Survival of VLSI design – coping with device variability and uncertainty

Kevin Nowka
Sr Mgr VLSI Systems
IBM Austin Research Laboratory

Acknowledgements:
Sani Nassif, Anne Gattiker (IBM Austin Research)
Chandu Visweswariah, David Frank (IBM Watson Research),
Lars Liebmann, Dan Maynard (IBM Server & Technology Group)

Motivation for overcoming variation (or at least coping)?

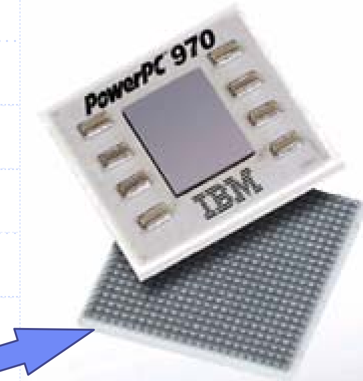
What is at stake? The VLSI economy

to make these..

Very Large Scale Integration is:

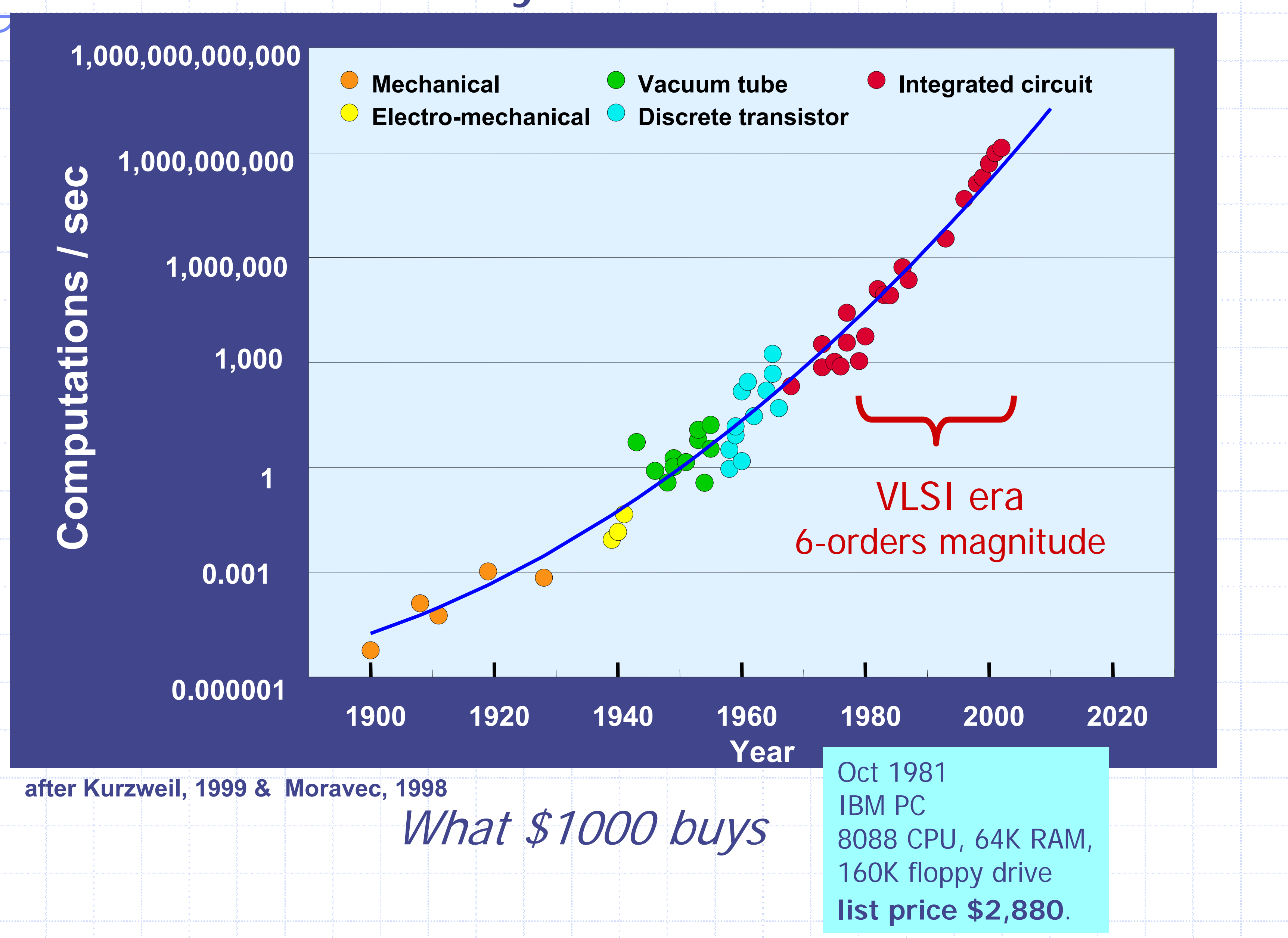
to make these..

Using greater than 10k of these..



to make these..

The VLSI Economy



after Kurzweil, 1999 & Moravec, 1998

What $1000 buys

Oct 1981
IBM PC
8088 CPU, 64K RAM,
160K floppy drive
list price $2,880.

The Secrets to this Success

- Resilient CMOS VLSI Devices & Interconnect
- Simple Design Processes
  - *Physical Abstraction* with small number of rules
    - Simple design and design migration
    - Composable designs
  - *Functional Abstraction*
  - *Resulting predictable functional & timing behavior*
    - Cell-based design, place & route, static timing
- Scaled Lithography (and Manufacturing Process Improvements)
  - Lithography improvements and the application of Dennard Scaling Rules enabling Moore's Law

65nm technology and beyond

- Is the VLSI Economy in jeopardy because of "variability?"
  - What is variability?
  - What are the important sources of variability?
  - What are the effects on VLSI design?
  - How are fundamental design processes impacted?
  - How can we cope?

What is "variability"

- Intending to build this....
- And sometimes (or someplaces) getting this..
- And sometime (or some places) getting this

Variability and Uncertainty

- Variability: known quantitative relationship between design behavior (eg. current, delay, power, noise-margin, leakage, …) and a source
  - Relationship can be accurately modeled, simulated, and compensated.
  - eg. Conductor thickness as function of interconnect density.
- Uncertainty: sources unknown or model too difficult/costly to generate or simulate
  - must be “budgeted” with some type of worst case analysis
  - eg. Vt as a function of dopant dose and placement
- Lack of modeling resources often transforms variability to uncertainty.
  - eg: deterministic circuit switching activity factor

Some Classes of VLSI Variability

Physical

- Changes in characteristics of devices and wires (manufacturing & aging). Time scale: 10^9 sec (years).

Functional

- Changes in characteristics due to application cycles or workload changes. Time scale: 10^7 to 10^{-6} sec (execution time)

Environmental

- Changes in supply voltage, temperature, local noise coupling. Time scale: 10^{-3} to 10^{-9} sec (clock tick).

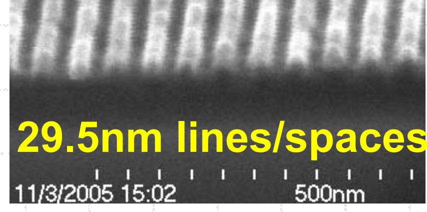
Informational

- Lack of knowledge about design due to inadequate modeling. Time scale: ignorance cannot be measured in units of time.

Lithography induced variability

Subwavelength lithography

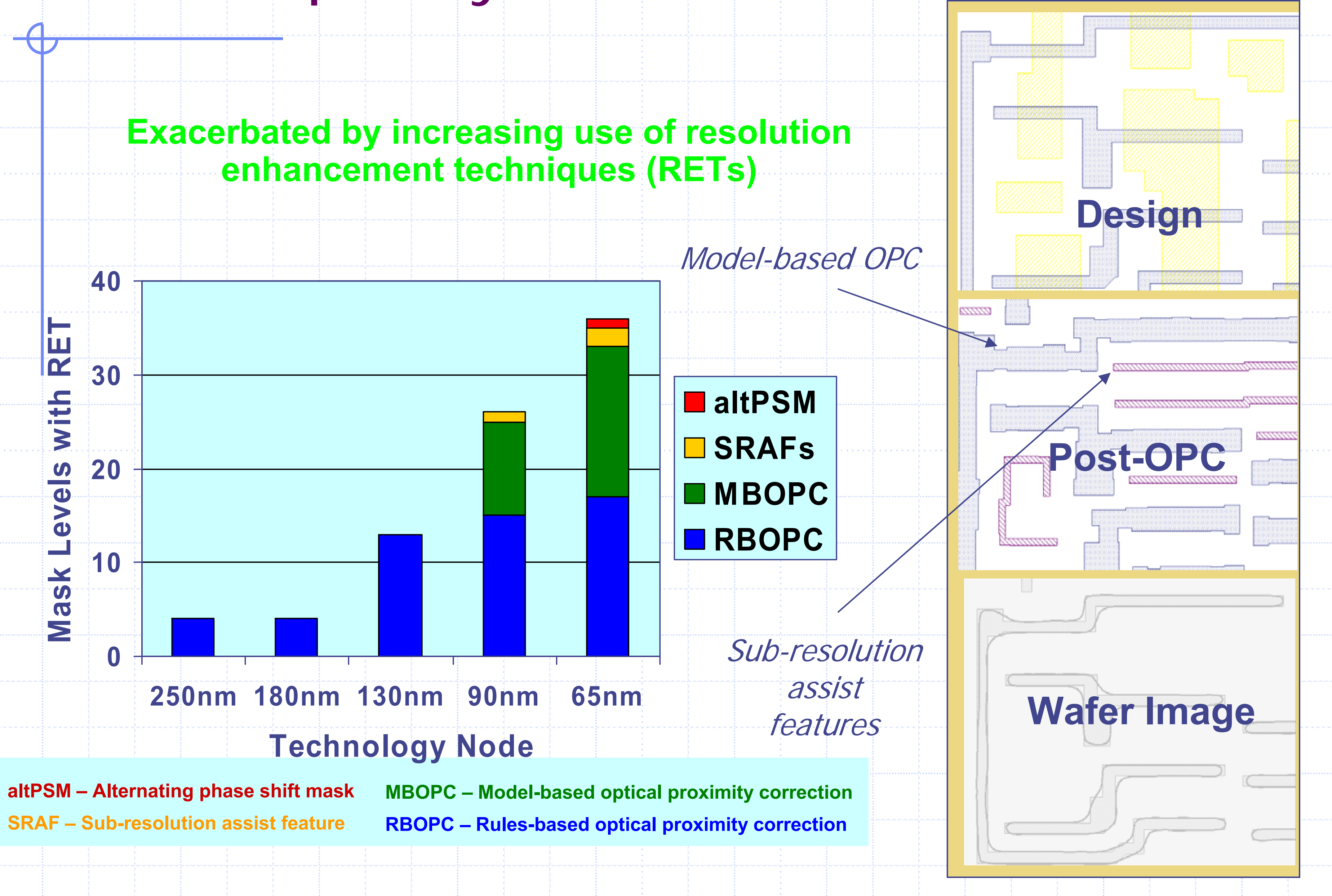
- Using 193nm light to create <30nm features



Imperfect Process Control

- Critical Dimensions are sensitive to:
  - focus
  - dose (intensity and time)
  - resist sensitivity (chemical variations)
  - layer thicknesses
- Intensity affected by interference
  - strongly dependent on layer thicknesses.
  - Anti-reflection coatings help
- Errors in Alignment, Rotation and Magnification:
  - Result in either global or local shape-dependent device variations.

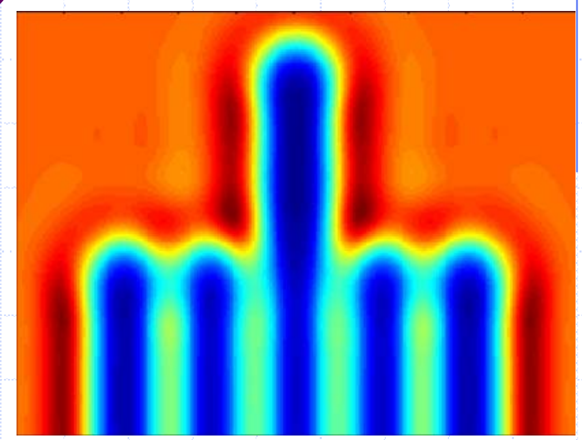
Mask Complexity Continues to Escalate



Lithography induced variability

Imperfect Process Control (cont'd)

- Pattern sensitivity.
  - Interference effects from neighboring shapes.
    - Predominantly in same plane
    - Some buried feature interference for interconnect

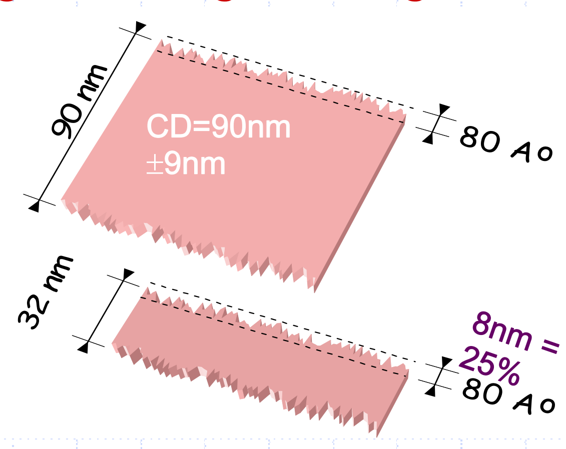


[T. Brunner, ICP 2003]

Line-edge roughness

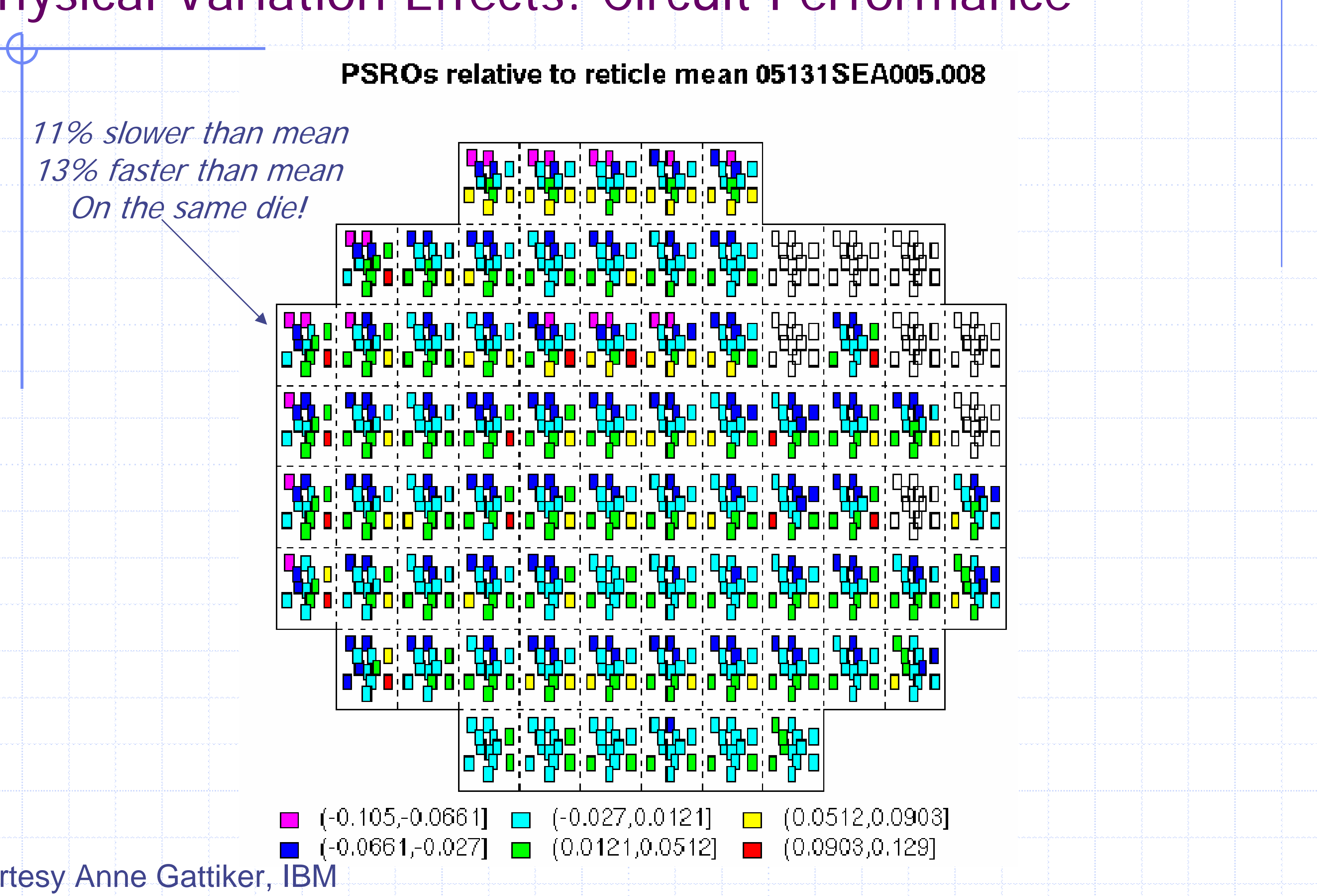
- Sources of line-edge variation
  - Fluctuations in the total dose due to finite number of quanta
    - Shot noise
- Fluctuations in the photon absorption positions
  - Nanoscale nonuniformities in the resist composition
- With decreasing feature size, a larger percentage of Lpoly has LER randomness
  - Impact delay and leakage power

