Time-Zero and Time-Dependent Variability in Advanced CMOS

Jeff Watt

October 11th, 2016

Disclaimer

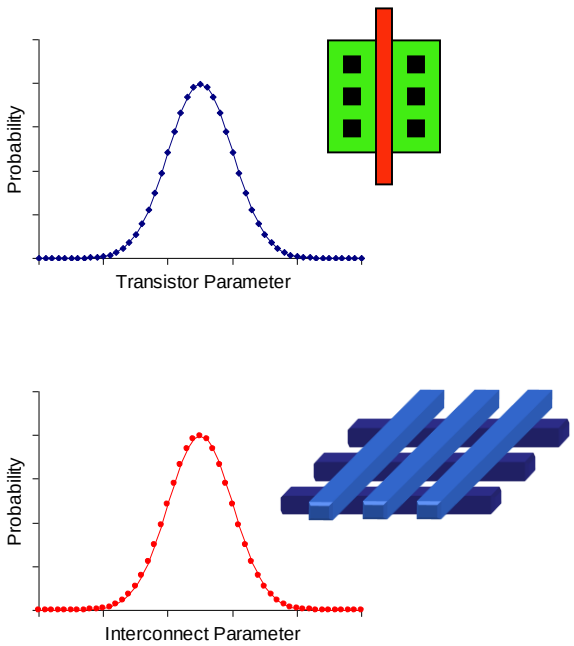
- All results presented here are based on R&D at Altera prior to acquisition by Intel and have been previously published
- Results and methodologies presented here do not represent Intel processes

Outline

- Introduction
- Variation components
- Array-based random variation extraction
- Time-zero and time-dependent results
- Summary

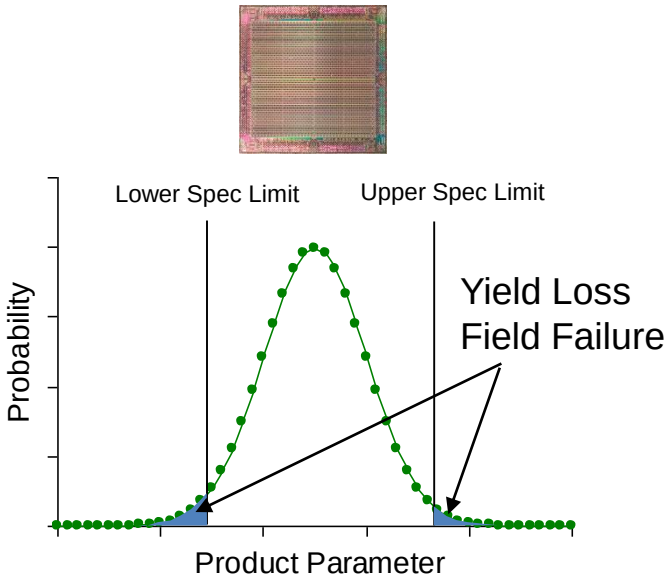
Introduction

Component variability



Variation classification
Characterization
Modeling
Simulation

Product variability



- Accurate modeling of component variability is critical for:
  - Predicting product variation
  - Limiting parametric yield loss
  - Ensuring no functional failures due to normal component variation

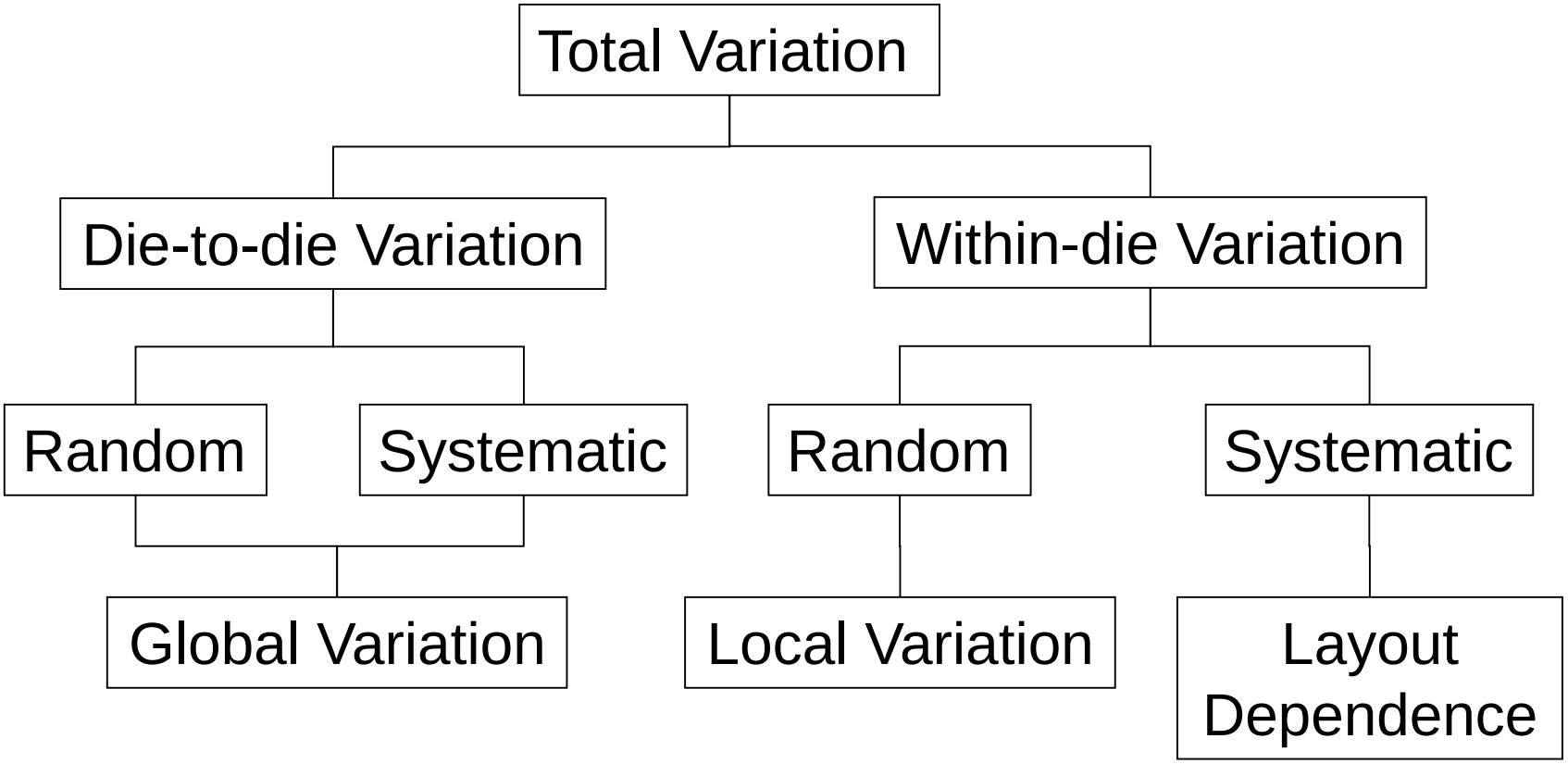
Components of Transistor Variation

- Process
  - Manufacturing
- Voltage
  - Customer power supply variation, IR drop
- Temperature
  - Ambient temperature, self heating, hot spots

Environmental

- time
  - Drift due to hot carrier injection (HCI), bias temperature instability (BTI)
  - Random telegraph noise (RTN)
  - Computational load dependent hot spots

Classification of Process Variation



Statistics Refresher

- If X_1 and X_2 are independent random variables, then:

$$\sigma^2_{a_1X_1+a_2X_2} = a_1^2\sigma^2_{X_1} + a_2^2\sigma^2_{X_2}$$

- And if they have identical standard deviations σ_X:

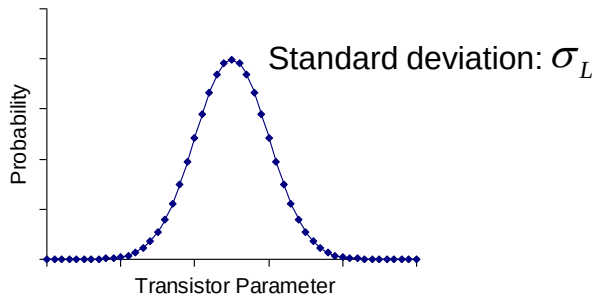
$$\sigma^2_{a_1X_1+a_2X_2} = (a_1^2 + a_2^2)\cdot\sigma^2_X$$

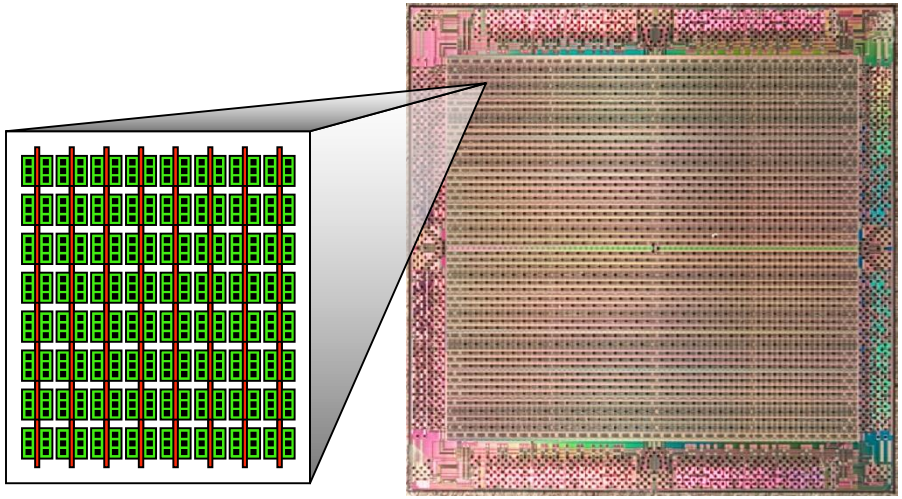
- Then:

$$\sigma^2_{X_1-X_2} = 2\sigma^2_X \Rightarrow \sigma_{X_1-X_2} = \sqrt{2}\sigma_X$$

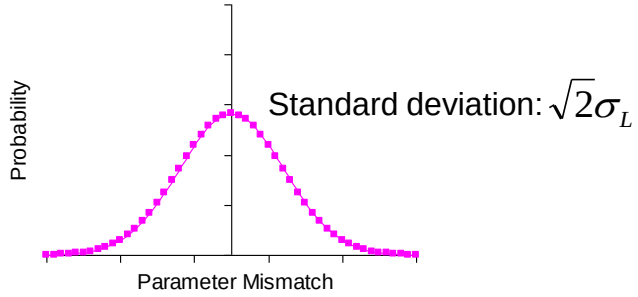
Within-die Random (Local) Variation

Distribution of "identical" transistors:





Mismatch between 2 "identical" transistors:



For independent random variables P_1 & P_2 with same standard deviation:

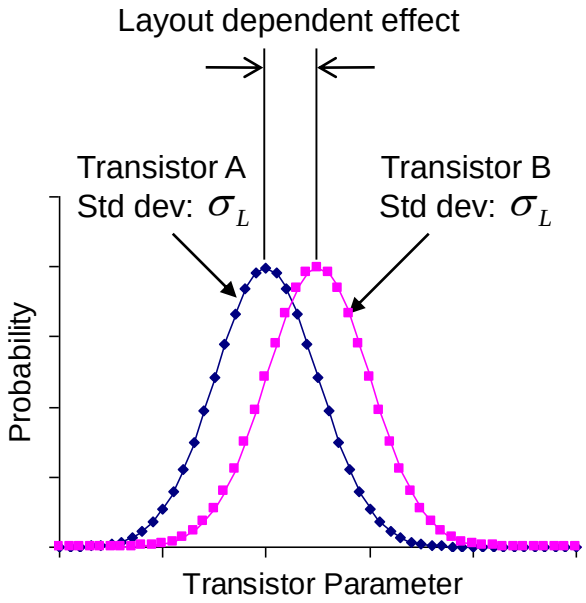
$$Var(P_2 - P_1) = Var(\Delta P) = \sigma_{P_2}^2 + \sigma_{P_1}^2$$

