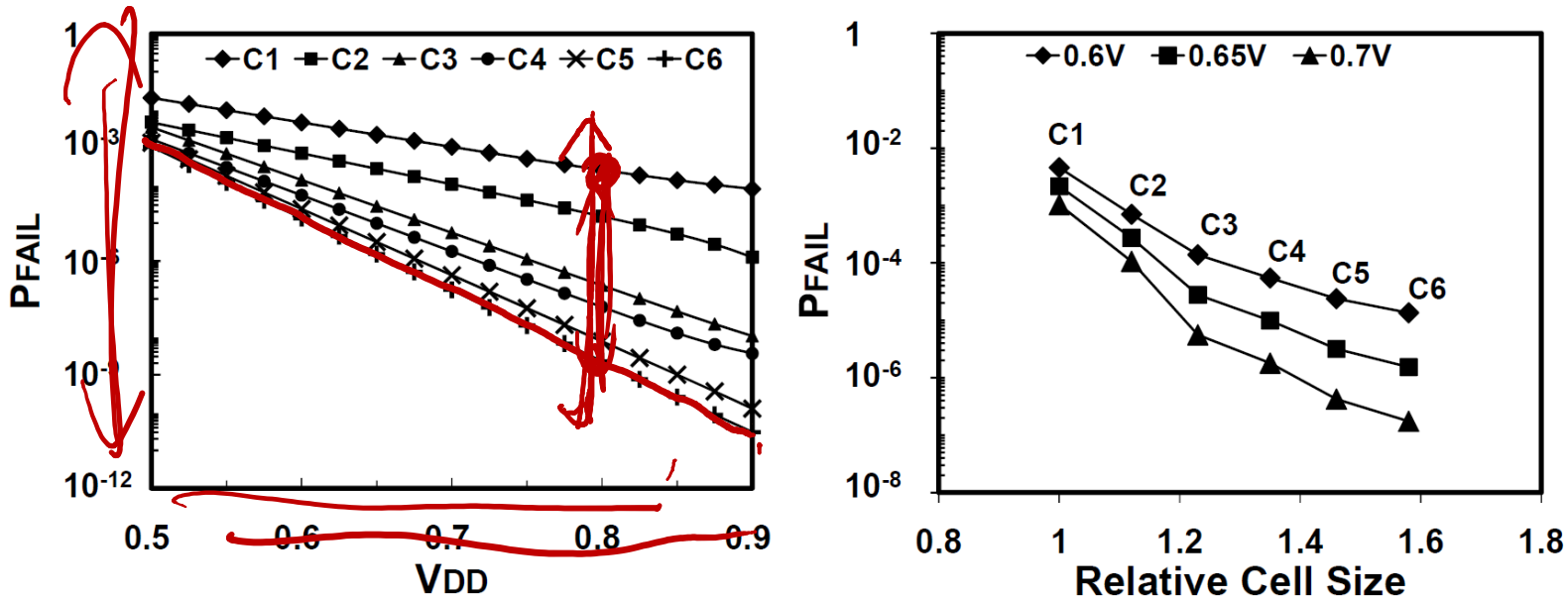
- shrinking voltage scaling window w/ technology scaling!
  - ✓ decreasing nominal V_{DD}
  - ✓ increasing V_{DDMIN}

[V_{DD} and SRAM P_{Fail} vs technology]



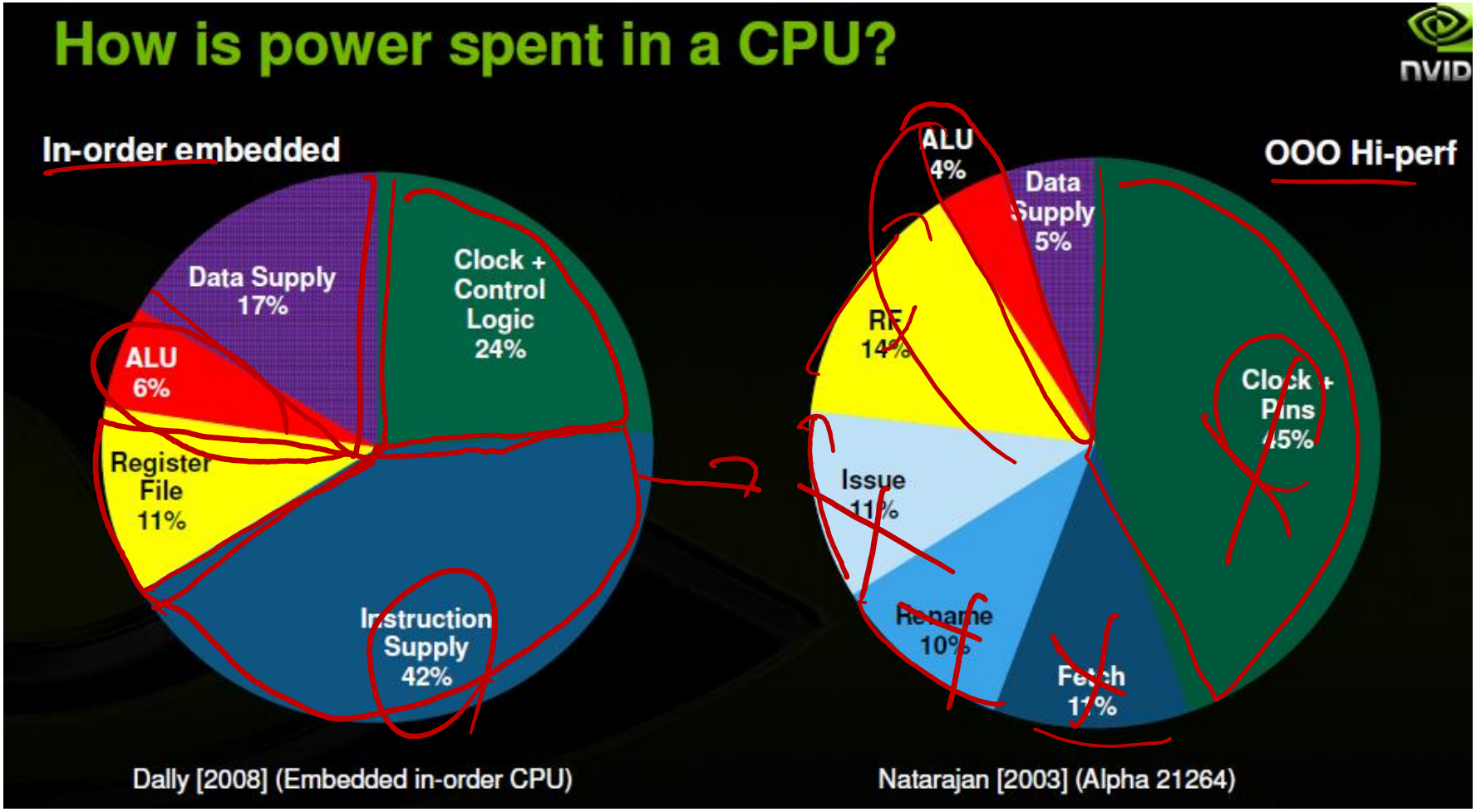
- current approaches
  - ✓ circuit-level techniques
    - 8T/10T SRAM cells [27-30]
    - read/write assisting [12, 13]
  - ✓ architecture-level techniques
    - cache-line disabling [3,4]
    - ECC [8,15]
  - ✓ either circuit or architecture-level solution alone will not be effective!
- circuit and architecture co-design approach for cost- and performance-efficient on-chip caches