Significant gate length uncertainty



Source: D. Frank, VLSI Tech 99

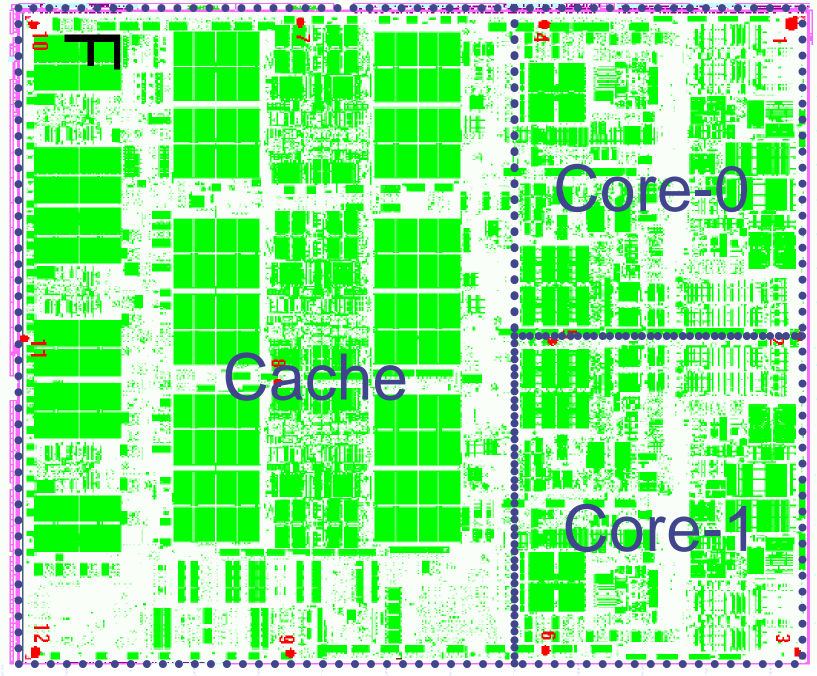
Physical Variation Effects: Circuit Performance



Courtesy Anne Gattiker, IBM

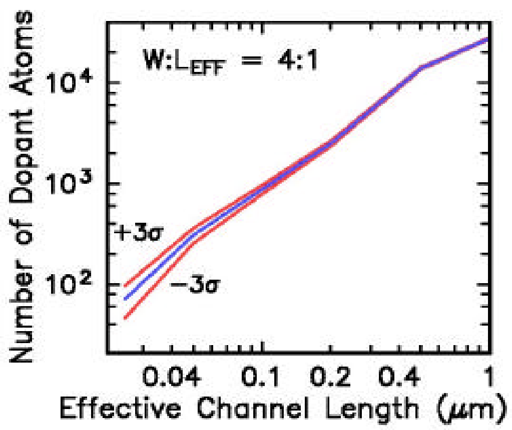
Variation Effects: Not just ring oscillators....Real Microprocessors

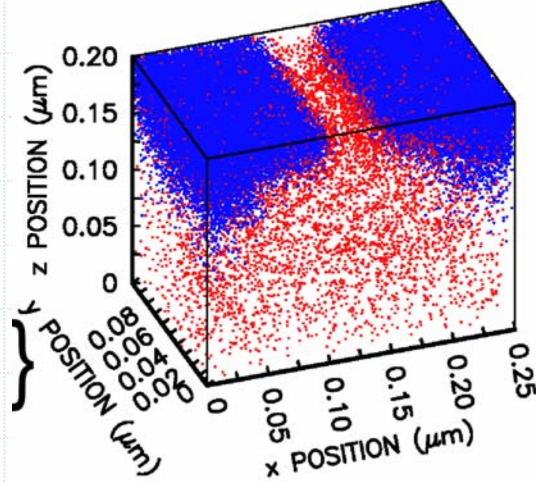
- Multicore design -- Core-0 was found to be ~15% slower than other parts.
- Models predict all parts of the design are identical.



Random Dopant Fluctuation

- Threshold Voltage is dependant upon the doping within a device channel area.
  - The number of dopant atoms in the depletion layer of a MOSFET has been scaling roughly as $Leff^{1.5}$.
  - Statistical variation in the number of dopants, N, varies as $N^{1/2}$, causing increasing Vt uncertainty for small N.



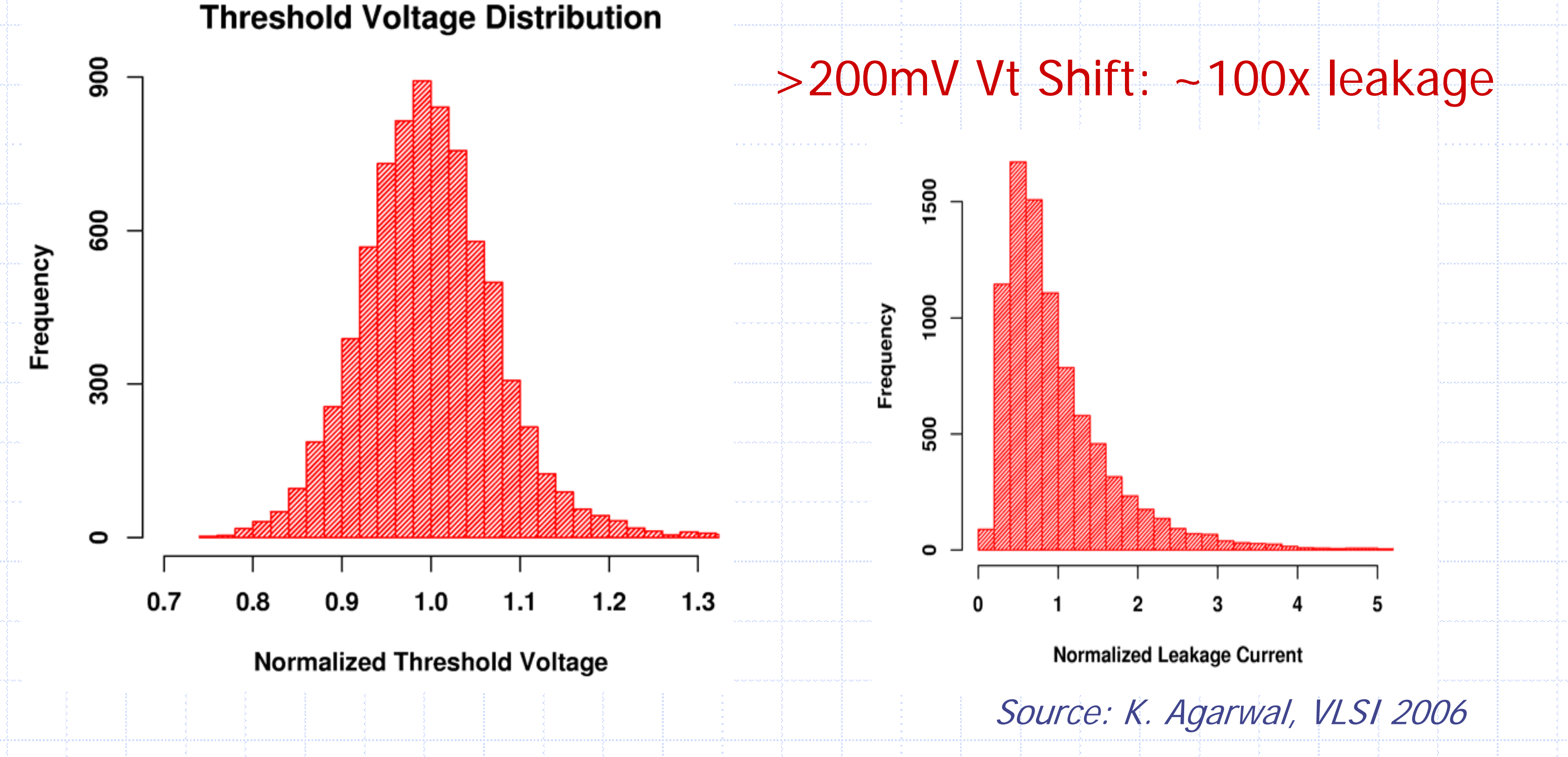


Source: D. Frank, et al, VLSI Tech 99, D. Frank, H. Wong IWCE, May 2000]

>200mV Vt Shift

Random Dopant Fluctuation Effect

- Performance, power, and leakage variation



Source: K. Agarwal, VLSI 2006

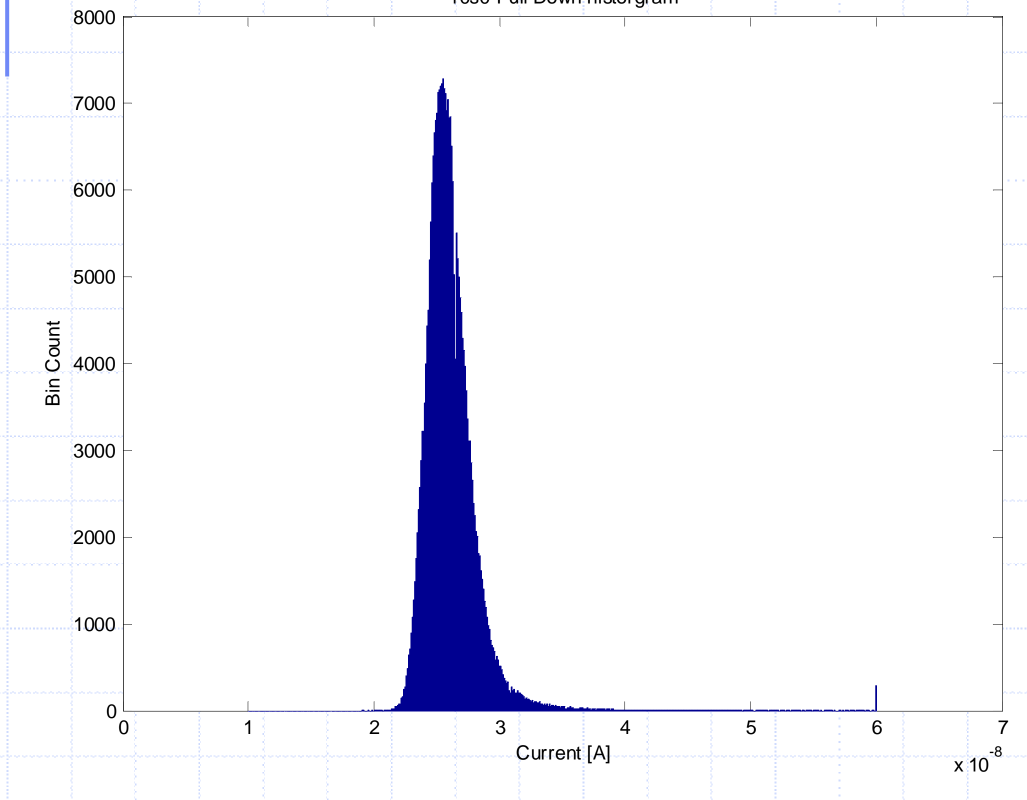
NBTI and Hot-carrier-induced Variation

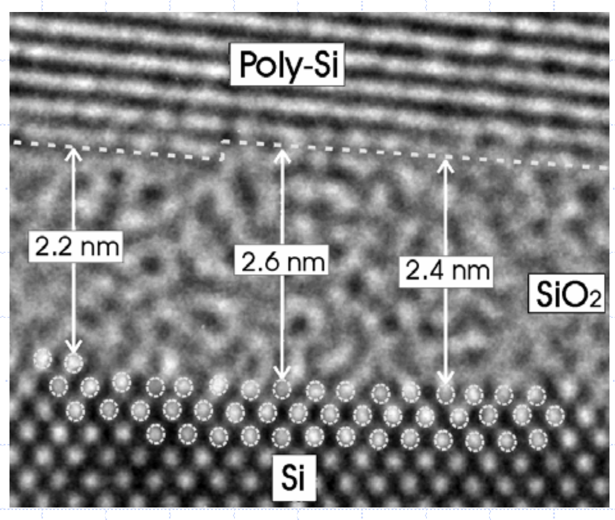
- Negative Bias Temperature Instability
  - At high negative bias and elevated temperature the pFET Vt gradually shifts more and more negative (reducing the pFET current).
    - The mechanism is thought to be the breaking of hydrogen-silicon bonds at the Si/SiO2 interface, creating surface traps and injecting positive hydrogen-related species into the oxide.
    - Associated with the average NBTI shift, there are also random shifts, which even for identical use conditions and devices, will cause mismatch shifts due to random variations in the number and spatial distribution of the charges/interface states formed.
- There are also other charge trapping and hot-carrier defect generation mechanisms that cause long-term Vt shifts in both nFETs and pFETs.
- Long-term Vt shifts are parameter variations that must be accounted for in the design of circuits.

