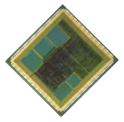
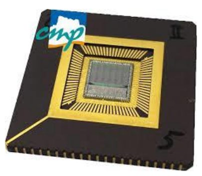
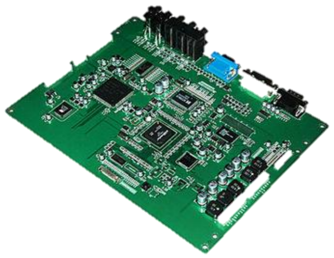
EE-429
Fundamentals of VLSI Design

Corners, Mismatch, and Yield

Andreas Burg, Alexandre Levisse

The Need for Conservative Design

- **Integrated circuits are at the foundation of complex systems**
  - These systems rely on accurate specifications of their components.
  - Deviation from these specifications may or may not lead to failure.

- **Cost of “repair" increases exponentially in each step of the integration chain**
  - Need to assume that any deviation from the specification will lead to a system failure and requires repair or discarding of the entire system.

Uncertainties in IC Design

- Despite good models, **integrated circuits are designed with many unknowns**

- **The most important unknown factors impacting circuit behaviour are:**

Process　　　**V**oltage　　　**T**emperature

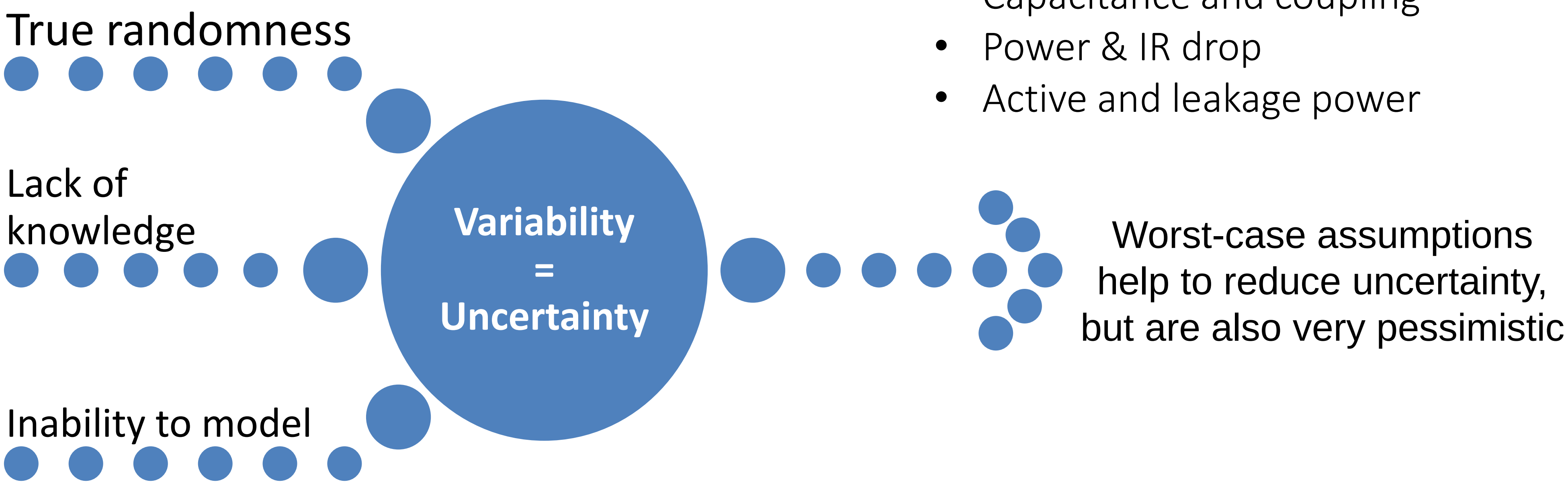
- **We often refer to these as PVT Conditions**

Uncertainty is the Designer's Worst Nightmare

- **Variability summarizes three different problems:**

Metrics and properties subject to variability

- Process, Voltage, Temperature
- Critical path length and activation
- Capacitance and coupling
- Power & IR drop
- Active and leakage power



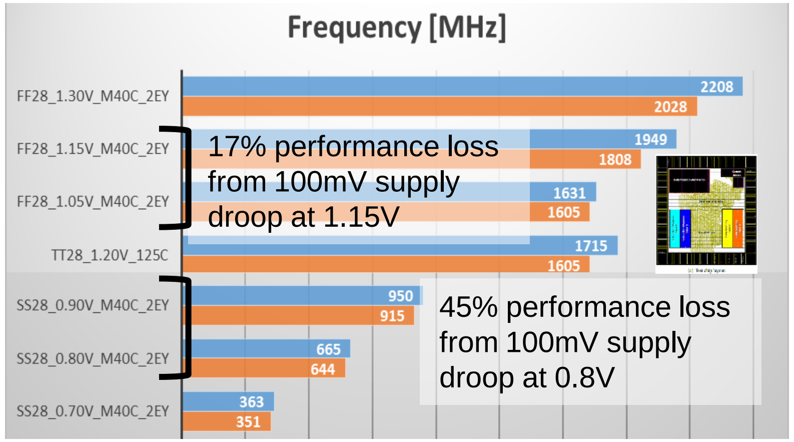
Worst-case design paradigm

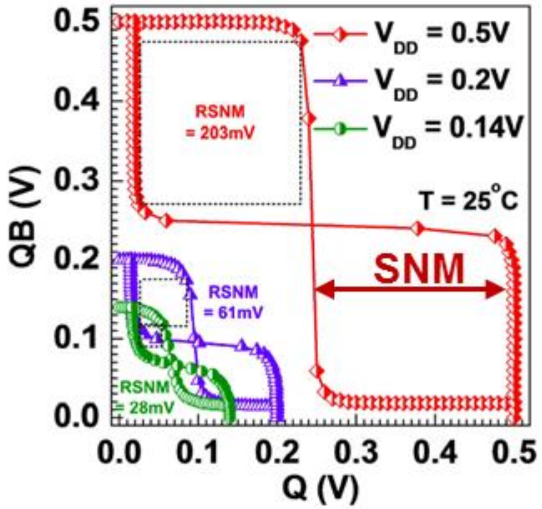
100% reliable operation, under all-worst-case assumptions

Uncertainty: Impact

Logic (Timing) | Memory (Stability)