6T SRAM Cell Size vs P_{FAIL}



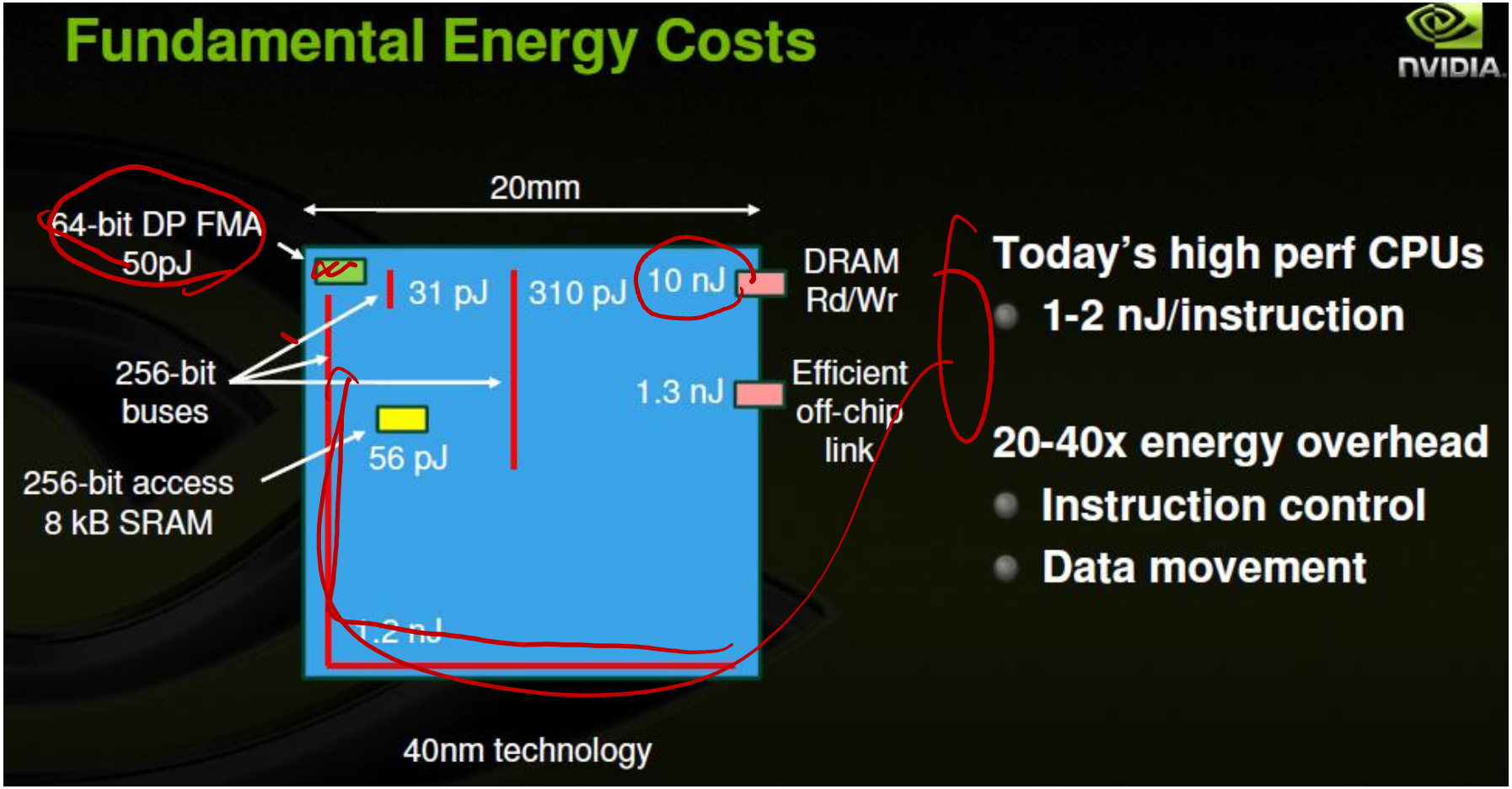
- random V_{th} and L_{eff} variations
  - ✓ main source of SRAM failure at low V_{DD}[1]:
  - ✓ $\sigma V_{th} \propto \frac{1}{\sqrt{L \times W}}$ [11]
- larger cell size (i.e., bigger transistor size)
  - ✓ less variations and lower P_{FAIL}, but …
  - ✓ not affordable for large size on-chip LLCs

Data Transfer Energy



Keckler Micro Keynote talk: “Life After Dennard and How I Learned to Love the Picojoule”