N. Rohrer, ISSCC 2006

Gate Oxide Thickness Fluctuation

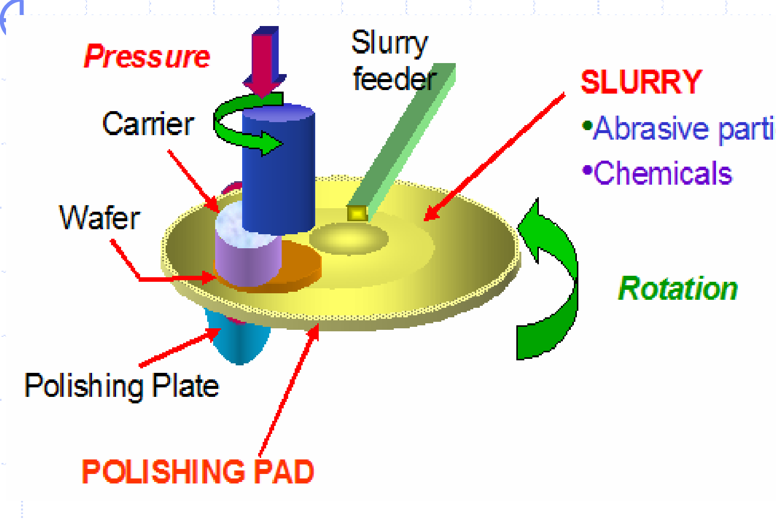
- Gate oxide variation
  - Exponential effect on gate tunneling currents
  - Affects device threshold, but significantly less important Vt variation factor than random-dopant fluctuation





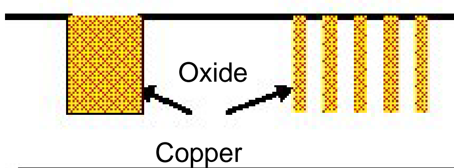
1.1nm oxide is ~6 atomic layers. across a 300mm wafer (>10^9 atomic layers)

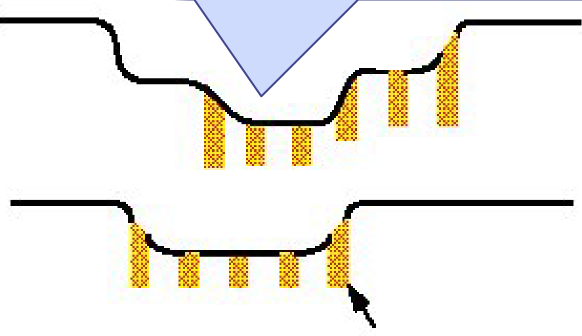
Back-end Variability -- CMP

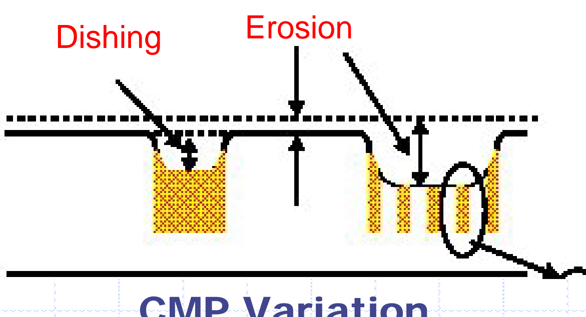


- Chemical/Mechanical polishing
  - Introduces large systematic intra-layer interconnect thickness
  - Additional inter-layer interconnect thickness effects as well

Topography variation translated into focus variation for lines which results in width variation



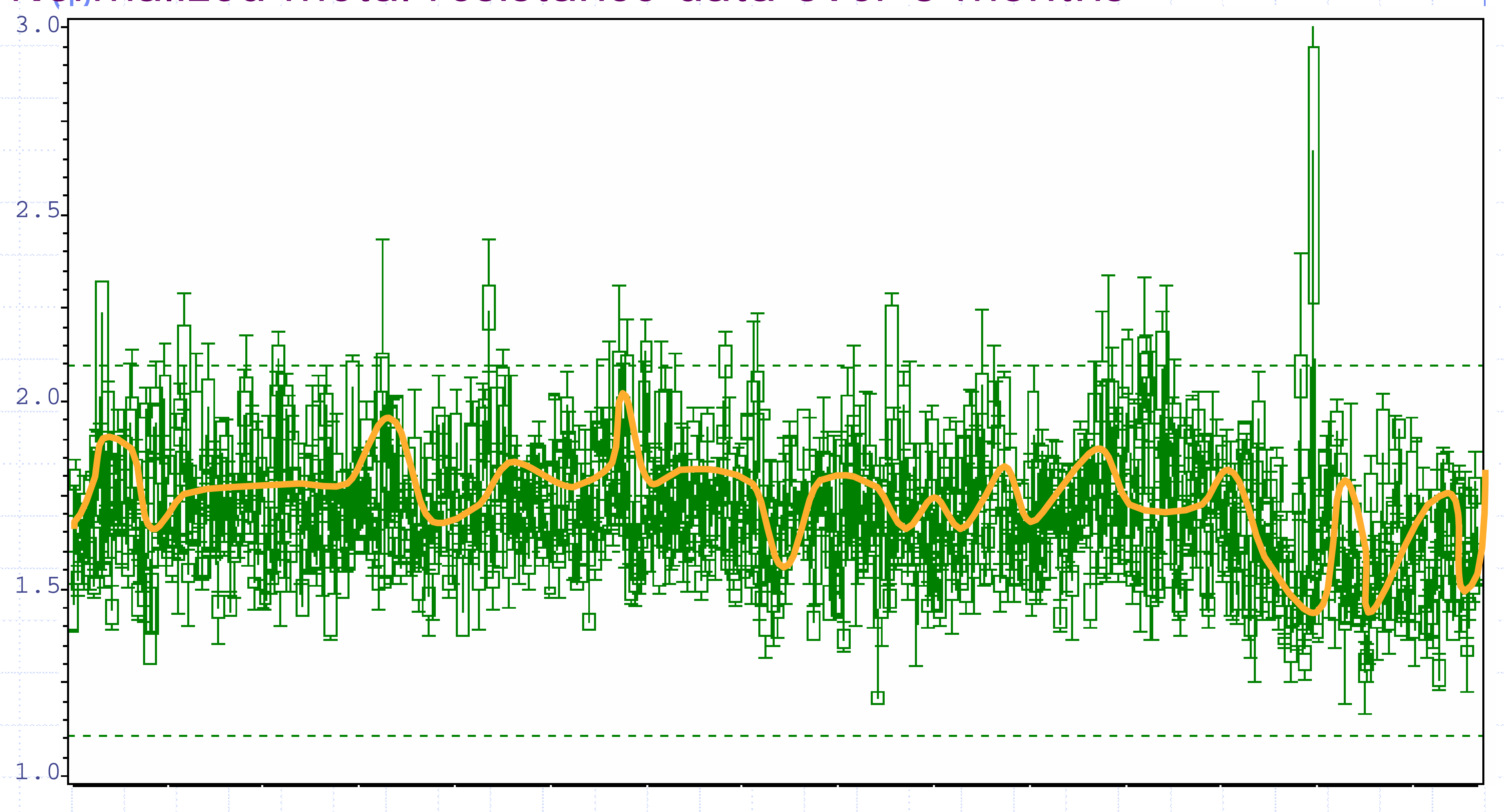




CMP Variation

Measured Variation: interconnect performance

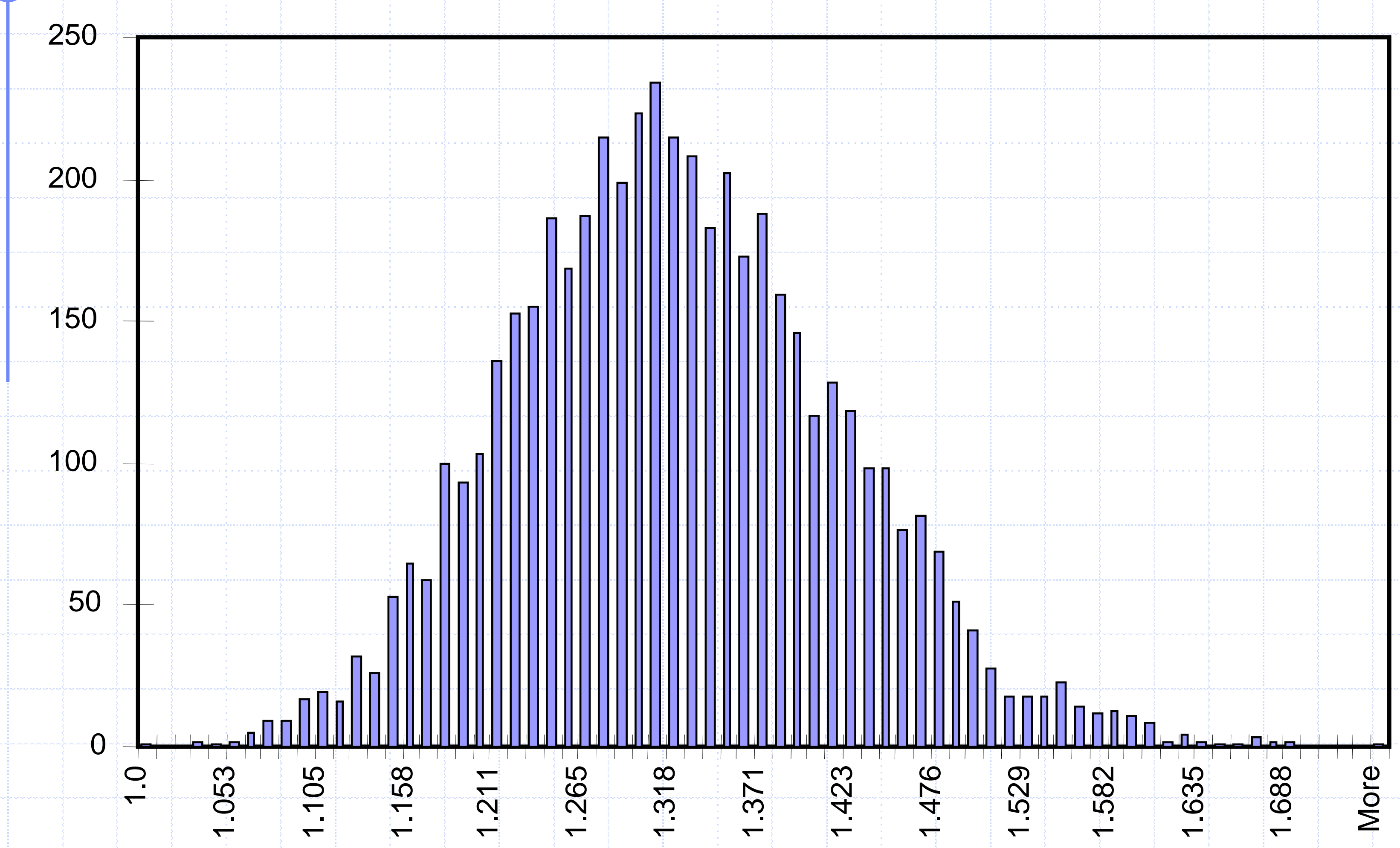
Normalized metal resistance data over 3 months



- Wafer means change over time
- Some real outliers

Source: Chandu Visweswariah, C2S2 Robust Circuits Wkshp, 7/28/06

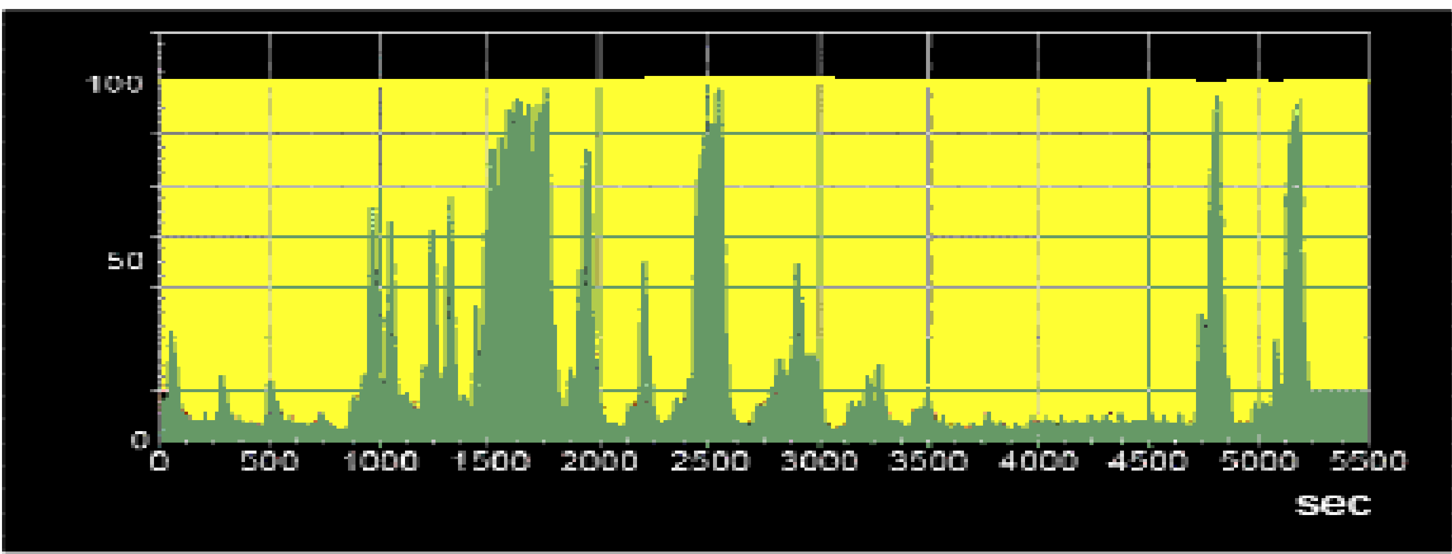
Normalized single-level capacitance distribution



- Variability is enormous!

Source: Chandu Visweswariah, C2S2 Robust Circuits Wkshp, 7/28/06

Functional Variation

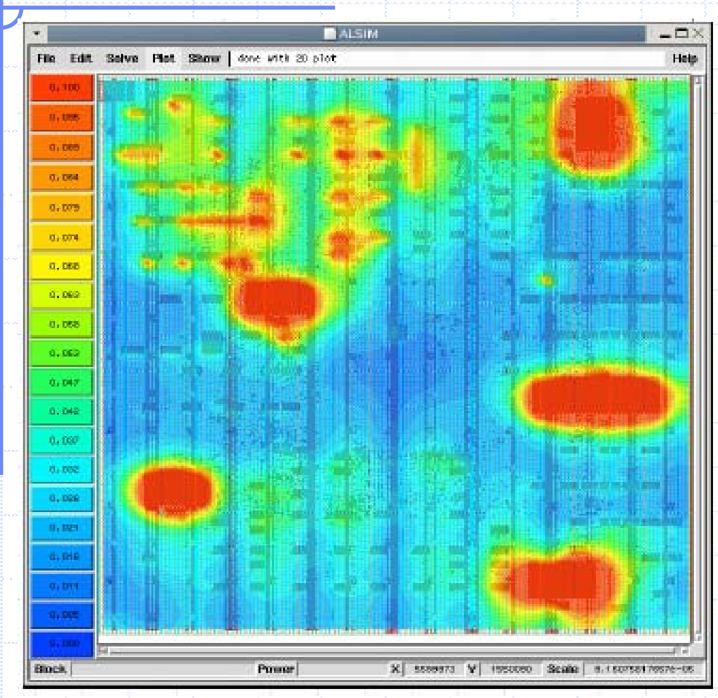


~5% to ~95% time

Source: J. Fredrich, ACEED '07

- Workload variability – utilization of design based on changing workload requirements

Environmental Variation – Supply voltage



Power supply droop map

Source: Sani Nassif, IBM

VDD

Chip Y (mm)