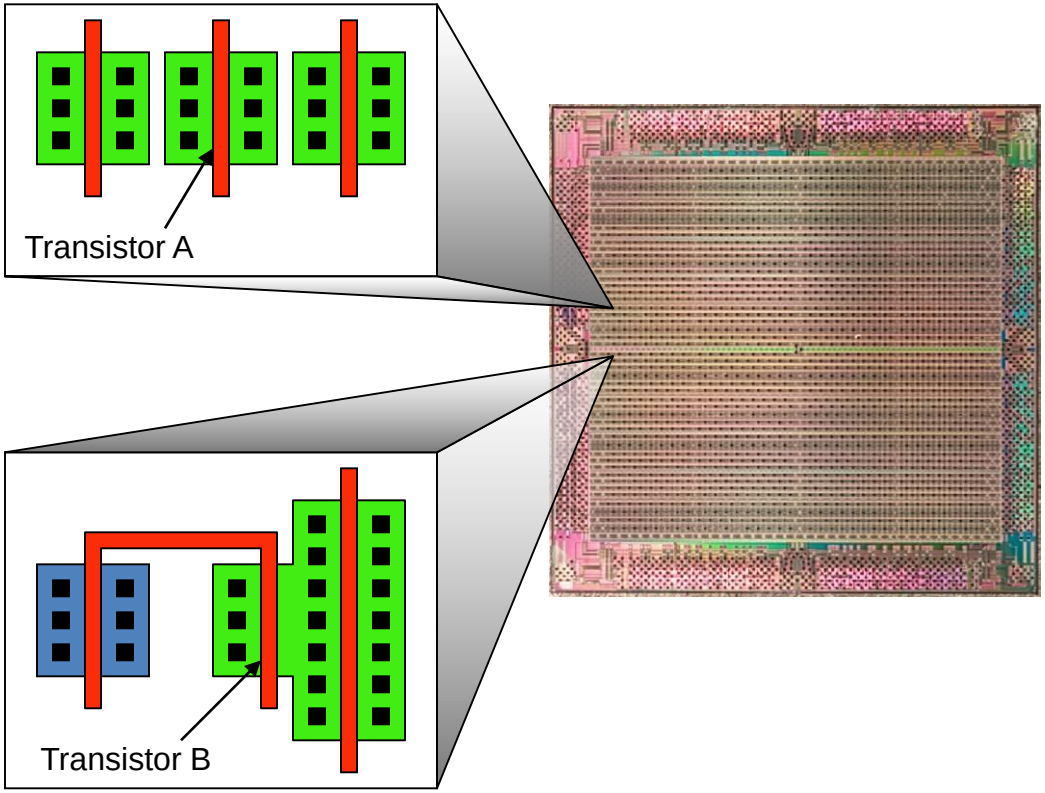
$$\sigma_{\Delta P}^2 = 2\sigma_P^2$$

$$\sigma_{\Delta P} = \sqrt{2}\sigma_P$$

- Transistors in close proximity with identical layout vary randomly with respect to each other and reference transistor
- Difference in parameter values (mismatch) between 2 "identical" transistors will vary randomly with standard deviation 1.4X of individual transistor standard deviation

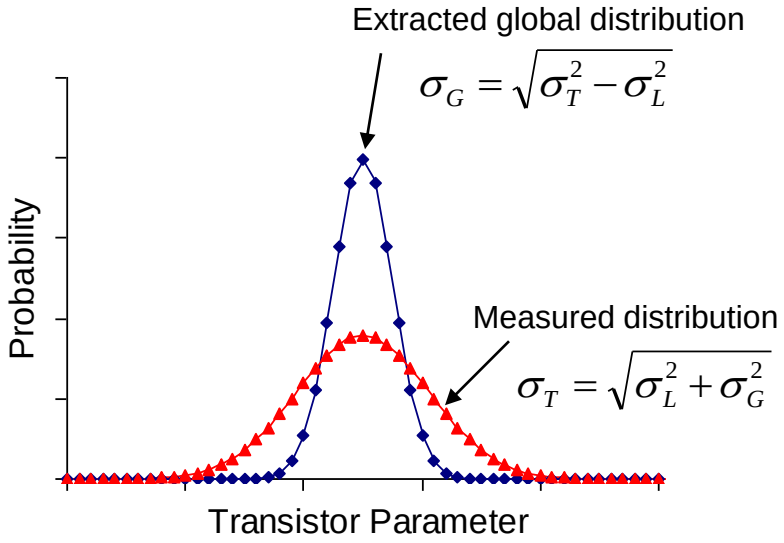
Within-die Systematic Variation

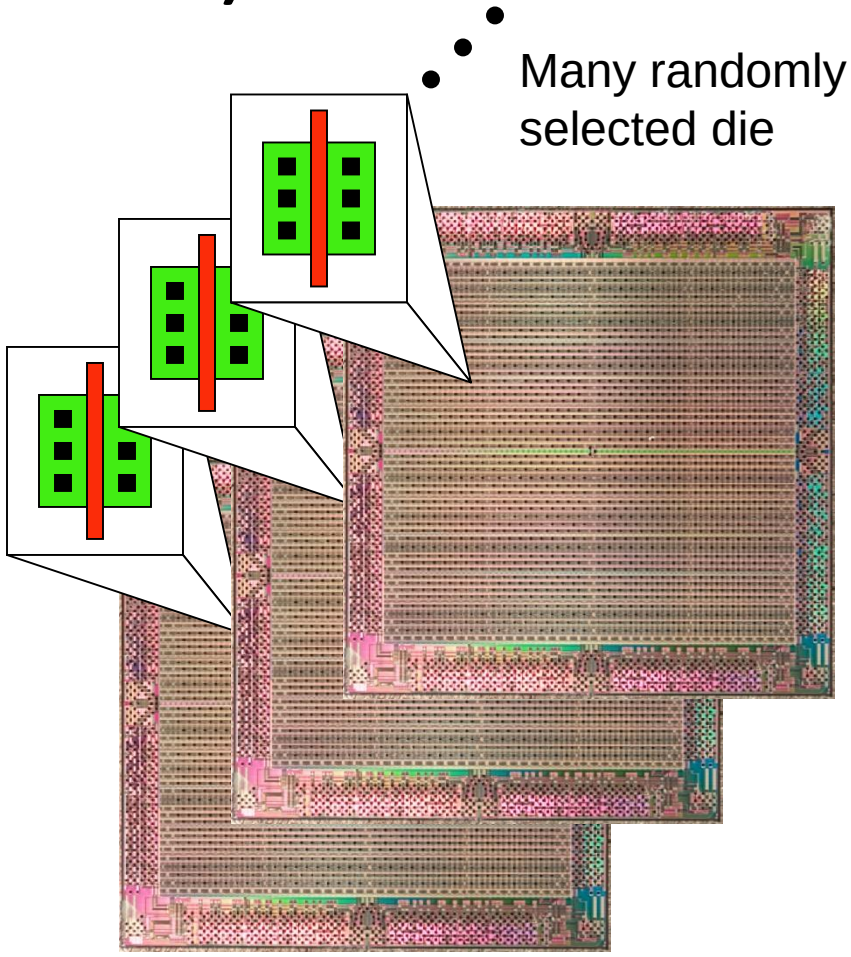




- Most within-die systematic variation due to layout differences, also known as “context”
- Each unique layout style has approximately the same amount of local variation
- Most systematic effects can be extracted from layout with appropriate models

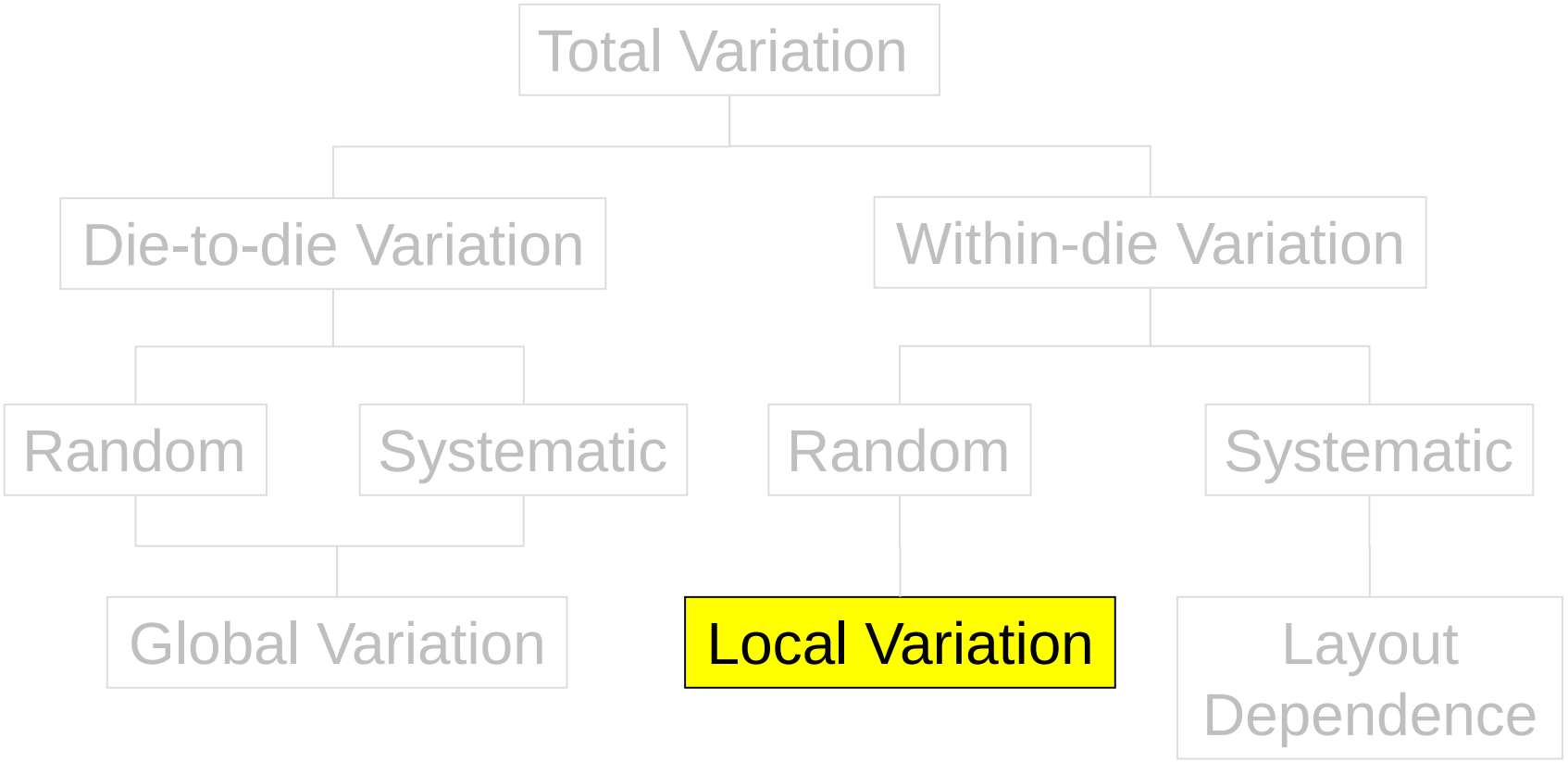
Die-to-die (Global) Variation



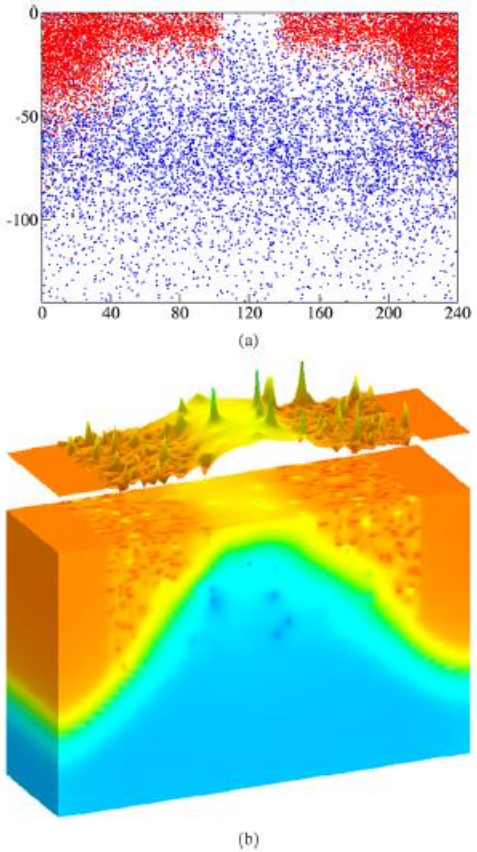


- Measurement of transistors with same context across multiple die gives total variation
- Global variation standard deviation is derived by removing local variation effect
- Die-to-die random and systematic effects pooled together using this approach

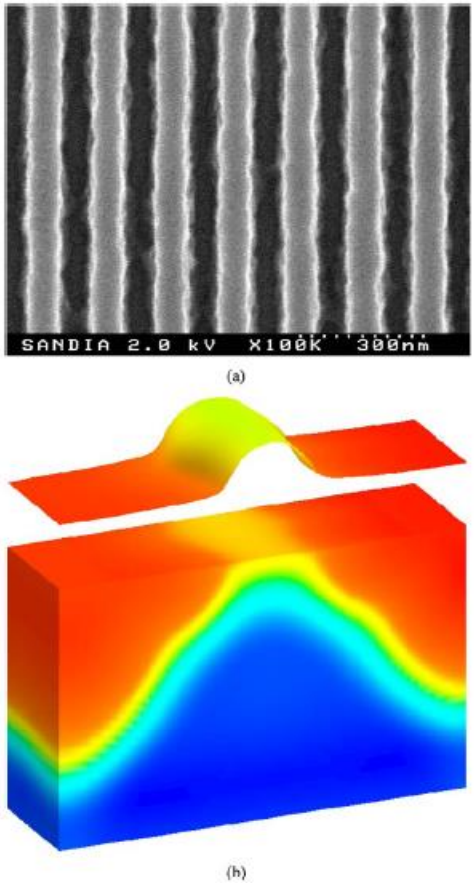
Classification of Process Variation



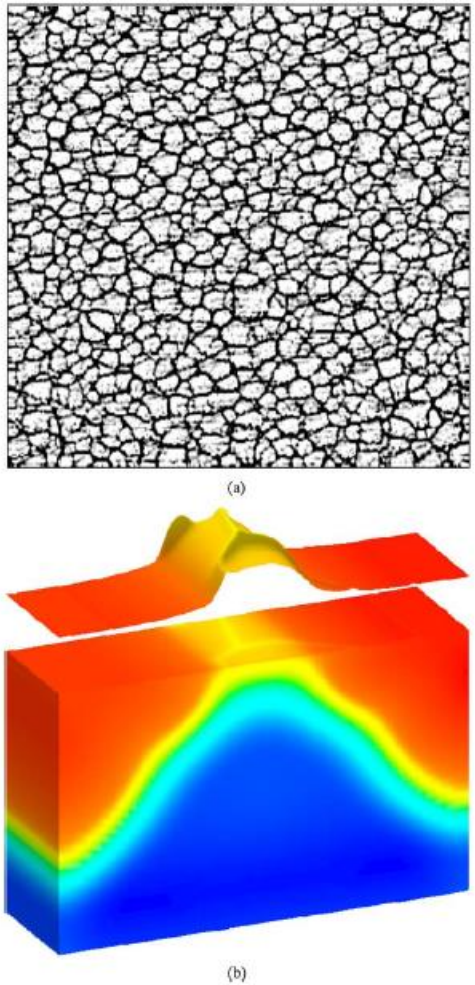
Random Local Variation Sources



Random Discrete Dopants (RDD)
aka Random Dopant Fluctuations (RDF)



