IC Manufacturing

Ramon Canal
NCD - Master MIRI

Departament d'Arquitectura de Computadors
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Sand to silicon wafer

Sand (SiO_2) Quartzite → (Arc furnace) → Metallurgical grade Si ($ 2 /kg) 95~98% pure → (Fluidized bed reaction w/HCl and distillation) → $SiHCl_3$ ($ 20 /kg) → (Vapor deposition) → Electronic grade Si (poly-Si) (> $ 80 /kg) → (Crystal growth) → Pure Si crystal ($ 400 /kg) → (Wafering & polishing) → Polished Si wafer ($ 1300 /kg)

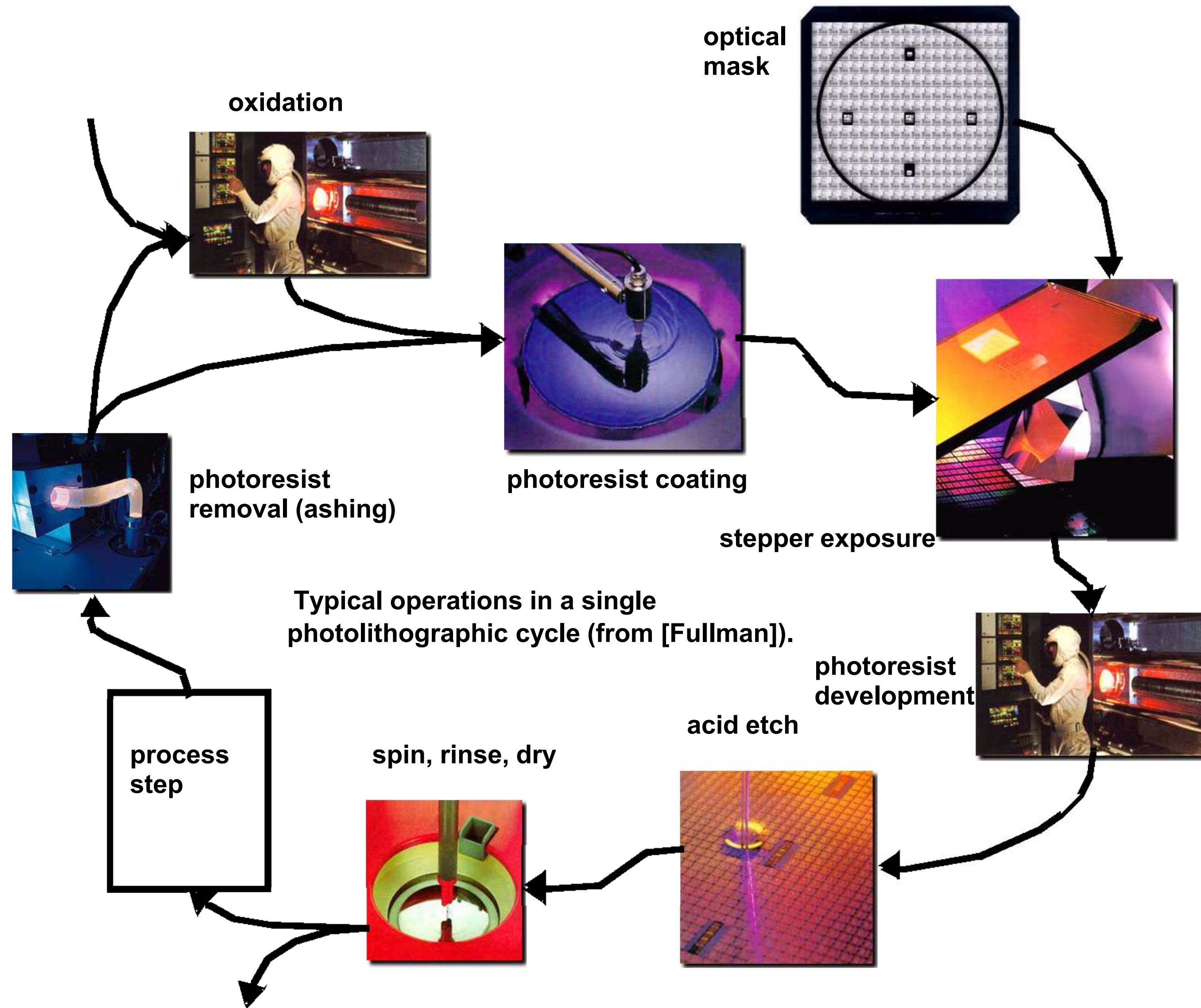
Single Crystal Silicon Ingot

Polysilicon Ingots

Source: http://www.fullman.com/semiconductors/semiconductors.html

UPC
de Computadors
UNIVERSITAT POLITÈCNICA DE CATALUNYA

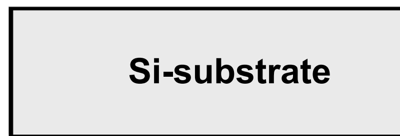
Photo-Lithographic Process



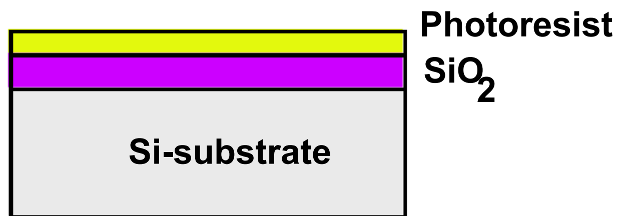
Typical operations in a single photolithographic cycle (from [Fullman]).

UPC
Departament d'Arquitectura de Computadors
UNIVERSITAT POLITÈCNICA DE CATALUNYA

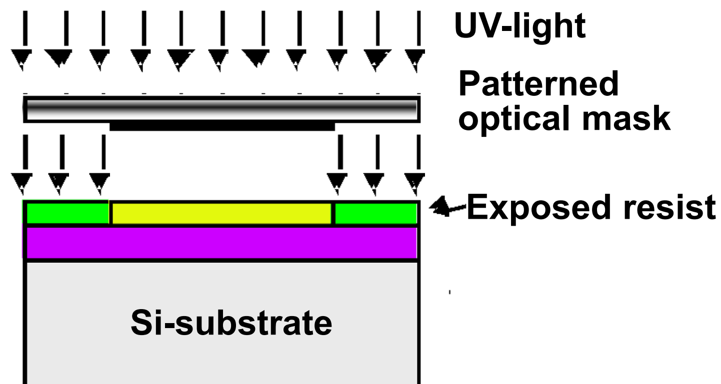
Patterning of SiO2



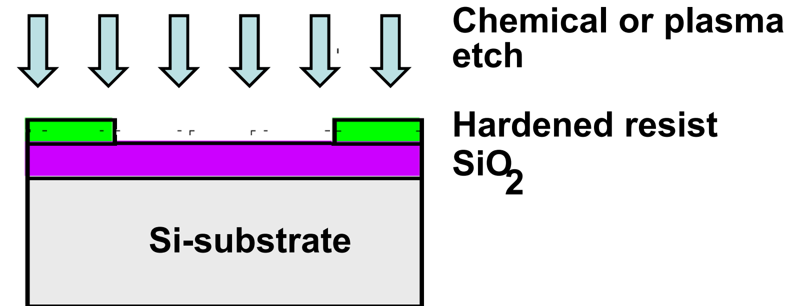
(a) Silicon base material



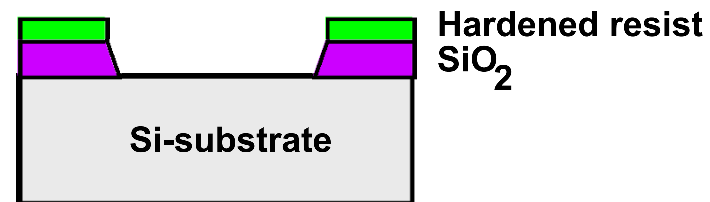
(b) After oxidation and deposition of negative photoresist



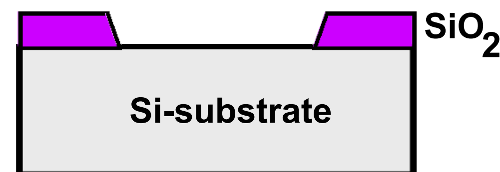
(c) Stepper exposure



(d) After development and etching of resist, chemical or plasma etch of SiO_2



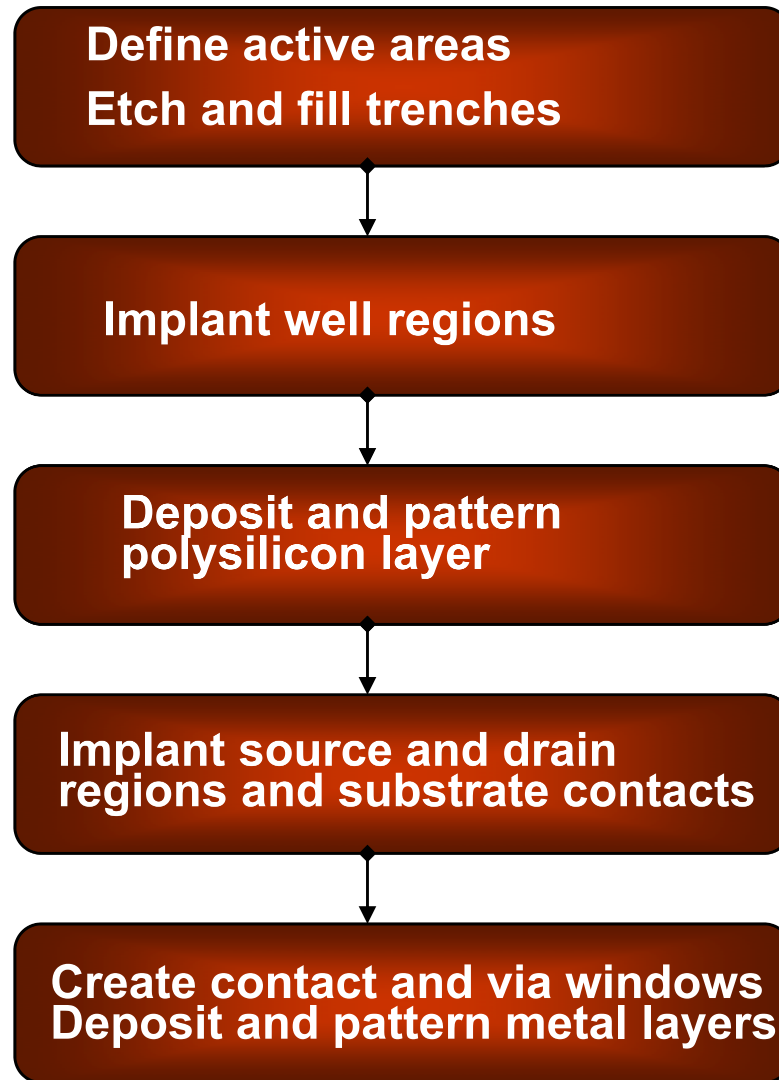
(e) After etching



(f) Final result after removal of resist

UPC
Departament d'Arquitectura de Computadors
UNIVERSITAT POLITÈCNICA DE CATALUNYA

CMOS Process at a Glance



UPC Departament d'Arquitectura de Computadors
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Manufacturing at a glance

http://am.renesas.com/company_info/v_factory/

Process Variations

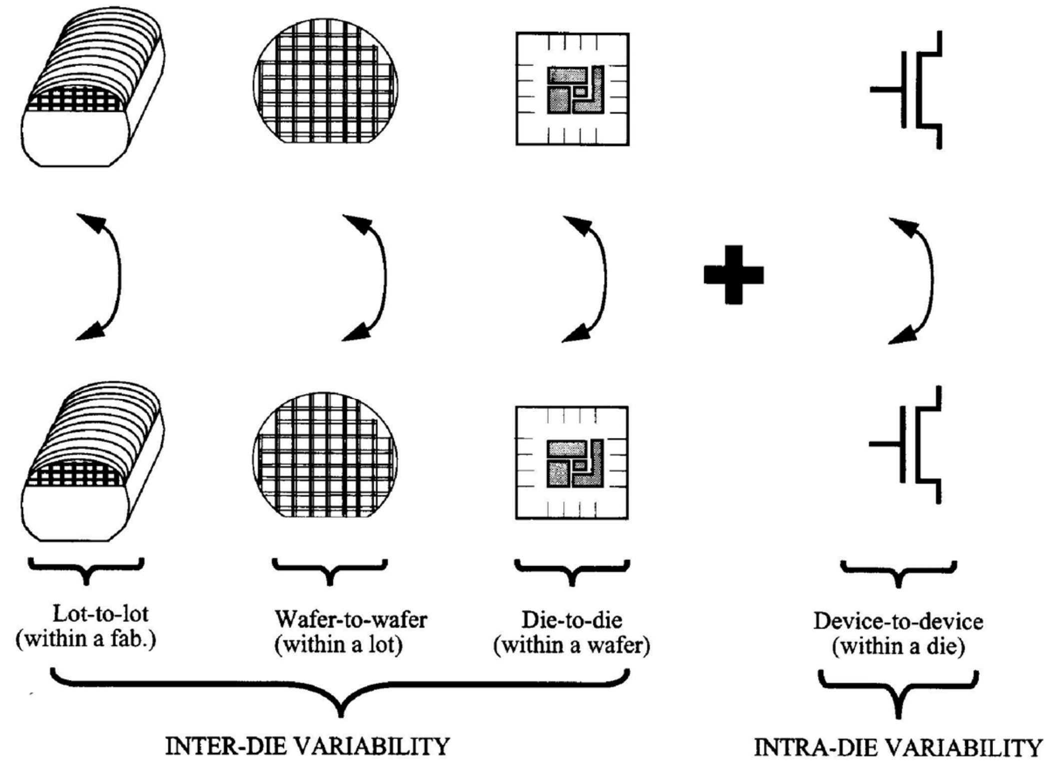
Ramon Canal

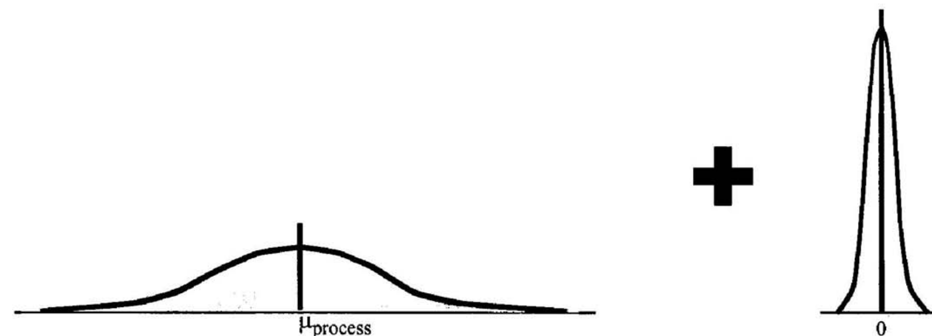
NCD - Master MIRI

Outline

- Variations overview
  - Process variations
  - Impact of design choices
  - Voltage and temperature variations
- Circuit techniques
  - Adaptive body bias
  - Adaptive supply voltage
- Future trends & conclusion