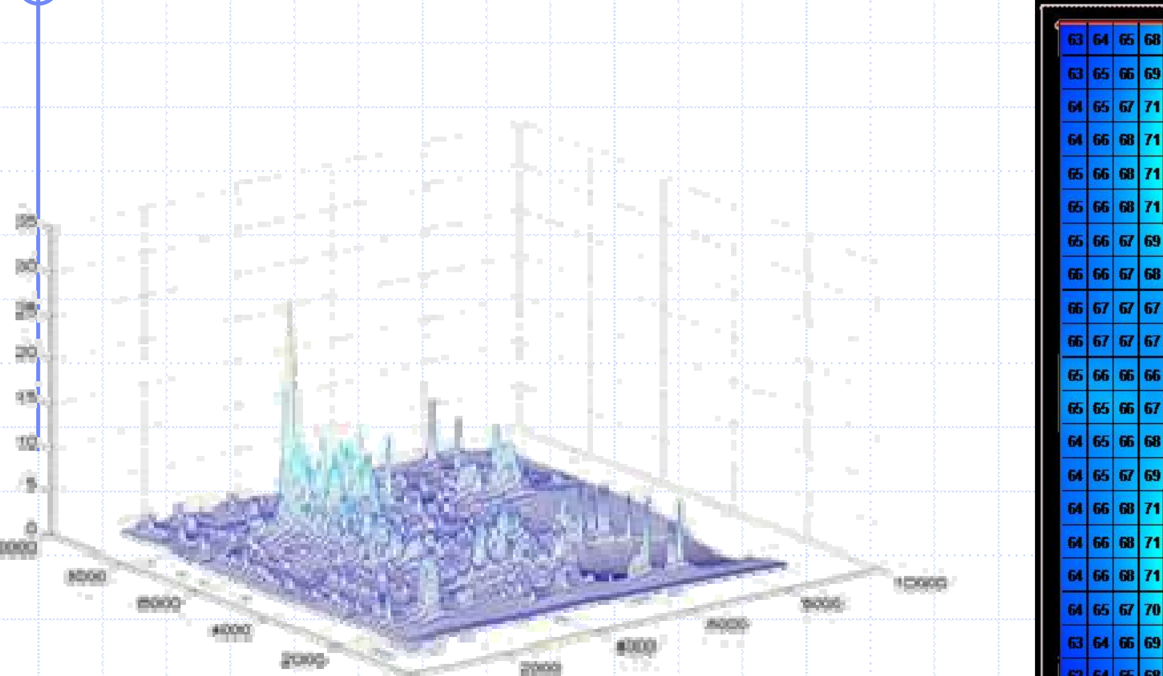
Chip X

Source: P. Restle, ICCAD06, IBM

>10% dynamic supply droop

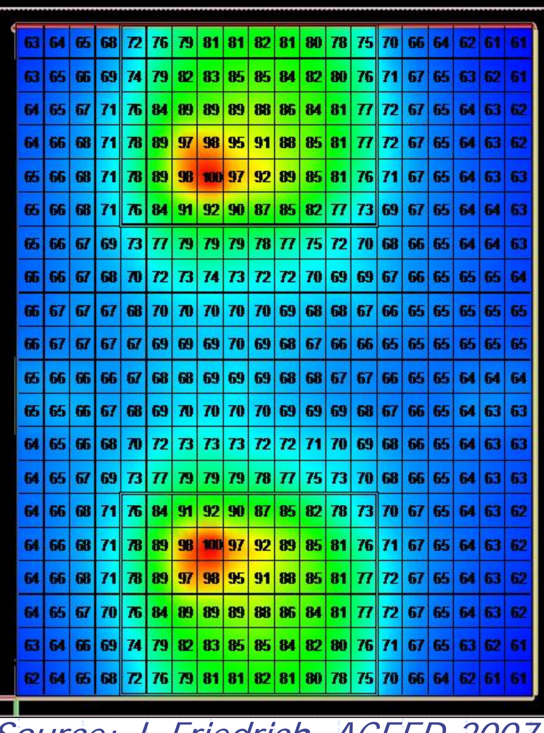
- Supply variation due to input variation (eg. battery lifecycle) and self-generated and coupled supply noise
- Supply variation affects performance, power, reliability

Environmental Variation – Thermal



Die thermal map

Source: Sani Nassif, IBM



Source: J. Friedrich, ACEED 2007

~30C dynamic temperature variation

- Thermal variation due to ambient fluctuation and self-heating
- Thermal variation affects performance, reliability

45nm technology and beyond

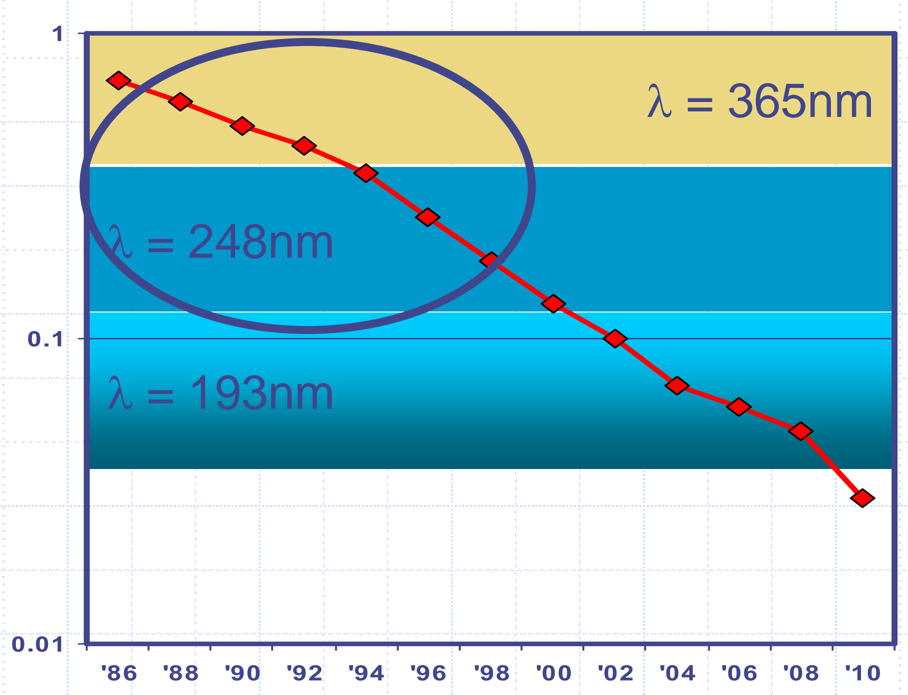
- Is the VLSI Economy in jeopardy because of "variability?"
  - What is variability?
  - What are the important sources of variability?
  - What are the effects on VLSI design?
  - How are fundamental design processes impacted?
  - How can we cope?

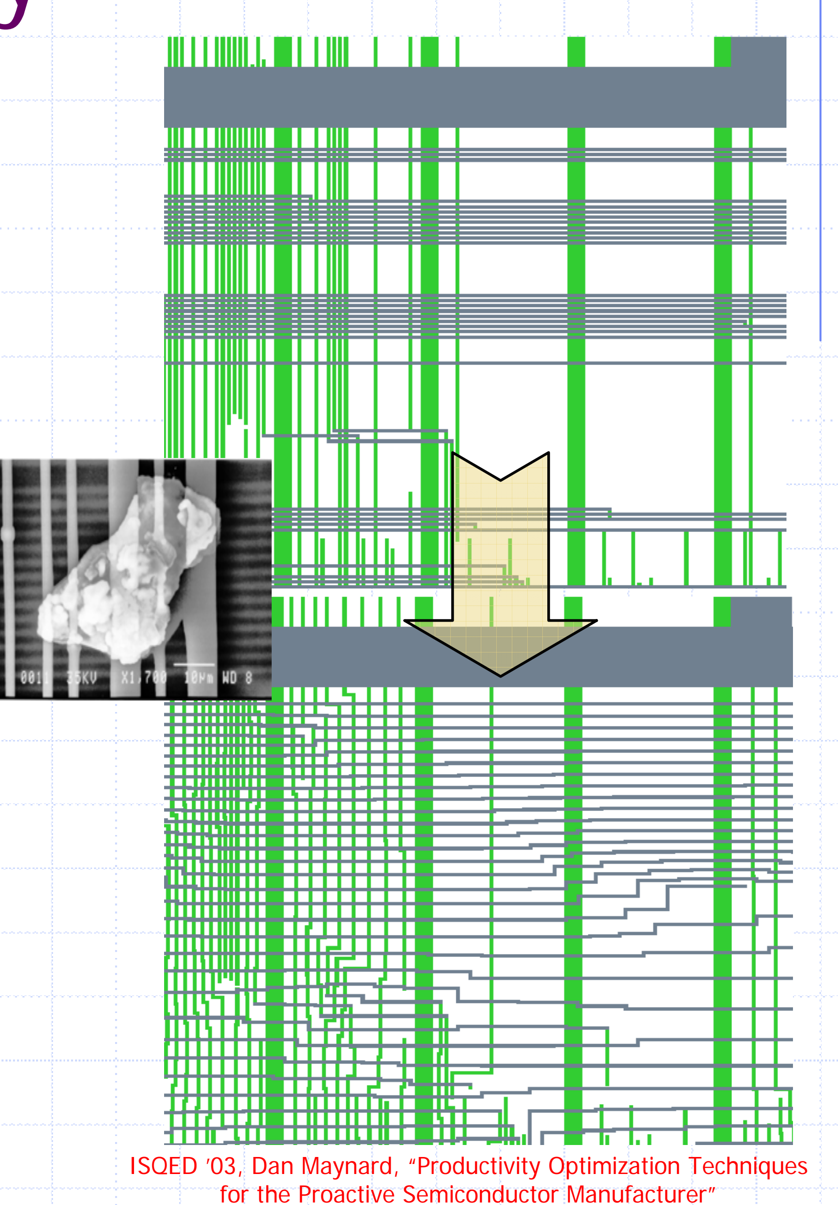
Revisiting....the Secrets to Success

- Resilient CMOS VLSI Devices & Interconnect
- Simple Design Processes
  - *Physical Abstraction* with small number of rules
    - Simple design concepts and design migration
    - Composable designs
  - *Functional Abstraction*
  - *Resulting predictable functional & timing behavior*
    - Cell-based design, place & route, static timing
- Scaled Lithography (and Manufacturing Process Improvements)
  - Lithography improvements and the application of Dennard Scaling Rules enabling Moore's Law

Technology Resiliency

Defects were the major yield detractors for technology in the early days, yield and area were the major tradeoffs.

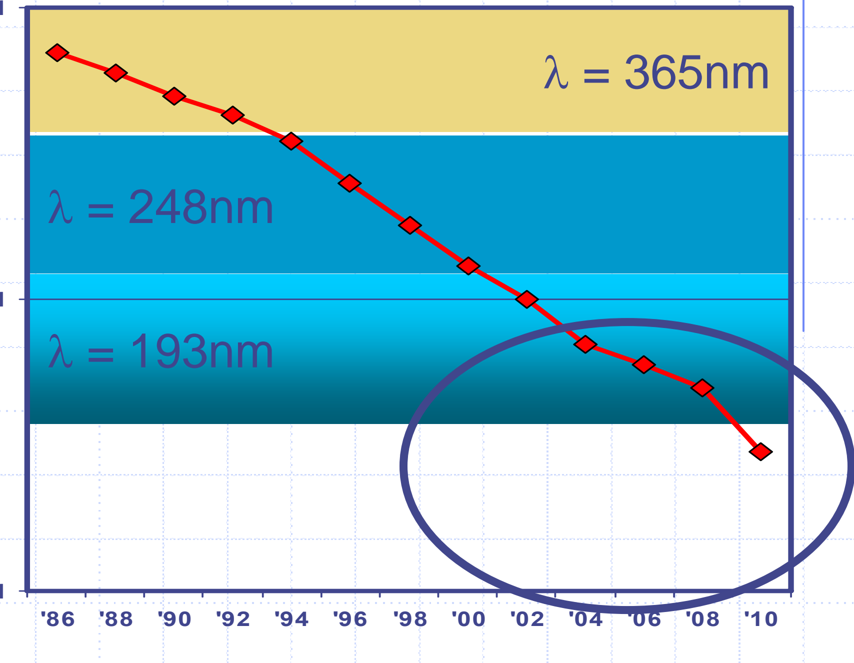




ISQED '03, Dan Maynard, "Productivity Optimization Techniques for the Proactive Semiconductor Manufacturer"

The Resiliency Problem

- With scaling, variability – both random and systematic – has emerged as a source of performance and yield loss.
  - This can be viewed as the merger of failure modes due to structural (topological), and parametric (variability) defects.
  - In the very near future, we will have to deal with circuits where a non-trivial portion of the devices simply do not work!



Distribution of "Good" contacts

Near futur distributio

Other factors, like the environment, make the failure region fuzzy and broad!

s that look e opens!

Distribution of "Bad" contact

1Ω 100Ω 1MΩ

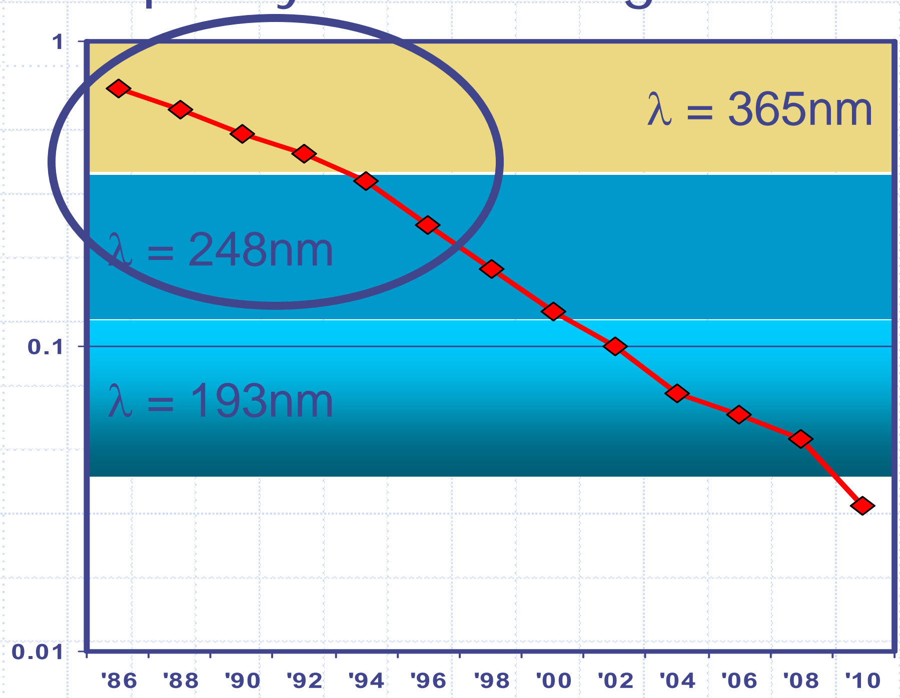
Contact Resistance

What has changed?

- Resiliency & redundancy cannot be **ignored**.
  - Need to start design assuming partial functionality!
- Key Factor: **Variability**

1980: Abstraction – the great enabler

With abundant performance, it became possible to abstract design to a few simple rules. Thus came the age of “chip computer science” and equality for all designers!