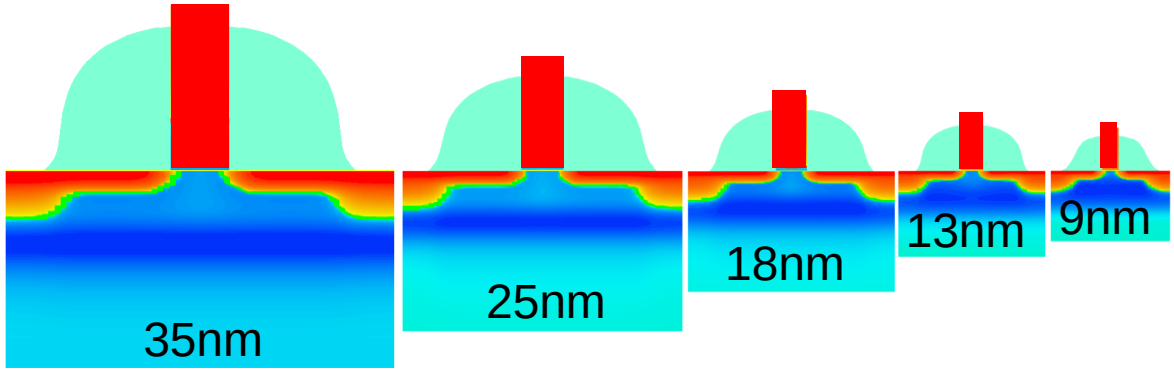
Line Edge Roughness (LER)

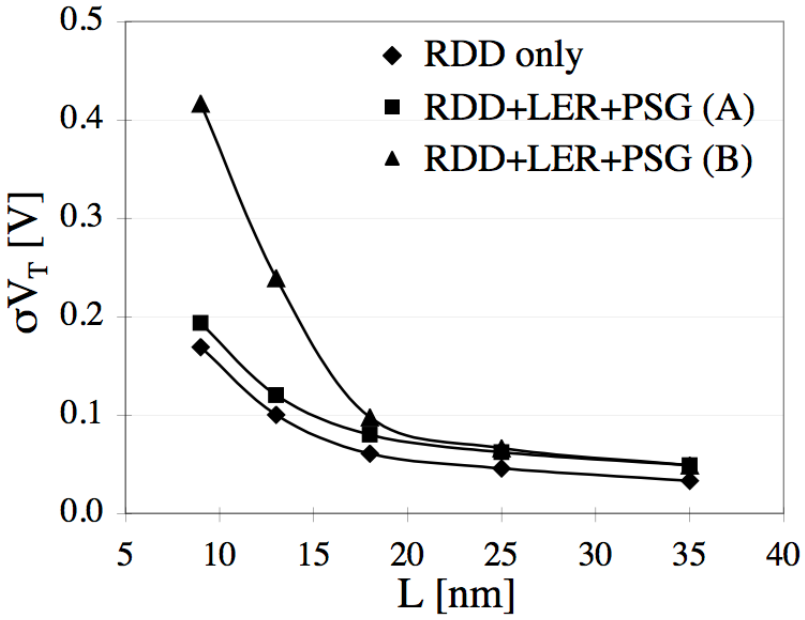


Metal Gate Granularity (MGG)

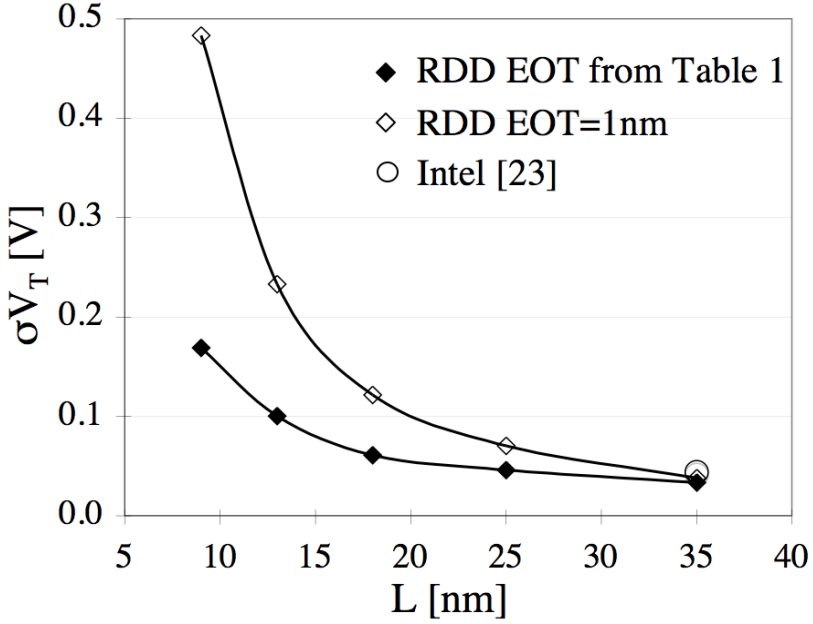
A. Asenov, J. Comput. Electron., 2009

Variability in Scaled MOSFETs





t_{ox} scales according to ITRS



t_{ox} remains constant

Extraction of Local Variation

IEEE JOURNAL OF SOLID-STATE CIRCUITS, VOL. 24, NO. 5, OCTOBER 1989

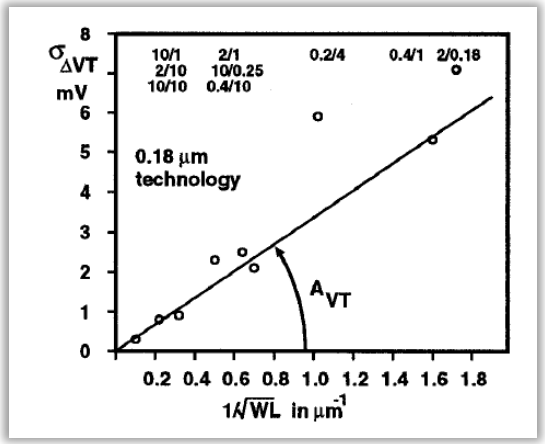
Matching Properties of MOS Transistors

MARCEL J. M. PELGROM, MEMBER, IEEE, AAD C. J. DUINMAIJER, AND ANTON P. G. WELBERS

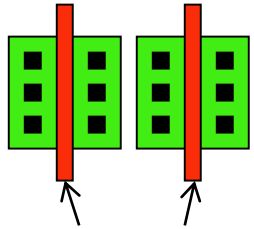
IEDM 98–915

Transistor matching in analog CMOS applications.

Marcel J.M. Pelgrom, Hans P. Tuinhout and Maarten Vertregt
Philips Research Laboratories, Bldg. WAY5, Prof. Holstlaan 4, 5656AA Eindhoven, the Netherlands,
FAX +31-40-2744657, e-mail: pelgrom@natlab.research.philips.com



$$\sigma_{\Delta VT} = \frac{A_{VT}}{\sqrt{WL}} = \frac{qt_{ox}\sqrt{2Nt_{depl}}}{\epsilon_0\epsilon_{ox}\sqrt{WL}}$$

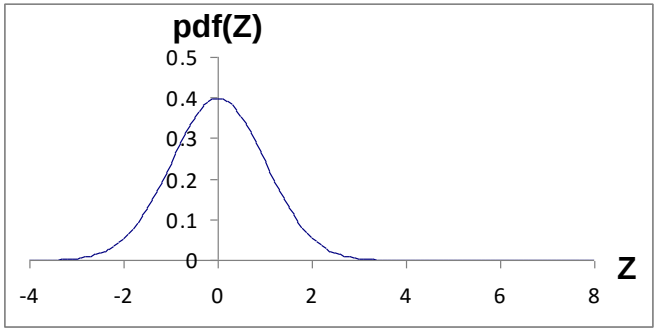


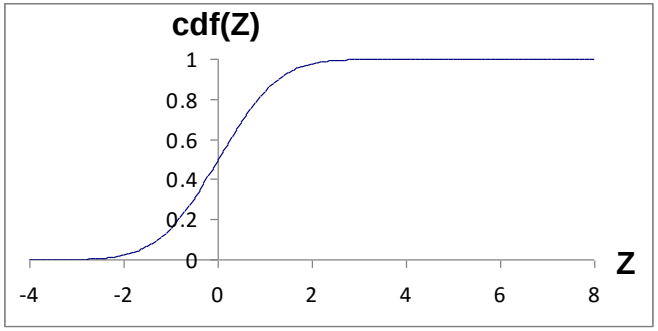
$$\Delta VT = VT_2 - VT_1$$

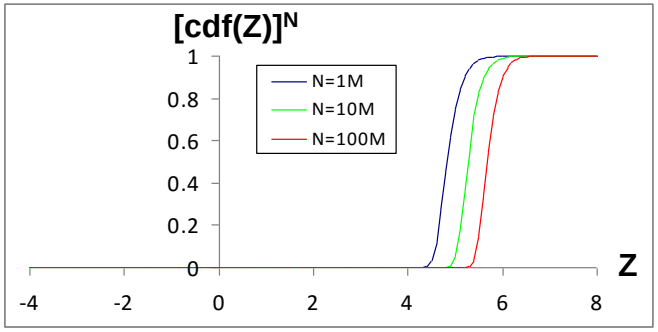
Local variation standard deviation is extracted from measurements of multiple matched pairs:

$$\sigma_{VT} = \frac{\sigma_{\Delta VT}}{\sqrt{2}} = \frac{A_{VT}}{\sqrt{2}\sqrt{WL}}$$

How Many σ's Are Enough?

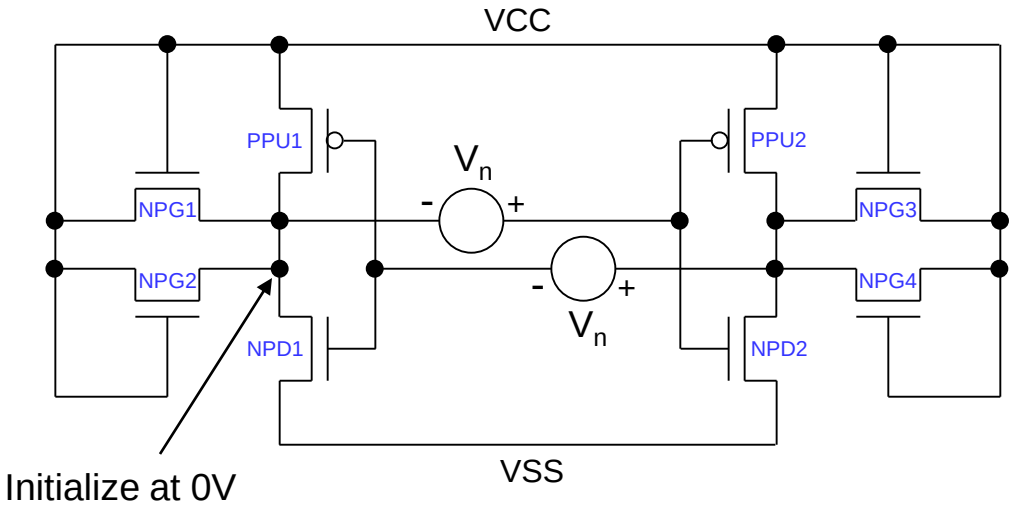






- 99% yield for one component per die
  - 2.3σ
- 99% yield for multiple components per die
  - 1M: 5.6σ
  - 10M: 6.0σ
  - 100M: 6.4σ
- Applies to any identical components which must simultaneously work within die (e.g., SRAM)

Dual-Port SRAM SNM Example



Static noise margin (SNM) = maximum V_n for stability

Test case	Worst-case cell transistor Vt shift (# of standard deviations)							
	NPD1	NPD2	NPG1	NPG2	NPG3	NPG4	PPU1	PPU2
SNM	3.5	-3.8	-1.9	-1.9	0	0	0	-1.5

6σ combined local variation for cell, but only 3.8 σ for worst-case transistor

A Better Way to Measure Local Variation

- Limitations of matched pair method:
  - Local Vt variation inferred from delta-Vt
  - Thousands of pairs required for high confidence
- High density transistor array enables:
  - Compact structure
  - Verification of normality of distribution
  - Measurement to high sigma
  - Evaluation of non-random local variations

Example of Transistor Array Structure

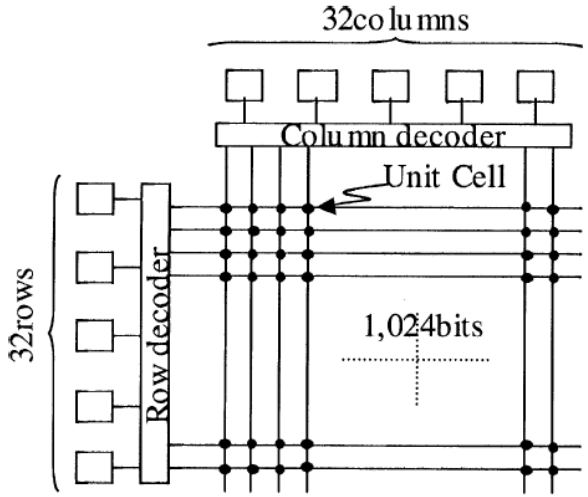


Fig. 1. Schematic illustration of 1024 bit transistor array test structure. 1024 MOSFETs are arrayed with each one being measured by selecting its address, permitting 1024 MOSFETs properties to be obtained.