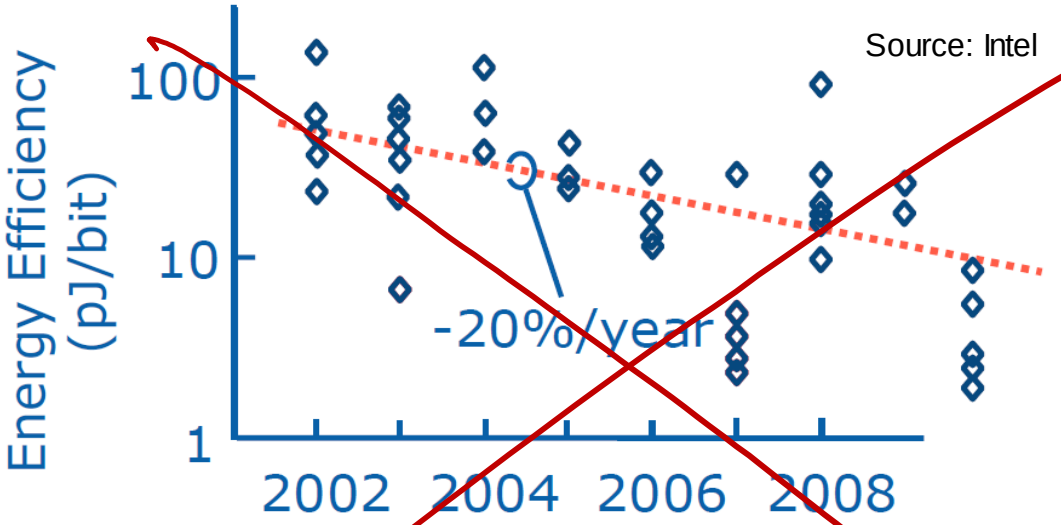
Data Transfer Energy

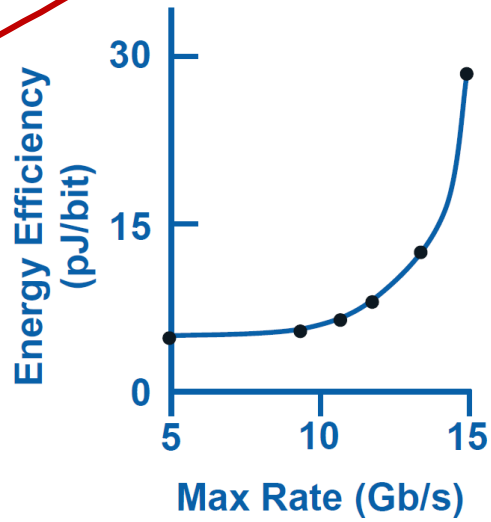


Keckler Micro Keynote talk: "Life After Dennard and How I Learned to Love the Picojoule"

Memory I/O Interface Power

- energy efficiency (pJ/bit) improving
  - ✓ driven by architecture, circuit, and process improvements

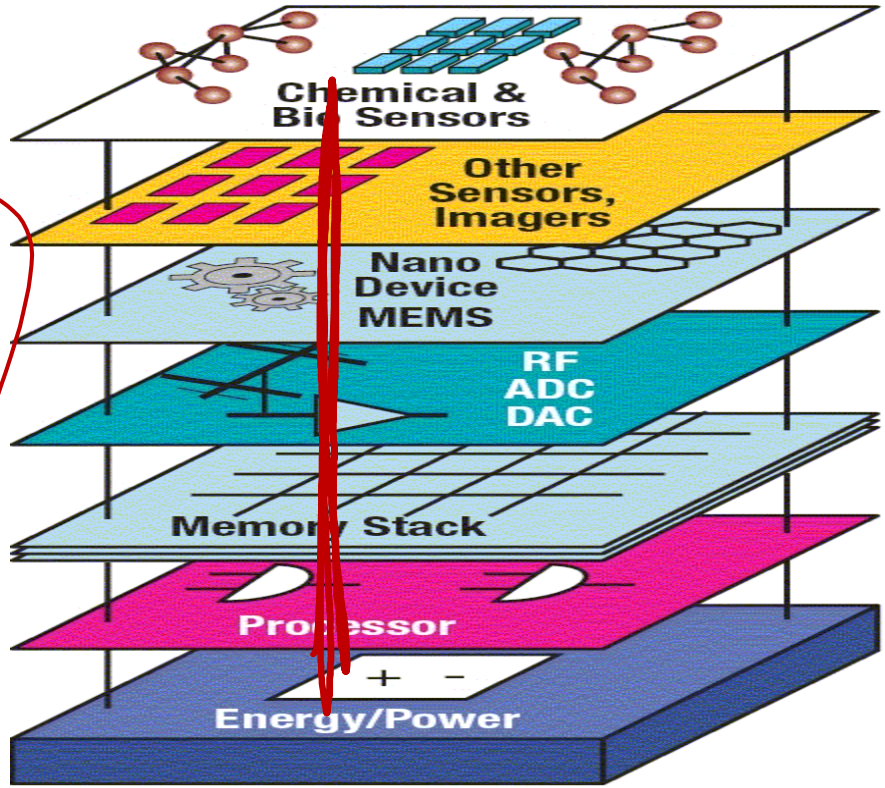




  - ✓ I/O power kept at a low level
- but in future …
  - ✓ super-linear degradation in energy efficiency
  - ✓ over-extending channel speed requires power-hungry I/O circuitry
  - ✓ 100 GB/s × 20 pJ/bit = 16 W !

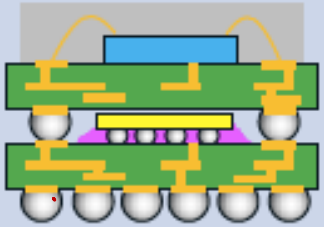
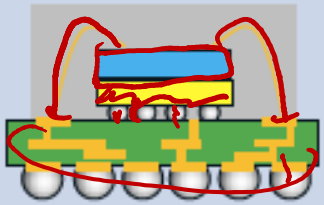
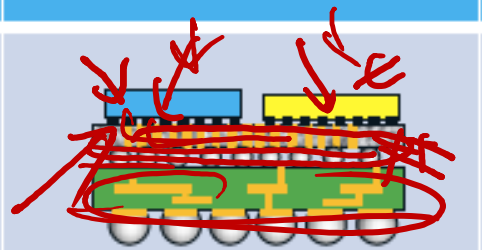
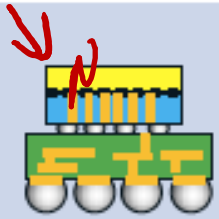
Heterogeneous Integration Technologies

- 3D-integration
  - ✓ 3D-integration enables far more than an alternative for increased integration and provides another dimension of design flexibility.
  - ✓ a well-known aspect of this flexibility is the ability to split the design into layers which could be processed and operated independently, and still be tightly interconnected



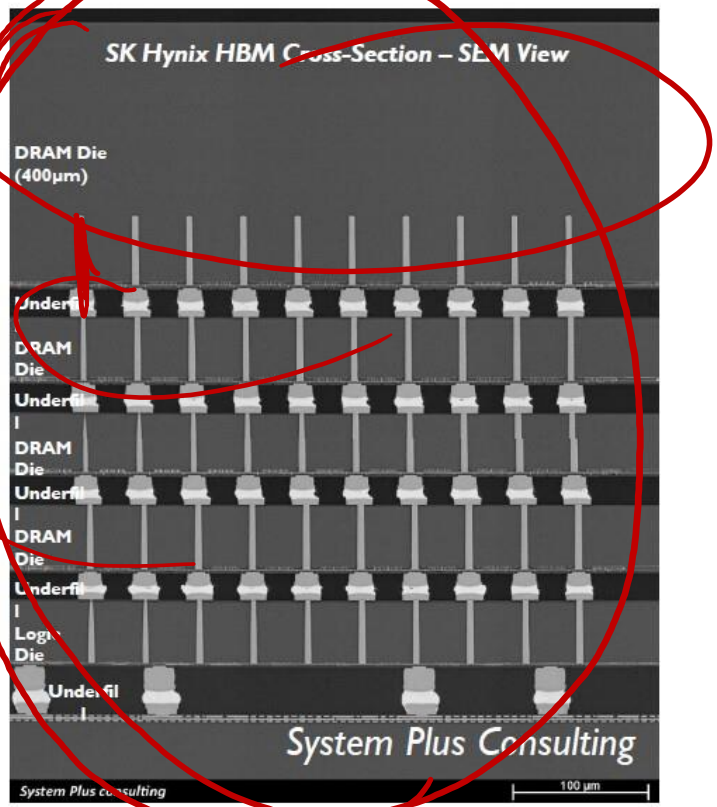
Heterogeneous Integration Technologies

- 3D and 2.5D-die stacking
  - ✓ https://en.wikipedia.org/wiki/Three-dimensional_integrated_circuit