Process Variations





Process Variation

- Device parameters vary between runs and even on the same die
  - Variations in process parameters caused by non-uniform conditions during fabrication
    - Sheet resistance ($R_{\square}$) due to variations in wire thickness
    - Threshold voltage (V_T) due to variations in oxide thickness, substrate, poly and implant impurity levels, surface charge
    - Process transconductance ($k'_n = \mu_n C_{ox} = \mu_n \varepsilon_{ox}/t_{ox}$) due to variations in oxide thickness
  - Variations in device dimensions caused by the limited resolution of the photolithographic process
    - Transistor widths (W) and wire widths due to variations in the field-oxide step
    - Transistor lengths (L_{eff}) due to variations in polysilicon widths and source/drain diffusion sizes
- Large number of the deviations are uncorrelated

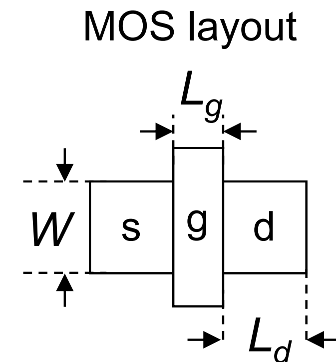
Important Device Variations

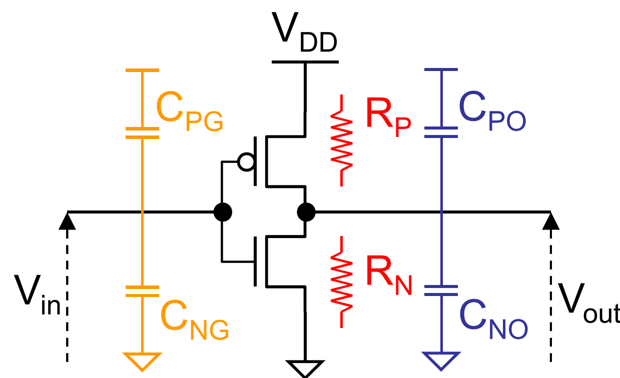
- Channel length L
  - Photolithography proximity effects
  - Optics deviations
  - Plasma etch dependencies
- Oxide thickness t_{ox}
  - Well-controlled -- only significant between wafers
- Threshold voltage V_t
  - Varying dopings
  - Annealing effects
  - Mobile Q in gate oxide
  - Discrete dopant variations (few dopant atoms in transistors)

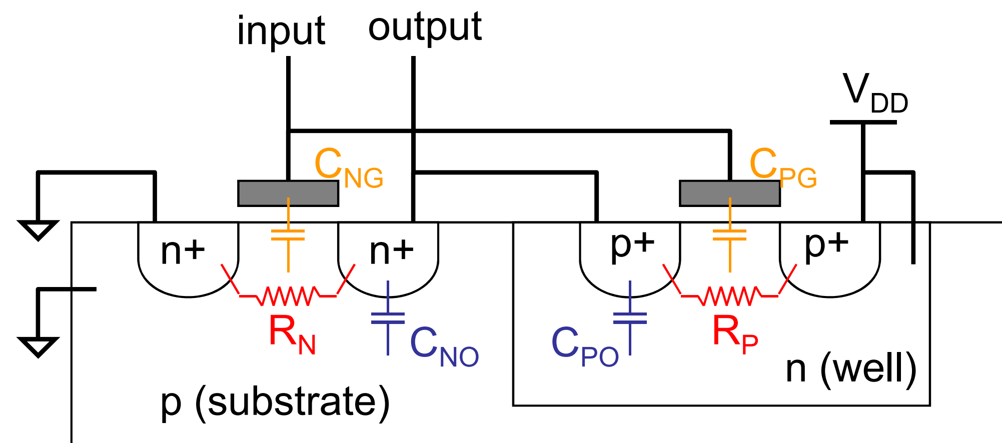
MOS gate delay

- R_N, R_P : ON-resistance of nMOS / pMOS
  → Proportional to L_g / W
- C_{NG}, C_{PG} : Gate capacitance of nMOS / pMOS
  → Proportional to $L_g \cdot W$
- C_{NO}, C_{PO} : Drain capacitance of nMOS / pMOS
  → Proportional to $L_d \cdot W$ and $L_d + W$

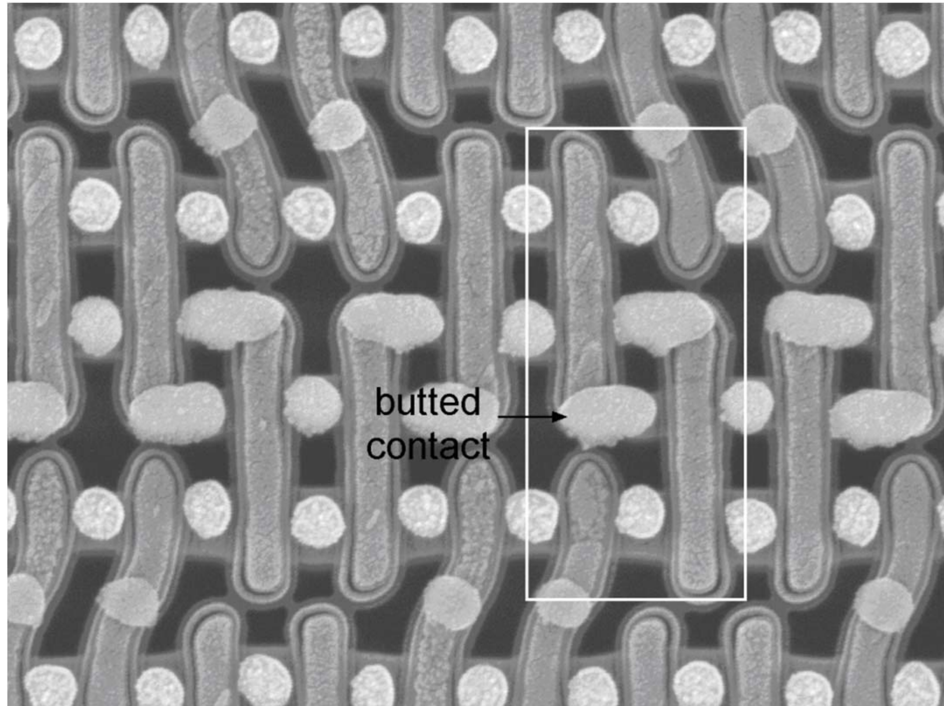
➢ *In reality, R_N, R_P, C_{NO}, C_{PO} are not constant, but dependent on the voltages at the gate and drain*



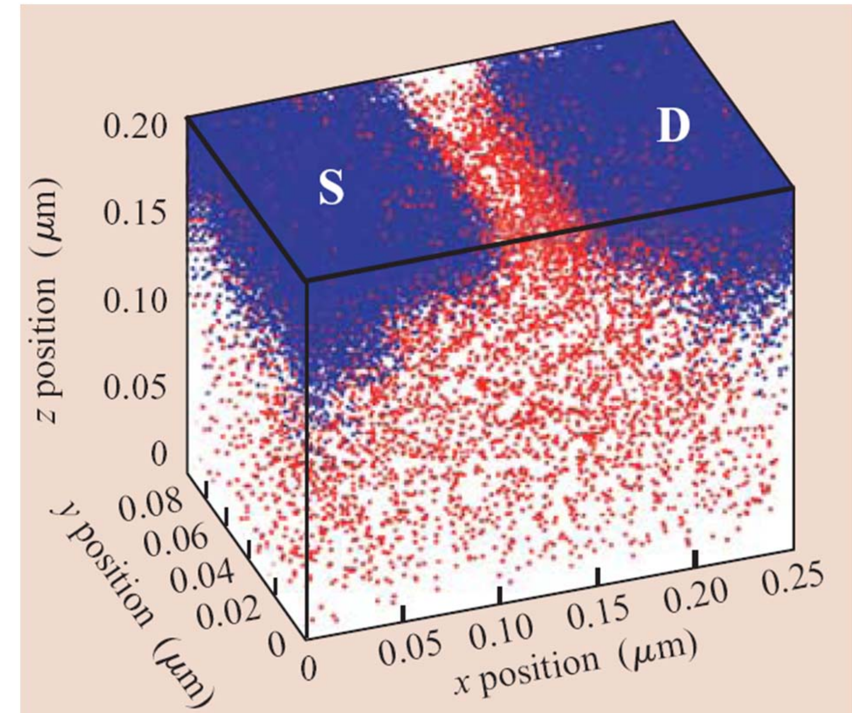




Random Variations



65nm SRAM photo

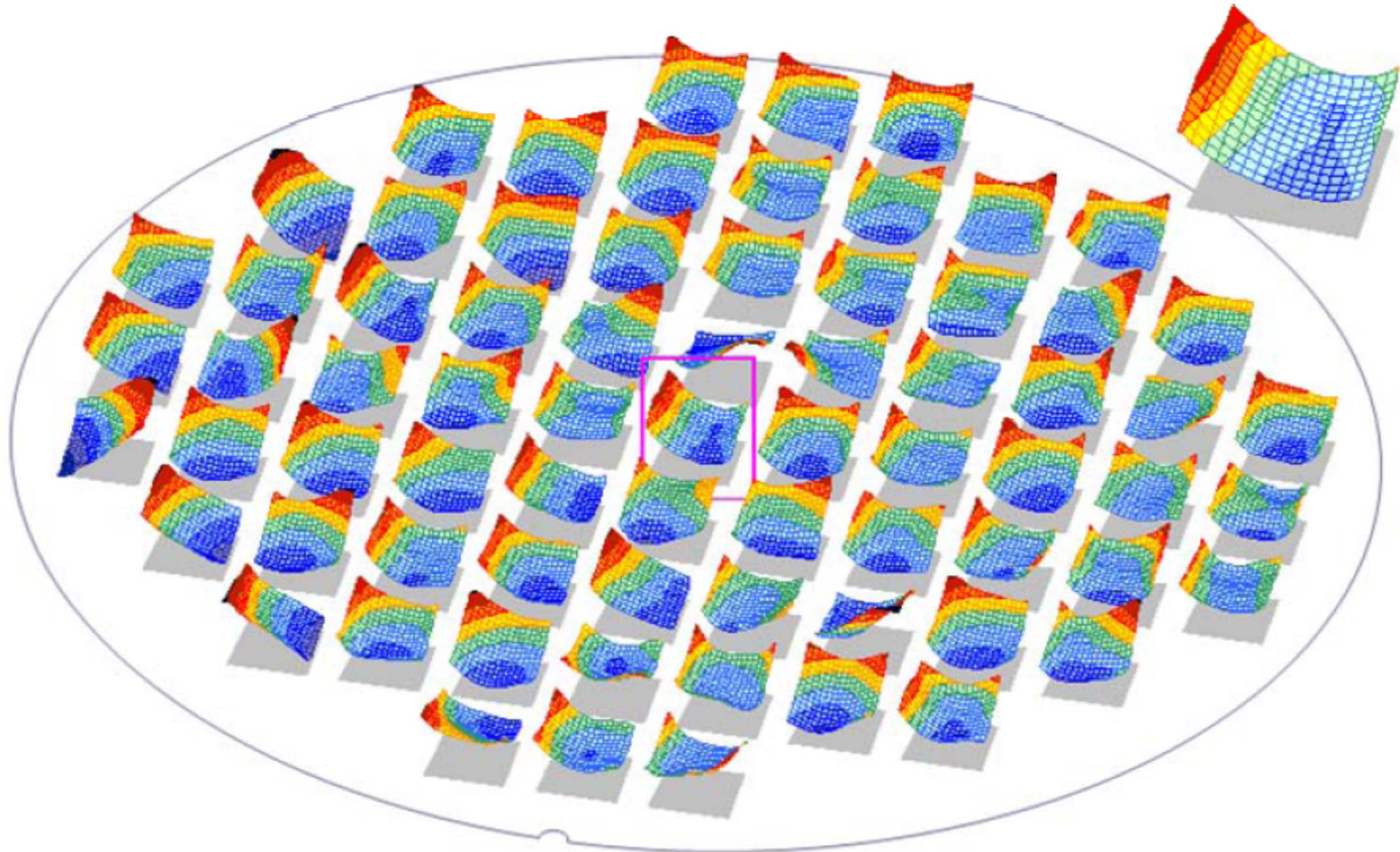


[Source: Bernstein, IBM J.R&D'06]

- Random line-width (LER) variations (left)
- Random dopant fluctuations (right)

13

Systematic WID Variations



[Source: Masuda et al., CICC 2005]

Departament d'Arquitectura de Computadors
UPC
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Process Variations

Sub-wavelength lithography

1 μm
100 nm
10 nm
1980 1990 2000 2010 2020
Lithography Wavelength
365nm
248nm
193nm
180nm
130nm
90nm
65nm
45nm
32nm
Generation
Gap
13nm EUV