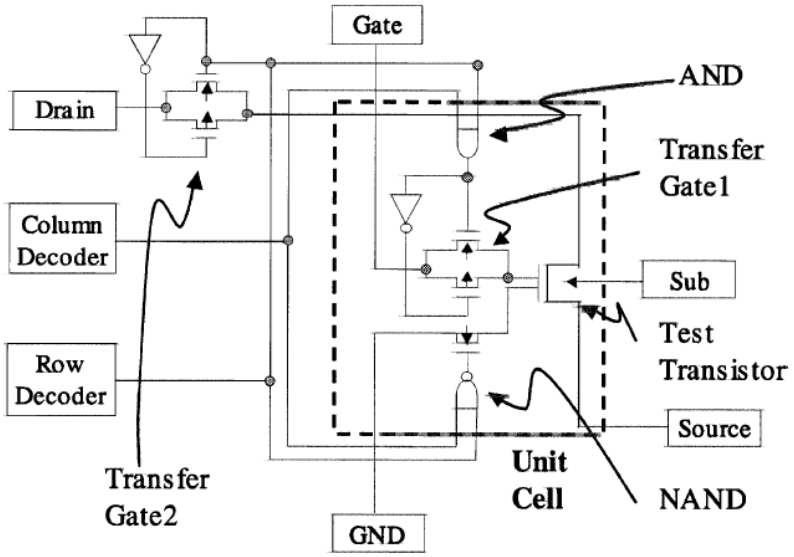


Fig. 2. Schematic image of a unit cell circuit. Unit cell consists of the test transistor, transfer gate 1, and AND and NAND circuits.

Izumi et al., IEEE Transactions on Semiconductor Manufacturing, Aug. 2004

- Complex unit cell required to select device under test and turn off unselected devices
- Very wide Drain transfer gate and Source/Drain metalization required to minimize IR drops

Enhanced Array with Sense Capability

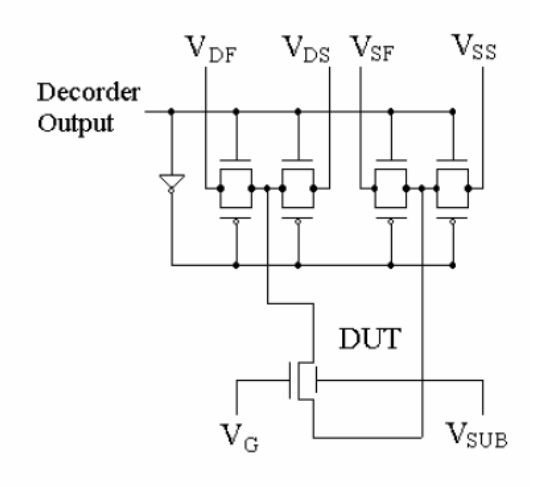


Fig. 2. Cell structure in the test circuit

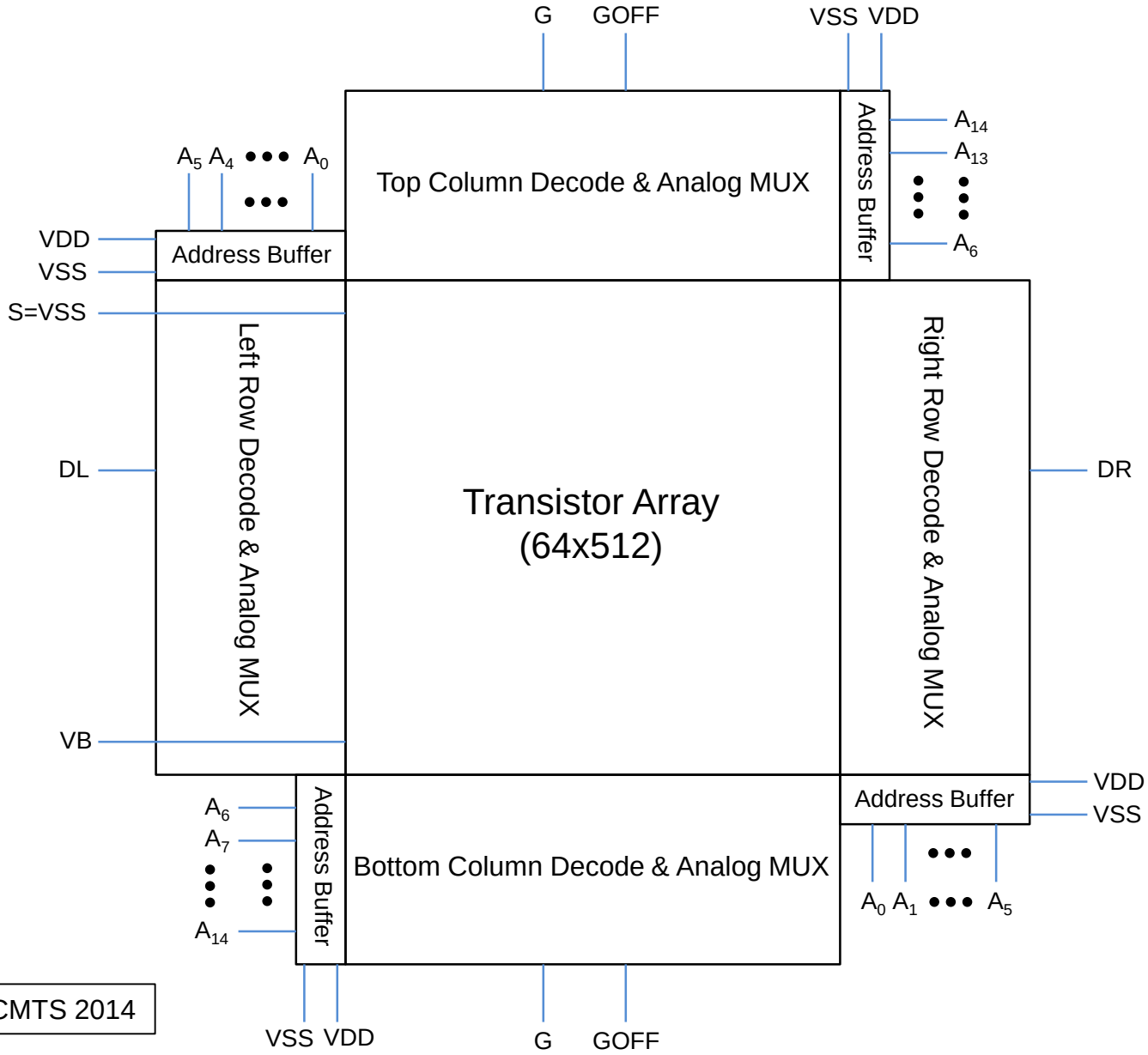
Chagawa et al., ICMTS 2008

- Voltage sense capability enables Kelvin measurement to directly measure and compensate for IR drop in transfer gate

Goals for Improved Transistor Array

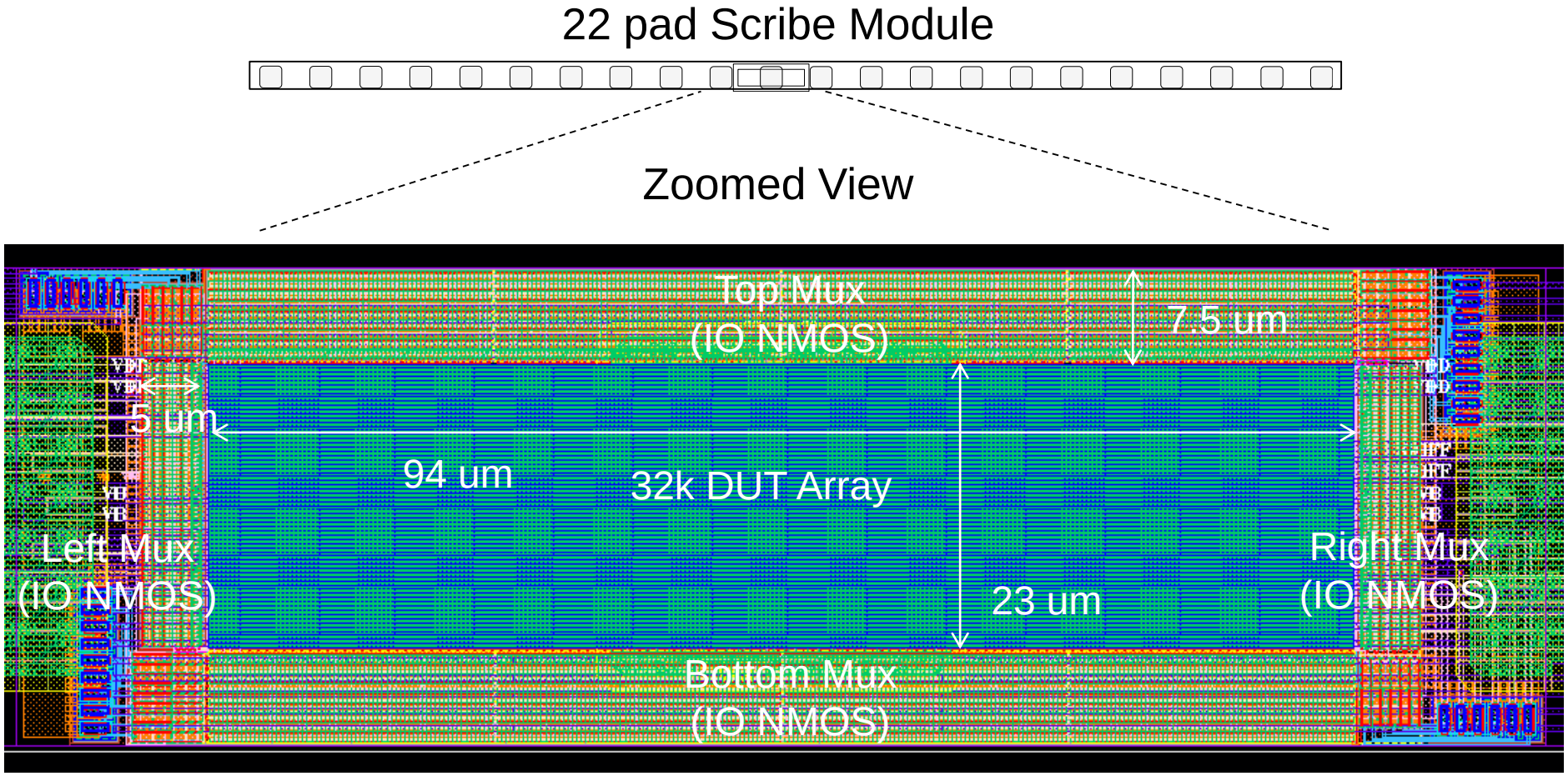
- Compact unit cell to enable high density array
- Design to fit in product scribe lane
- Simple addressing/decoding scheme
- Reduction of unselected transistor leakage
- Sense capability to measure/compensate voltage drop