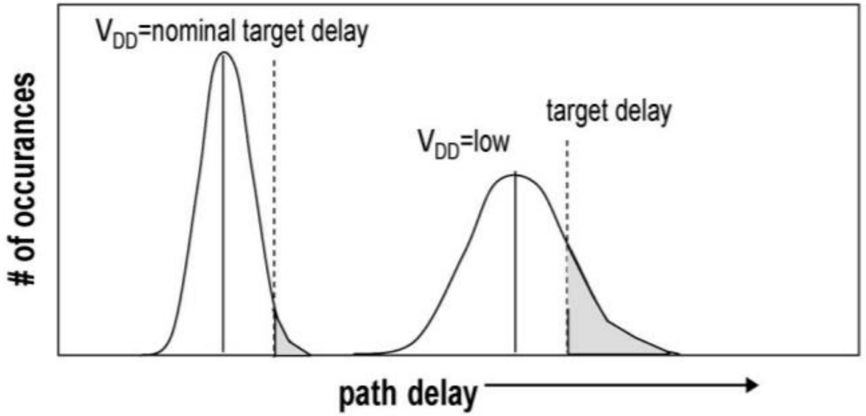
Global Variations

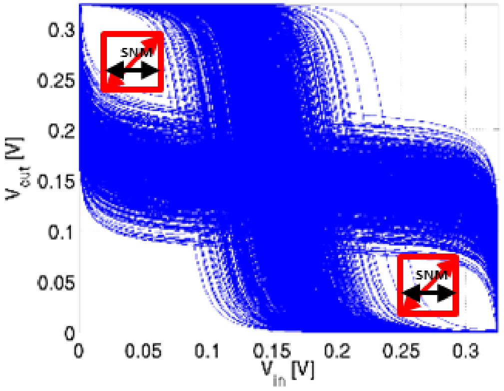




[Sarfraz, JSSC, 2017]

Local Variations

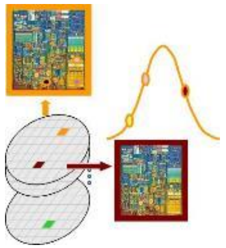




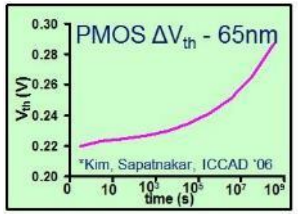
SNM: Margin for noise to ensure reliable operation

Uncertainty: Across Very Different Time Scales

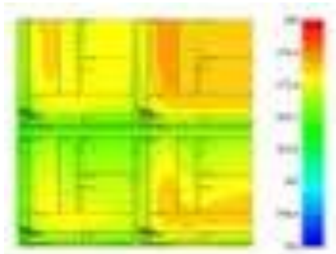
- **General misconception: “Variations are random and unpredictable”**
  - Variations are often deterministic consequences of unknown conditions
  - The impact of a given condition is typically not so difficult to predict
  - **Variations appear on very different** time scales



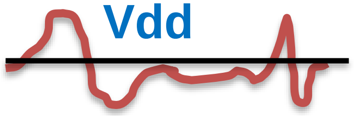
Process variations



NBTI



Thermal



Voltage variation

1010010
0100100
0100100

Data dependencies

Toward shorter time-scale

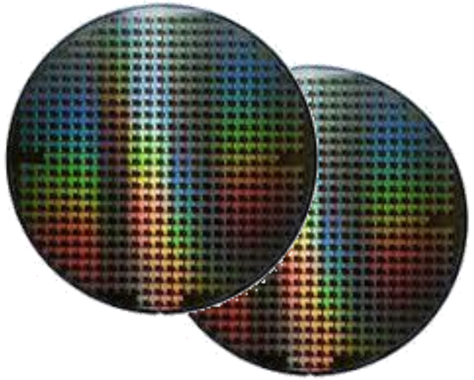
Increasingly difficult to adapt

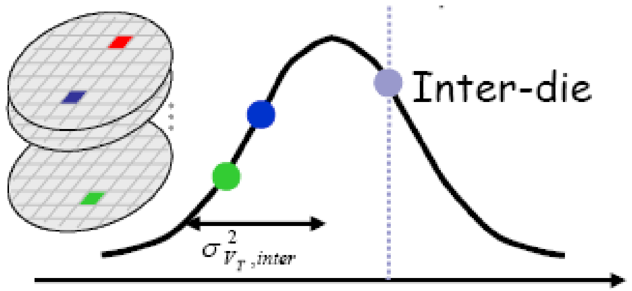
Years — Months — Seconds — Microseconds — Nanoseconds

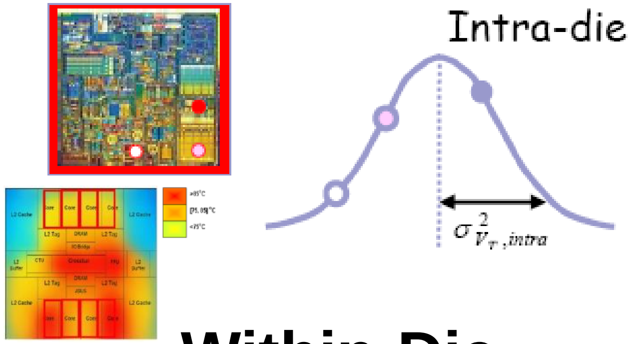
Uncertainty: Spatial Correlation

- **Variations are rarely fully independent across space (elements of a chip)**
  - PVT conditions can be highly correlated in space