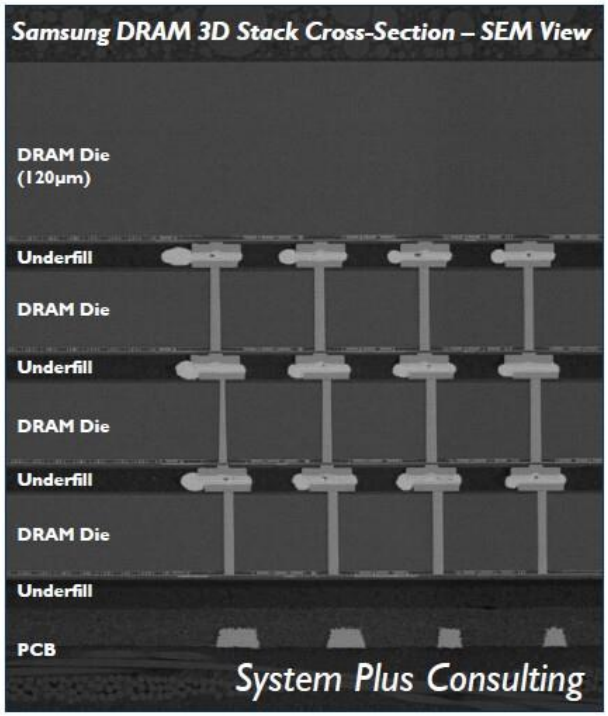
Package Stacking	Traditional Stacking	2.5D Interposer	3D TSV Stack
			
Package	Flip Chip & Wire Bond	2.5D side-by-side integration	Vertical Stacking

Heterogeneous Integration Technologies

- 3D stacking
  - ✓ https://www.microarch.org/micro46/files/keynote1.pdf



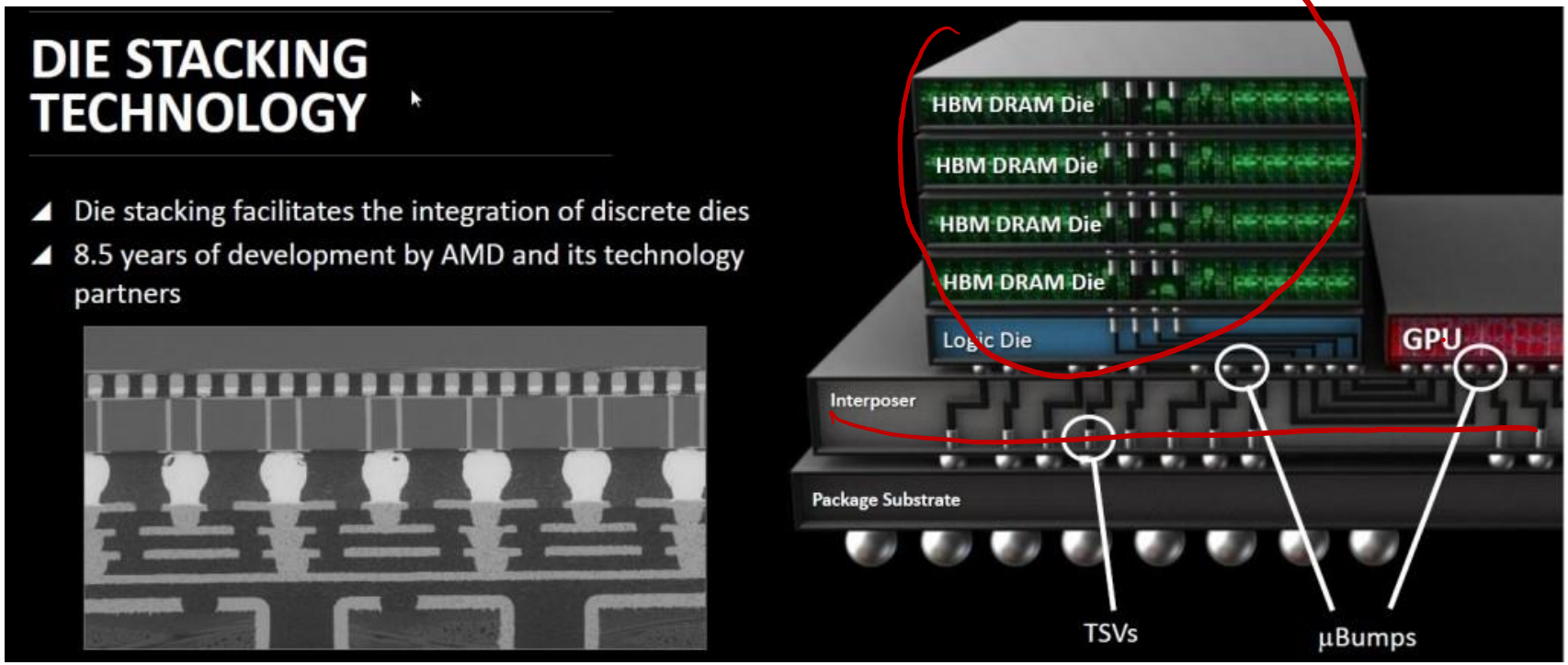
- Top die: 400µm
- TSV included in top die