Block Diagram of New Test Structure



Chen et al., ICMTS 2014

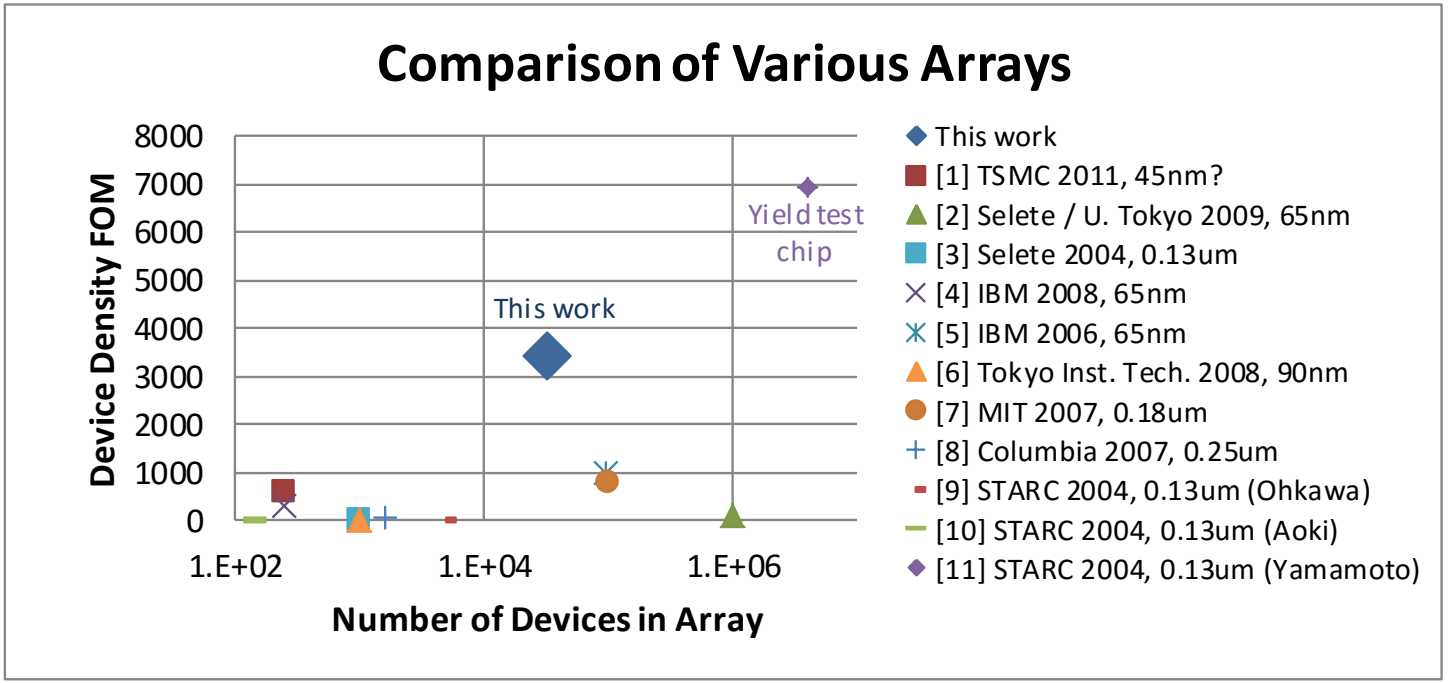
32k NMOS Array in Scribe Module



- About 30% of array area situated beneath a probe pad

Chen et al., ICMTS 2014

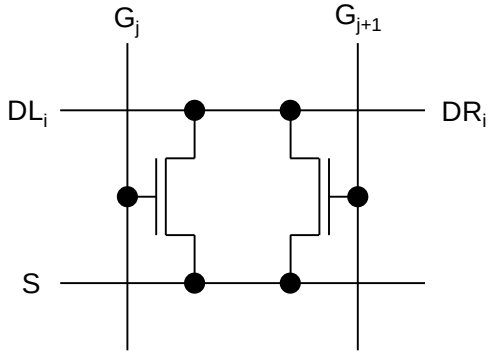
Transistor Array Landscape

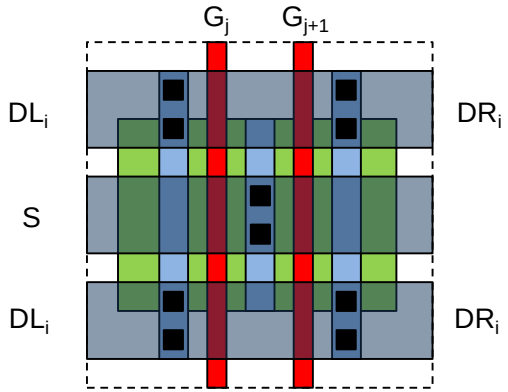


- Created figure of merit (FOM) to benchmark density
  - Number of DUTs per total area, after normalizing for technology scaling
  - Many add extra transistors to DUT and decoder basic cells for added functionality
  - STARC yield test chip is only for open/short (not parametric) test
- New array structure: >3X better area efficiency than similar arrays
  - Single scribe module testable using standard parametric tester

Chen et al., ICMTS 2014

Simple DUT Unit Cell

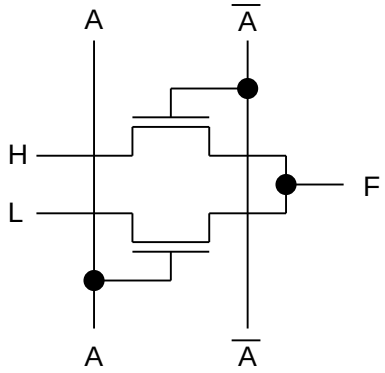


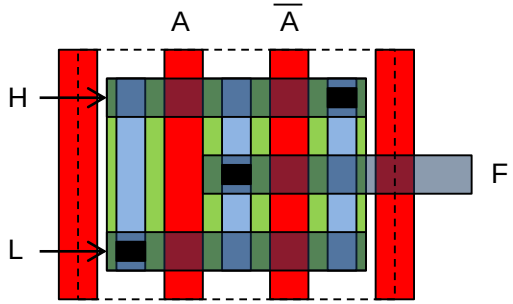


- Compact cell contains only 2 DUTs
  - No access or control devices within unit cell
  - Height/width pitch matched to mux
  - Left/right drain (DL/DR) simultaneously access same row for Kelvin sense
  - One gate connected to top mux, other to bottom mux
- Low resistance terminal connections
  - Drain strapped horizontally; source strapped vertically to reduce resistance
  - Common bulk connection at periphery of array

Chen et al., ICMTS 2014

Simple 2:1 Analog Mux Cell

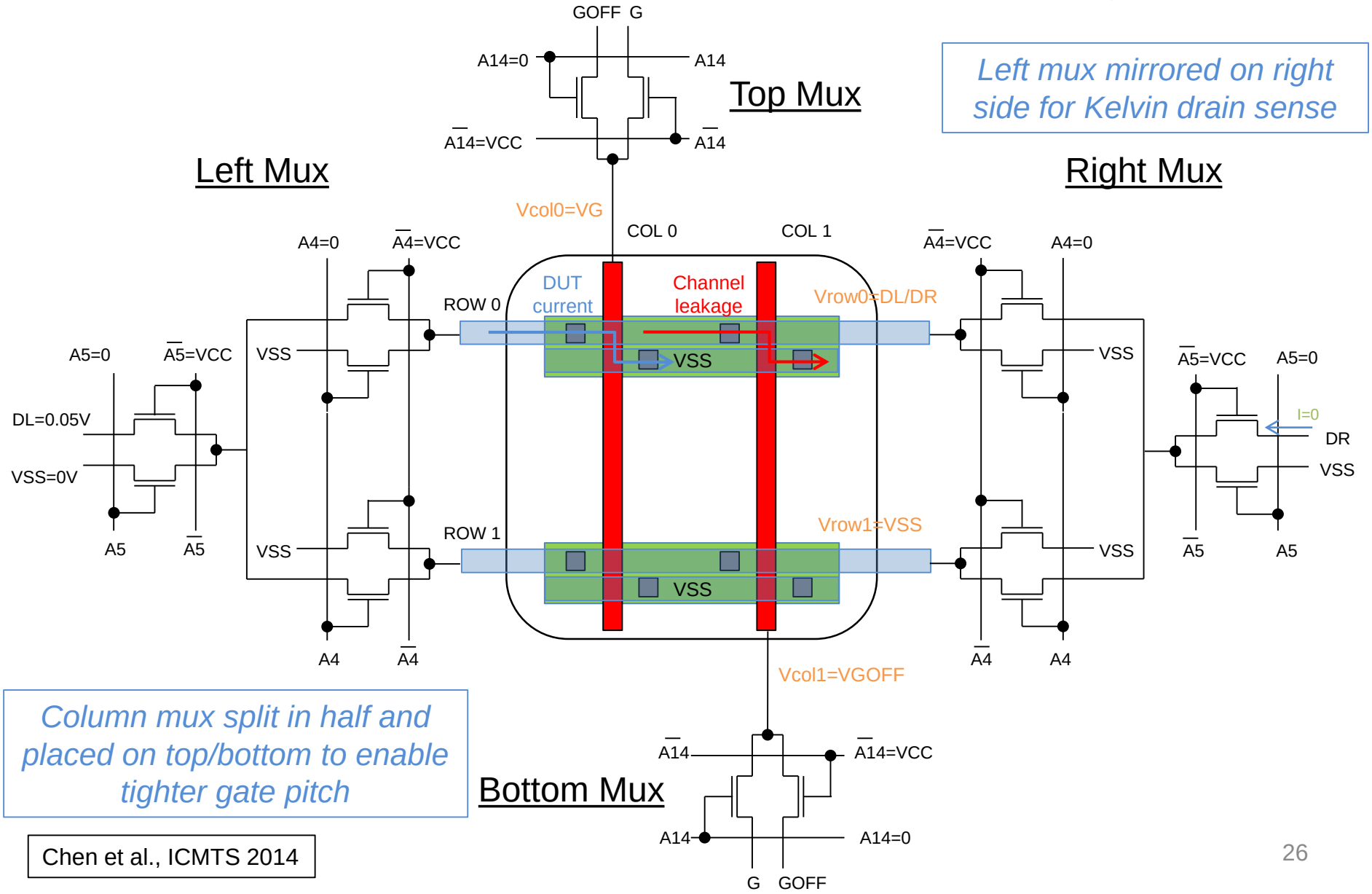




- Compact cell uses only 2 IO transistors for 2:1 mux
  - Cell is rotated 90° and used for top/bottom decoder
- Tree decoder implemented with tiled 2:1 analog mux unit cell
  - Active signal (drain/gate) passed through to one row/column
  - All inactive rows/columns connected to Vss/Vgoff
- Overdrive can be used to minimize mux resistance
  - I/O transistors enable core voltage to be passed through pass transistor mux