- **Uncertainties exist on different scales**







Wafer-to-Wafer **Die-to-Die** **Within-Die**

Global **Local**

Accounting for Global Variations

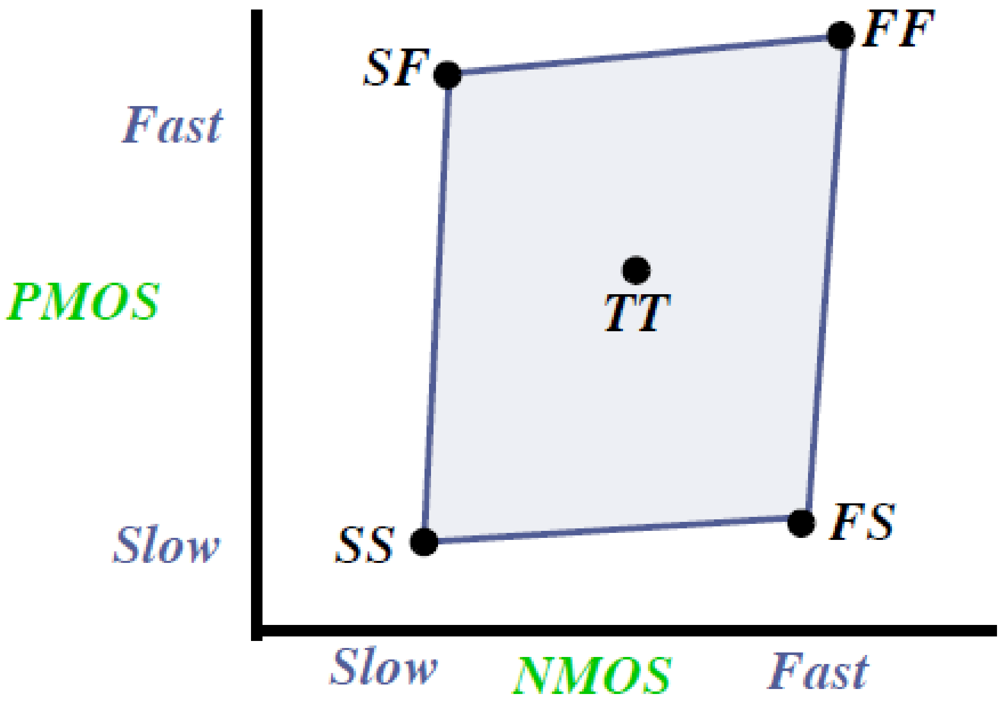
Corners Anticipate Global Uncertainty

- **Global uncertainties:** process corner, chip supply voltage, chip temperature
  - **Global uncertainties are usually common to all components on a die**

- **Common uncertainties can be anticipated by defining operating corners**
  - Operating corners are a combination of the most important factors that influence circuit behaviour
    - **P**rocess
    - **V**oltage
    - **T**emperature

PVT corner

Environmental corners (1.8V process)		
Corner	*Voltage*	*Temperature*
Fast (F)	1.98	0°C
Typical (T)	1.8	70°C
Slow (S)	1.62	125°C



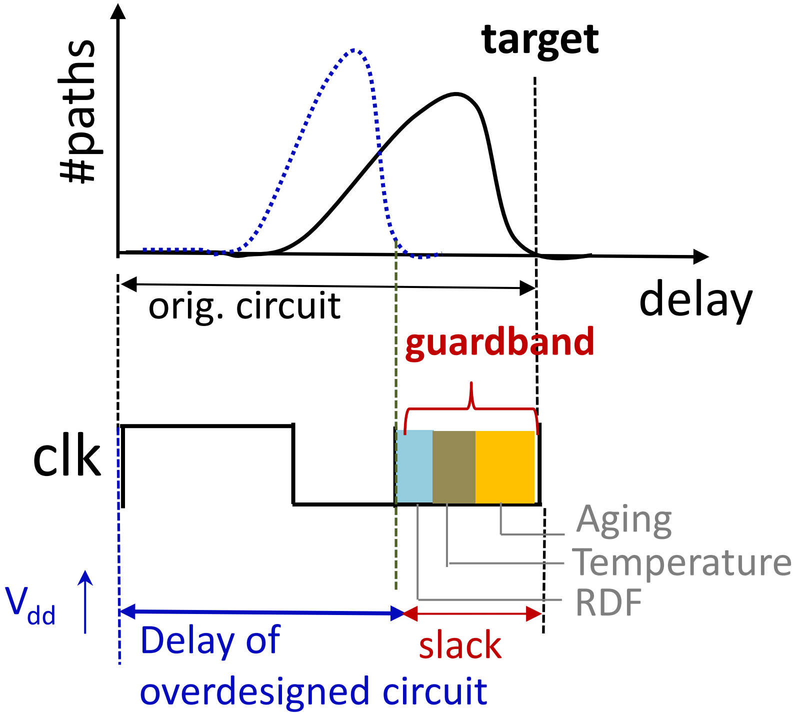
Verification Against Global Operating Corners

- **Different requirements must be verified in different corners**
  - Consider only the extreme conditions for each parameter
  - Number of possibilities is usually limited
  - Combine individual extreme conditions to further reduce the number of corners
- **Complex designs require verification in multiple corners and which corner to use is not always obvious**

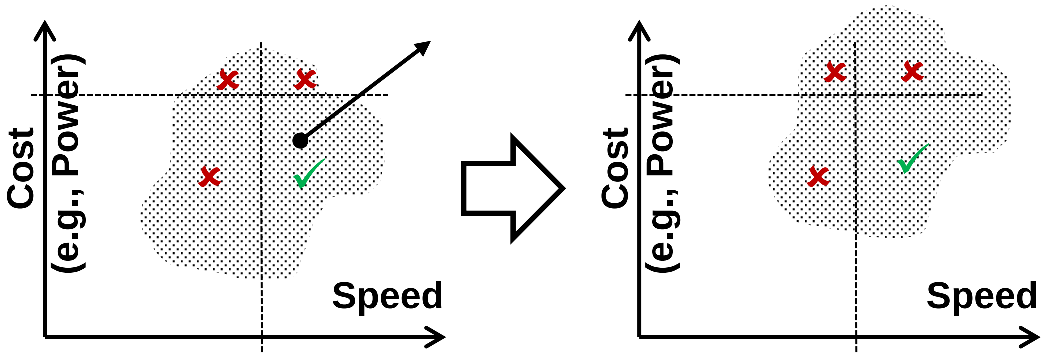
Corner					*Purpose*
NMOS	*PMOS*	*Wire*	V_{DD}	*Temp*	
T	T	T	S	S	timing specifications (binned parts)
T	S	S	S	S	timing specifications (conservative)
F	F	F	F	F	DC power dissipation, race conditions, hold time constraints, pulse collapse, noise
F	F	F	F	S	subthreshold leakage noise, overall noise analysis
S	S	F	S	S	races of gates against wires
F	F	S	F	F	races of wires against gates
S	F	T	F	F	pseudo-NMOS & ratioed circuits noise margins, memory read/write, race of PMOS against NMOS
F	S	T	F	F	ratioed circuits, memory read/write, race of NMOS against PMOS

Worst Case Design: Stay on the Safe Side

- **Fixed worst-case specifications drive the design process**



Design with large margins
to meet specifications under
all worst-case conditions

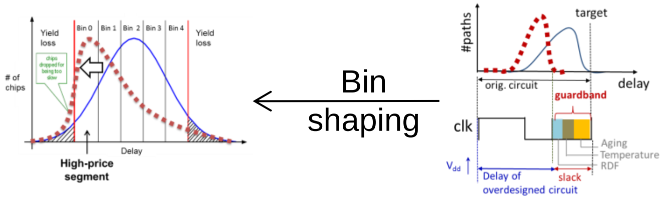


- Large spread in performance characteristics: Margins unnecessary for the majority of the design corners
- Multiple performance criteria: Taking margins on one criterion typically worsens performance for other criteria.