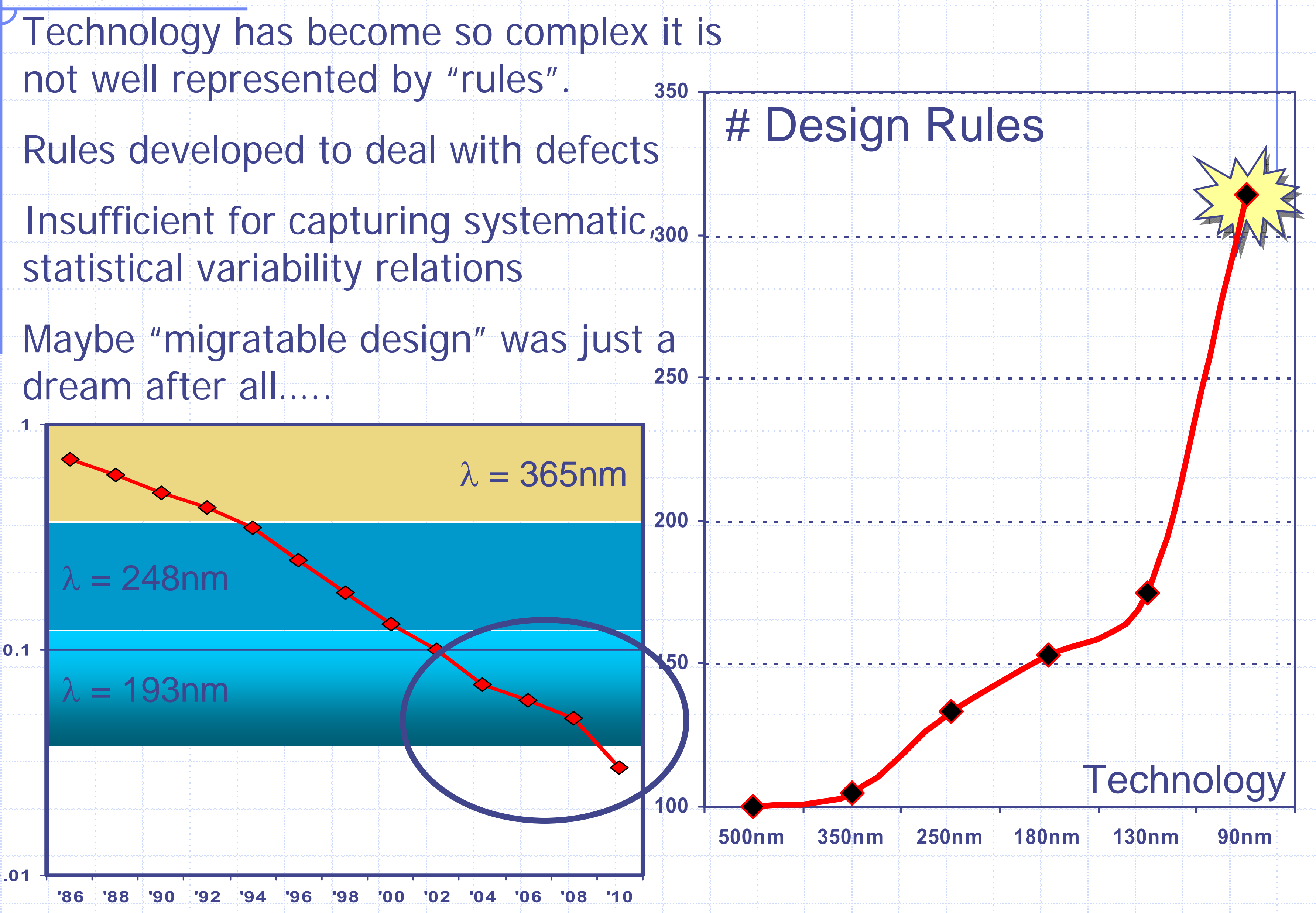
OCTOBER 20, 1981
ANNUAL TECHNOLOGY UPDATE ISSUE
Programmable VLSI forces software to the forefront, altering strategies in production, test, and equipment design / 114

Electronics

1981 AWARD FOR ACHIEVEMENT
LYNN CONWAY · CARVER MEAD

Physical Abstraction 2003: Abstract this!

- Technology has become so complex it is not well represented by "rules".
- Rules developed to deal with defects
- Insufficient for capturing systematic, statistical variability relations
- Maybe "migratable design" was just a dream after all.....

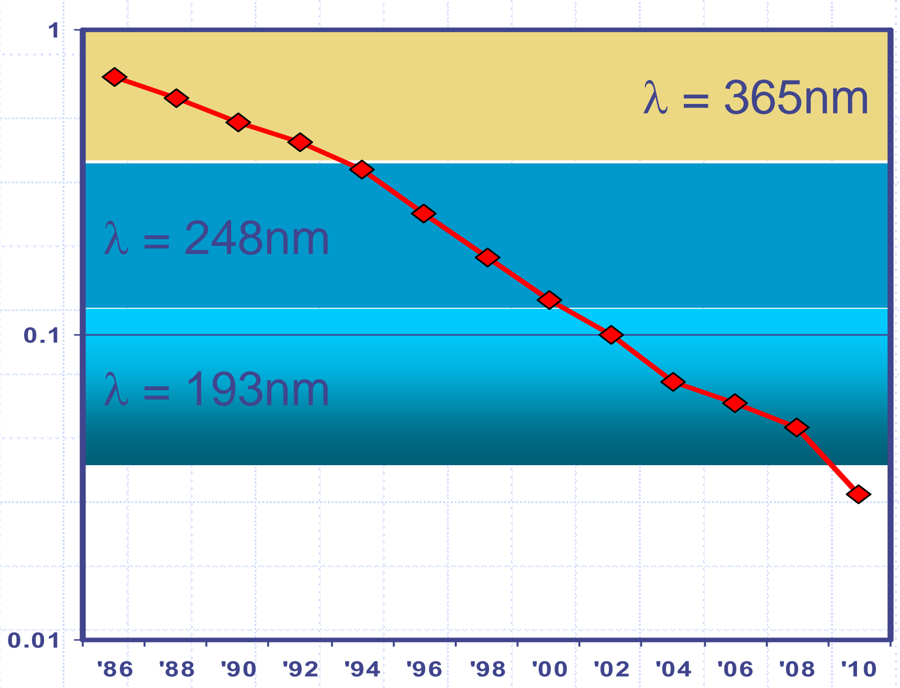


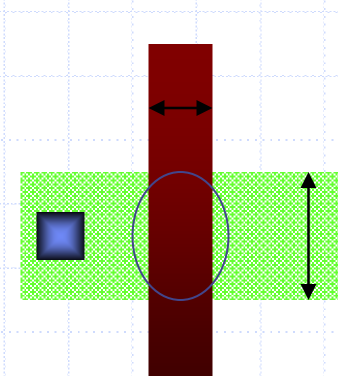
What has changed?

- Resiliency & redundancy cannot be **ignored**.
  - Need to start design assuming partial functionality!
- Mead-Conway design is **dead**…
  - Physical abstraction is broken – ground-rule explosion
  - Physical abstraction is broken – composability in jeopardy
  - Functional abstraction is broken – increasingly difficult to treat these as "logic devices"
  - Transistor performance determined by new features and phenomena, ∴ large variety in behaviors (not easily bounded).
- Key Factor: **Variability**

Litho and Physical Abstraction ca. 1990

Before the advent of deep sub-wavelength lithography, the salient properties of a transistor were determined by geometries very local to the device itself!

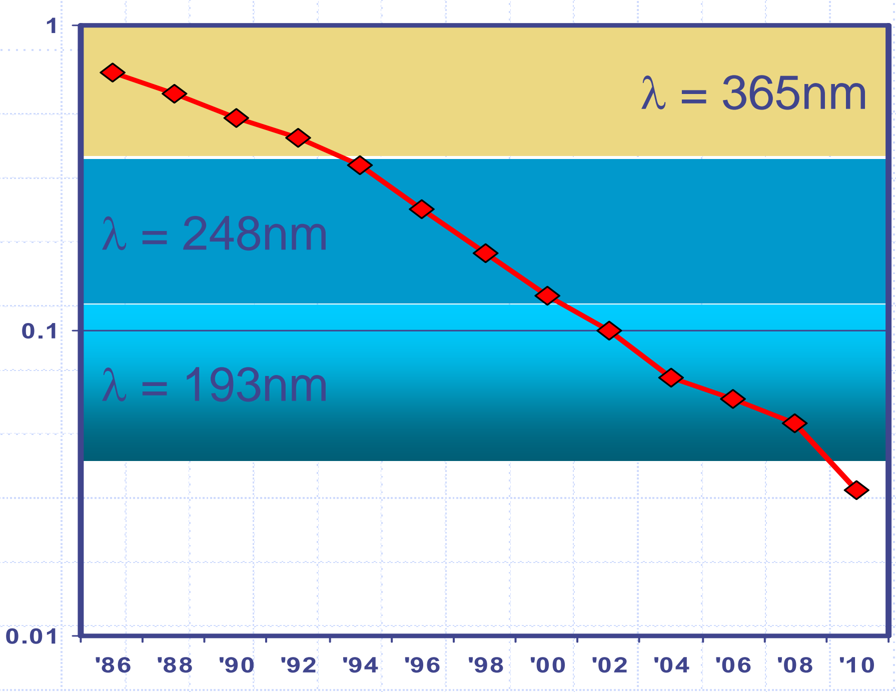


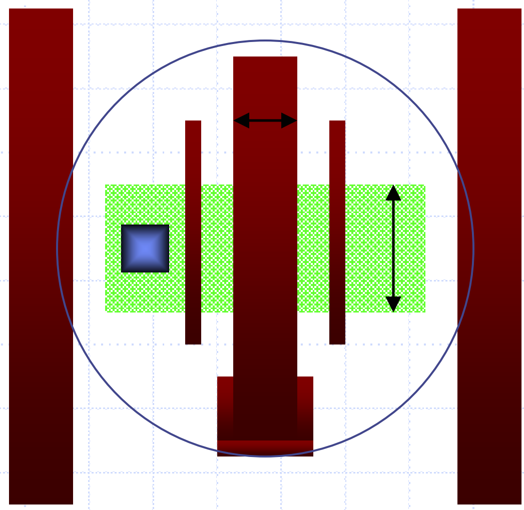


1990

Litho and Physical Abstraction ca. 2000

As scaling required resolution enhancement and optical proximity correction, the number of shapes that determine the final outcome increased.

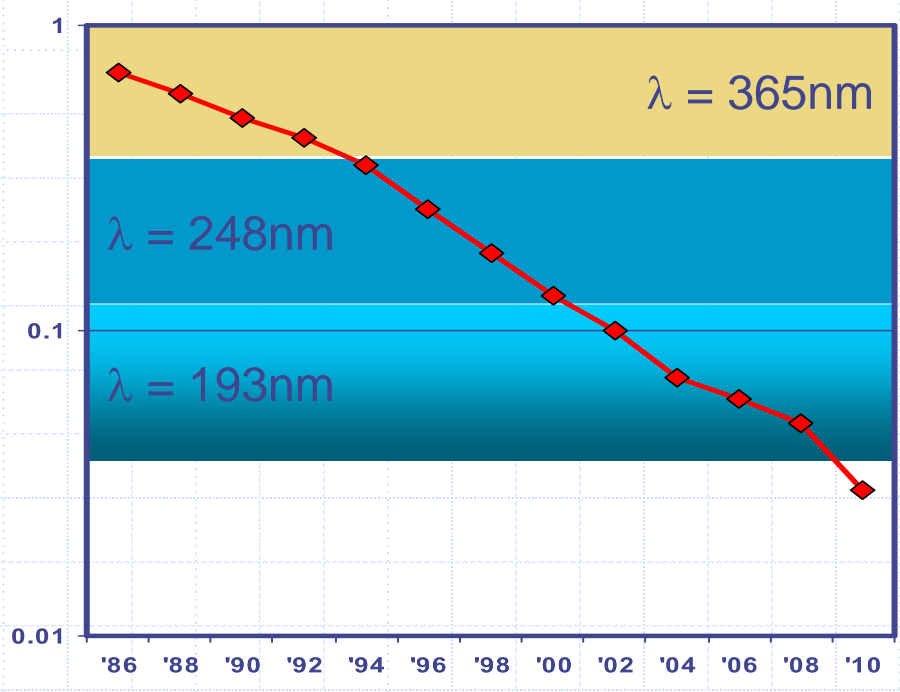


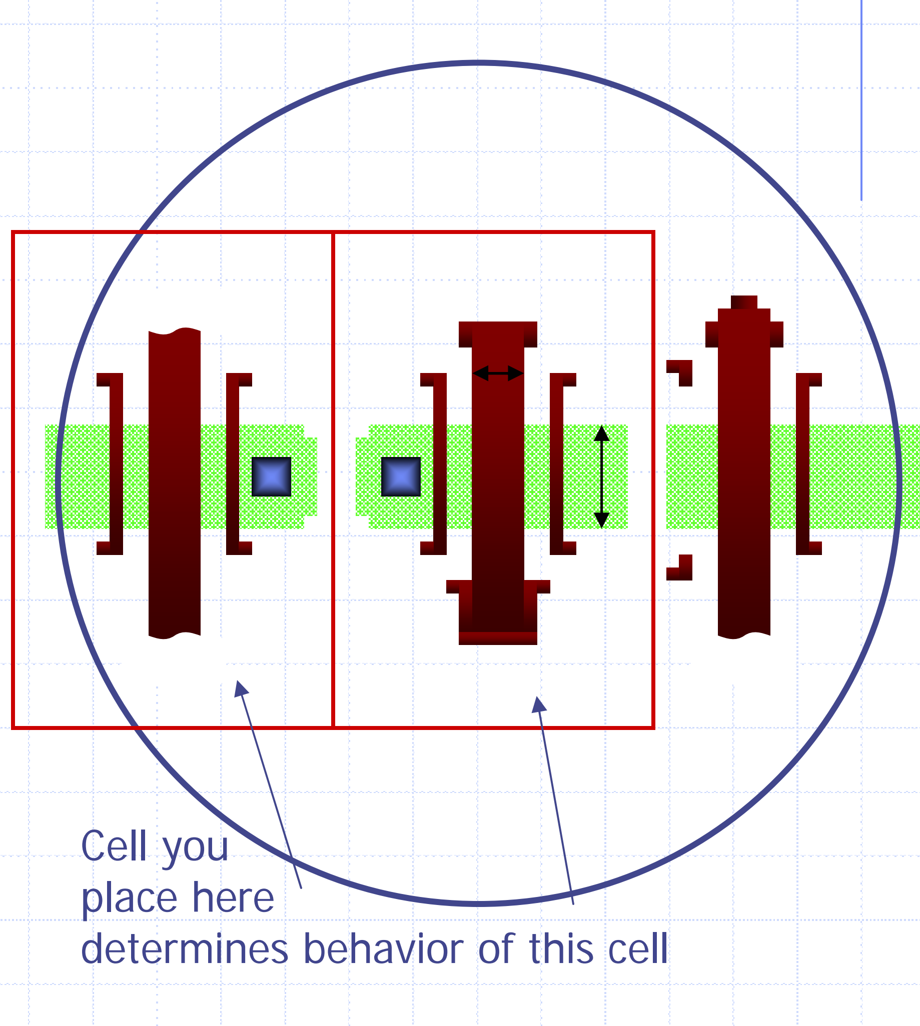


2000

Litho and Physical Abstraction ca. 2010

In the very near future, so much of what is around the device is needed that the notion of arbitrarily composable design is not valid any longer!

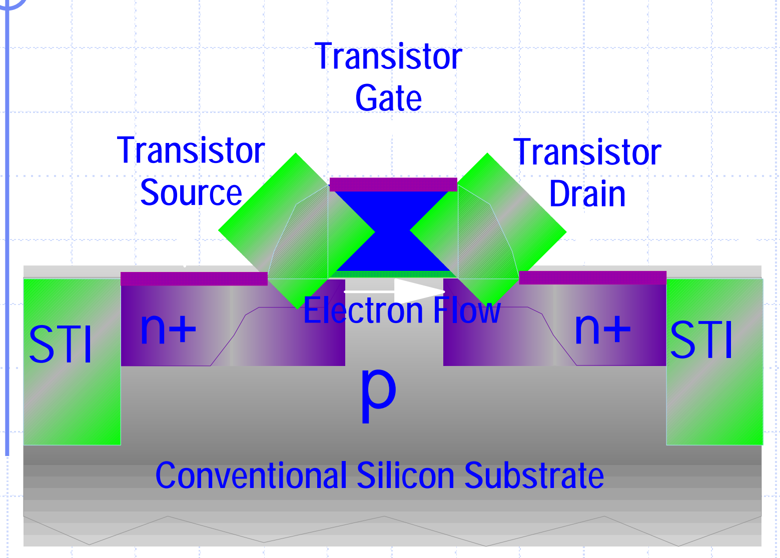




What has changed?