Chen et al., ICMTS 2014

Schematic of Mux + Array

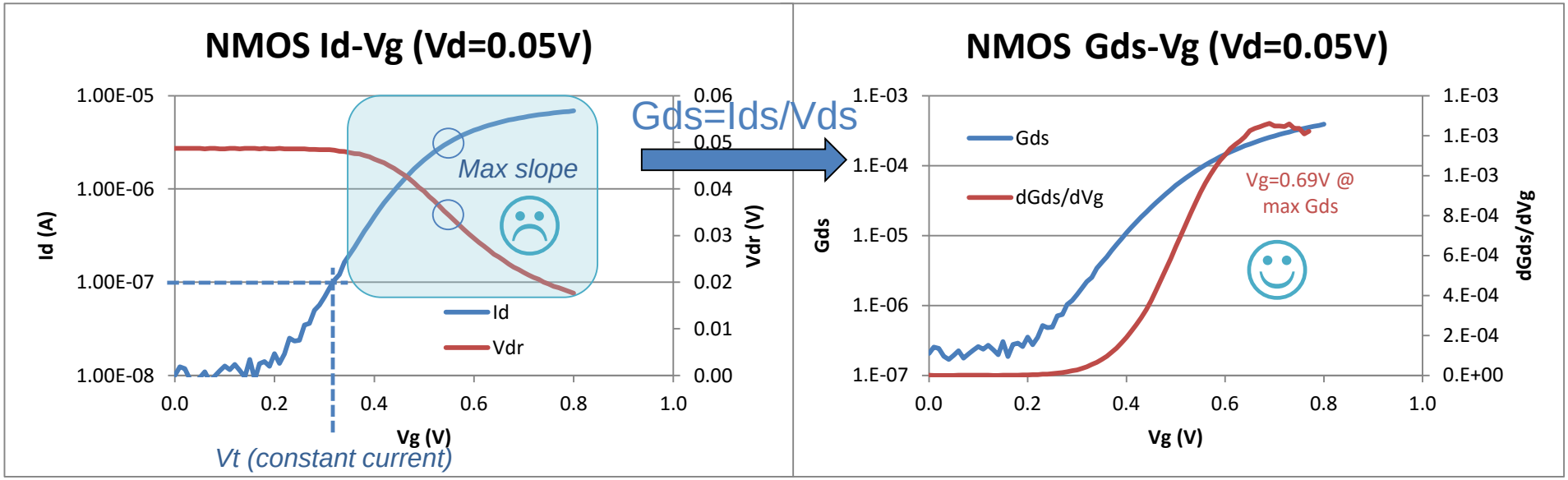


Left mux mirrored on right side for Kelvin drain sense

Column mux split in half and placed on top/bottom to enable tighter gate pitch

Chen et al., ICMTS 2014

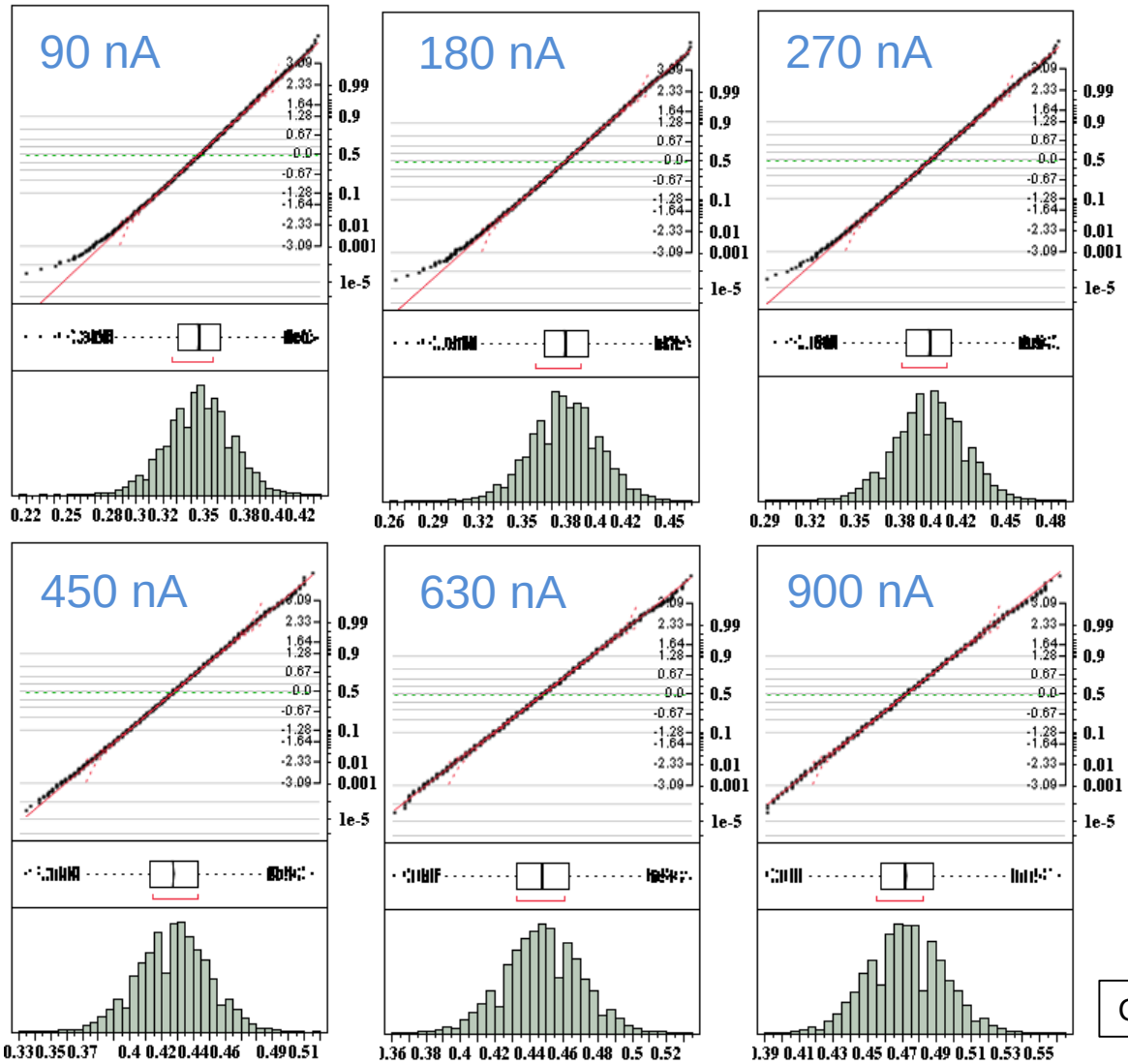
Parametric Measurements



- Constant current Vt can be accurately measured (low Id)
  - Leakage from unselected columns suppressed using VGOFF=-0.2V
- High IR drop at high drain currents
  - Can't extract Idlin, Idsat, Vtgm due to series resistance of mux, wiring, etc.
- To mitigate series resistance, extract Vtgm from Gds-Vgs
  - Gds = Ids/Vds ← Vds measured using Kelvin connection to transistor drain
  - Channel conductance is independent of series resistance

Chen et al., ICMTS 2014

Vtlin Distributions at Various Current Levels



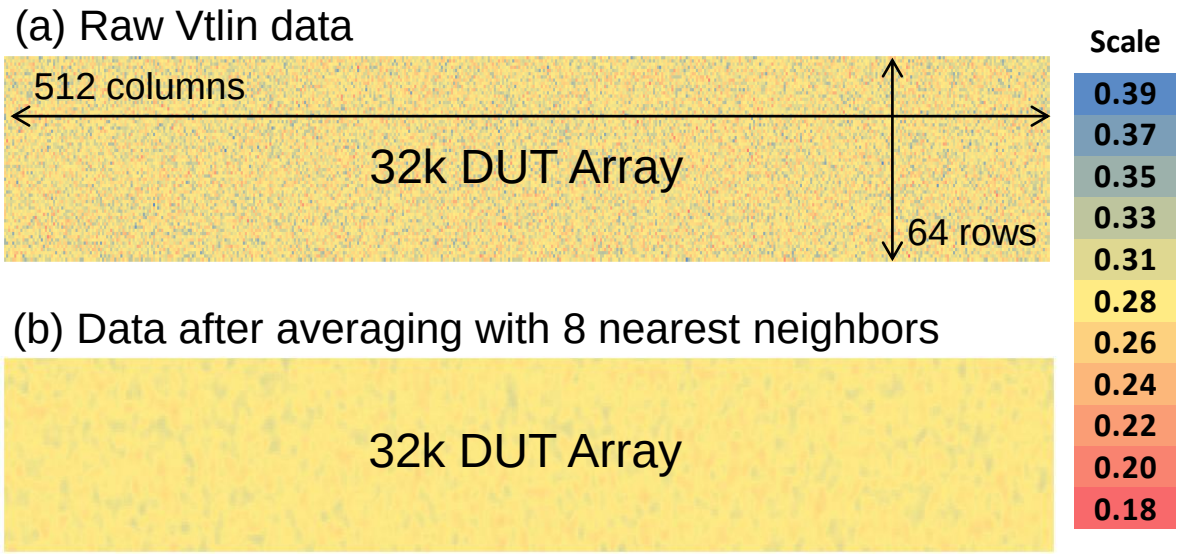
Gaussian distribution up to ~4σ for higher current levels used in memory circuits

Chen et al., ICMTS 2014

- Distribution tail diminishes at higher currents

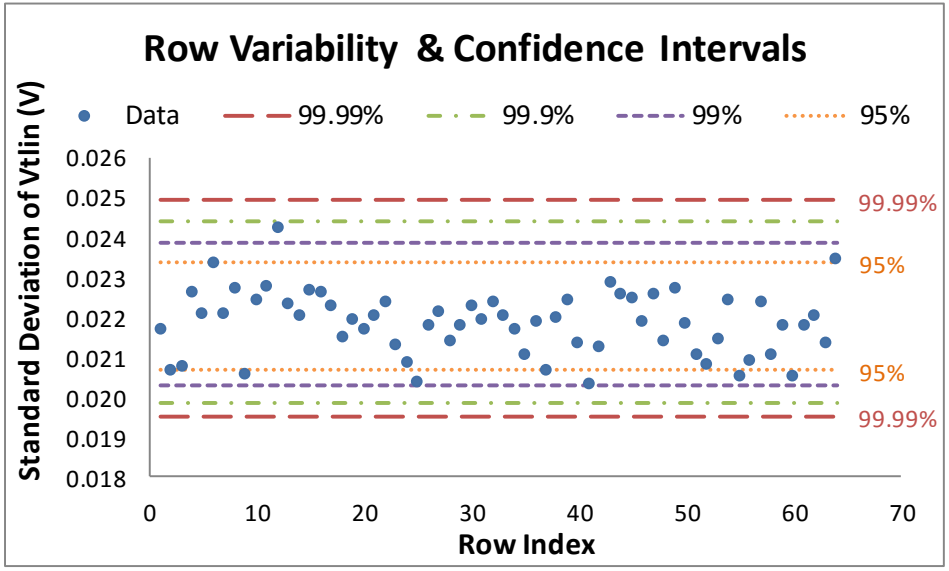
Spatial Variations

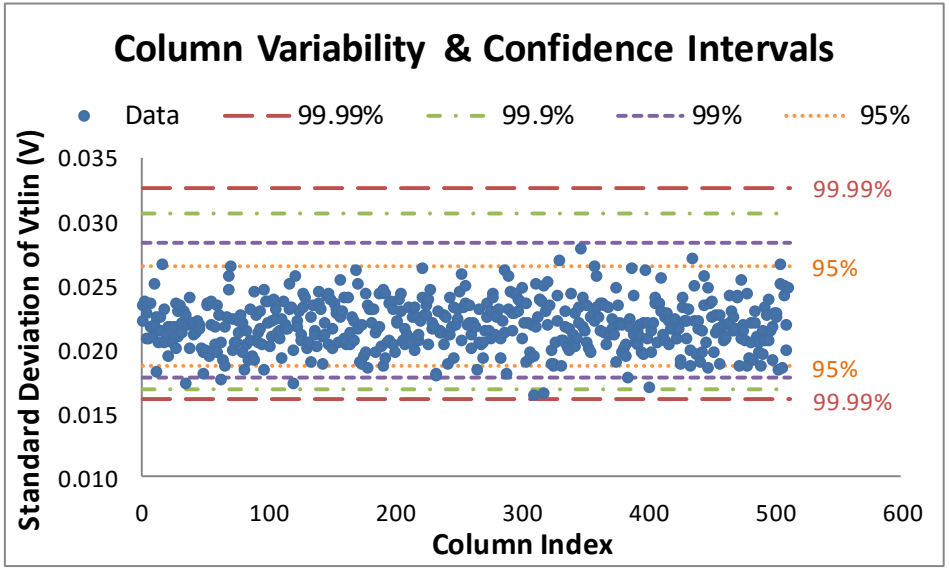
- Contour plots confirm that random variation dominates
  - Mismatch analysis shows that 99% of total variation is random
- Threshold voltage shows no spatial variations
  - Local variation overshadows effects of probe pressure, layout, etc.



Chen et al., ICMTS 2014

Variability per Row and Column

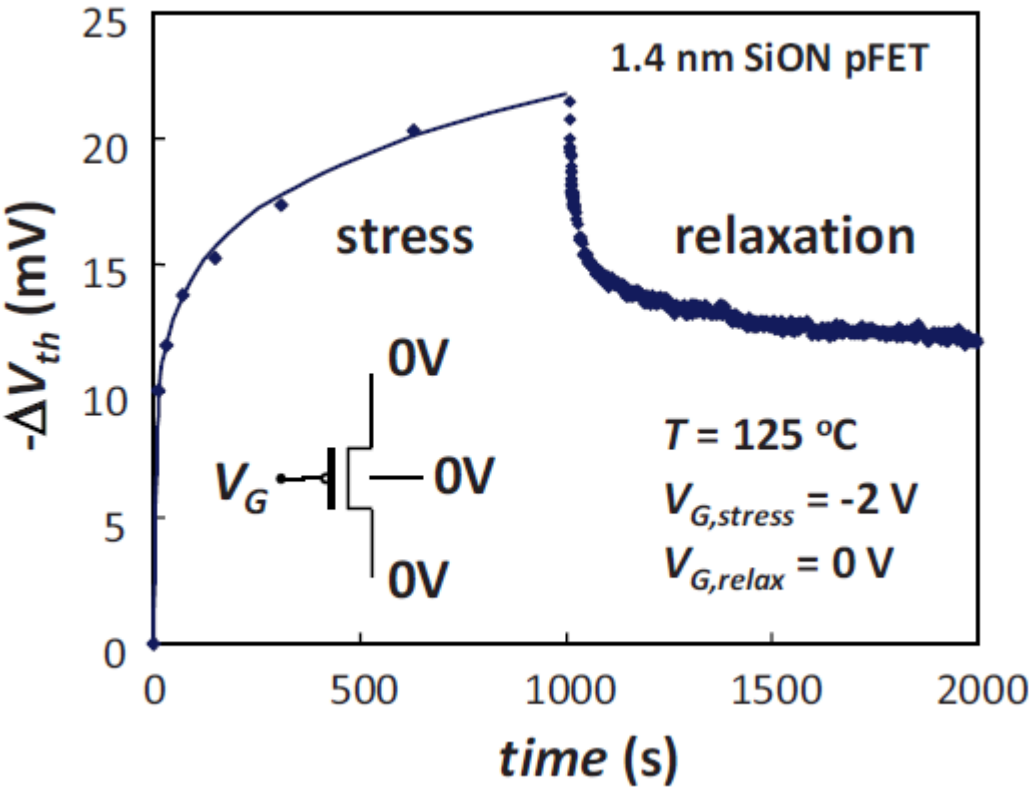




- No increase in variability at edges of array
  - Data is close to population standard deviation
- Analysis of row and column averages looks similar
  - Data not shown