Binning is Limited by Application Requirements

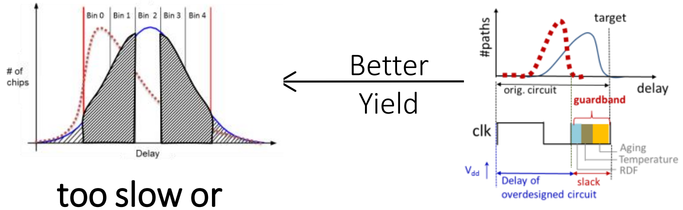
Binning for General Purpose Computing

- No stringent real-time requirements
- Reduced clock results gracefully degrades QoS (speed) of the system



Optimized Signal Processing

- Stringent real-time requirements
- Reduced clock results in complete system failure (e.g., dropped samples)



too slow or
too fast fast

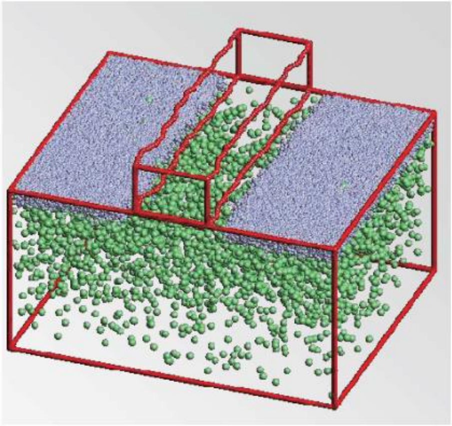
- Common practice for microcontrolers, microprocessors, and GPUs for general purpose computing platforms such as sensor nodes, PCs, and data centers
- *BUT* not well applicable to many dedicated circuits and applications with fixed frequency/throughput requirement such as video and audio or communications

Accounting for **Local Variations** (=**Mismatch**)

Manufacturing Variations: RDF (Bulk Process)

- **Discrete number of dopants in the channel depletion region**
  - Implantation is a random process that leads to statistical fluctuation of the number of dopants N in a given volume (channel)
  - Variance of dopands follows Poisson distribution: $\sigma_N = \sqrt{N}$
  - Example: $W = L = 90\text{nm}$, $D = 350\dot{\text{A}}$, $N_a = 10^{18}\text{cm}^{-3}$

$$N = N_a \cdot W \cdot L \cdot D = 284 \rightarrow \sigma_N = 17$$



Miyamura, M., et al.

- **Number of dopants determines threshold voltage**
  - Threshold voltage variation is Gaussian with variance

$$\sigma_{V_{th}} = \sqrt[4]{2q^3\varepsilon_{Si}N_a\phi_B}\,\frac{T_{Ox}}{\varepsilon_{Ox}}\,\frac{1}{\sqrt{3WL}}$$

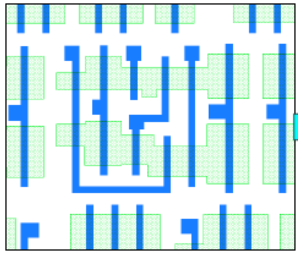
Impact of RDF decreases with increasing transistor size (WL)
- Upsizing helps
- Large impact on min. size SRAM

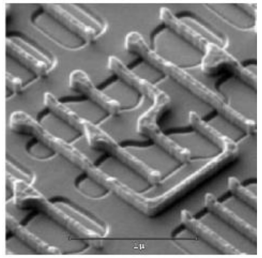
Mizuno, Tomohisa, J. Okumtura, and Akira Toriumi. "Experimental study of threshold voltage fluctuation due to statistical variation of channel dopant number in MOSFET's." Electron Devices, IEEE Transactions on 41.11 (1994): 2216-2221.

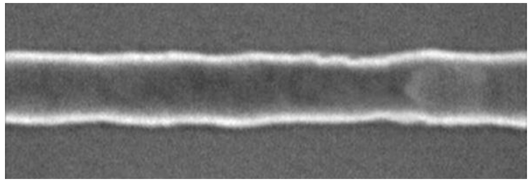
Miyamura, M., et al. "SRAM critical yield evaluation based on comprehensive physical/statistical modeling, considering anomalous non-Gaussian intrinsic transistor fluctuations." VLSI Technology, 2007 IEEE Symposium on. IEEE, 2007.

LER and Proximitry Effects

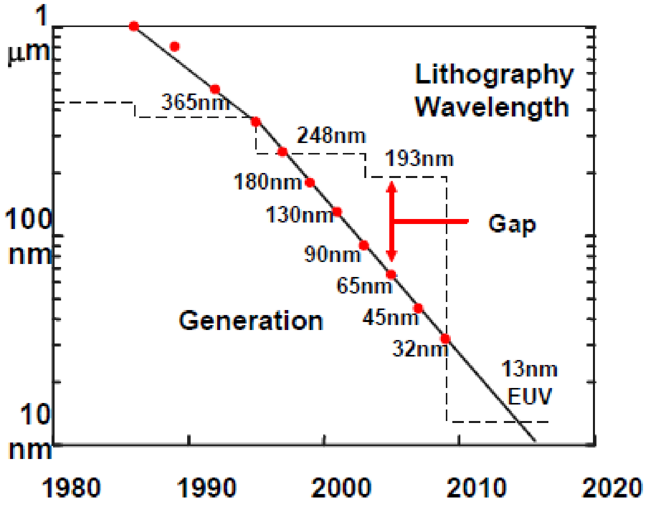
- **Optical lithography: feature size far below wavelength of the light**
  - Sub-wavelength lithography with optical proximity correction (OPC)
  - Systematic variation of dimensions (gate and interconnect)
  - Hard to predict, but deterministic
- **Line-edge roughness (LER)**
  - Caused by un-isotropic edging
  - Generally random but impact is small







Mack CA, Conley W; Special section guest editorial: line-edge roughness. J. Micro/Nanolith. MEMS MOEMS.