Effect of process variations

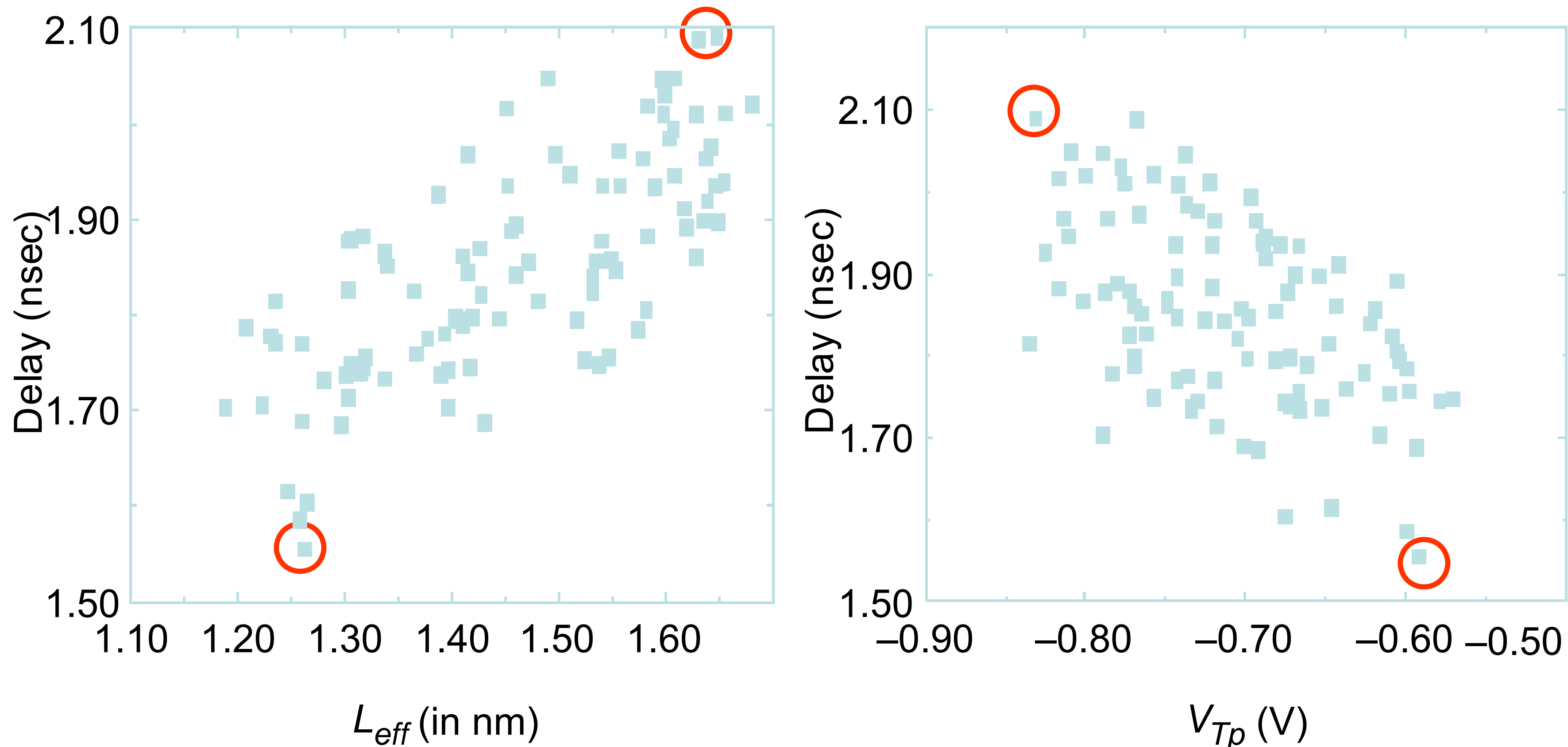
Normalized Frequency
1.4 1.3 1.2 1.1 1.0 0.9
0 5 10 15 20
Normalized Leakage (Isb)
0.18 micron
~1000 samples
30%
20X

- Functionality & Yield
- Performance
- Power

Departament d'Arquitectura de Computadors
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Impacts of Device Variations

- Process variations can cause a substantial deviation in circuit behavior from the nominal
  - the question for the designer is how much margin to provide

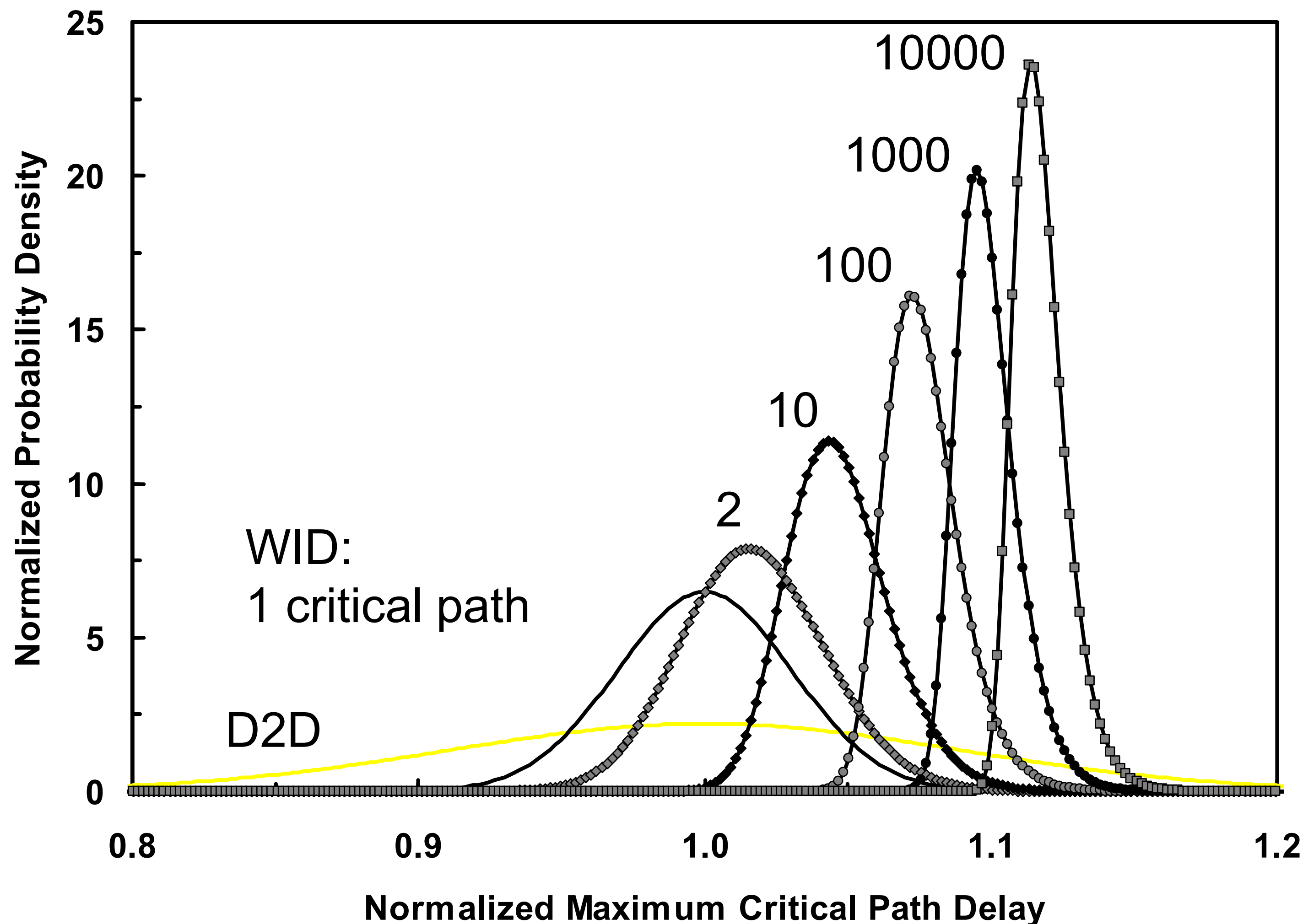


Delay of adder circuit as a function of variations in L_{eff} and V_T

UPC
Departament d'Arquitectura de Computadors
UNIVERSITAT POLITÈCNICA DE CATALUNYA

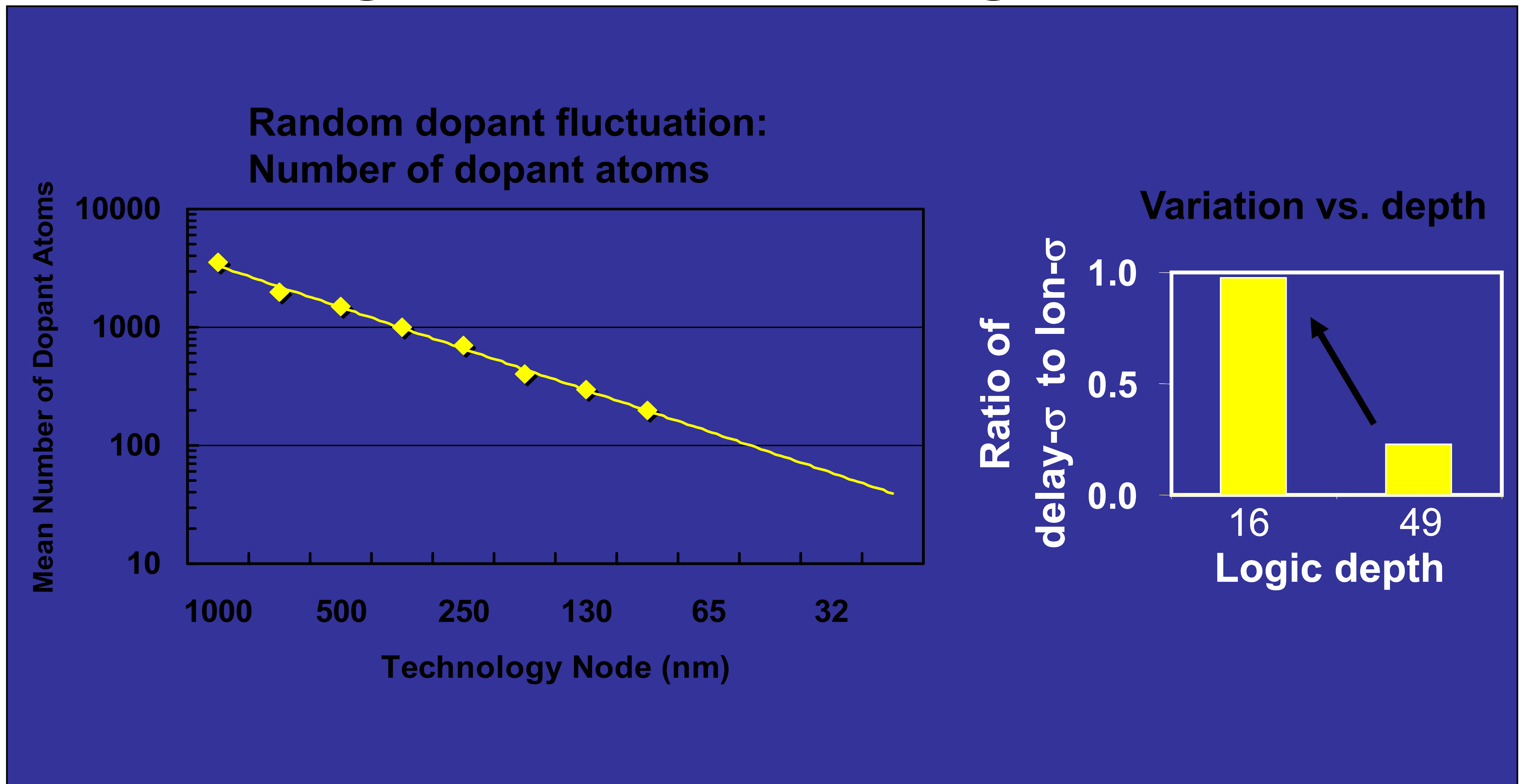
From Boskin, UCB and ATMEL Corp.

Design Impacts: Number of Paths



Large number of critical paths → reduces mean frequency

Design Impacts: Logic Depth



Reducing logic depth increases impact of random variations

Effect of Within-Die Variations

Cumulative Distribution (%)

100
80
60
40
20
0

-4 -3 -2 -1 0 1 2 3 4

Number of FMAX Standard Deviations

Model: Only WID Variations

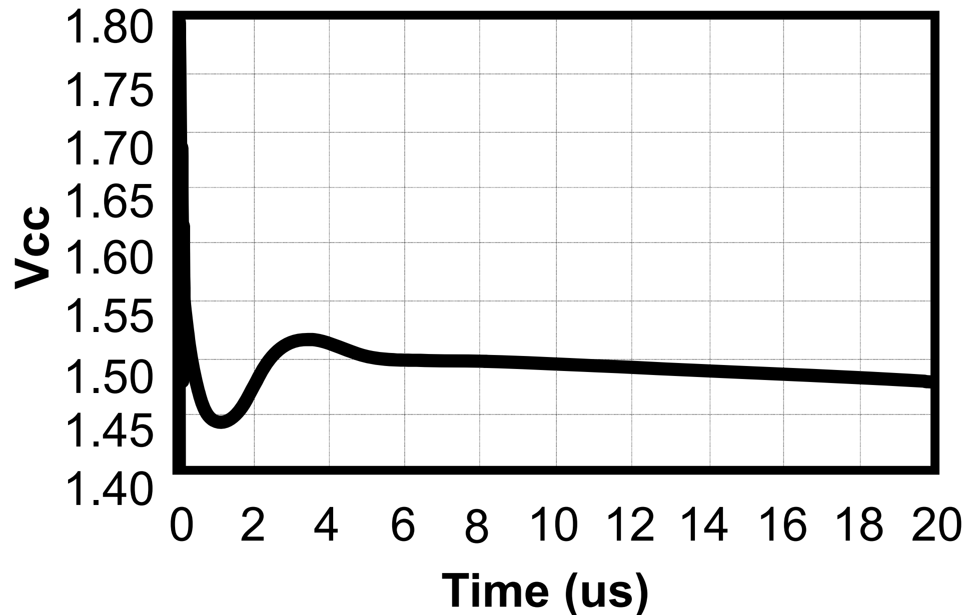
Model: Only D2D Variations

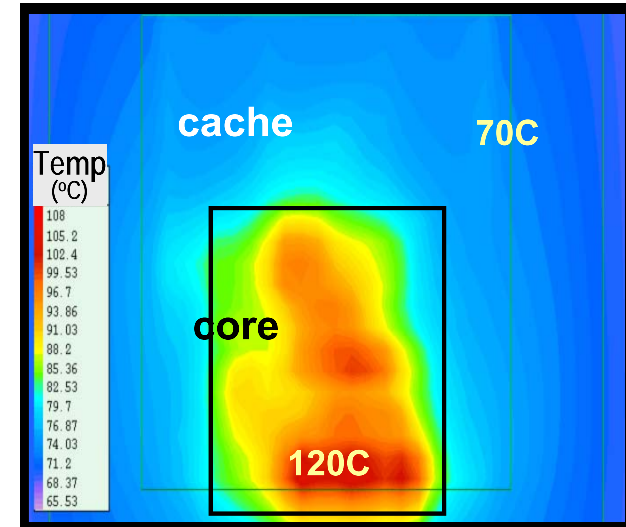
Model: D2D & WID Variations

- Measured Data
- Model: D2D & WID
- Model: D2D
- Model: WID

D2D variations: impact variance
WID variations: reduce the mean FMAX

Voltage and Temperature Variations





<u>Variations are both static and dynamic</u>

<u>Static</u>: Tools required for prediction

<u>Dynamic</u>: Design margining or adaptation techniques

UPC
Departament d'Arquitectura de Computadors
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Design Margining