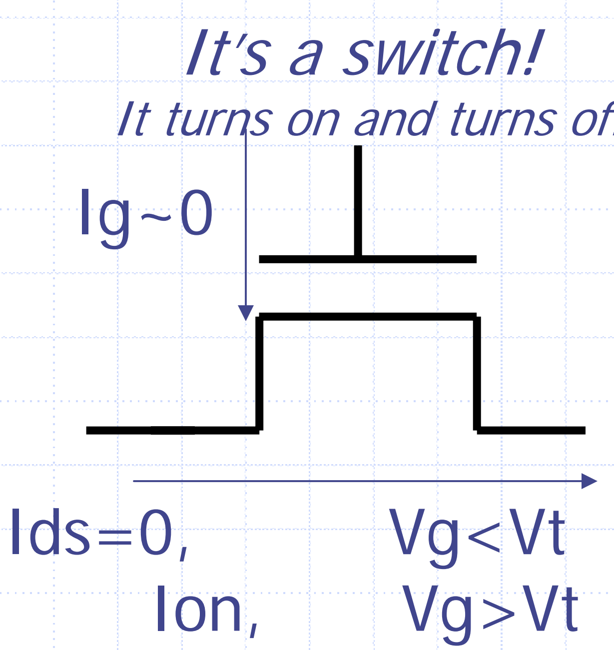
- Resiliency & redundancy cannot be **ignored**.
  - Need to start design assuming partial functionality!
- Mead-Conway design is **dead**…
  - Physical abstraction is broken – ground-rule explosion
  - Physical abstraction is broken – composability in jeopardy
  - Functional abstraction is broken – increasingly difficult to treat these as "logic devices"
  - Transistor performance determined by new features and phenomena, ∴ large variety in behaviors (not easily bounded).
- Key Factor: **Variability**

Functional Abstraction: 0.25um



It's a switch!

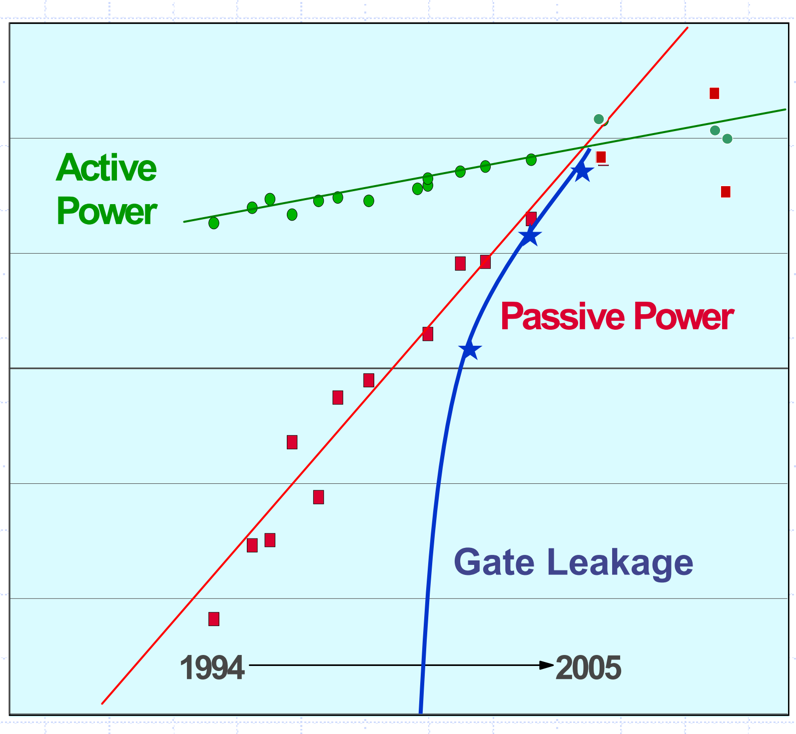
It turns on and turns off



- Resulting rather simple timing delay relations – static timing
  - Tout = Tin + delay(cell output load, interconnect, Vdd, Temp, Process).
  - Modest number of corner analysis cases required (long-path – SS, hold-time -- FF, power corner, noise corner, reliability/electomigration corner)
- Stupidity screens – functional verifications, slew-violations, x-talk…
- *footnote: analog and array designers exempt from simple abstraction*

Functional Abstraction Broken:

Just when is it off?



Power Density (W/cm²)

Gate Length (microns)

Source: E. Nowak, et al

-Vt variability increasing passive power contribution

-Thin oxide and high supply driving gate leakage power

Functional Abstraction: 65nm

$$J_g = A \cdot \left(\frac{T_{oxref}}{t_{ox}}\right)^{ntox} \cdot \frac{V_g \cdot V_{aux}}{t_{ox}^2} \cdot e^{-B(\alpha - \beta|V_{ox}|)(1+\gamma|V_{ox}|)t_{ox}}$$

$$A = q^2 / 8\pi h \phi_b \;,\; B = 8\pi\sqrt{2qm_{ox}}\phi_b^{3/2} / 3h$$

$$I_{leakage} = \mu_0 \cdot C_{OX} \cdot \frac{W}{L} \cdot e^{b(V_{dd}-V_{dd0})} \cdot v_t^2 \cdot \left(1 - e^{\frac{-V_{dd}}{v_t}}\right) \cdot e^{\frac{-|v_{th}|-V_{off}}{n \cdot v_t}}$$

Now add variability – like Vt shifts…

$$\delta V_t \sim \frac{1}{\sqrt{N}} \sim \frac{1}{\sqrt{w\, L_{poly}}}$$

And everything is a distribution!

What has changed?

- Resiliency & redundancy cannot be **ignored**.
  - Need to start design assuming partial functionality!
- Mead-Conway design is **dead**...
  - Physical abstraction is broken – ground-rule explosion
  - Physical abstraction is broken – composability in jeopardy
  - Functional abstraction is broken – increasingly difficult to treat these as “logic devices”
  - Transistor performance determined by new features and phenomena, ∴ large variety in behaviors (not easily bounded).
- Key Factor: **Variability**

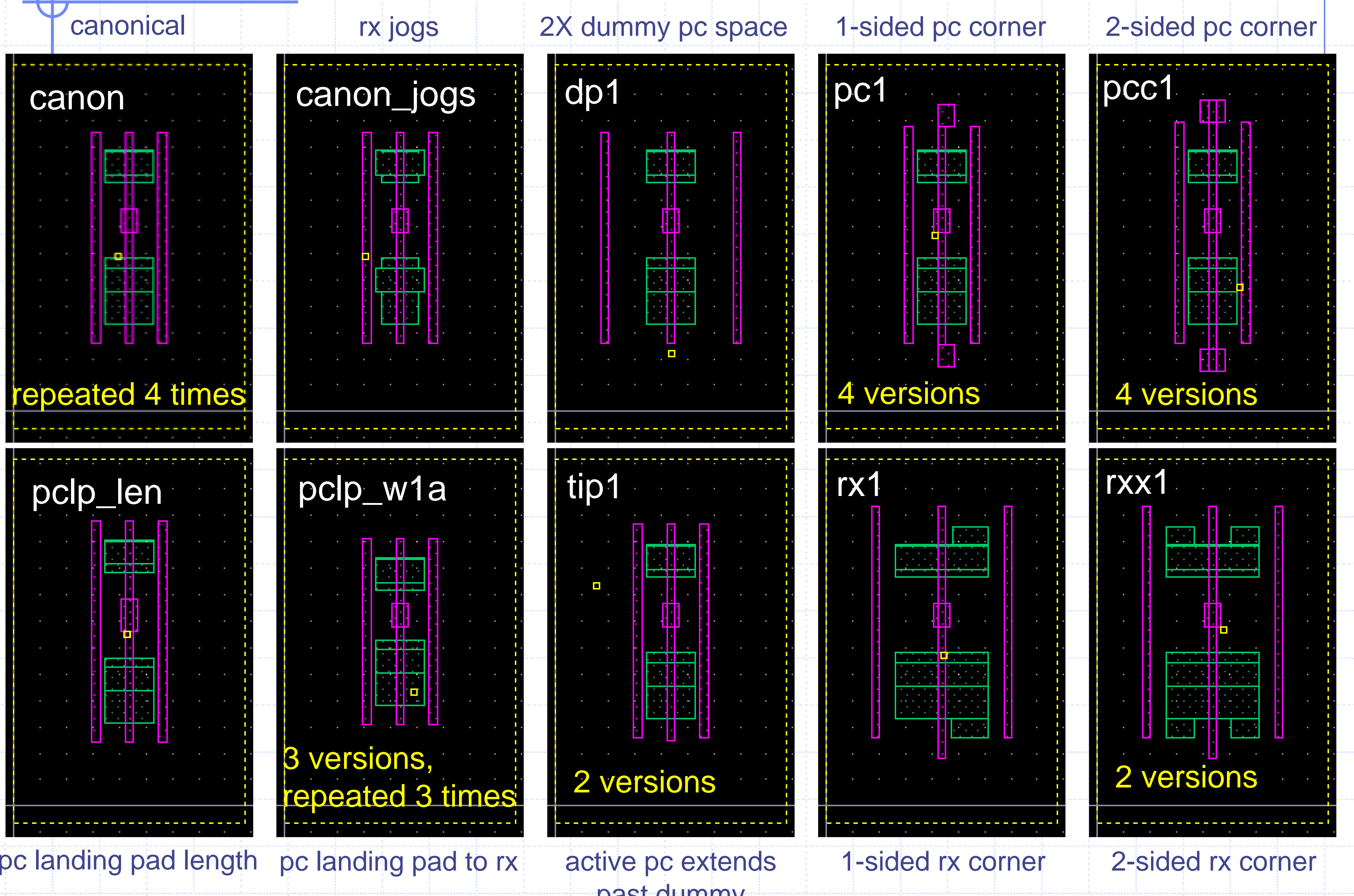
So now what?

- Back to days of the “Hero Designer?”
- Or Cope? -- fix the incomplete technology specification, modify the abstractions, validate the models, and change the design practices.
- *Just how many Hero Designers are there in VLSI?*

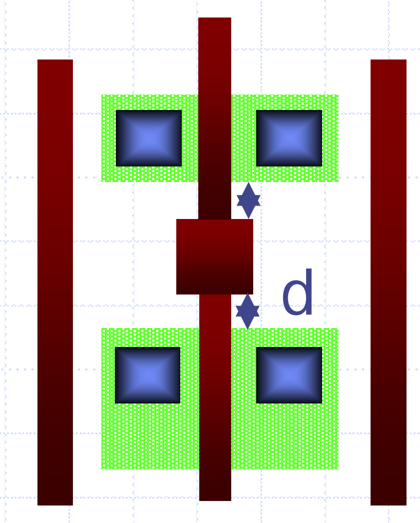
Coping – part 1

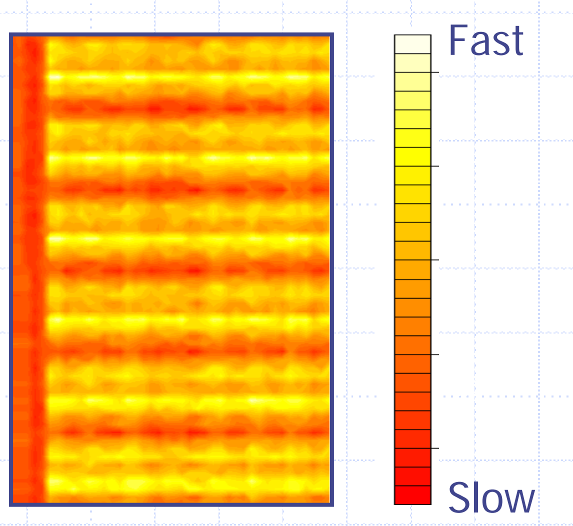
- Know thine enemy: "You can fix what you can't measure"
  - Build structures to measure variation effects and causes – density & pattern sensitivities, CAA, threshold variation, matching….
  - Capture significant variation effects in models
  - In-situ variation sensing thru on-die monitor circuits – thermal, performance ROs, supply, aging, …

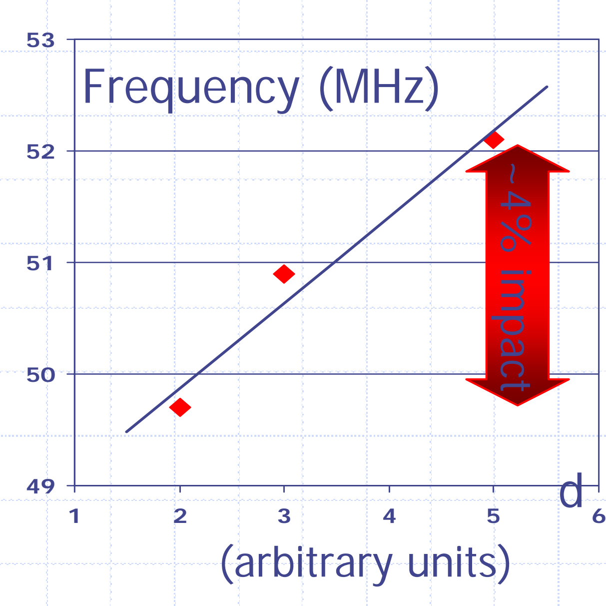
Variations for Design Rule Exploration



Example Results

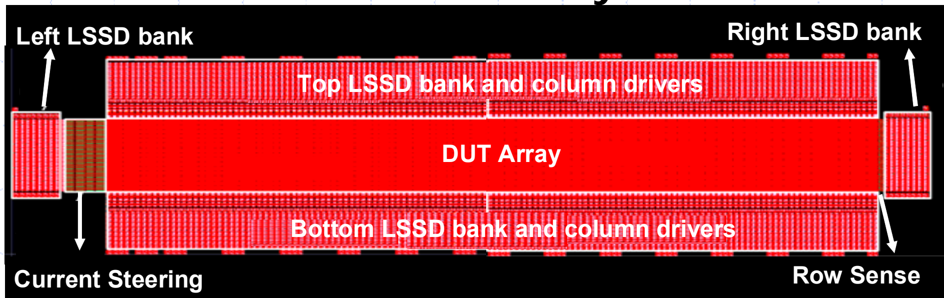






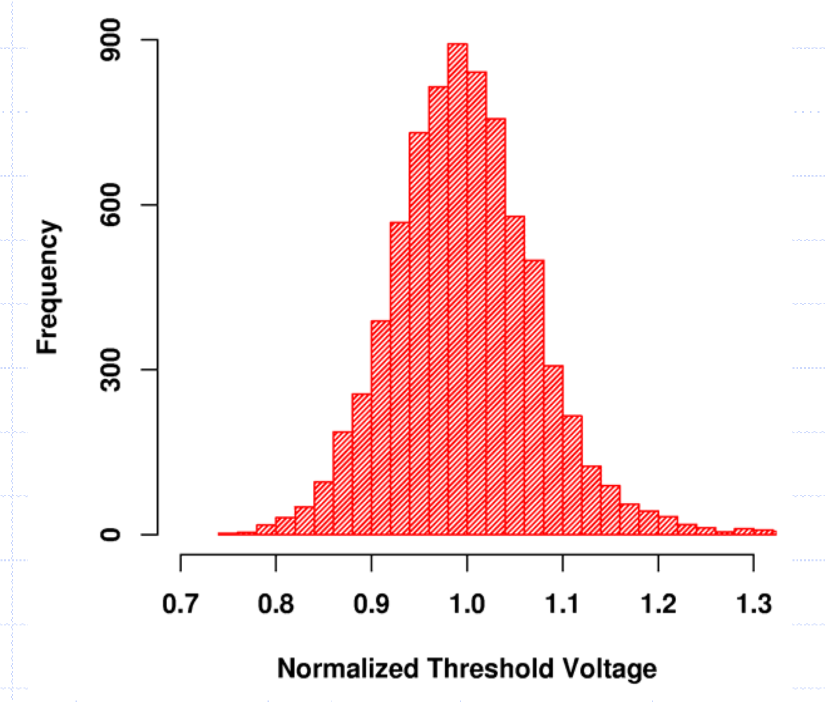
- Experimental hardware
- Schematically identical
- Layout variation
- 20% frequency variation