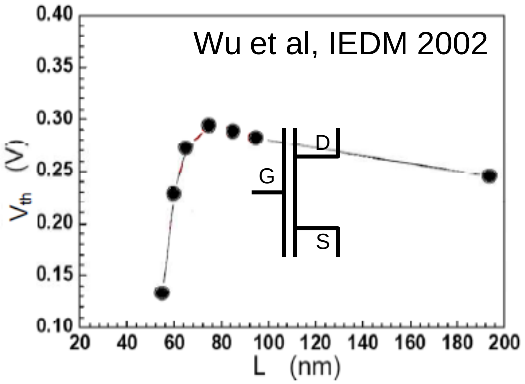


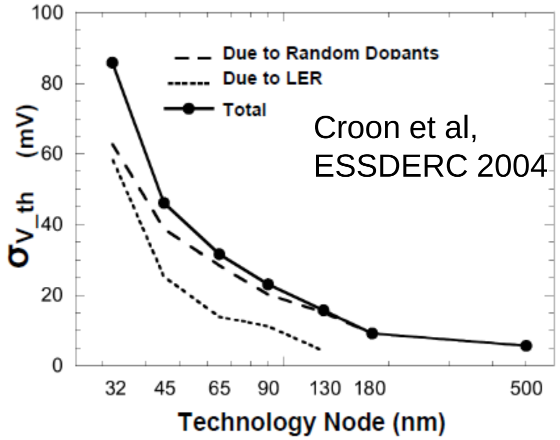
Line-edge roughness becomes relevant beyond 45-nm nodes

Impact of channel-length variation

- Threshold voltage through drain induced barrier lowering (DIBL)
- Directly on drain current through channel length

$$I_D \propto 1/L \quad V_{th} \propto V_{th0} - (\zeta + \eta V_{DS})e^{-L/\lambda}$$





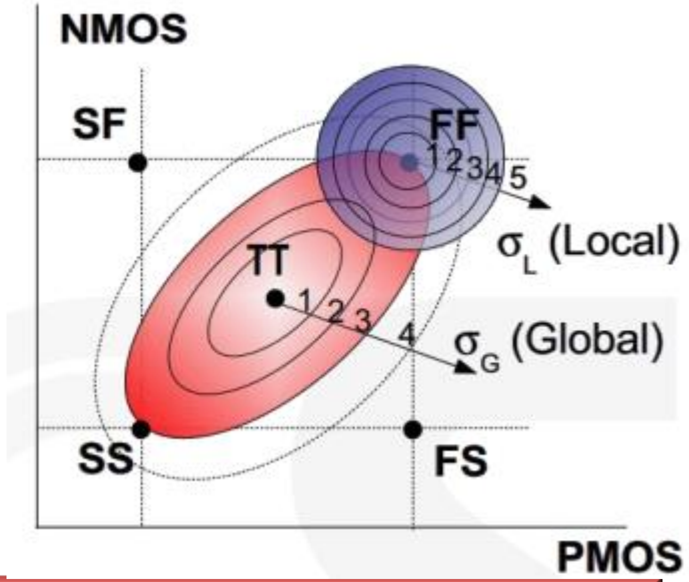
J. Tschanz, K. Bowman, and V. De, \Variation-tolerant circuits: circuit solutions and techniques," in DAC '05: Proceedings of the 42nd annual conference on Design automation, 2005, pp. 762{763.

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Dealing with Local (Within-Die) Variations

- **Within die or device-to-device variations are more difficult to deal with**
  - Millions of devices on a die lead to an uncountable number of combinations
  - The worst-case combination of possible conditions is either
    - Difficult to clearly identify OR
    - Very unlikely to occur if not covered by global variations

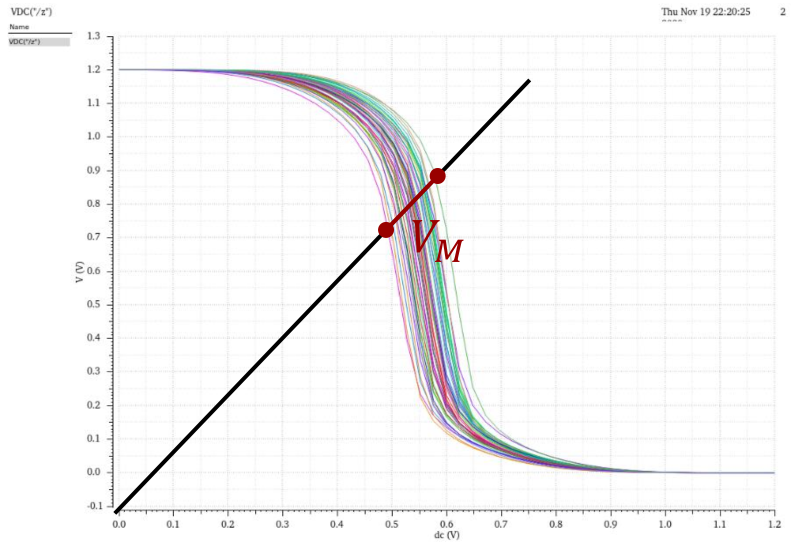
- **Monte-Carlo simulations:** explore possible realizations of a random process
  - Draw a random set of parameters based on a given distribution
  - Evaluate performance characteristics for each realization
  - A global corner provides the mean for the distribution
  - Parameter variations (drawn randomly) model device mismatch



Evaluating Outcome of Monte-Carlo Simulations

- **Monte-Carlo simulations yield a set of simulation results**
  - Each result corresponds to a specific instance of the circuit (an instance can be an instance on a chip OR an entire chip)
  - Non-scalar results (plots) result in a “family” of plots (one per realization)