- Other factors affecting circuit behavior
  - Operating temperature: As the temperature increases, drain current is reduced (commercial parts are spec'ed to perform from 0°C to 70°C)
  - Supply voltage: Variations on supply voltage are normally ± 10%
- Boundary cases of performance:
  - nominal
  - fast
  - slow

which map to the CMOS boundary conditions:

fast-n/fast-p	fast-n/slow-p
slow-n/slow-p	slow-n/fast-p
slow-n/low V_{Tp}	low V_{Tn}/slow-p

Departament d'Arquitectura de Computadors
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Design Corners

Boundary	ºC	V_{DD} (±10%)	L_{eff} (±10%)	Tests
fast-n/fast-p	0ºC	2.5V + .25V = 2.75V	.25μ - .025μ = 0.225μ	peak power dissipation (DC), clock skew, contamination delay
slow-n/slow-p	70ºC	2.5V - .25V = 2.25V	.25μ + .025μ = 0.275μ	max clock rate, flipflop setup and hold times

- Example: $V_{GS} = V_{DS}$ = 2.5V for our 0.25 μm process

  I_d nominal = 220 μA

  For L_{eff} and W_{eff} of ±10%, V_T of ±60mV, t_{ox} of ±5%

  I_d fast = 265 μA – a 20% improvement

  I_d slow = 182 μA – a 17% degradation

  For a run time variation in V_{DD} of ±10%

  V_{DD} = 2.75V, I_d fast = 302 μA – a 37% improvement

  V_{DD} = 2.25V, I_d slow = 155 μA – a 30% degradation

Technology Scaling Models

- Full scaling (constant electrical field)
  - ideal model - both dimensions and voltage scale together by the same factor S (S > 1)
  - goal is to keep the electrical field patterns in the scaled device identical to those in the original device (ensures the physical integrity of the devices)
- Fixed voltage scaling
  - most common model five years ago - only dimensions scale, voltages remain constant
- General scaling
  - most realistic for today - voltages scale by factor U while dimensions scale by factor S (normally S > U > 1)

Scaling Effects (Short-Channel)

Parameter	Relation	Full	General	Fixed Voltage
W, L, t_{ox}		1/S	1/S	1/S
Area/Device	WL	$1/S^2$	$1/S^2$	$1/S^2$
V_{DD}, V_T		1/S	1/U	1
N_{SUB}	V/W_{depl}^2	S	S^2/U	S^2
C_{ox}	$1/t_{ox}$	S	S	S
C_{gate}	C_{ox} WL	1/S	1/S	1/S
k_n, k_p	C_{ox} W/L	S	S	S
I_{sat}	C_{ox} WV	1/S	1/U	1
Current Density	I_{sat}/Area	S	S^2/U	S^2
R_{on}	V/I_{sat}	1	1	1
Intrinsic Delay	$R_{on}C_{gate}$	1/S	1/S	1/S
P_{av}	$I_{sat}V$	$1/S^2$	$1/U^2$	1
Intrinsic Energy	$C_{gate}V^2$	$1/S^3$	$1/SU^2$	1/S
Intrinsic Power	E/Delay	$1/S^2$	$1/U^2$	1
Power Density	P/Area	1	S^2/U^2	S^2

Technology Directions: SIA Roadmap

Year	2001	2003	2005	2007	2010	2013
Feature size (nm)	90	65	45	32	25	18
Gate oxide t_{ox} (nm)	2.4	2.0	1.9	[illegible]	[illegible]	[illegible]
Power supply V_{DD} (V)	1.2	1.0	0.9	[illegible]	0.6	[illegible]

S = 1.38, U = 1.2

S = 1.4, U = 1.17

ITRS Technology Roadmap, 2001

Technology Directions: SIA Roadmap

Year	2001	2003	2005	2007	2010	2013
Feature size (nm)	90	65	45	32	~~25~~ 28	~~18~~ 22
Gate oxide t_{ox} (nm)	2.4	2.0	1.9	1.4	1.2	1.0
Power supply V_{DD} (V)	1.2	~~1.0~~ 1.2	~~0.9~~ 1.1	~~0.7~~ 1.1	~~0.6~~ 1.0	~~0.5~~ 0.86

S = 1.38, U = 1.2

S = 1.38, U = 1.0

S = 1.4, U = 1.17

S = 1.14, U = 1.09

10x leakage per generation

Technology Roadmap:

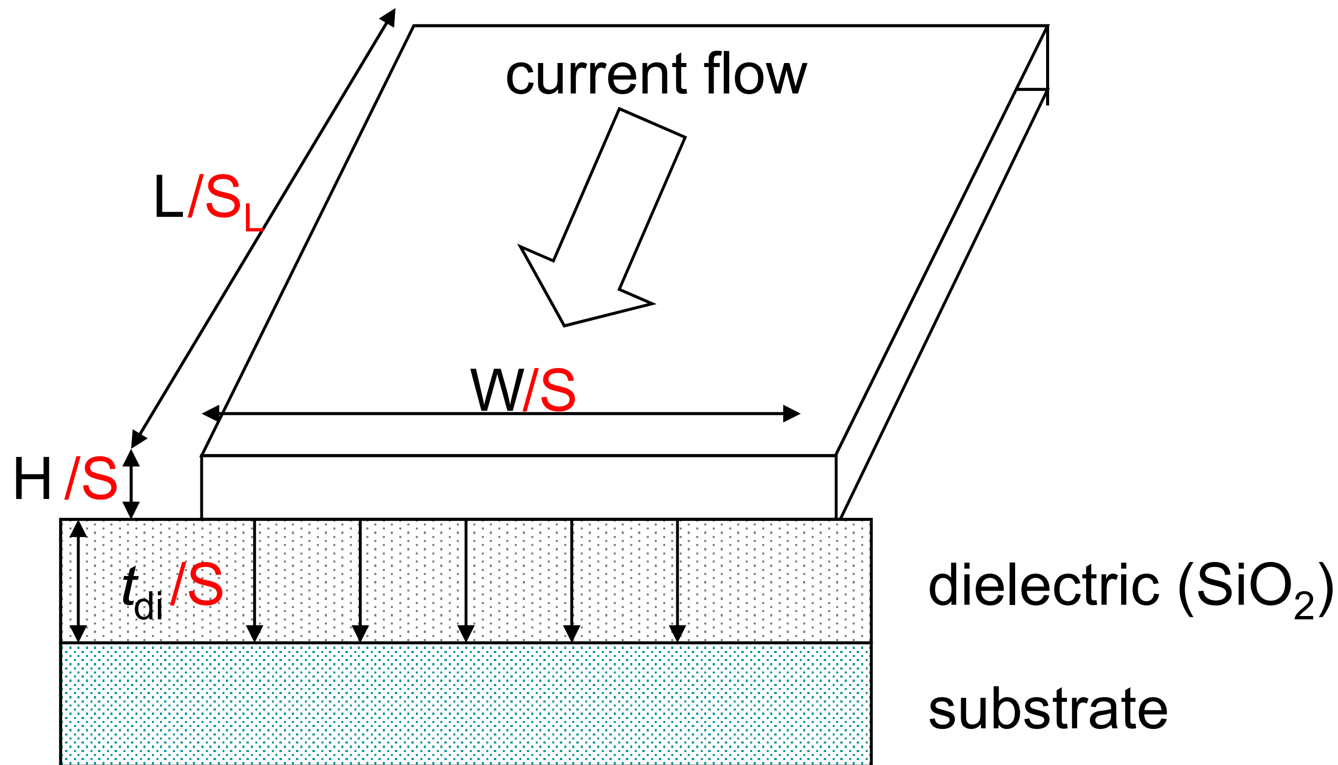
http://www.itrs.net/Links/2013ITRS/Home2013.htm

Interconnect Variations

- Line width and line spacing
  - Photolithography
  - Etching proximity effects
- Metal and dielectric thickness
  - Chemical Mechanical Polishing
- Contact resistance
  - Contact dimensions
  - Etch and clean steps

Departament d'Arquitectura
de Computadors
UNIVERSITAT POLITÈCNICA DE CATALUNYA

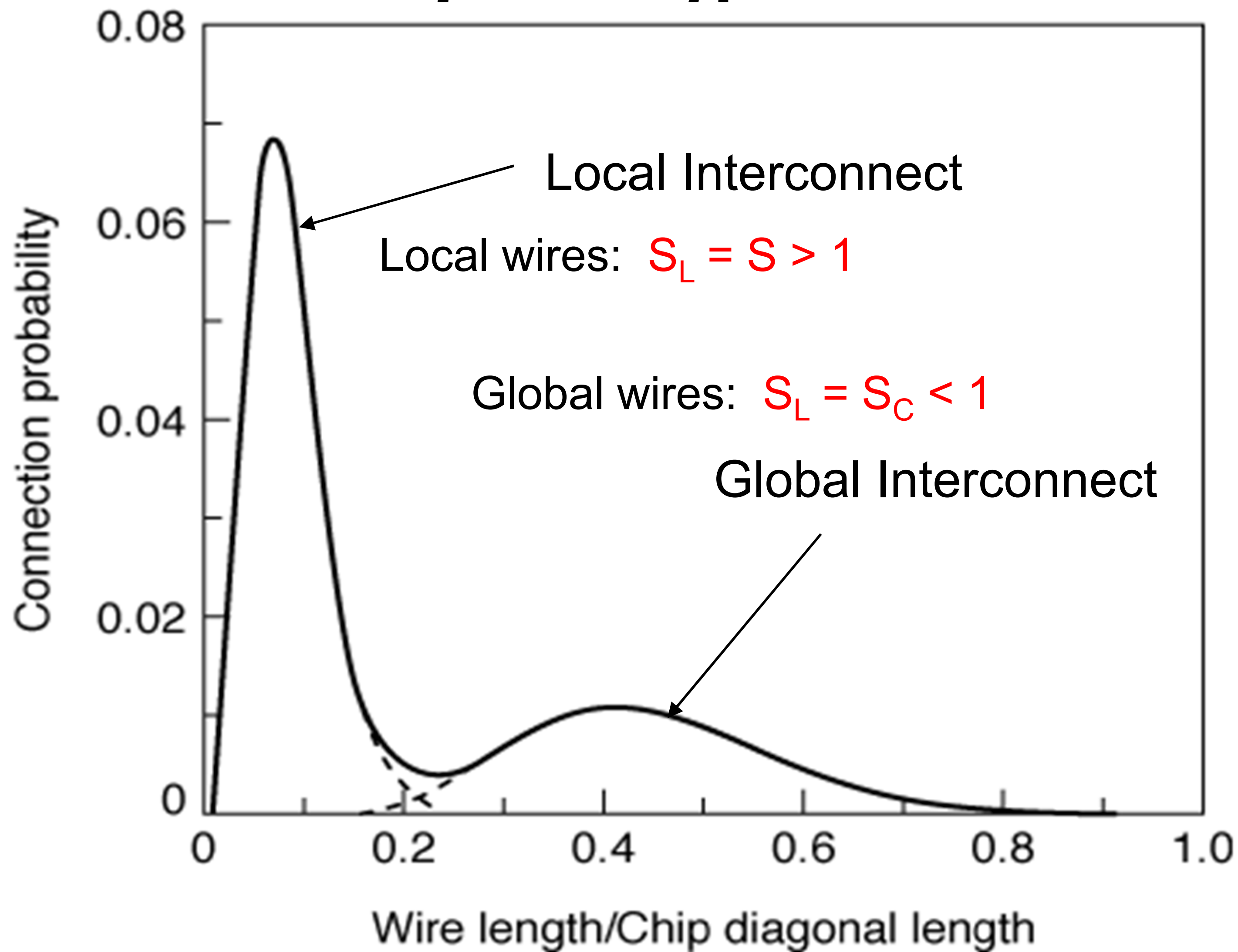
Scaling of Interconnect



$C_{pp} = (\varepsilon_{di}/t_{di})\, WL$

$S_{C,wire} = (S \times S_L)/S = S_L$

Scaling of Interconnect



From Kang, 87

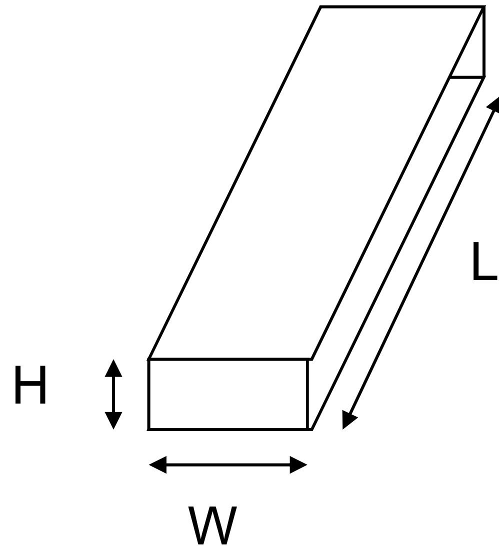
Wire Length Scaling Effects

Parameter	Relation	Local	Constant	Global
W, H, t_{ox}		$1/S$	$1/S$	$1/S$
L		$1/S$	1	$1/S_C$
C	WL/t_{ox}	$1/S$	1	$1/S_C$
R	L/WH	S	S^2	S^2/S_C
RC	L^2/Ht_{ox}	1	S^2	S^2/S_C^2

- Constant delay for local wires and increase in delay for global wires
  - for $S = 1.15$ and $S_C = 0.94$, global wire delay goes up 50%

- Why wire delays are now playing a predominate role in design!