Chen et al., ICMTS 2014

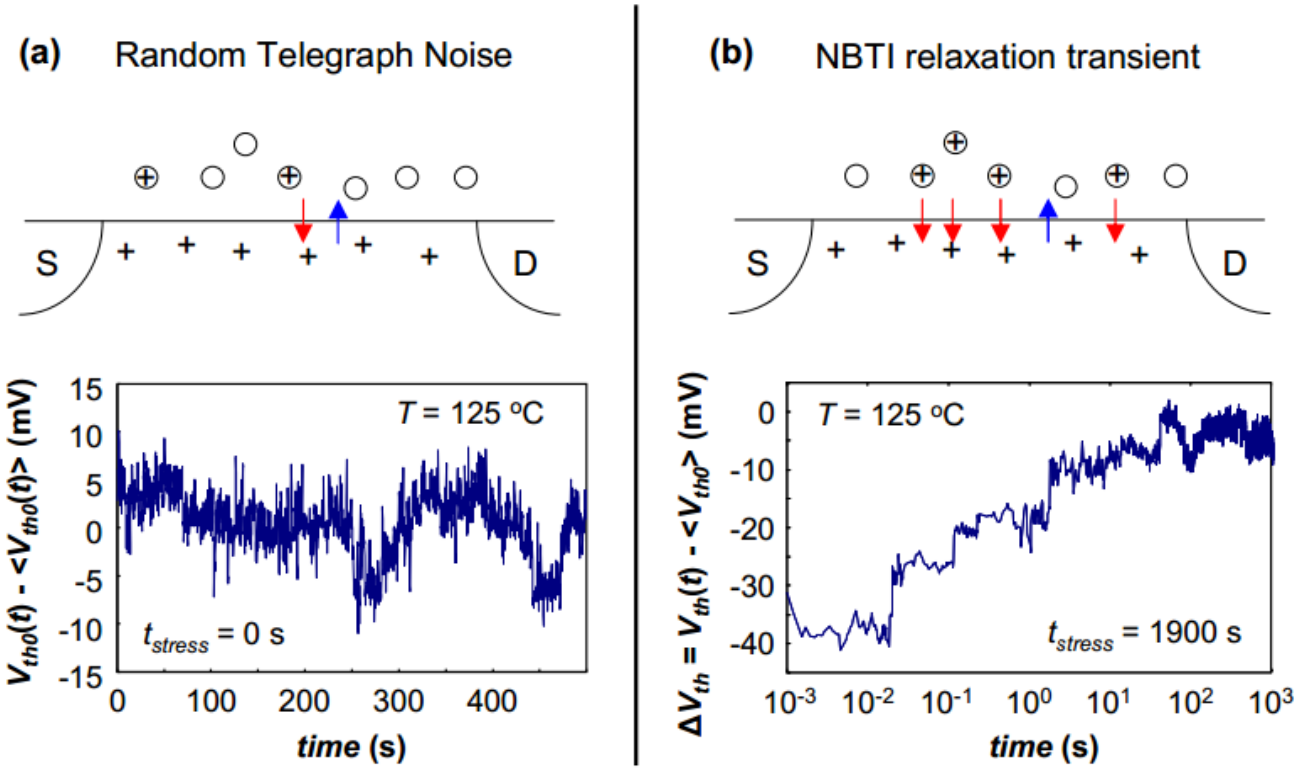
Time Dependent Variation



Kaczer et al., J. Vac. Sci. Tech., 2011

- BTI causes time-dependent variation
  - Systematic shift of Vt
  - Increase in local variation

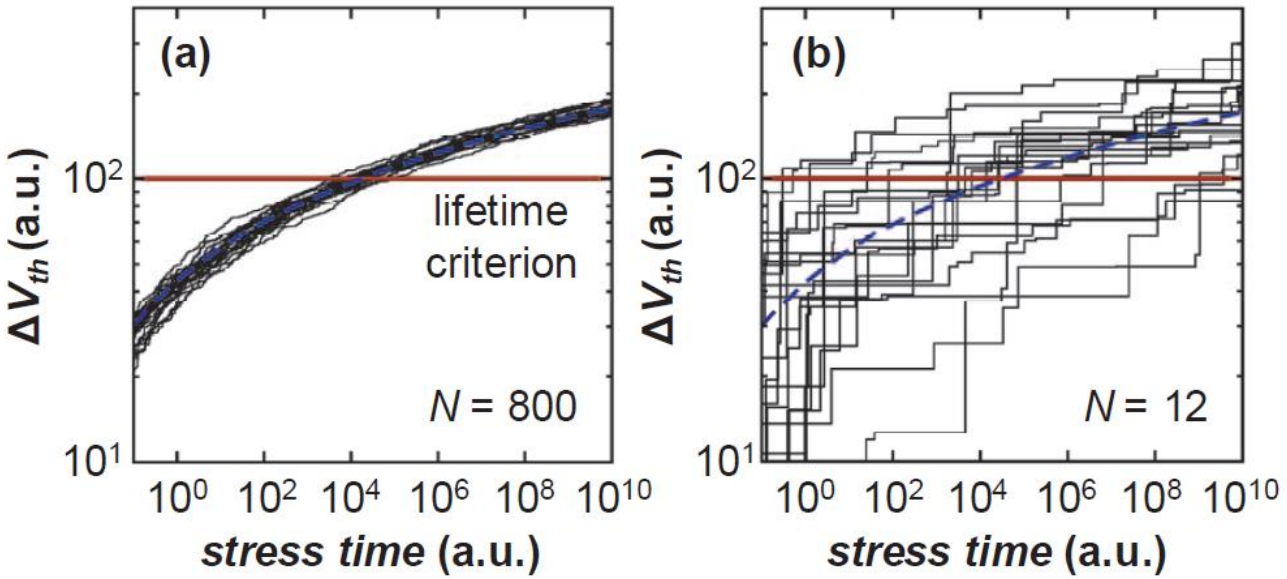
Defect-Centric Model for RTN and BTI



Kaczer et al., IRPS 2010

- RTN: Channel/dielectric in equilibrium
- BTI relaxation: Perturbed system returning to equilibrium

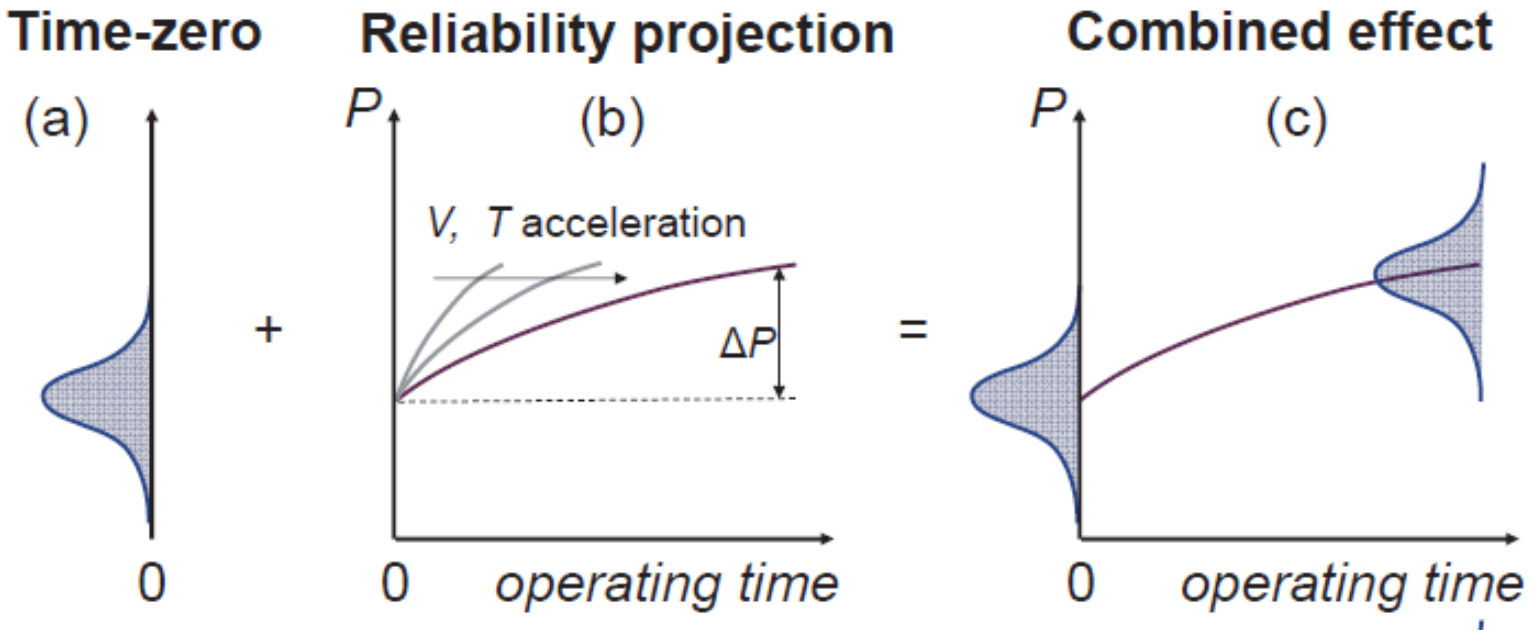
Increased Variability in Small Devices



Kaczer et al., ESSDERC 2015

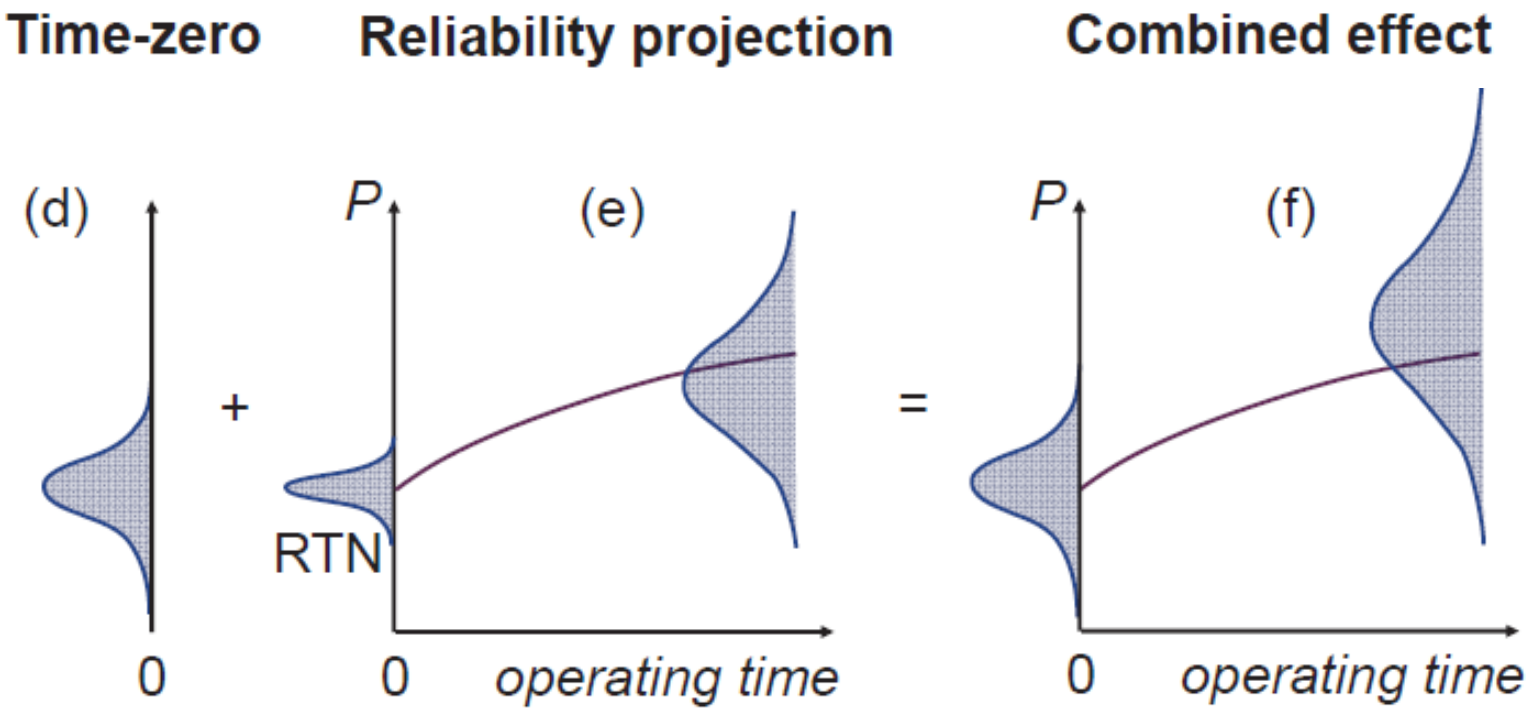
- NBTI/PBTI induced Vt variation increases in small devices due to reduced number of oxide defects