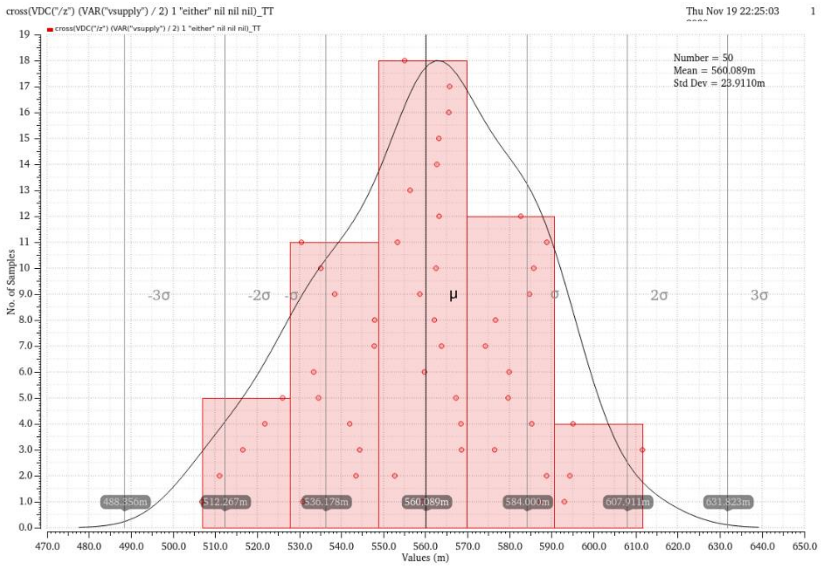
- **Interpretation of results often requires scalar metrics for each run**
  - Scalar metric can often easily be derived from plots (e.g., delay, noise margin, VM, …)
  - Scalar metrics reflect the quality of a circuit



- **Sometimes, a single scalar metric is not sufficient**
  - **Multiple quality metrics** are acceptable, but reducing it helps to avoid complex tradeoffs with outcomes that **are difficult to compare**

Reminder of some Probability Theory

- Metric **histogram is an empirical representation of the probability distribution function (PDF)** of the chosen performance metric
  - The PDF $f(x)$ specifies how likely it is that a random variable x assumes a given value
  - Since x is a continuous variable, $f(x)$ is usually defined indirectly $\Pr[a \leq x \leq b] = \int_a^b f(x)dx$
  - The histogram provides $\Pr[a \leq x \leq b]$



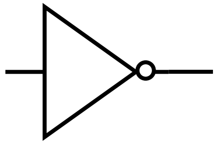
- **The PDF is related to the cumulative distribution function (CDF)**
  - Provides an indication how likely it is that a metric (e.g., delay) does not exceed a given value

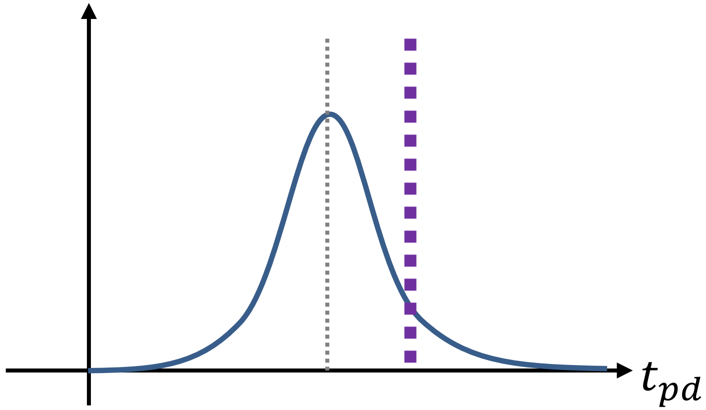
$$F(x) = Pr[X < x] = \int_{-\infty}^{x} f(u)du$$

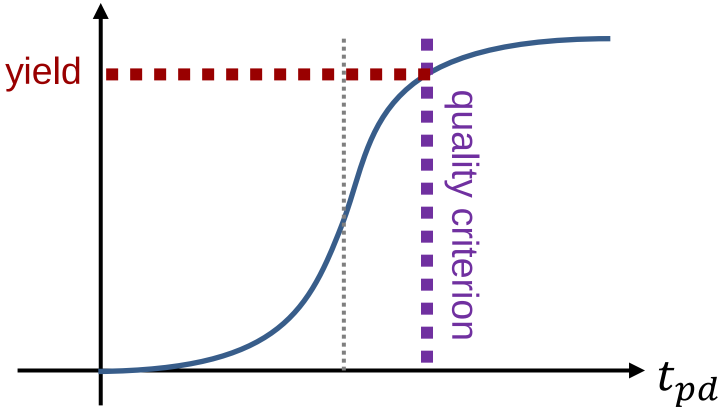
$$f(u) = \left.\frac{\partial F(x)}{\partial x}\right|_{x=u}$$

Estimating Parametric Yield

- **The parametric yield defines the probability that a circuit meets a given quality criterion**
- **Need to define a quality metric X (could be a vector) and a quality criterion for "sufficient quality"**
- **For a scalar quality metric, the CDF $F(x) = Pr[X < x]$ or he inverse CDF $1 - F(x) = Pr[X > x]$ define the parametric yield**
- **Example:** inverter delay





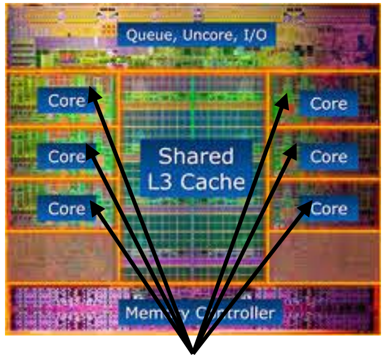


- **Simulations allow to directly obtain the yield from the "empirical" CDF**

What is an Acceptable Yield? – It Depends

- **What matters in the end is the percentage of functional chips: chip-yield**
  - **However,** each chip contains many copies of the same component
  - Chip yield should be >85%, depending on many factors

- Need to **formulate the chip yield as a function of the component yield**
  - **Assumption**: the **component** under consideration is instantiated N times on the chip (e.g., cores, SRAM cells, …)



N cores

  - The chip or system works correctly only if all sub-systems work correctly

$$Chip-yield = Pr[chip = OK] = (Pr[component = OK])^N$$
$$Component-Yield = Pr[component = OK] = \sqrt[N]{Chip-Yield}$$

  - Example 1kBit memory: 1024 bit-cells, $Chip-Yield = 90\%$ ➔ $Component-Yield = 99.989\ \%$

Analysing Very High Yield with MC Simulations

- Many instances require a **very good component-yield for** achieving just an **acceptable chip-yield**
- **Estimating very high yield, requires a huge number of MC simulations!**