Scaling of Wire Resistance



$$R_{wire} = \frac{\rho L}{A} = \frac{\rho L}{H W}$$

Material	Sheet Res. (Ω/❐)
n+, p+ diffusion	50 to 150
n+, p+ diffusion with silicide	3 to 5
polysilicon	150 to 200
polysilicon with silicide	4 to 5
Aluminum	0.05 to 0.1

$S_{R,wire} = S_L / S^2$ if H is scaled

$S_{R,wire} = S_L / S$ if H is not scaled

UPC
Departament d'Arquitectura de Computadors
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Constant Resistance Wire Scaling Effects

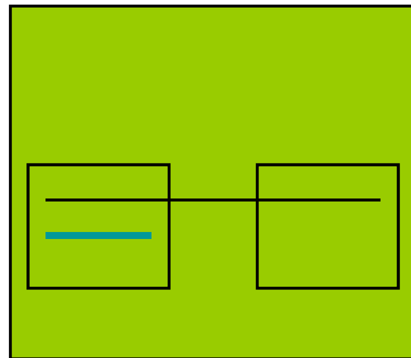
- Capacitance scaling factor ε_C (> 1) that captures the "horizontal" nature of wire capacitance

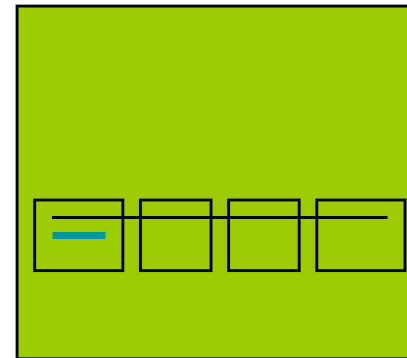
Parameter	Relation	Local	Constant	Global
W, t_{ox}		1/S	1/S	1/S
H		1	1	1
L		1/S	1	$1/S_C$
C	$\varepsilon_C WL/t_{ox}$	ε_C/S	ε_C	ε_C/S_C
R	L/WH	1	S	S/S_C
RC	L^2/Ht_{ox}	ε_C/S	$\varepsilon_C S$	$\varepsilon_C S/S_C^2$

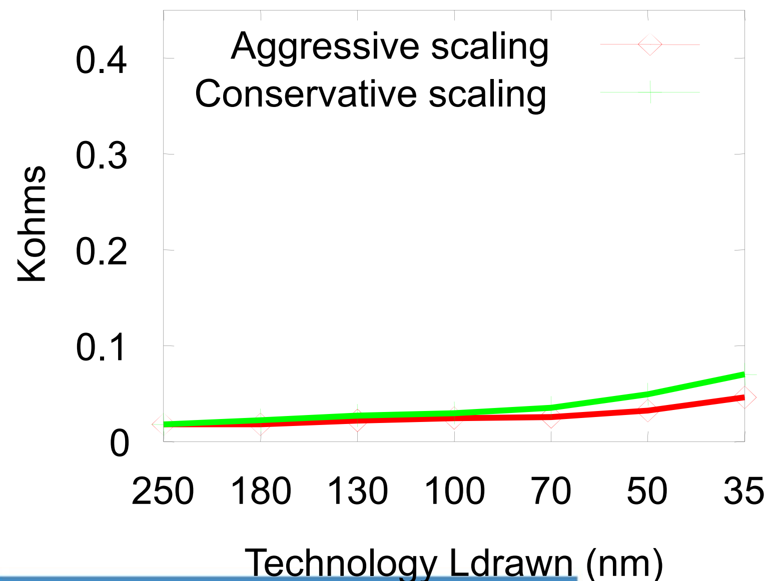
- Slightly more optimistic results (assuming ε_C < S)

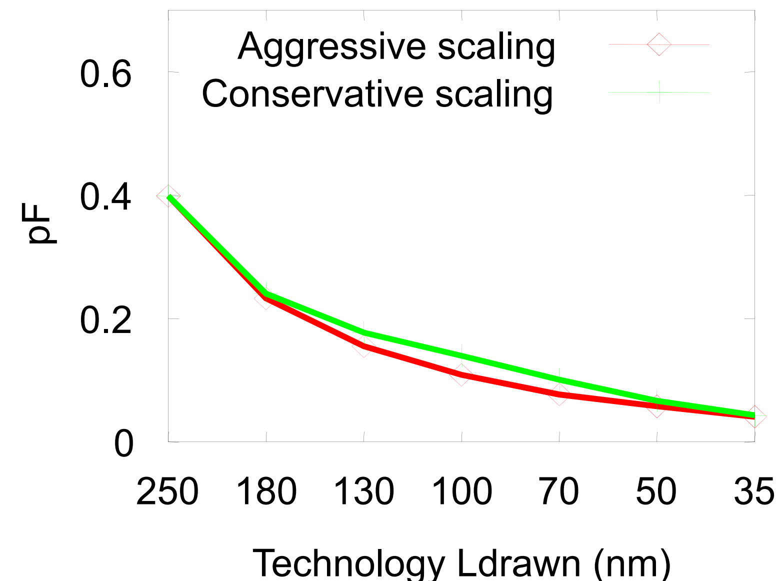
Scaling Module (Short) Wires

- R is basically constant; C falls linearly with scaling





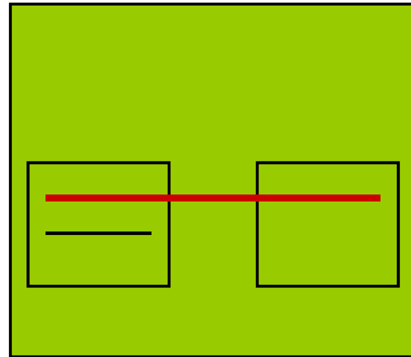


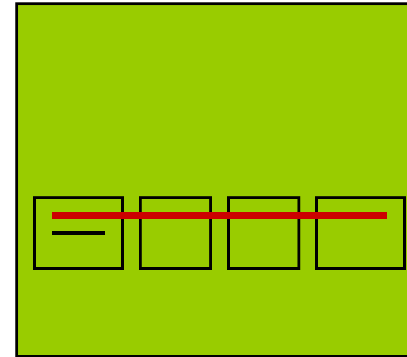


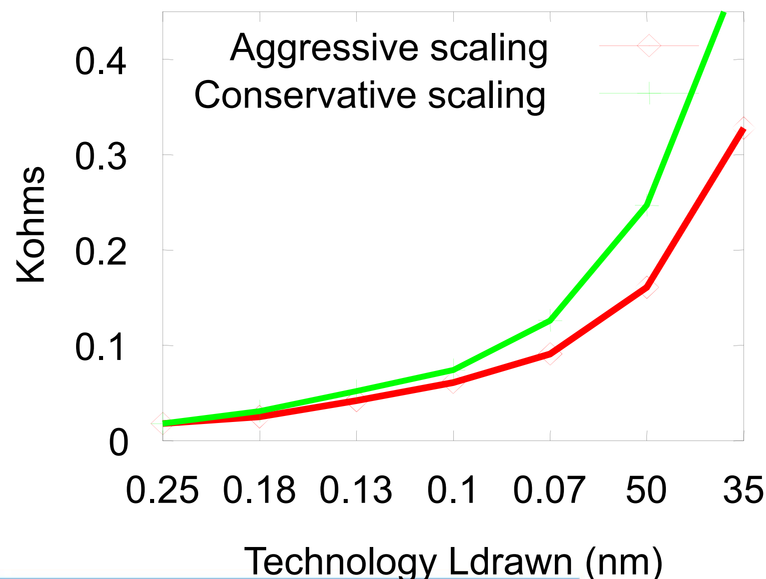
Source: ISCA'01 Panel

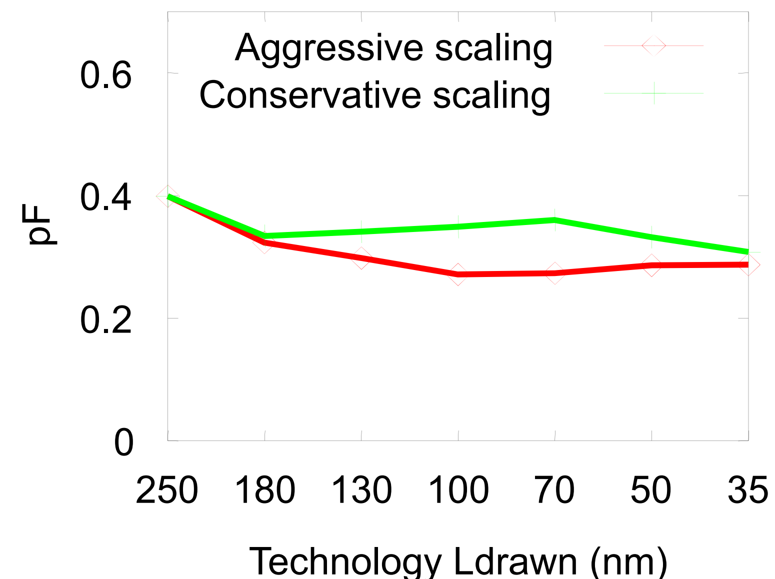
Scaling Global (Long) Wires

- R gets quite a bit worse; C is basically constant





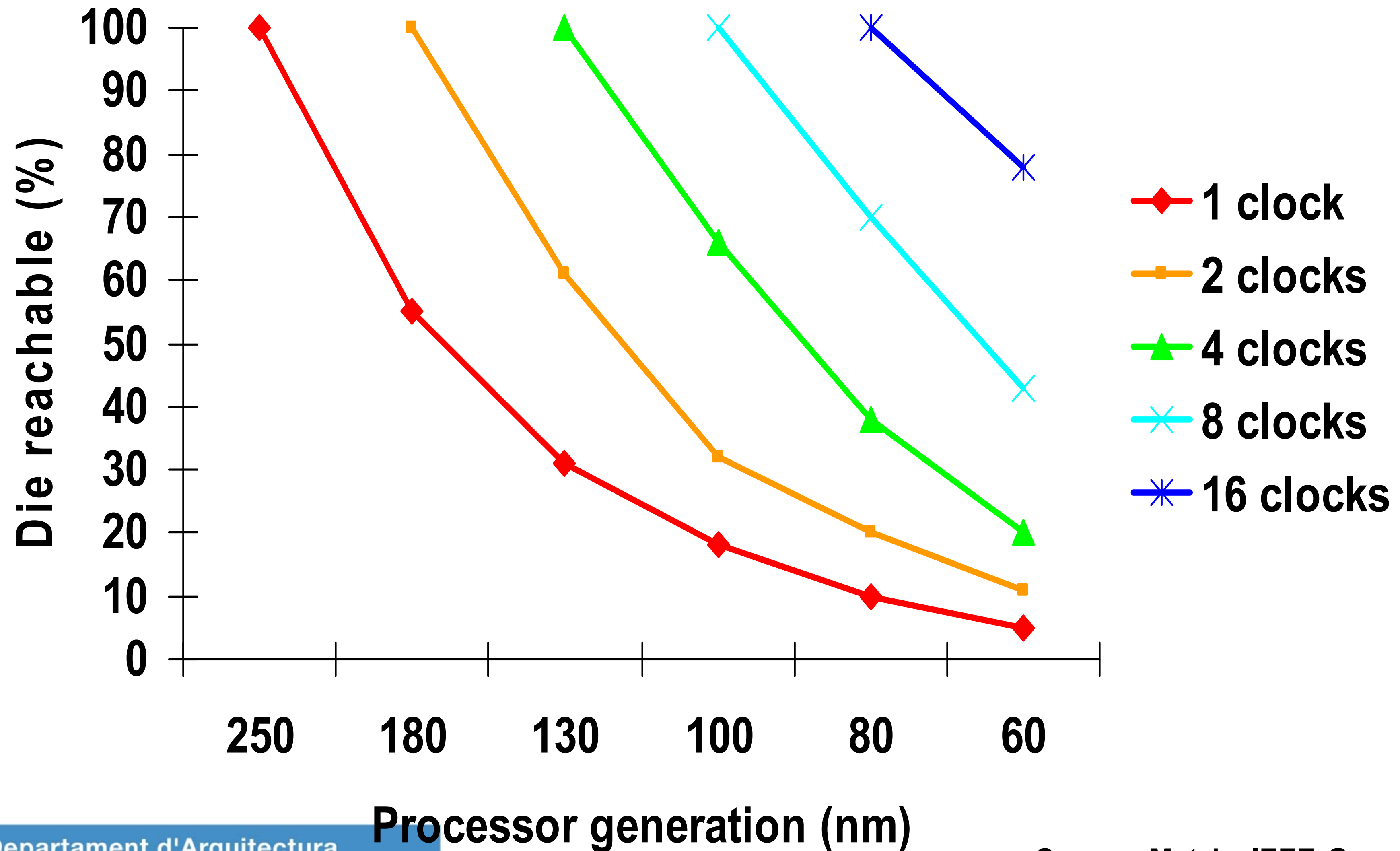




Source: ISCA'01 Panel

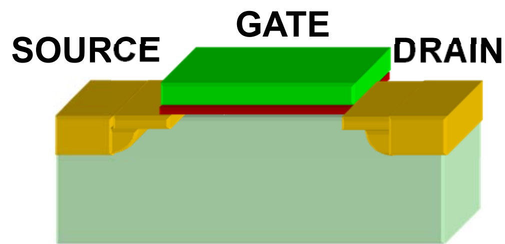
Clock Distribution Limits

- Clocks are **long** wires with **big** RC's that have rigid clock skew constraints



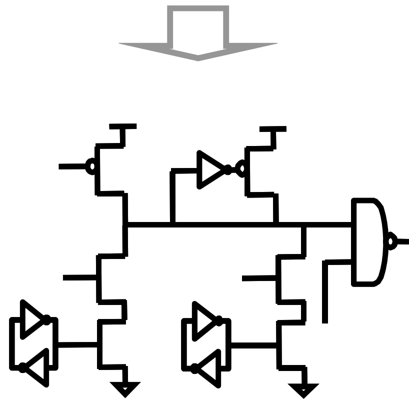
Source: Matzke IEEE *Computer*

Variation-Tolerant Design



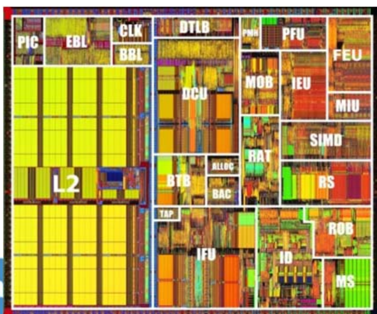
Reduce <u>SOURCE</u> of variation

- Multi-Le and multi-Vt insertion
- Circuit styles and logic decisions
- Power delivery and thermal design



Reduce <u>EFFECTS</u> of variation: design

- Leakage-reduction techniques
- Variation-tolerant circuits
- Dynamic compensation circuits



Reduce <u>EFFECTS</u> of variation: post-silicon

- Clock tuning
- Adaptive body bias
- Adaptive supply voltage

Variations and solutions

Ramon Canal
NCD - Master MIRI

Overview

- Faults, errors and failures. Who is who?
- Classification and countermeasures
  - Permanent faults
  - Transient faults
- Fault prevention and detection
- Conclusions

Faults, errors and failures

- Fault: A fault is a defect that may trigger an error or stay dormant.
  - Bit-flips, stuck-at-0, stuck-at-1, …

- Error: A wrong computation/data which changes the system behavior
  - Different addition results, different memory address to access, etc.