- Top die: 120µm
- TSV not included in top die

Heterogeneous Integration Technologies

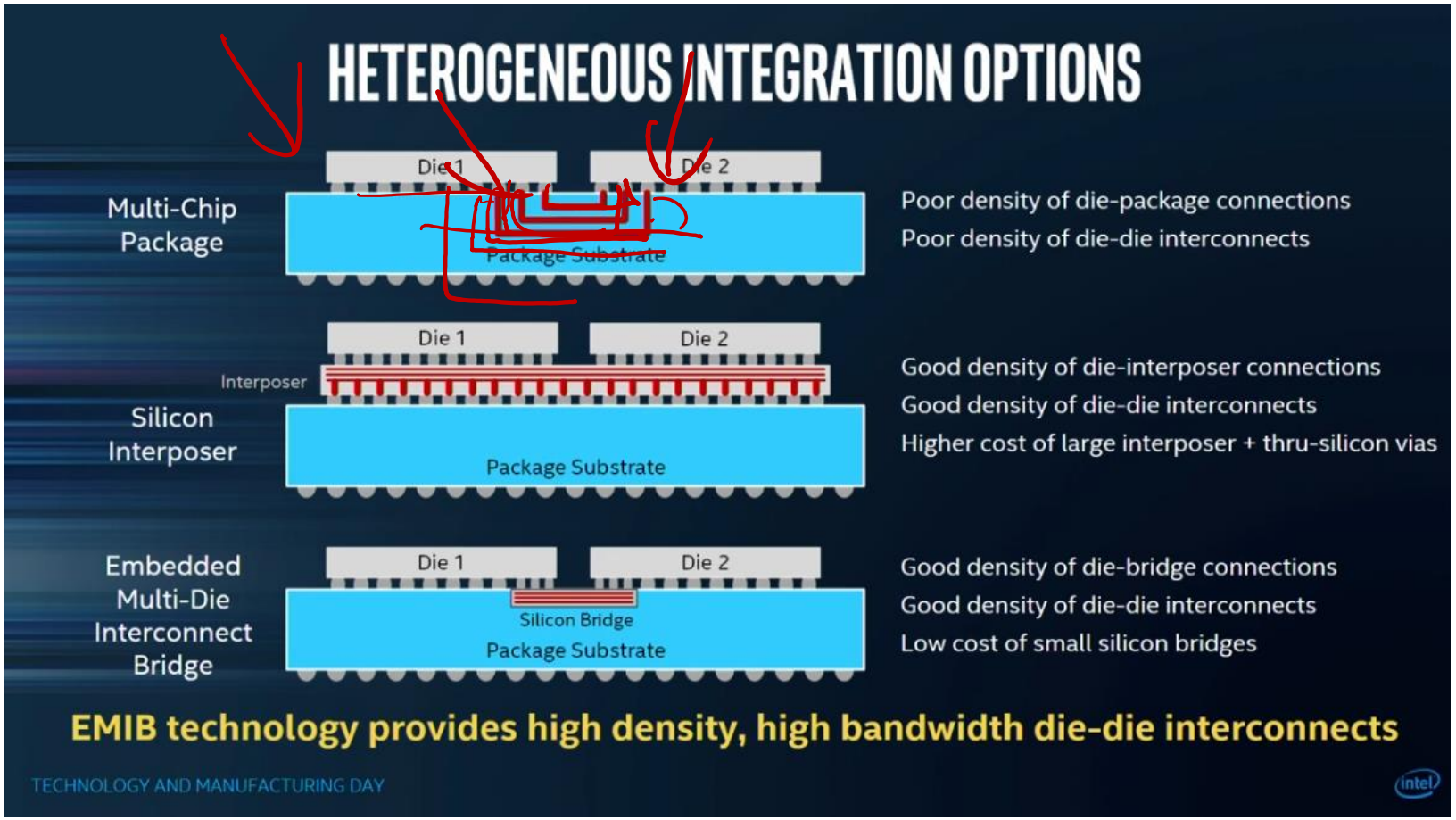
- 2.5D-die stacking
  - ✓ http://electroiq.com/blog/2011/06/silicon-interposers-building-blocks-for-3d-ics/
  - ✓ An alternative approach to a full 3D-IC stack is to place active dies on a passive silicon interposer, which in turn is placed on the package substrate. Silicon interposers with TSVs offer a way for designers to achieve the benefits of chip-scale connected configurations, without having to confront the issues currently presented by a full 3D-IC implementation through active silicon. The use of a silicon interposer is often referred to as a 2.5D-IC.



Heterogeneous Integration Technologies

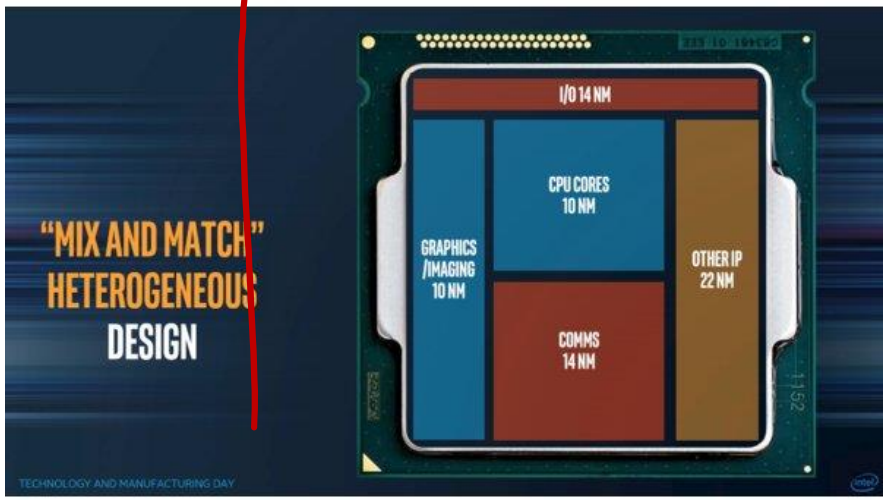
- package-level integration technologies