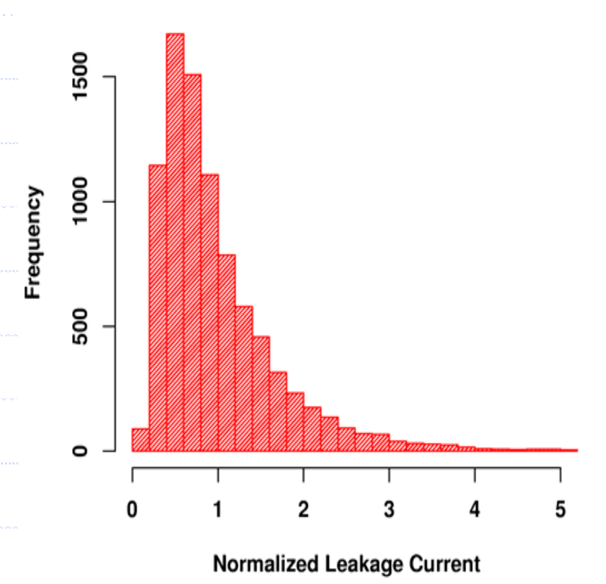
Targeted structures -- IV spatial variation



Spatial Leakage Distribution

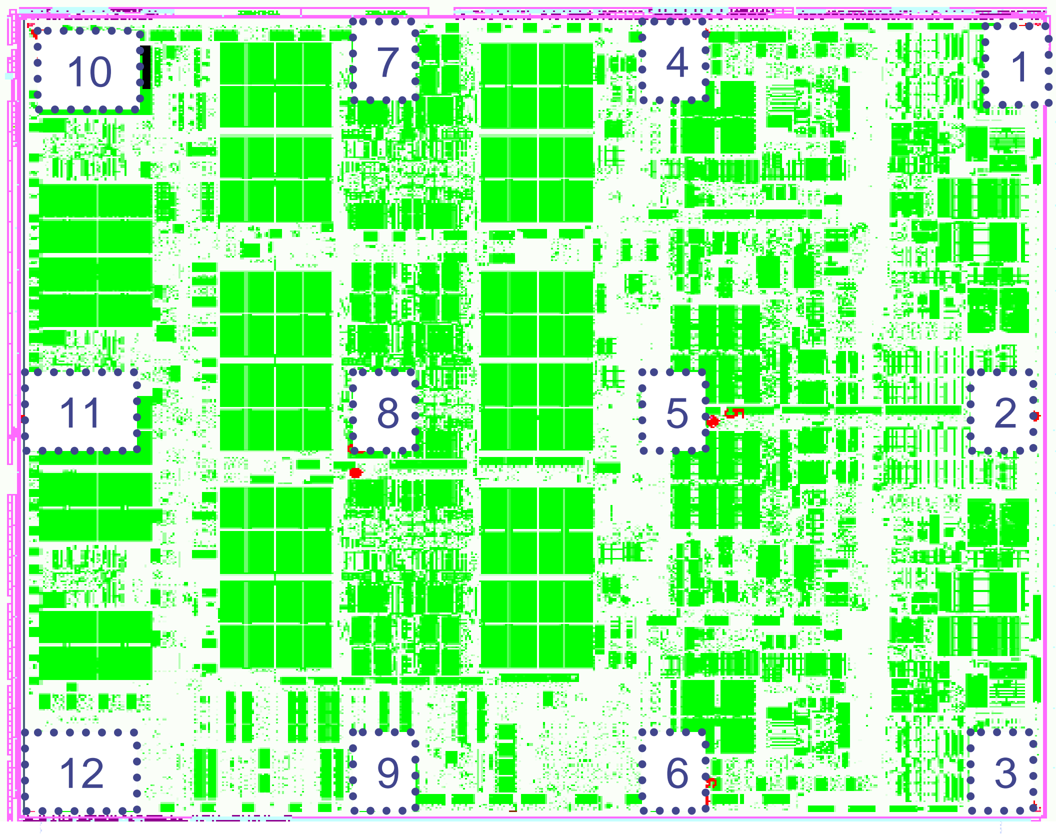
Columns





Use of In-situ ring-oscillator structures

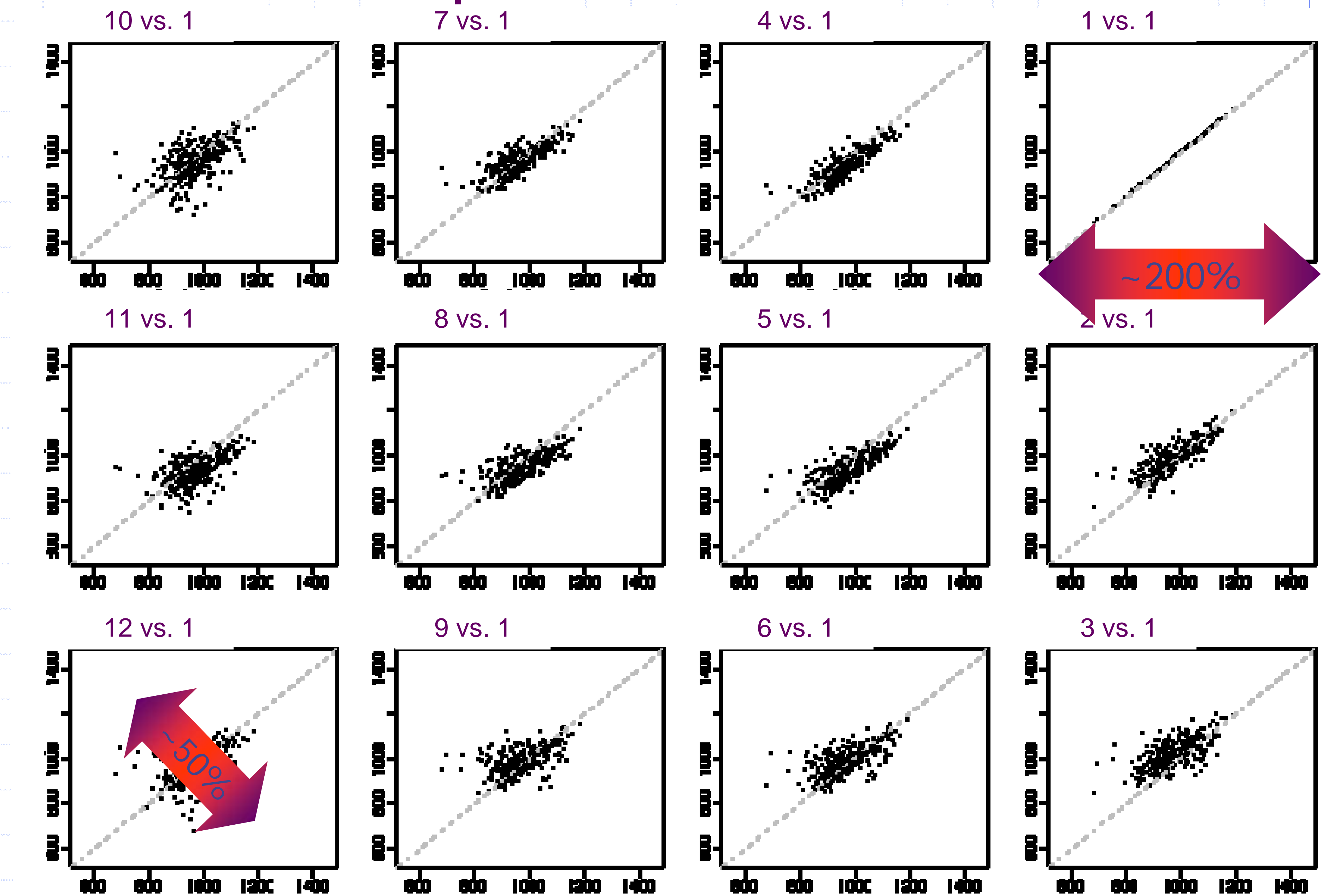
- 12 ring oscillators distributed across the die.



Chip map with Ring Oscillator locations

Courtesy Anne Gattiker, IBM

Measured Speed Variations

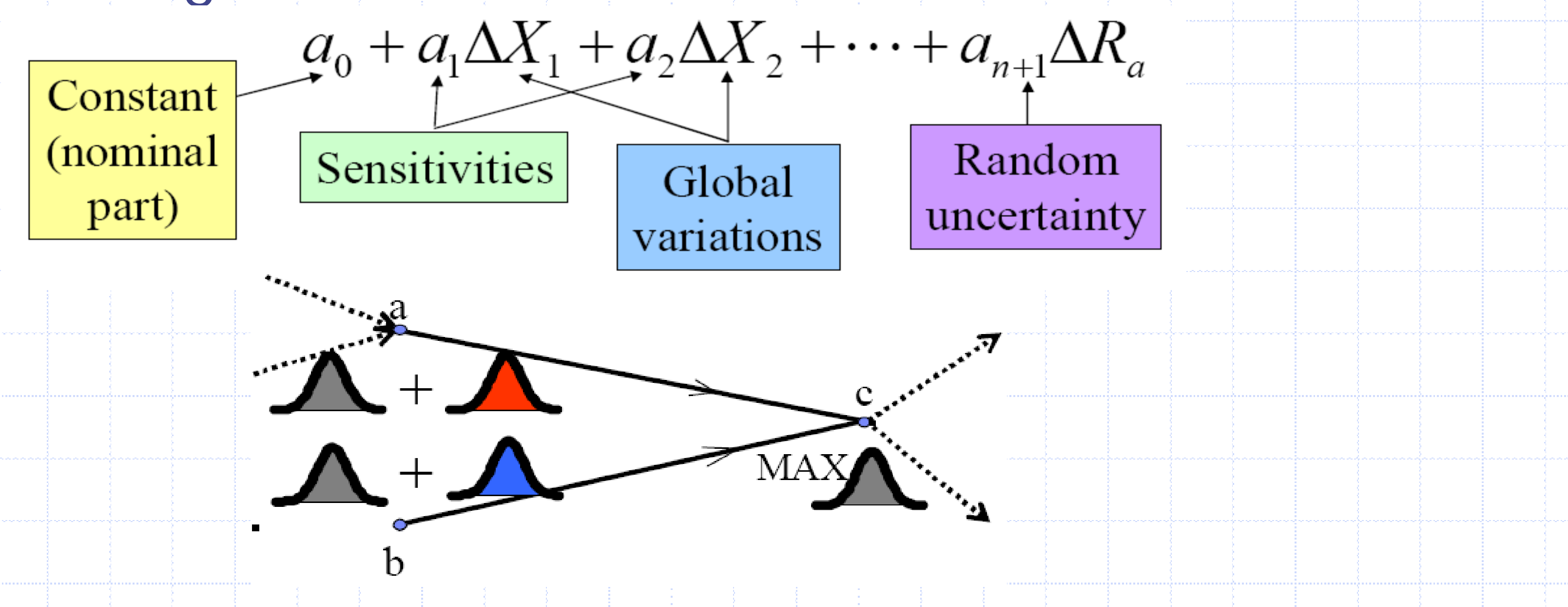


Coping – part 2

- Fix the abstraction and the design process
  - Use modeled behavior to drive physical and functional abstraction
    - Incorporate sensitivities into physical abstraction – eg. Raise the level of physical abstraction for cells
    - Incorporate sensitivities into timing abstraction – eg. Statistical Static Timing
  - Variation aware DA (placement, routing, buffer insert…)
  - Recognize that rampant variability = defective
    - Test for the tails – At Speed Scan Tests
    - Cut out the tails – eg. SRAMs with Vt-induced Vmin issues should be mapped out with redundant row/columns
    - With 80 cores can't you just turn the worst one or two into decoupling capacitors?

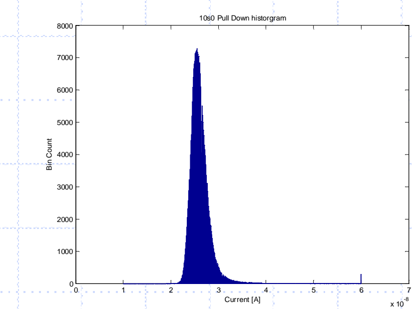
Statistical Static Timing

- Path-based SSTA
  - Conduct a nominal timing analysis
  - Select a representative set of critical paths
  - Model the delay of each path as a function of random variables (the underlying sources of variation)
  - Predict the parametric yield curve, as well as generate diagnostics (integration of a feasible region in parameter space)
- EinsStat (IBM tool) models all timing arcs and produces all timing results in the canonical 1st order form:



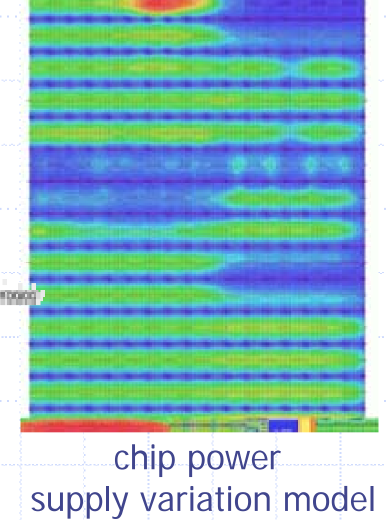
Current Modeling Environment

- Lots of variability characterization data
- Numerous variability modeling tools
- Little commonality!

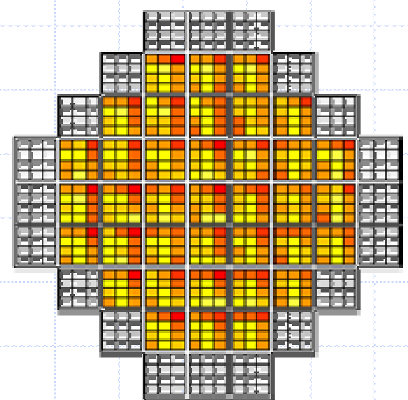


Measurement of spatial variations of device performance

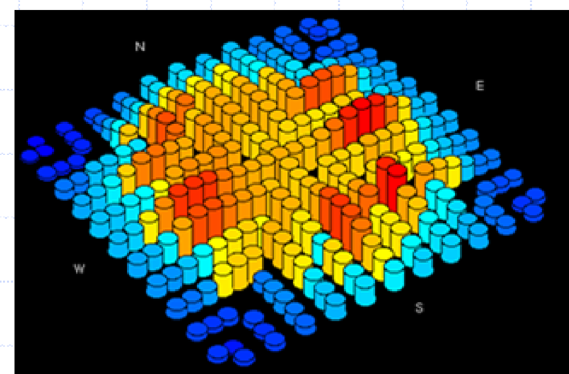
die temperature variation modeling



chip power supply variation model



leakage estimate & modeling



Package electrical variation model

Coping – part 3

- "Bob and weave" – Adapt design for variation
  - If it's functional then adapt… to spatial/temporal variation
    - split/multiple supplies
    - body bias
    - DVFS
    - thermal throttling
    - power and performance efficiency-based job scheduling
  - Does variation-induced timing variation warrant fundamental shift from synchronous systems to inherent timing adaptation?
    - Is 2X die-to-die, 50% within die variation sufficient?
    - If half of this is systematic and nullible, where do we spend our effort?