- Failure: Inability to provide the function defined

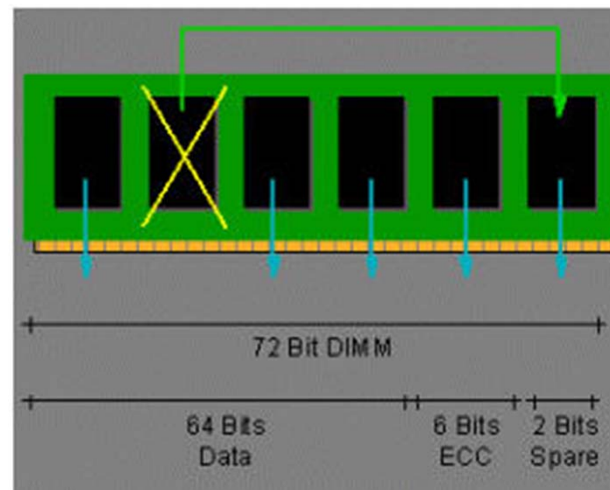
Classification & Countermeasures

- Permanent faults
  - Disable too faulty blocks
  - Redundancy:
    - Replace blocks (i.e. Need to provide spares)
    - Dual-modular/Triple-modular redundancy
  - Error detection and correction codes
  - Wordline Boosting
  - Adaptive Body Biasing
  - Adaptive Supply Voltage

- Transient faults
  - Strengthing or increasing design margins
  - Error detection and correction codes
  - Bit/Line interleaving
  - Scrubbing

Permament faults

- Disable faulty blocks
  - Disable cache lines/columns or any block that produces too many faults.
  - Usually referenced as “chipkill[1]” for main memory
    - Similar to RAID 3 for hard drives

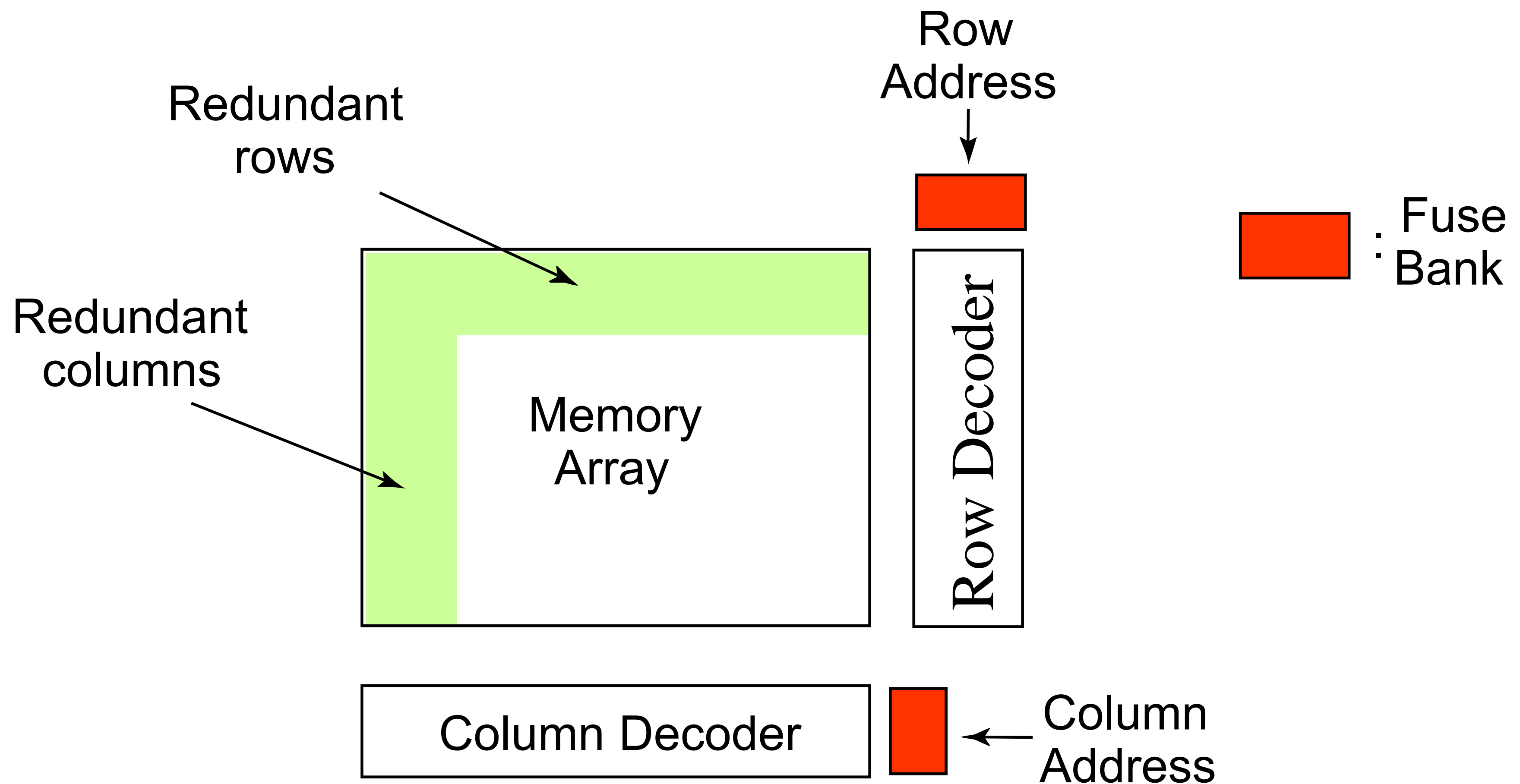


[1]Trademark from IBM back in the time of mainframes (70’s)

UPC
Departament d'Arquitectura de Computadors
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Permament faults

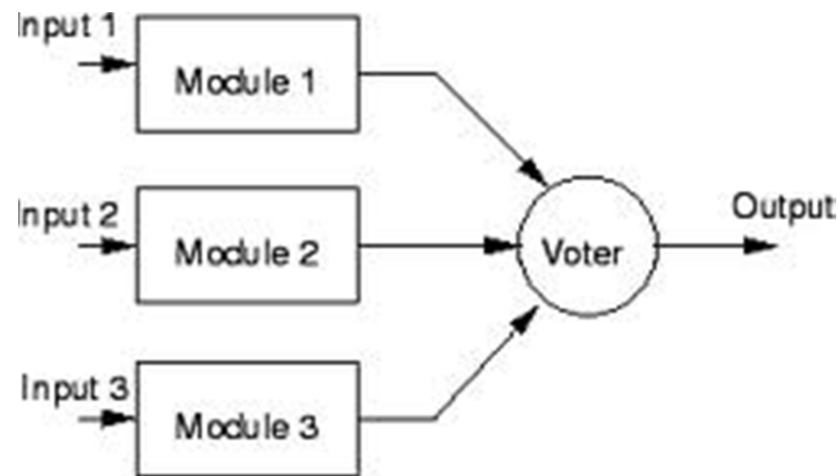
- Redundancy
  - Built-in spare blocks (e.g. cache lines, ALUs, etc)



UPC
Departament d'Arquitectura
de Computadors
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Permament faults

- Redundancy
  - Double/triple-modular
    - Usually a majority vote
    - Can add weights to leverage modules



  - TMR first implemented in the ALUs of SAPO (Samočinný počítač). Built between 1950-1956 in Czechoslovakia.

Departament d'Arquitectura de Computadors
UPC
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Permament faults

- Error detection and correction codes
  - Add extra-bits to detect bit-flips

  - Started by parity (CDC7600 in 1969)
  - Usually SECDED, now in high-performance DECTED
  - Large variety of codes
    - Hamming codes
    - Hsiao codes
    - BCH codes
  - Present in all memory structures
  - Problem: area overhead and code computation delays as error coverage increases

Permament faults

- Error detection and correction codes

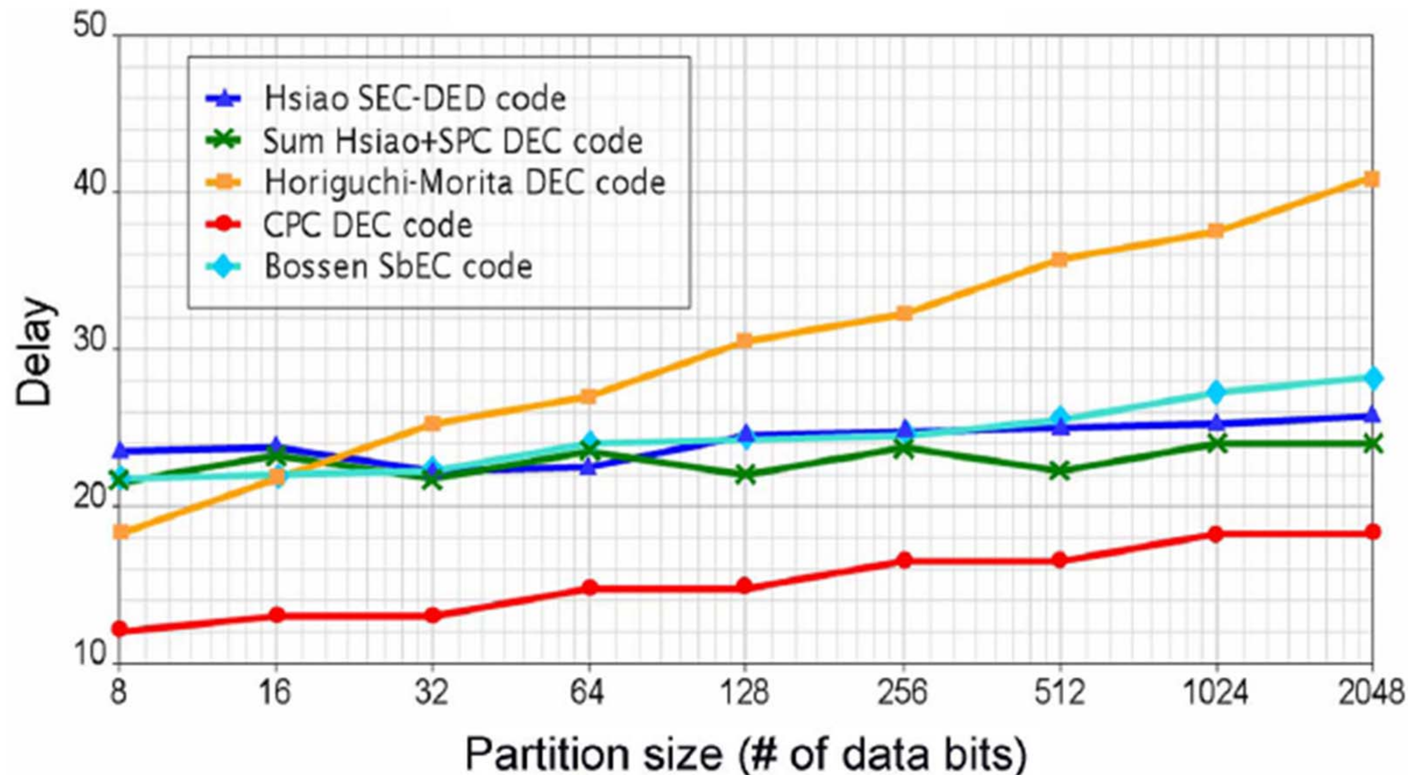
Tab. 1. Hsiao SEC/DED code: Code Size Analysis.