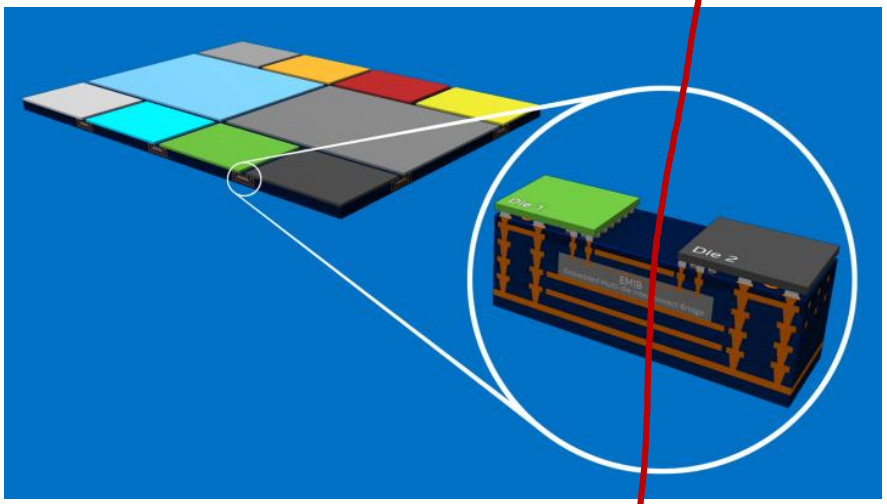
You can have a look at this later



Heterogeneous Integration Technologies

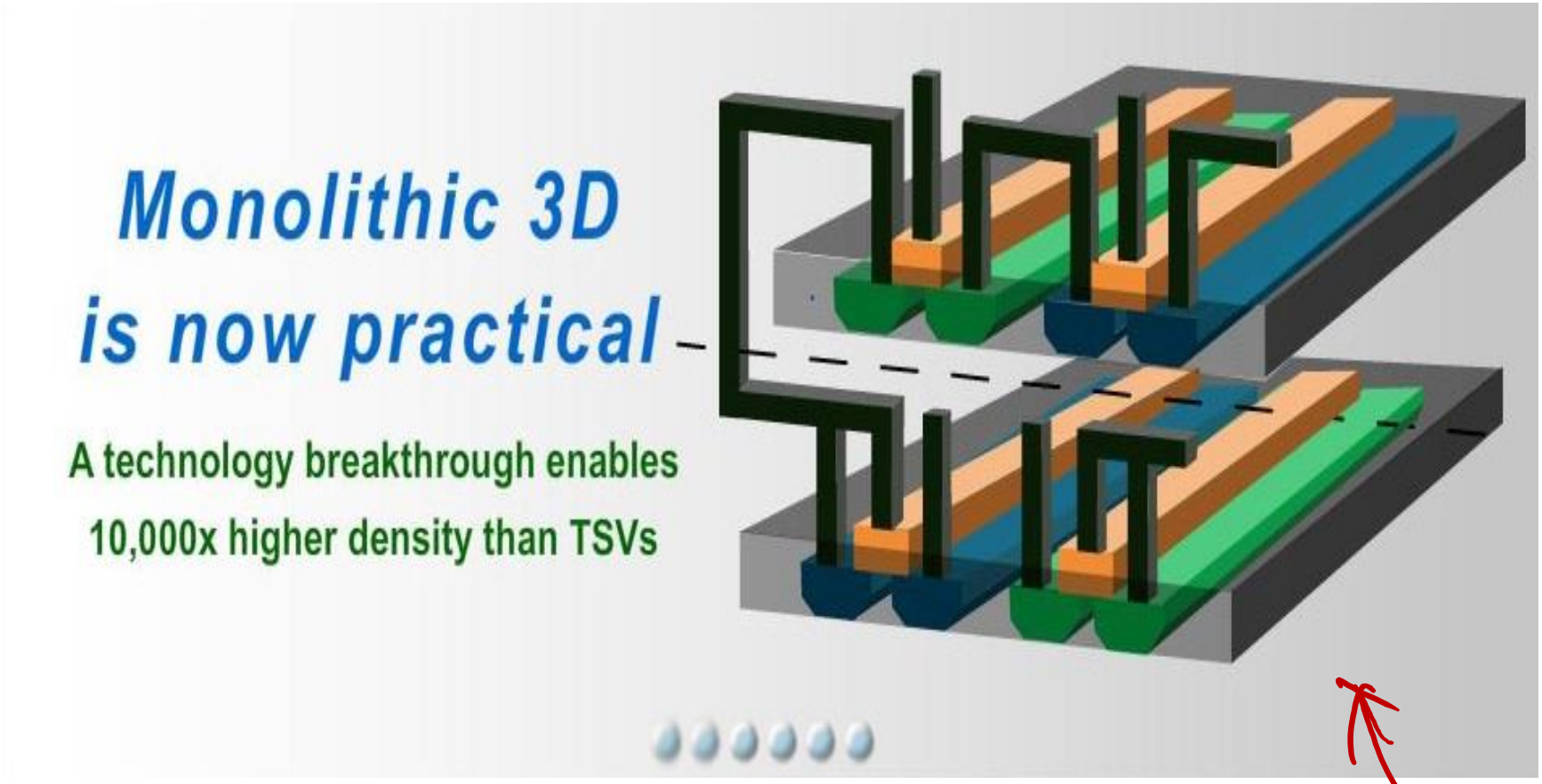
- Intel embedded multi-die interconnect bridge
  - ✓ https://www.intel.com/content/www/us/en/foundry/emib.html