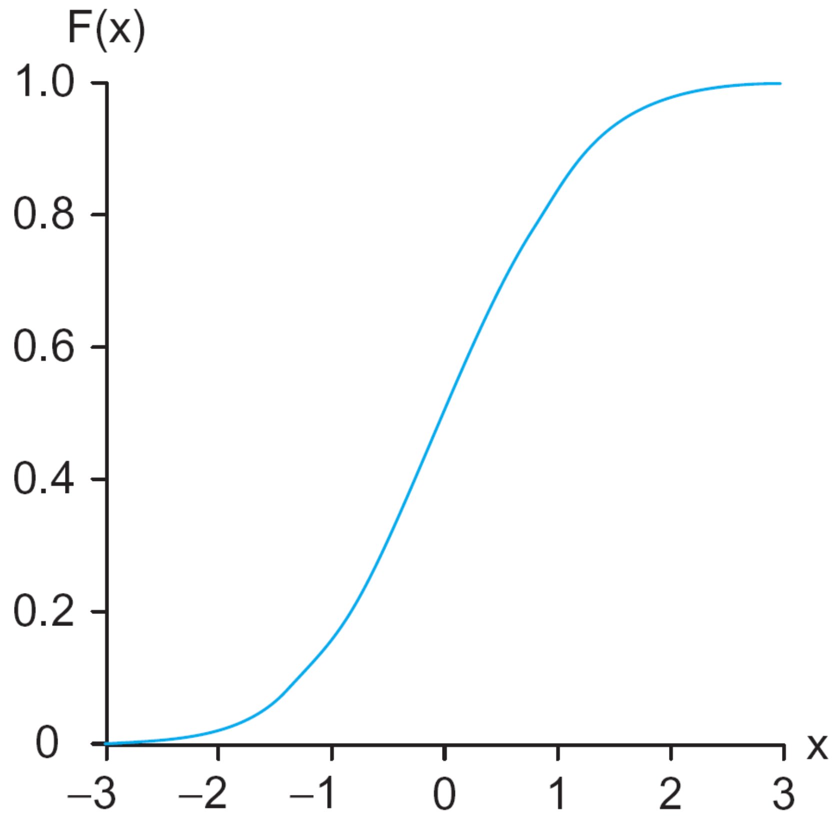
- To solve this issue lets **assume that the metric distribution is Gaussian** with zero-mean ($\mu = 0$) and variance one ($\sigma = 1$)
  - Analytical expressions for PDF and CDF are available

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \qquad F(x) = \frac{1}{2}\left[1 + erf\left(\frac{x}{\sqrt{2}}\right)\right]$$

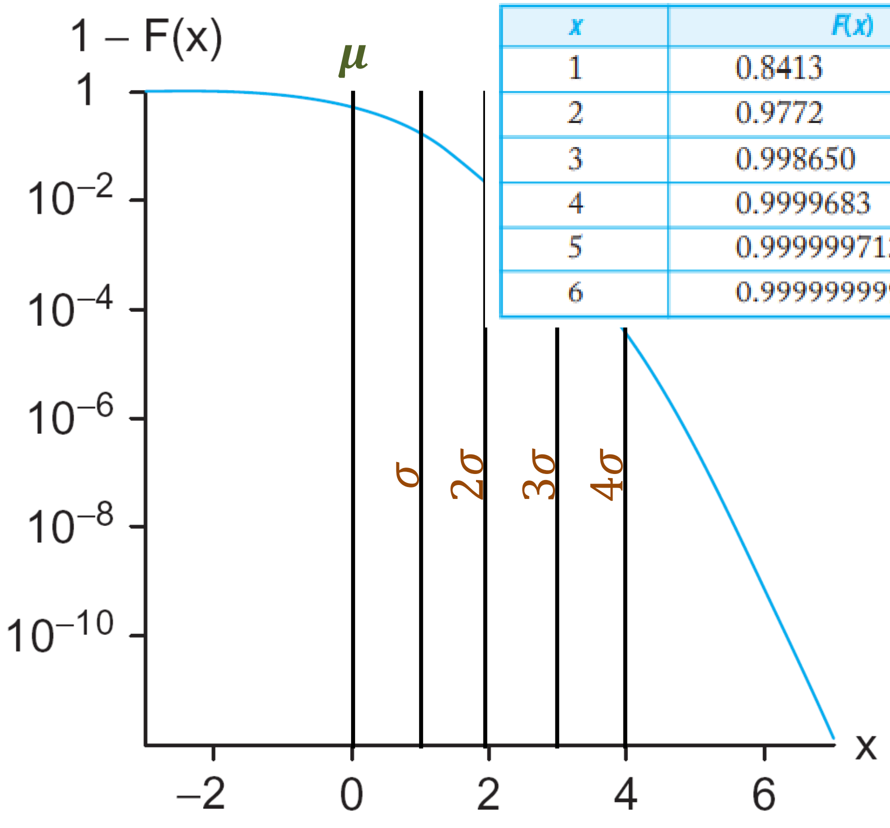
- **We can now easily compute the yield for any quality criterion x**

Analysing Very High Yield with MC Simulations

- **Yield for a gaussian quality metric with** $\mu = 0$ and variance $\sigma = 1$



(a)



(b)

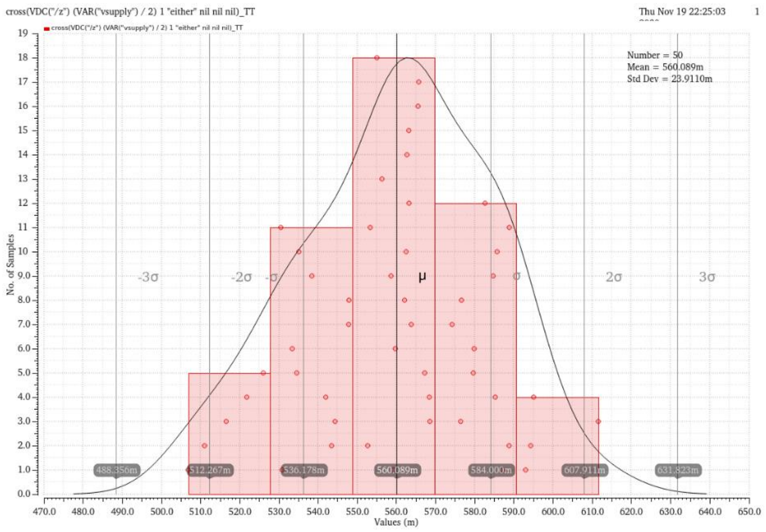
x	$F(x)$	$1 - F(x)$
1	0.8413	1.59×10^{-1}
2	0.9772	2.28×10^{-2}
3	0.998650	1.35×10^{-3}
4	0.9999683	3.17×10^{-5}
5	0.999999713	2.87×10^{-7}
6	0.999999999013	9.87×10^{-10}