Data Bits	Check Bits	CA (%)	Delay (EqG)	AO (%) 32kB	AO (%) 6MB
8	5	74.2	13.5	62,55%	62,50%
16	6	73.8	16.0	37,59%	37,50%
32	7	72.6	18.5	22,04%	21,88%
64	8	70.1	19.5	12,82%	12,50%
128	9	67.7	23.0	7,71%	7,03%
256	10	57.9	25.5	5,35%	3,91%
512	11	44.8	28.0	5,20%	2,16%
1024	12	27.6	29.5	7,60%	1,21%
2048	13	10.4	35.0	15,00%	0,71%

Source: Daniele Rossi, N. Timoncini, M. Spica, Cecilia Metra: Error correcting code analysis for cache memory high reliability and performance. DATE 2011: 1620-1625

Permament faults

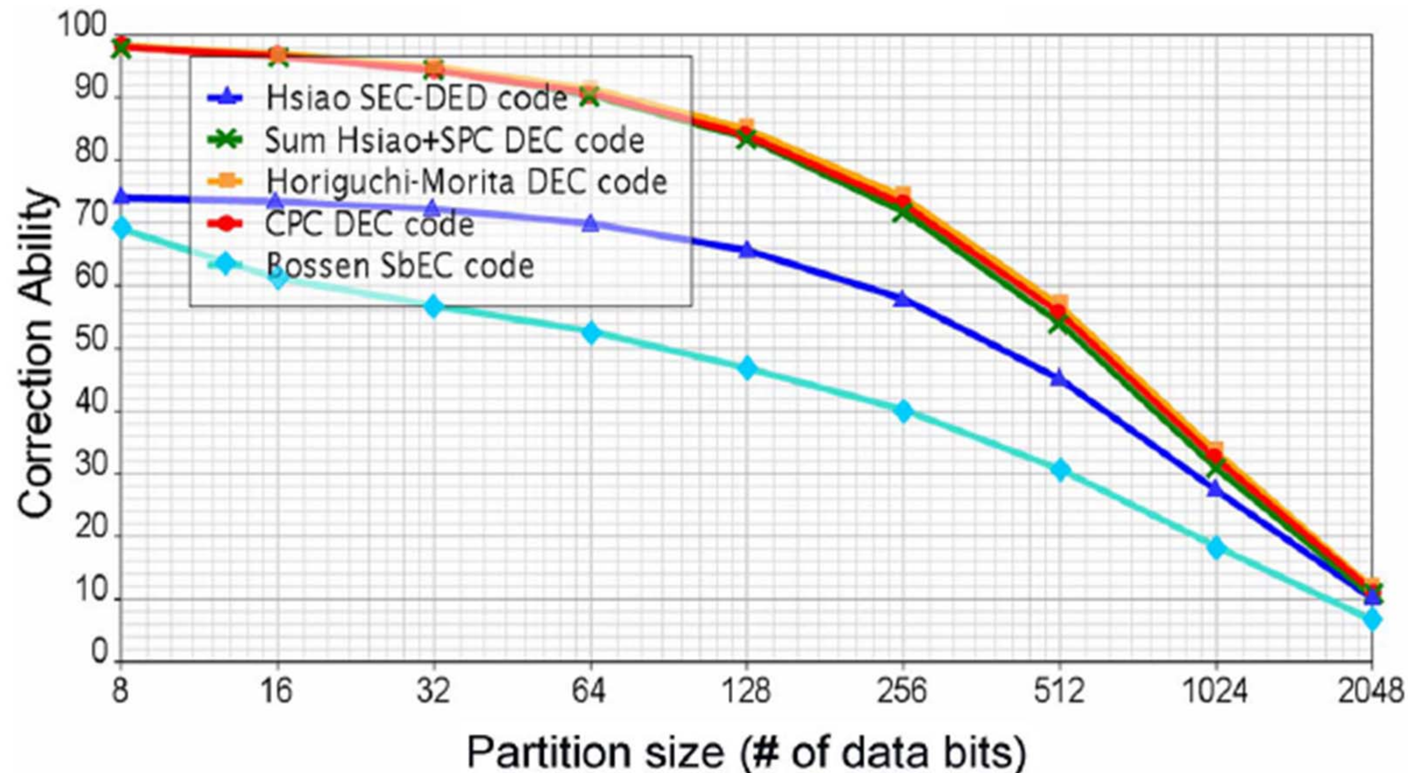
- Error detection and correction codes



Source: Daniele Rossi, N. Timoncini, M. Spica, Cecilia Metra: Error correcting code analysis for cache memory high reliability and performance. DATE 2011: 1620-1625

Permament faults

- Error detection and correction codes

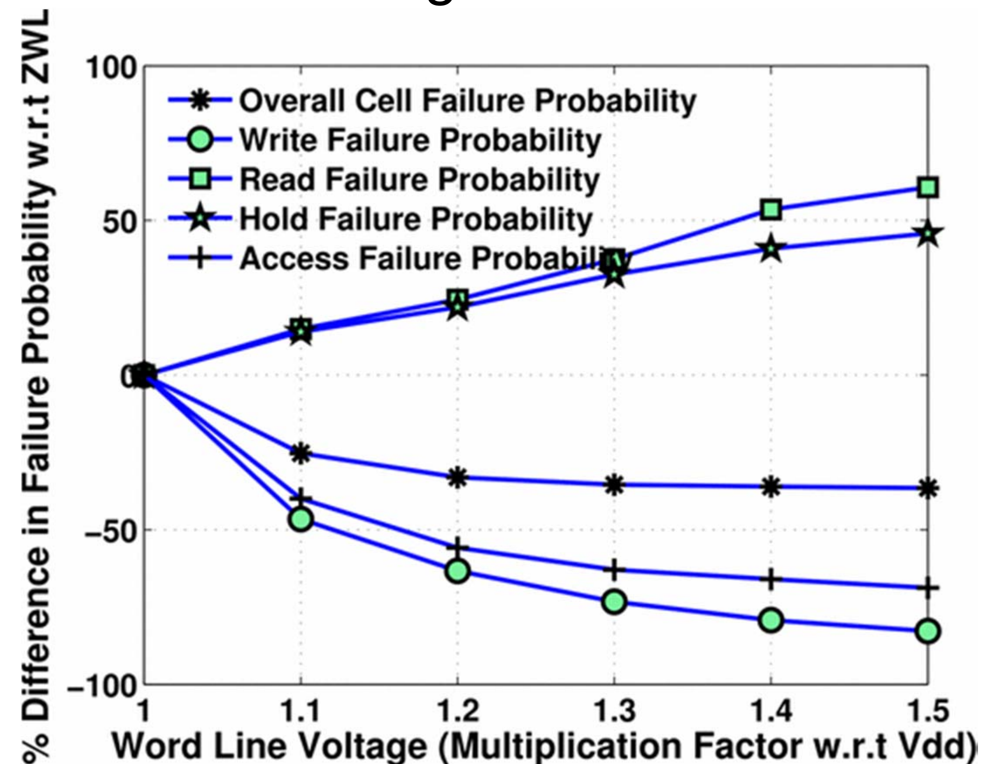


Source: Daniele Rossi, N. Timoncini, M. Spica, Cecilia Metra: Error correcting code analysis for cache memory high reliability and performance. DATE 2011: 1620-1625

Permament faults

- Wordline Boosting
  - Boosting increases the Vgs of the access transistor and hence increases its drive strength
  - The increased drive strength of the access transistor aids significantly in flipping the bit cell making it easier to write

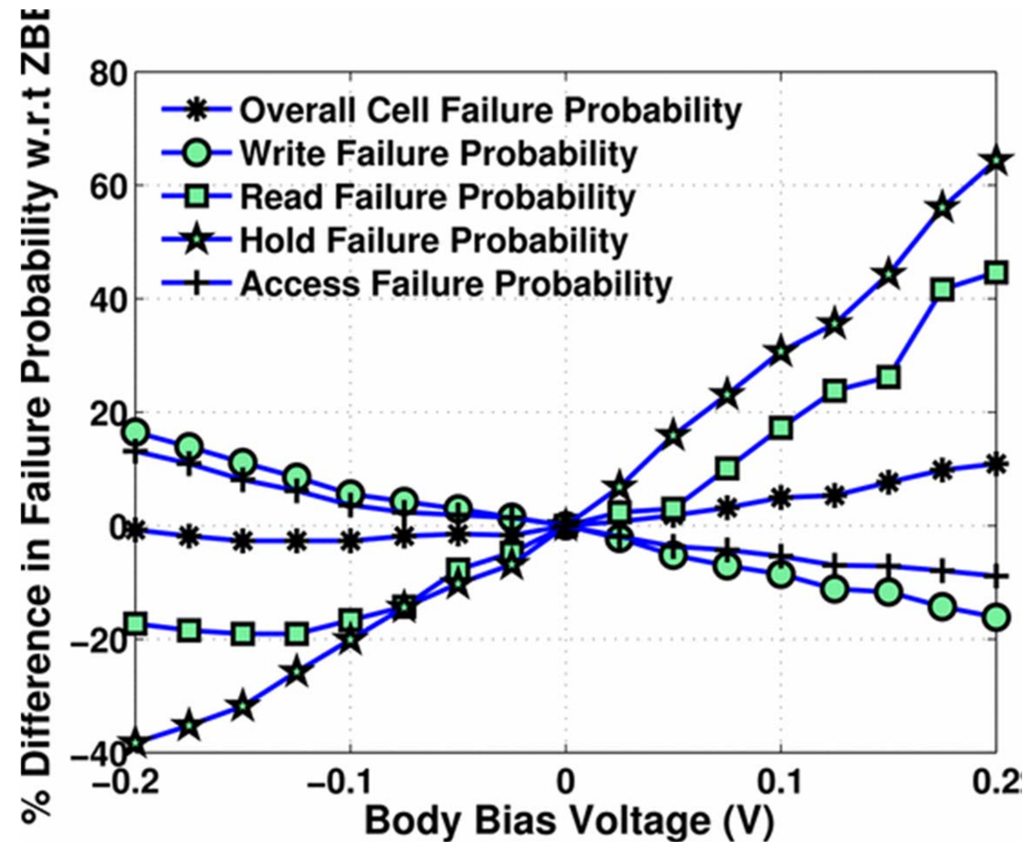
Use an extra capacitor to store charge and then release it (i.e. charge pump)



Source: Ganapathy et al. ISQED 2013

Permament faults

- Adaptive Body Biasing
  - Same concept as for leakage reduction
  - Change drive capabilities of transistors



UPC
Departament d'Arquitectura de Computadors
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Source: Ganapathy et al. ISQED 2013

Classification & Countermeasures

- Permanent faults
  - Disable too faulty blocks
  - Redundancy:
    - Replace blocks (i.e. Need to provide spares)
    - Dual-modular/Triple-modular redundancy
  - Error detection and correction codes
  - Wordline Boosting
  - Adaptive Body Biasing
  - Adaptive Supply Voltage

- Transient faults
  - Strengthing or increasing design margins
  - Error detection and correction codes
  - Bit/Line interleaving
  - Scrubbing

Departament d'Arquitectura de Computadors
UNIVERSITAT POLITÈCNICA DE CATALUNYA

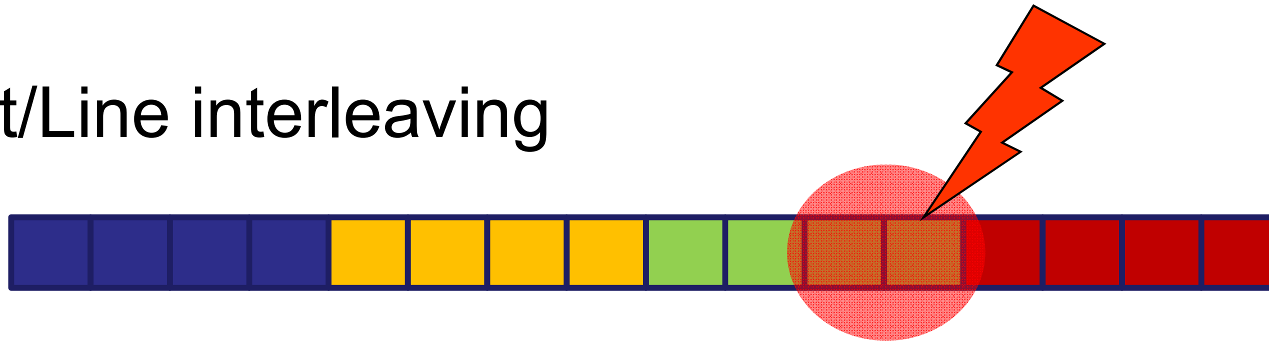
Transient Faults

- Causes:
  - Particle strikes: neutron (sun radiation), alpha particles (lead – Pb- in package)
    - Spacecraft Cassini–Huygens, launched in 1997, @Jupiter at the moment measured –in the vicinity of the Earth- 280 bitflips on a quiet day and >3000 on a weak solar flare. (~300MB of memory)
    - Google reported 5 single bit-flips in 8 Gigabytes of RAM per hour.
  - Voltage droops
  - Current droops
  - Temperature

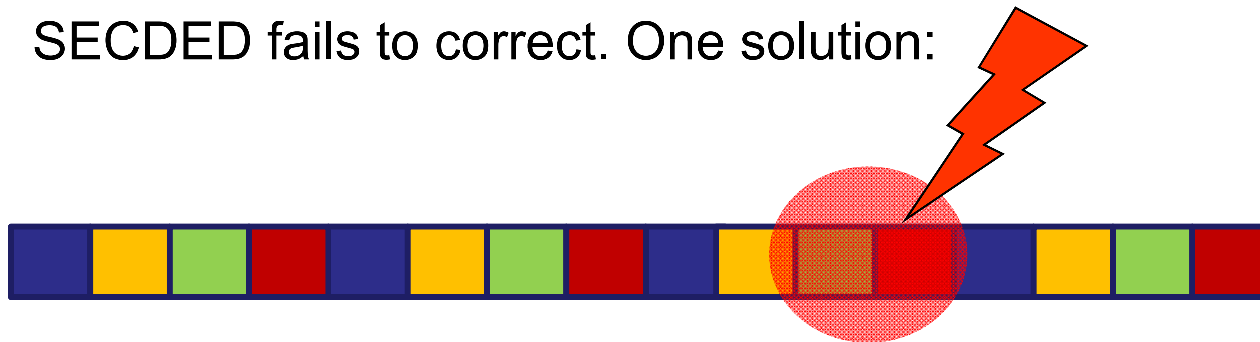
Transient Faults

- Strengthening or increase design margins
  - Make transistors bigger (Moore's Law slowdown)
  - Insert extra capacitors (extra area/power…)

- Bit/Line interleaving



SECDED fails to correct. One solution:

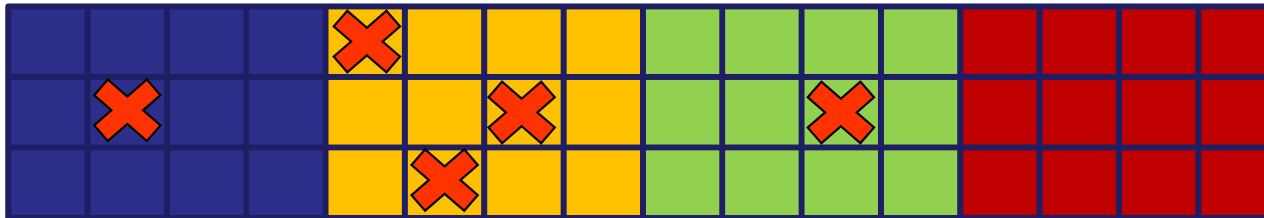


SECDED works!! ... but routing...