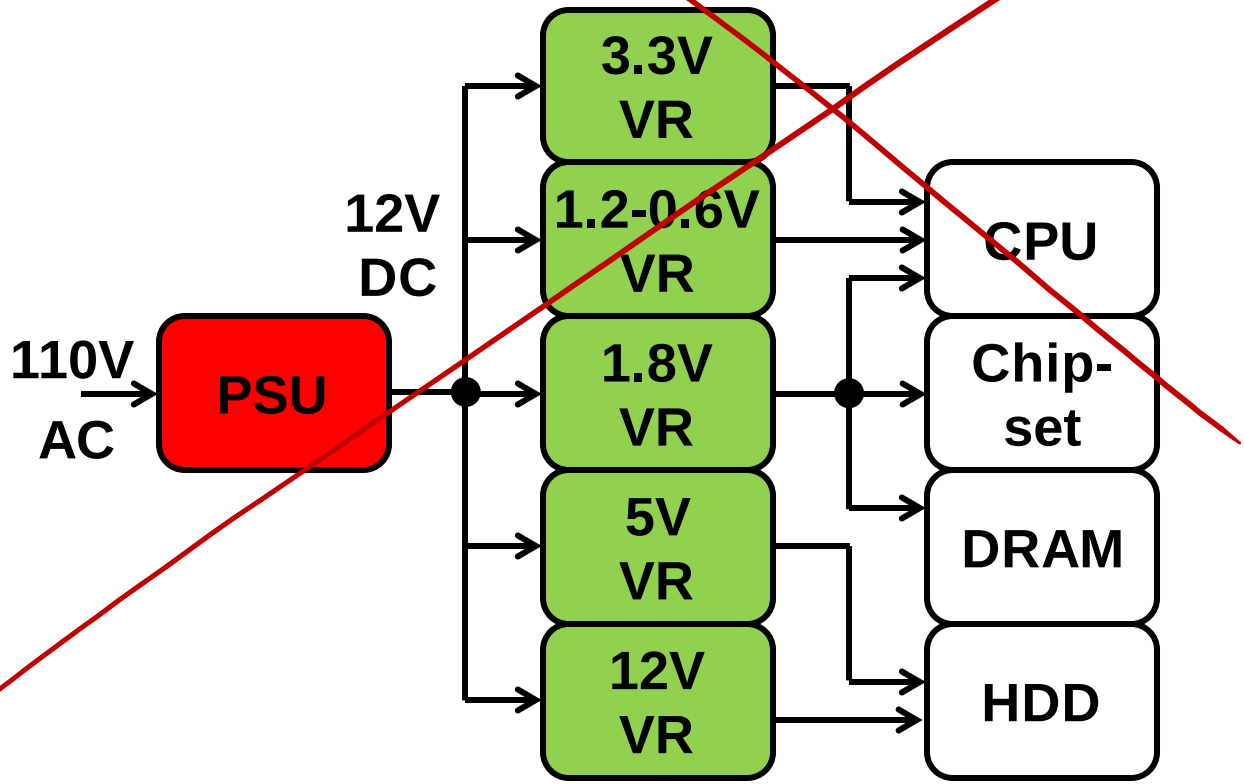
Heterogeneous Integration Technologies

- monolithic 3D integration



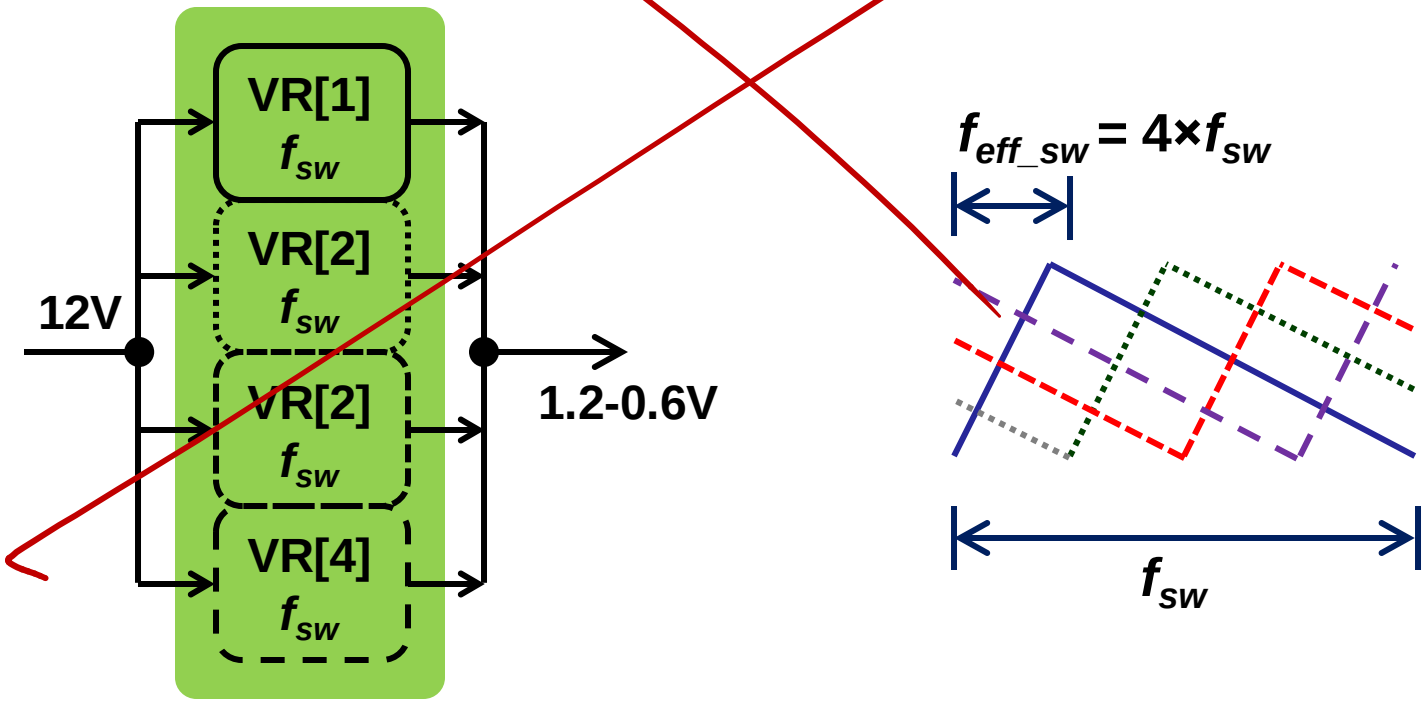
Platform Power Delivery Architecture

- hierarchical power delivery to maximize efficiency
  - 110V AC to 12V DC
  - 12V DC to various levels of DC using multiple VRs



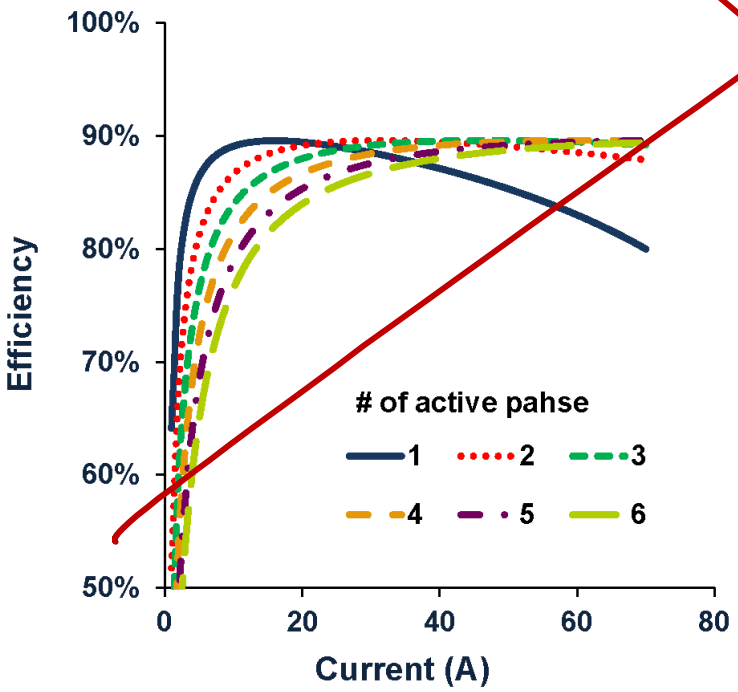
Multi-Phase VRs

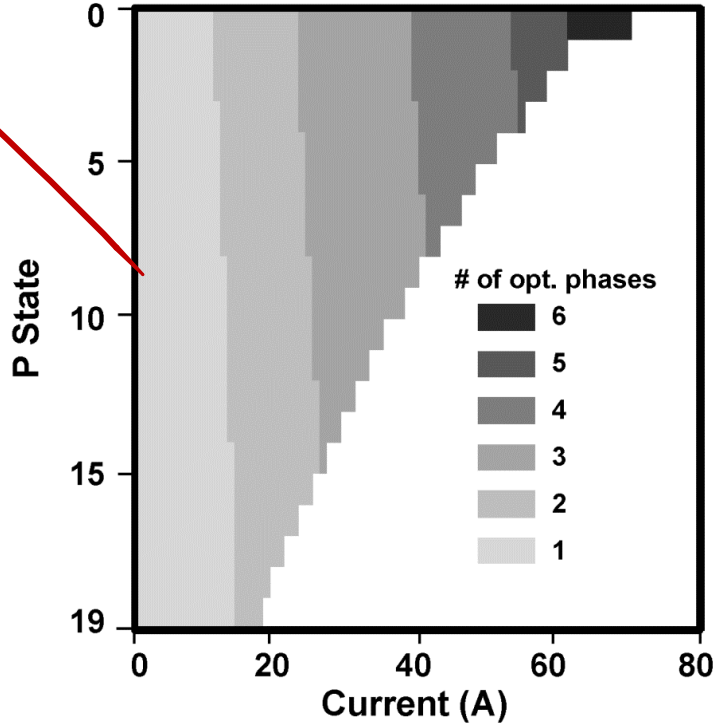
- comprised of n phases
  - each of n phases is turned on at $1/f_{sw}$ interval.
  - lowers switching loss
- high-performance processors use as many as 6-8 phases.