Analysing Very High Yield with MC Simulations

- In reality, **metric distributions are** almost **never zero-mean with variance one, BUT they often have an almost-Gaussian distribution**

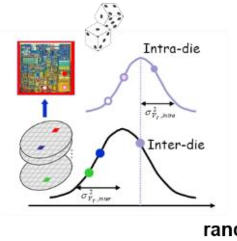
$$f_y\left(\frac{y-\mu_y}{\sigma_y}\right)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\bigg|_{x=\frac{y-\mu_y}{\sigma_y}}$$

- **Trick to extract a yield estimate from few simulations**
  1. Run a reasonable number of MC simulations
  2. Fit a Gaussian distribution to the simulation results
  3. Estimate μ_y and σ_y from the available data
  4. Formulate the quality criterion in multiples of σ_y $(\mathrm{n}\cdot\sigma_y)$ and get the yield from the reference Gaussian at $\sigma_x=\mathrm{n}$ **OR** find n from the desired component yield and check the minimum quality deviation from the mean you need to tolerate

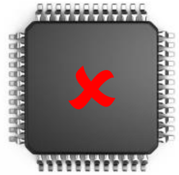


Yield Analysis for Qualification

- **Before a new technology (complex or fundamental circuit concept) is widely used, it must be "qualified"**
- **"Qualification" means the "demonstration that a technology can achieve high yield"** in high-volume manufacturing (proof of concept beyond MC-simulations).



100'000x



81x

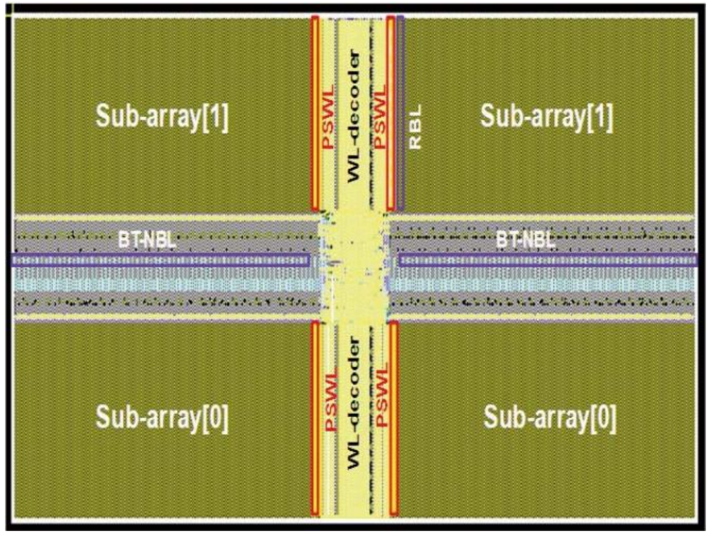
99'929x

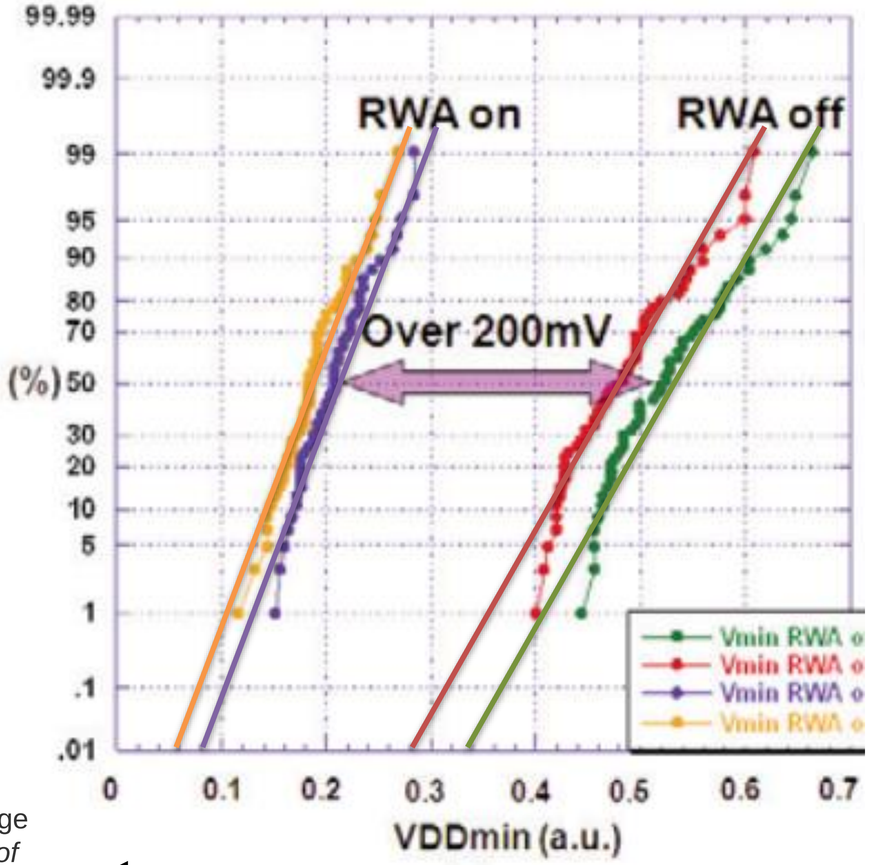
YIELD

- Typical expected yield: >95%

Qualification Example: SRAM

- **Silicon samples provide a yield assessment against operation parameters.**
- **Extrapolation of statistics for high-volume assessment**





Requires a lower offset

Probability for an offset to be lower than x*sigma

x	F(x)
1	0.8413
2	0.9772
3	0.998650
4	0.9999683
5	0.999999713
6	0.999999999013

J. Chang *et al*., "A 20nm 112Mb SRAM in High-κ metal-gate with assist circuitry for low-leakage and low-VMIN applications," *2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, San Francisco, CA, USA, 2013, pp. 316-317, doi: 10.1109/ISSCC.2013.6487750.