Adaptation Required vs. Variation Sigma

Delay

perf spec

OK | A | Adapt | B | Degrade | C | Fail

Sigma

Circuit delay exceeds specification

Circuit delay beyond fixing by adaptation

Circuit does not invert any longer

output voltage

nominal

time

output voltage

nominal

time

output voltage

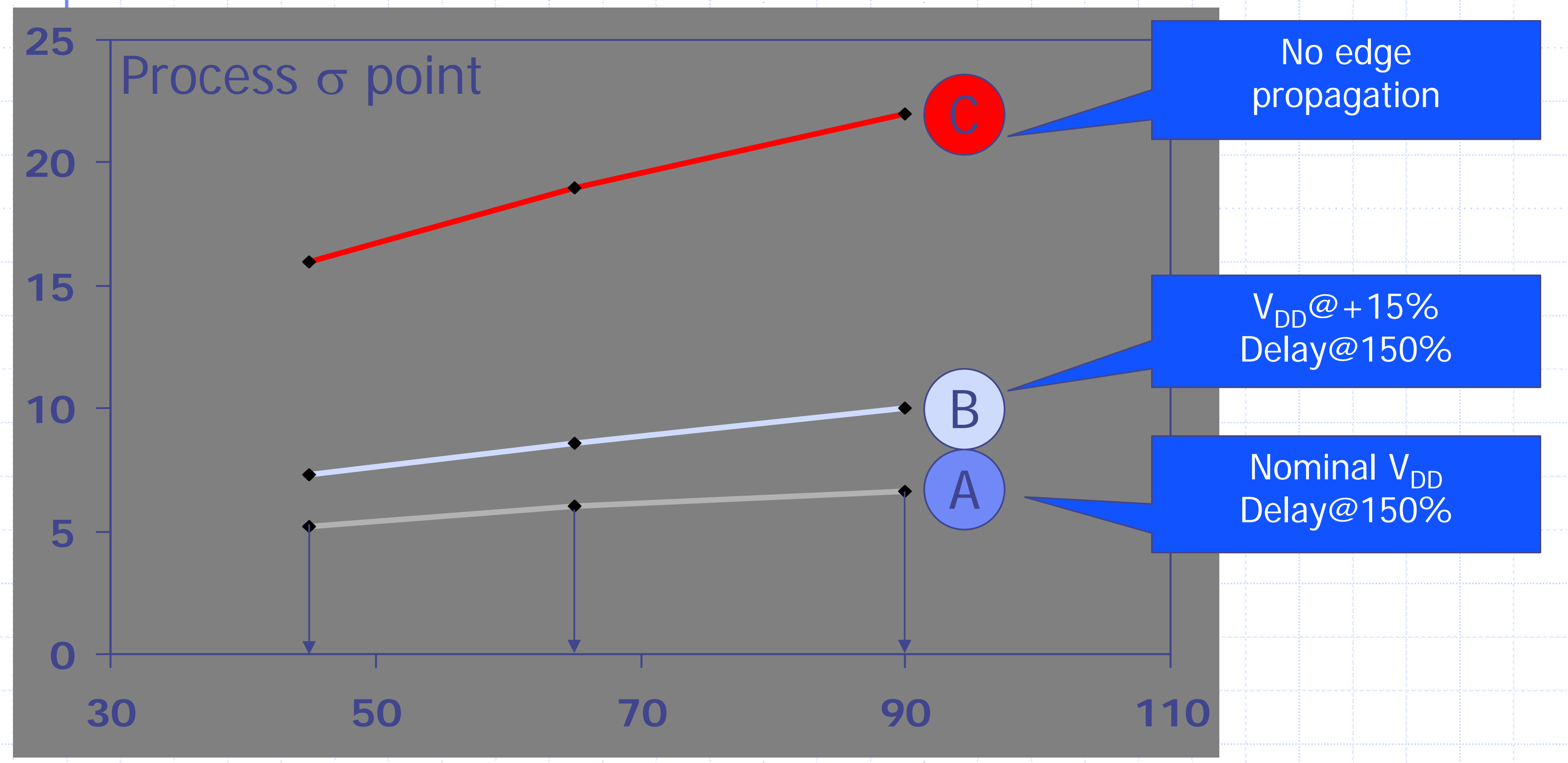
nominal

time

Performance indistinguishable from a "stuck at 1" fault!

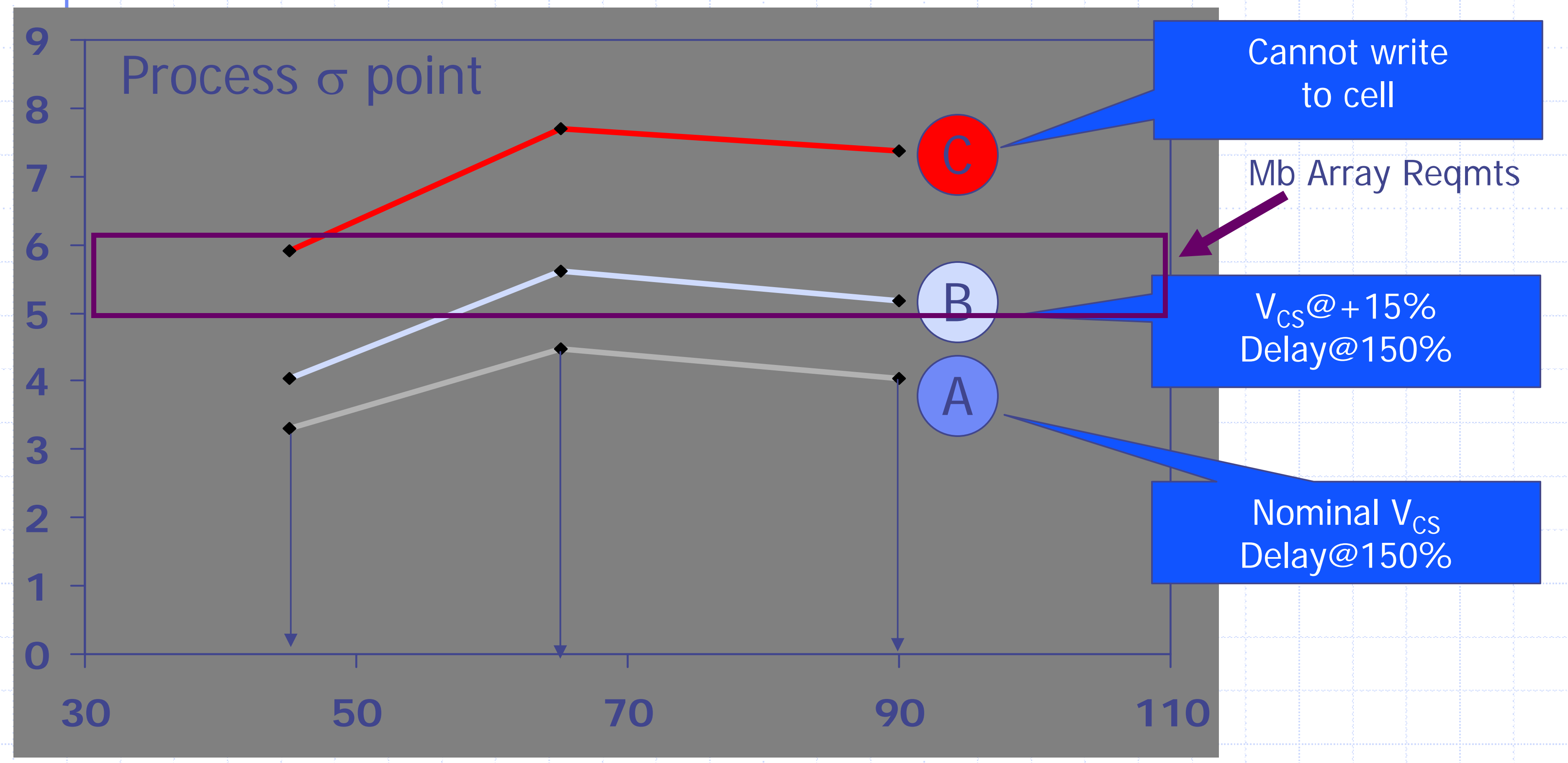
Technology Trend For a Simple Buffer

- Simplest possible circuit (if this fails, everything else will).
- Performed analysis for 90nm, 65nm and 45nm.
- Clear trend in sigma!



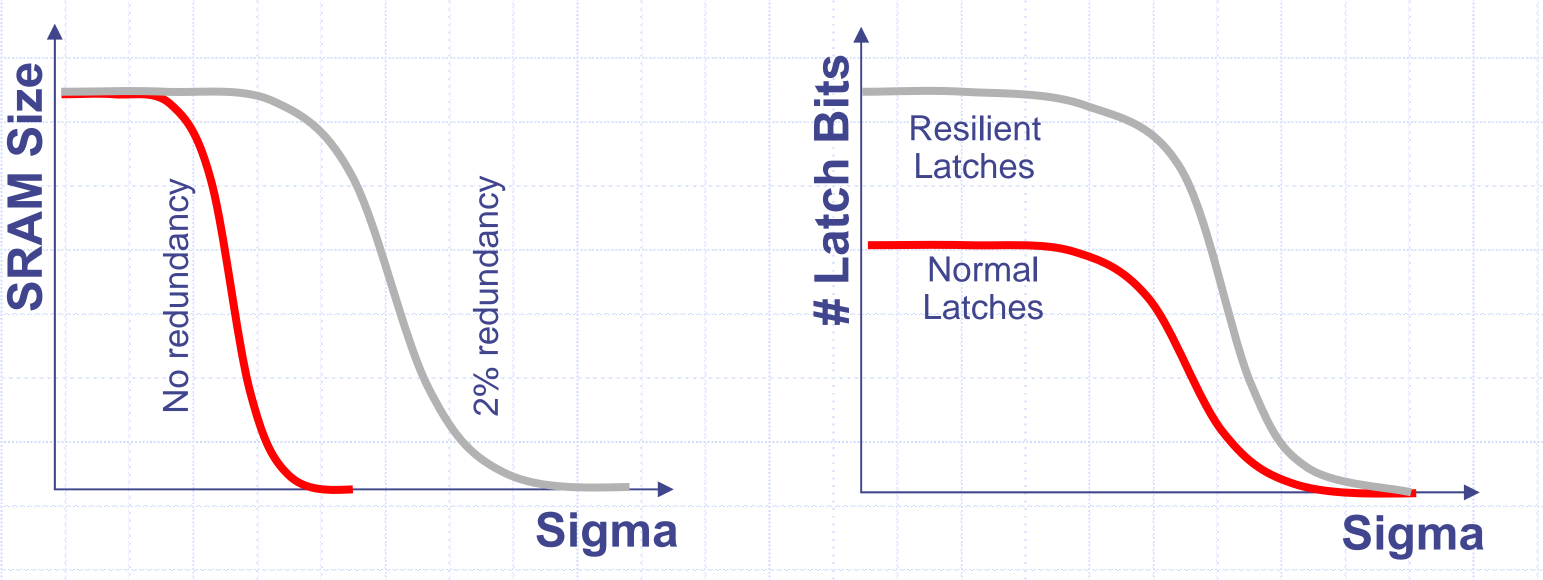
Technology Trend for an SRAM

- SRAM is known to be a more sensitive circuit… (lower σ).
- But, circuit optimized for each technology. (No redundancy included)
- Much lower σ values + similar trend in sigma!



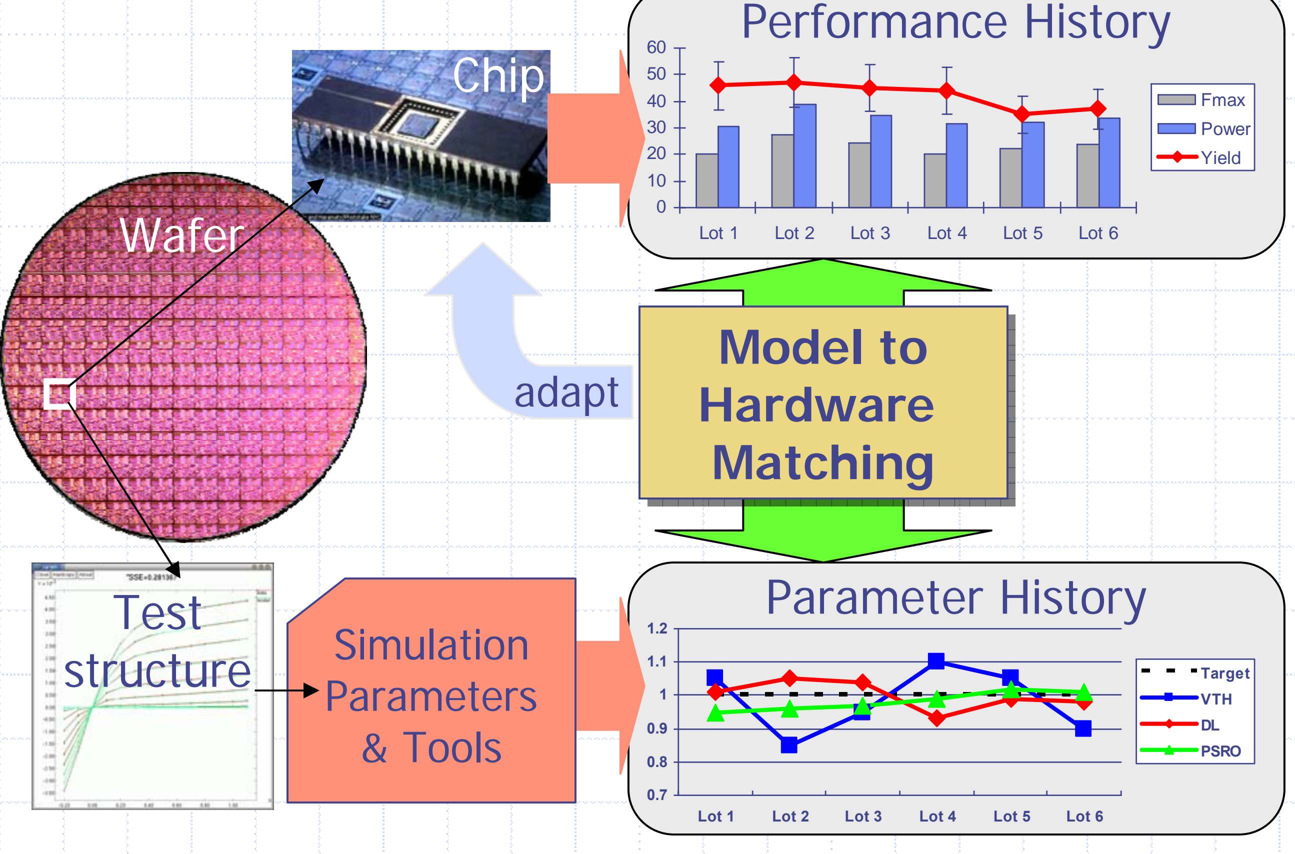
Impact of A/B/C Sigma on Chip Design

- The values of sigma determine:
  - Whether to build adaptation into the chip
  - Whether to include redundancy in the chip
  - The size of “yieldable” components on the chip
- Such activities are already routine in the design of SRAM.
  - But such techniques are not well developed for standard logic design…
  - Different technology sensitivities of SRAM vs. logic make the problem difficult



Ultimate Vision

Get to the point where site-specific hardware-derived models are ubiquitously available… Enable accurate model to hardware correlation and sophisticated design adaptation.



Summary Trends and Challenges

- Trends/Challenges
  - Variability increasing as Design/Manufacturing interface complexity rising.
    - More design rules, more 2nd order effects, more systematic variations, more correction steps…
  - Current techniques are insufficient
    - Abstractions no longer good enough
    - Predictability is poor
      - Ability to confidently bound performance is degrading.
      - Frequent model/hardware mismatch.
- Required Action
  - Better, targeted measurements through characterization structures
  - Hardware-driven variation-enabled modeling
    - Corners not sufficient any more – statistical timing
  - Technology aware circuit and PD tools
    - Variation tolerance in design
    - Technology aware physical design, redundancy, adaptation.