UPC
Departament d'Arquitectura de Computadors
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Transient Faults

- Scrubbing
  - Periodically read-correct-write data to avoid the accumulation of errors



After scrubbing:

Classification & Countermeasures

- Permanent faults
  - Disable too faulty blocks
  - Redundancy:
    - Replace blocks (i.e. Need to provide spares)
    - Dual-modular/Triple-modular redundancy
  - Error detection and correction codes
  - Wordline Boosting
  - Adaptive Body Biasing
  - Adaptive Supply Voltage

- Transient faults
  - Strengthing or increasing design margins
  - Error detection and correction codes
  - Bit/Line interleaving
  - Scrubbing

Technology Consequences

- Variation-centric design
  - Probabilistic timing and optimization
  - Variation-tolerant circuit styles
  - Microarchitecture support
- Dynamic adaptation
  - Reduce design margins required
  - Supply voltage droops
  - Temperature changes
  - Reliability effects

Getting practical

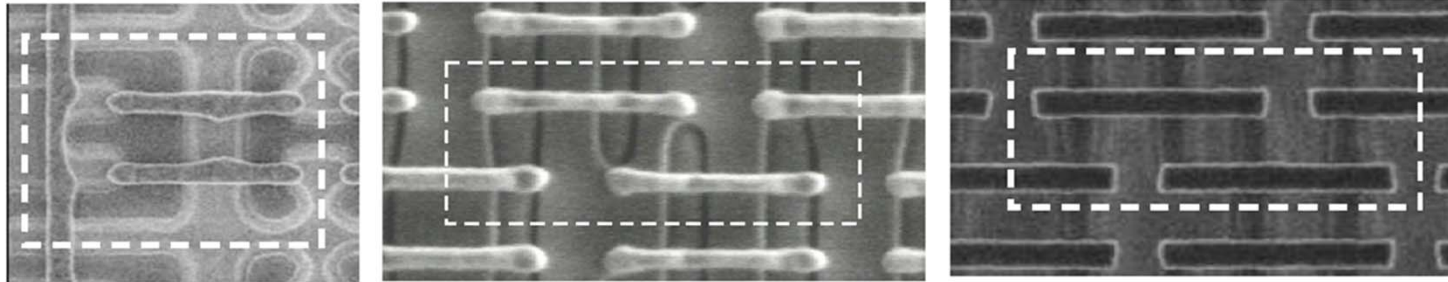
- Parameter distribution (Monte-Carlo)
  – Assume a percentage of variation
    - L~10% and Vt~10%
  – Each chip will be one set of combinations
- Correlation (variations come in “groups”)
  – Yes (pattern)
  – No (Random)
- Run Spice simulations to get characteristics of the modules

Conclusion

- Impact of variations is increasing
  - Aggressive feature size scaling
  - Increased die size and number of paths
  - Design margins are increasing
- Variation-aware design needed
  - Variation-tolerant circuits
  - Post-silicon tuning techniques
  - Supported by device, circuit, and microarchitecture innovations

Case study: Effect of process and temperature variations in SRAMs

PV Impact on 6T SRAMs



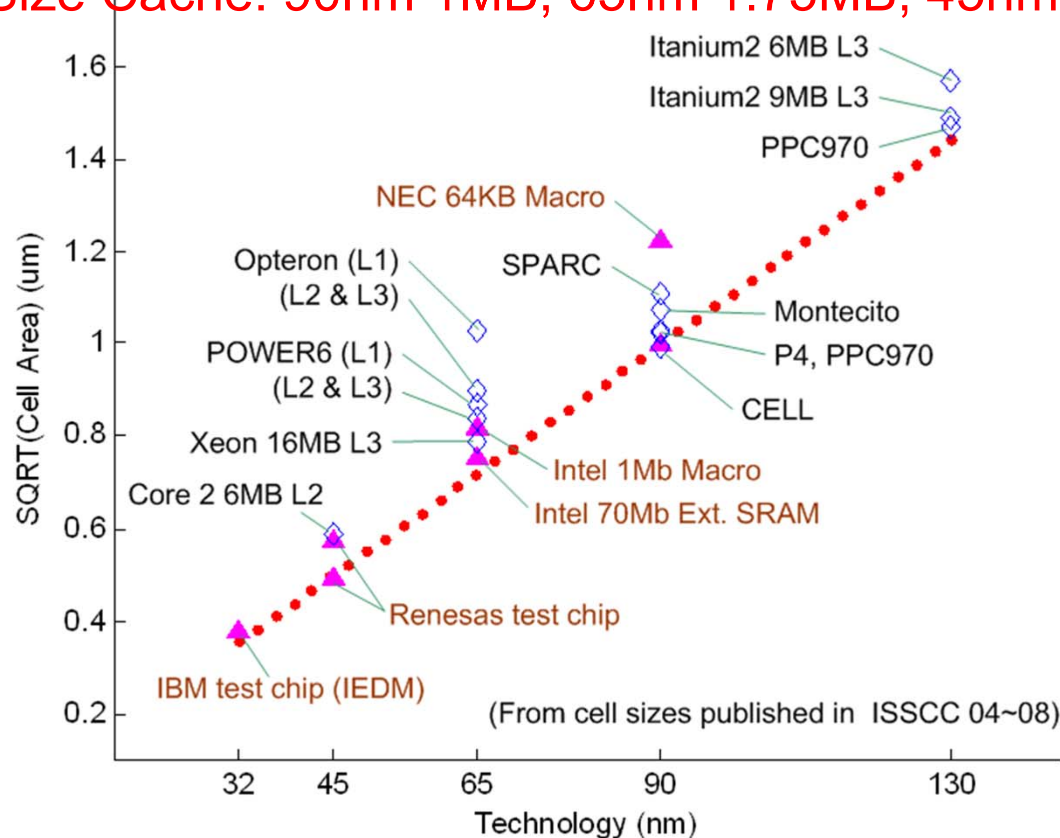
Intel SRAM cell size scaling (IEDM'07)

90nm – tall
1.0 μm²

65nm – wide
0.57 μm²

45nm – wide
w/ patterning enhancement 0.346 μm²

Deviation from Ideal Scaling: 90nm-> 0%, 65nm->14%, 45nm->39%
Equal Die Size Cache: 90nm 1MB, 65nm 1.75MB, 45nm 2.89MB



PV Impact on 6T SRAMs

SQRT (Cell Area) (um)

1,8
1,6
1,4
1,2
1
0,8
0,6
0,4
0,2
0

Itanium 2 6MB L3
Itanium 2 9MB L3
PPC970
NEC 64KB Macro
SPARC
Opteron (L1)
Opteron (L2 & L3)
CELL
Montecito
P4, PPC970
POWER6 (L2 & L3)
POWER6 (L1)
Xeon 16MB L3
Intel 1MB Macro
Intel 70Mb Ext.
Core2 6MB L2
Reneseas
Test Chip
IBM Test Chip (IEDM)
TSMC
Intel
AMD
Ivy Bridge
IBM
TSMC
Haswell
Intel

22
32
45
65
90
130

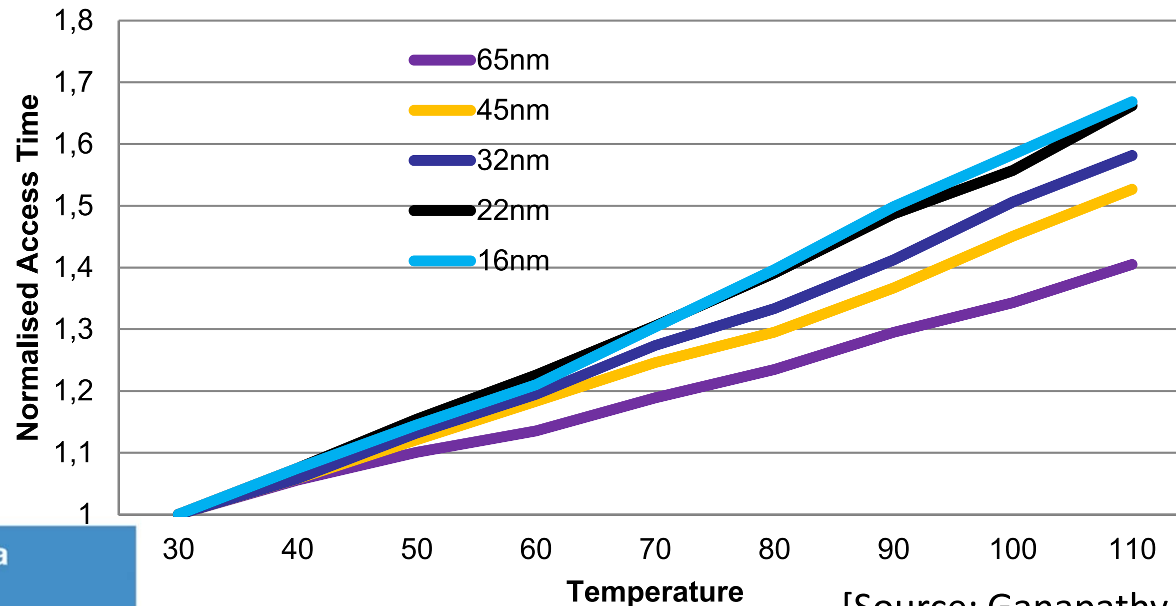
Technology (nm)

Departament d'Arquitectura
de Computadors
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Impact of Temperature Variations

- Sudden temperature shoot-ups can result in functionality and reliability problems.
- Their value dependent on
  - Spatially – Hotspot generation
  - Temporally – Computed workload and dissipated power
- Threshold voltage degradation and mobility reduction with increase in temperature

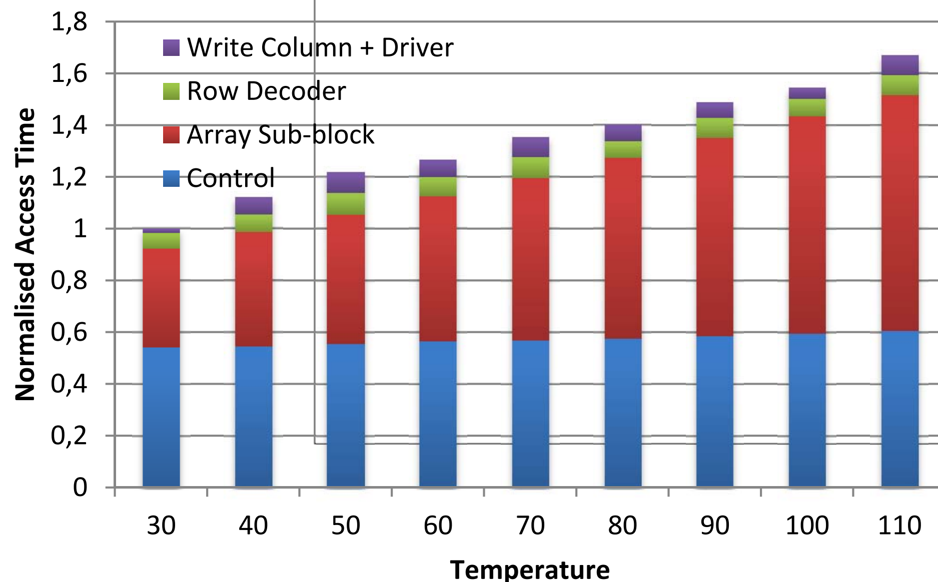
Cache Access Time Vs Temperature

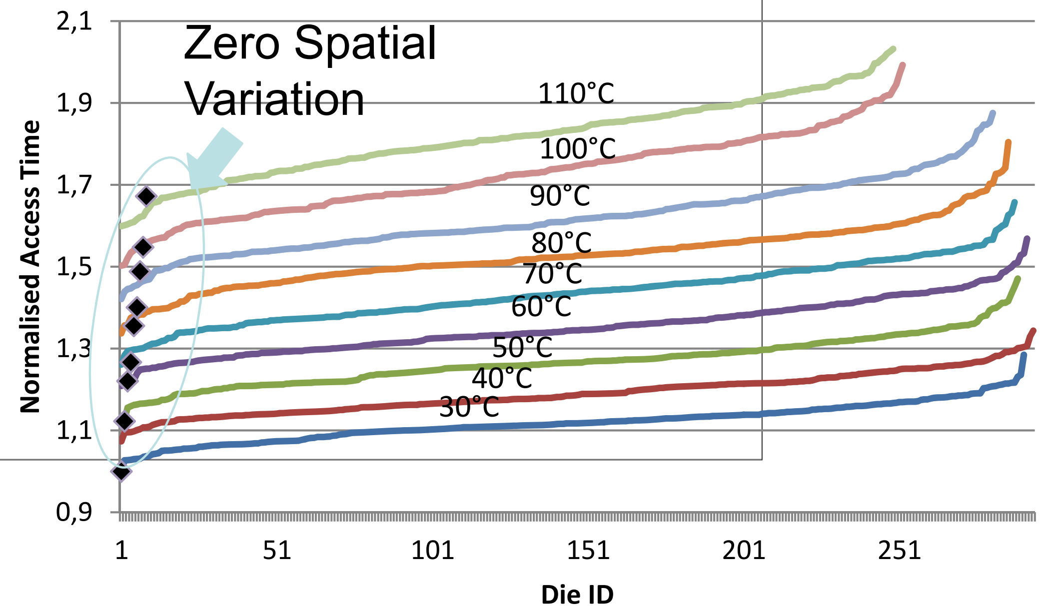


[Source: Ganapathy et al., DATE 2010]

Variation in Cache Memories

- **With reduction in feature size, memory cells will be affected the most from Spatio-Temporal variations**
  - **Intrinsic device level variations will dominate over other forms of variation**
  - **Reducing size means reducing Q_{crit}, making it susceptible to Hard Errors**
- **In addition to Memory cells, Periphery which form the core of the system will be affected**





[Source: Ganapathy et al., DATE 2010]

Conclusions

- Multi-level approaches to variations
  - Manufacturing: preciseness, materials, tools, ...
    - No definitive solution → move to next level
  - Circuit: Body Biasing, Source Biasing
    - Robust circuits (i.e. reduce scaling)
    - No definitive solution → move to next level
  - Architecture: Error correcting/detecting structures. Resilient architectures.
    - Extra hardware for correctness, not for performance.
    - So far good enough…but…