VR Efficiency vs Load Current

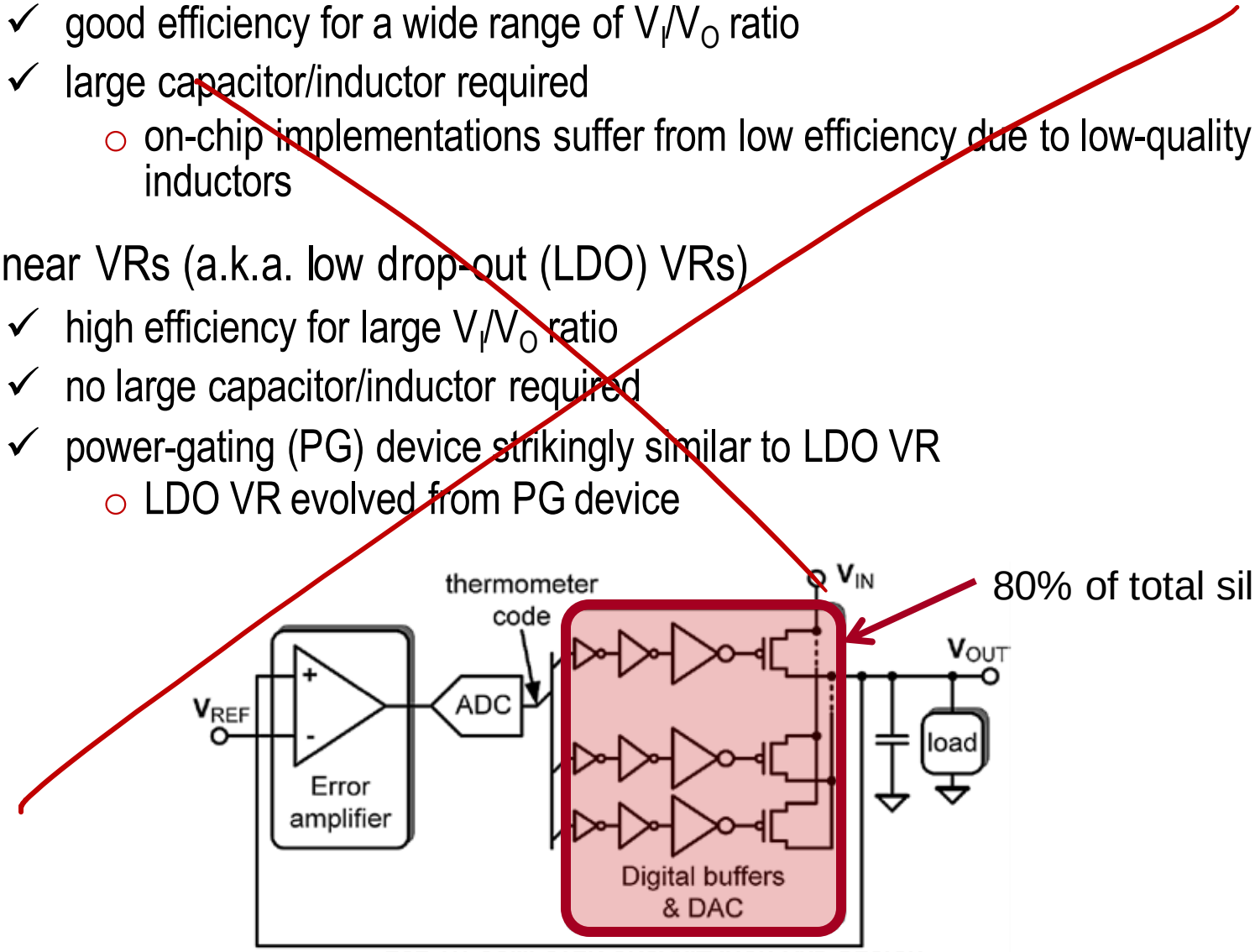
- function of number of active phases, current, and voltage
  - 5a load current 64% w/ 6 phases vs 86% w/ 1 phase at 1.2v
- optimal # of active phases
  - vary as a function of voltage (p state) and current
  - improve power efficiency 30-50% at low load current





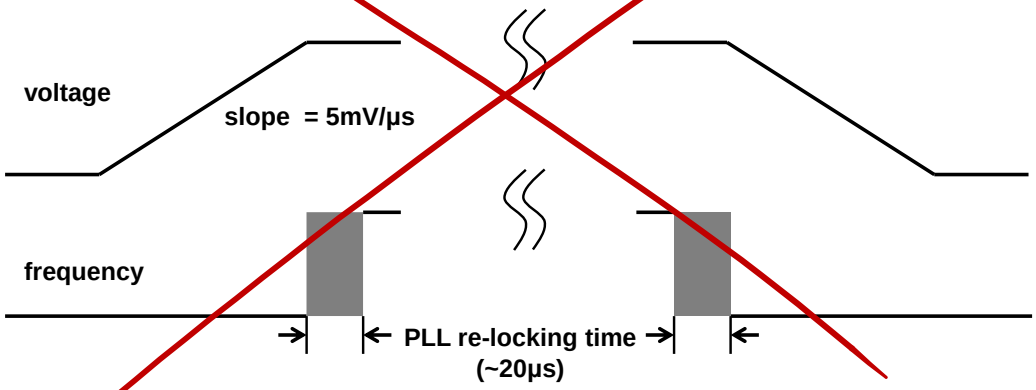
Linear vs. Switching VR

- switching VRs (a.k.a. buck converter)
  - ✓ good efficiency for a wide range of V_I/V_O ratio
  - ✓ large capacitor/inductor required
    - on-chip implementations suffer from low efficiency due to low-quality inductors
- linear VRs (a.k.a. low drop-out (LDO) VRs)
  - ✓ high efficiency for large V_I/V_O ratio
  - ✓ no large capacitor/inductor required
  - ✓ power-gating (PG) device strikingly similar to LDO VR
    - LDO VR evolved from PG device



Impact of Changing V/F

- V/F scaling
  - ✓ lowering V/F of cores executing spinning threads
    - ○ requiring per-core V domain
    - ○ perf. penalty associated w/ F change



Announcement

- next lectures on 4/24 and 4/26
  - ✓ project presentations
- final exam reviews on May 1
  - ✓ regular lecture hour: lectures 12 –20
  - ✓ evening special session at ECE 1002 between 7:15-8:45pm: lectures 1-11