Time-Dependent Variability Modeling: Traditional Approach



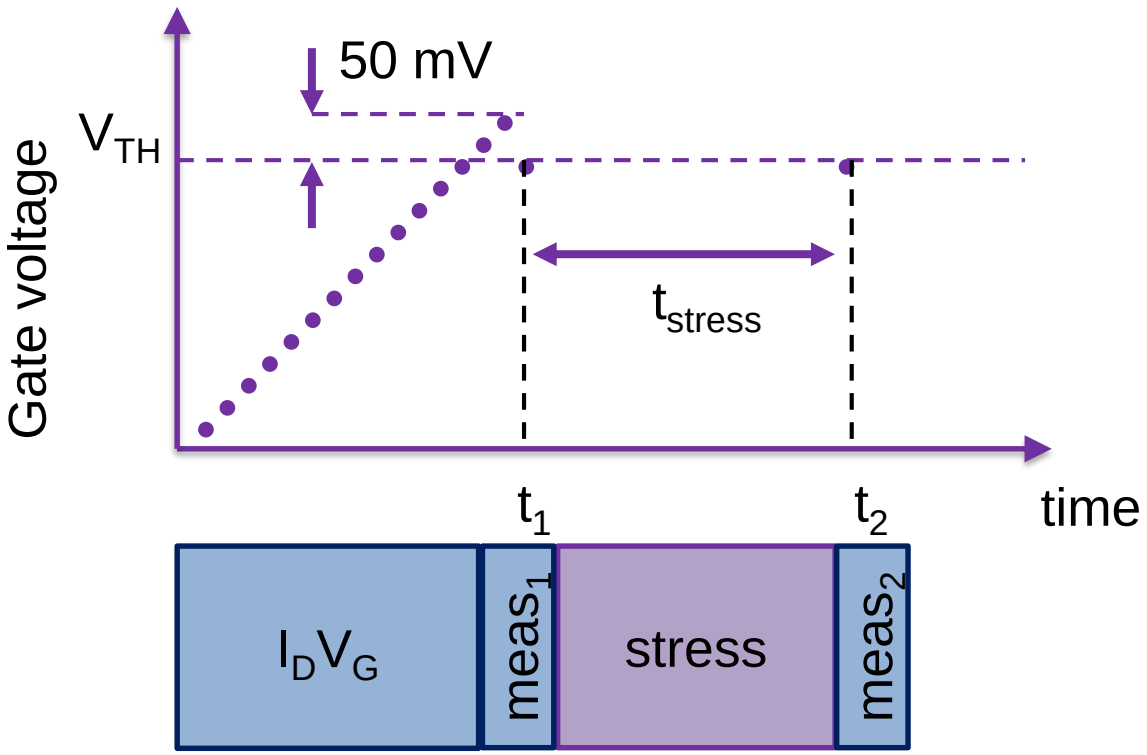
Kaczer et al., ESSDERC 2015

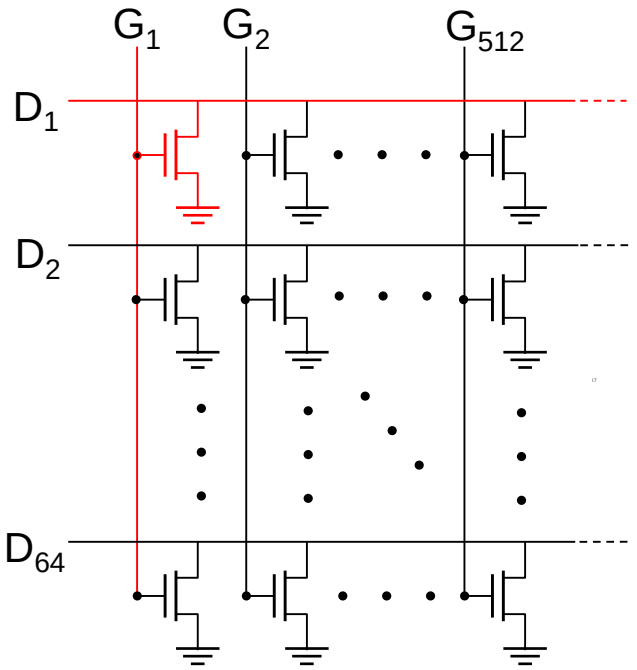
Time-Dependent Variability Modeling: Improved Approach



Kaczer et al., ESSDERC 2015

Time-Zero and Time-Dependent Variability Measurement

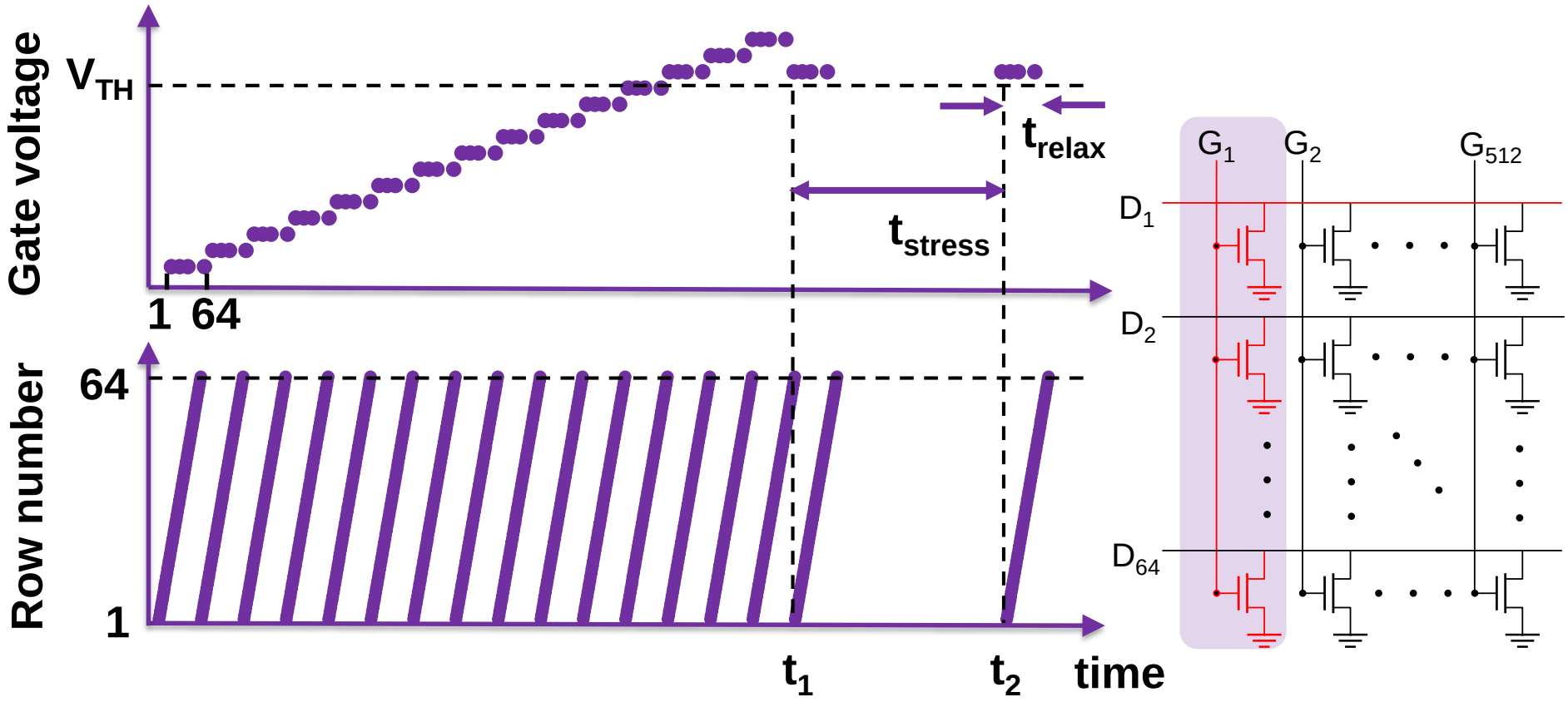




- Initial I_D-V_G with V_D=50mV and V_G swept to $|V_{TH}|$ + 50mV
- After I_D-V_G, two consecutive I_D current measurements around $V_G = V_{TH}$
- Timing window between t_1 and t_2 can be used to apply stress on DUTs, stressing 64 devices in parallel

Weckx et al., IRPS 2015

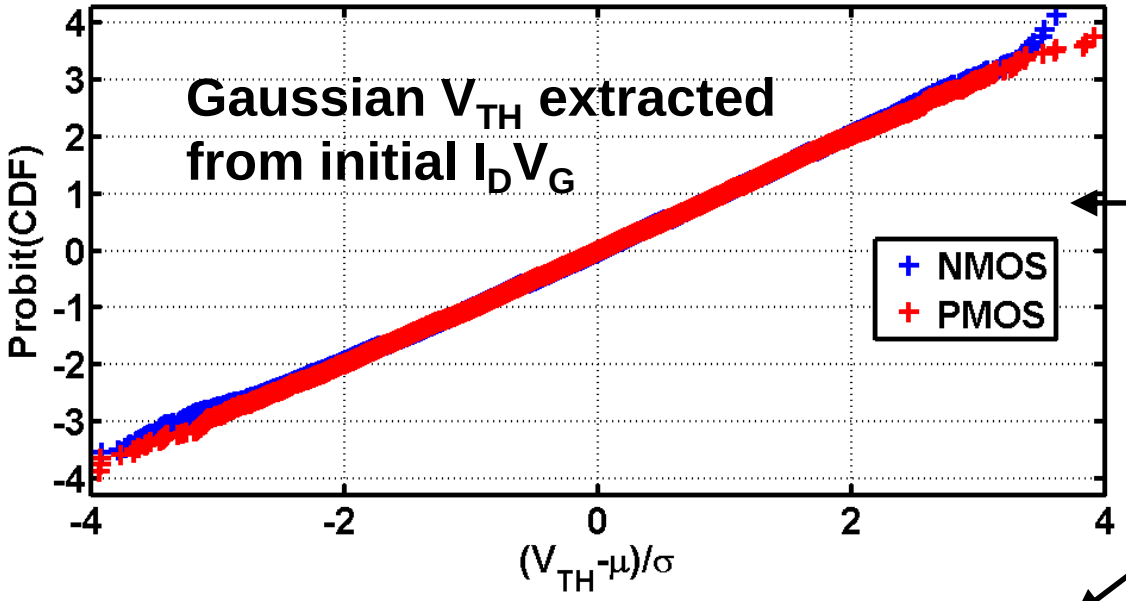
Multiplexing to Increase Throughput

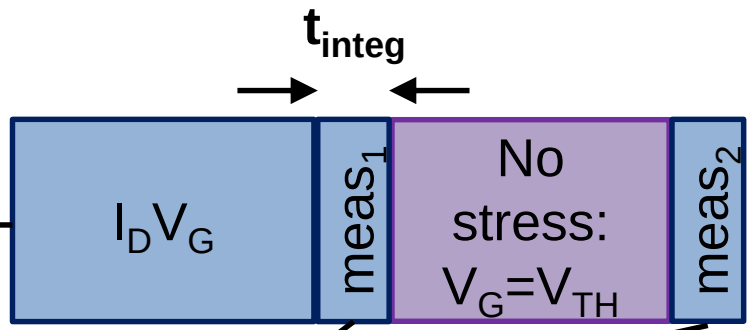


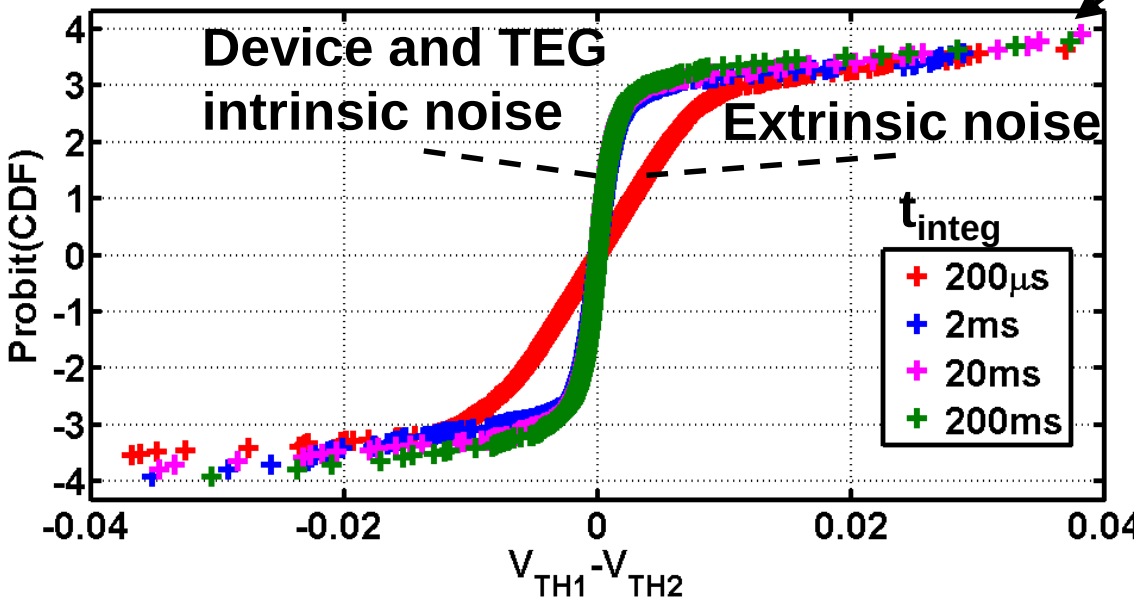
- Measure-stress-measure for entire array one column at a time
- Fast serial measurements of each row DUT after the removal of stress
- All rows are measured out serially for each gate voltage
- Current traces are multiplexed

Weckx et al., IRPS 2015

Time-zero and RTN Variability Extraction



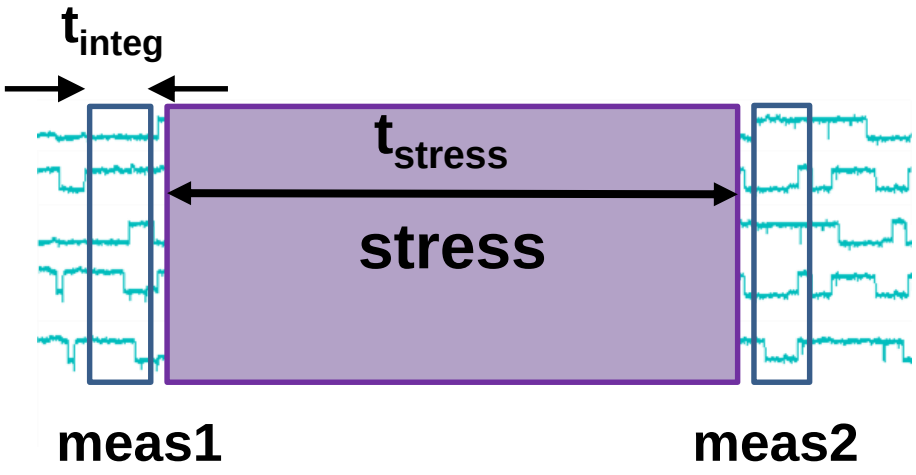


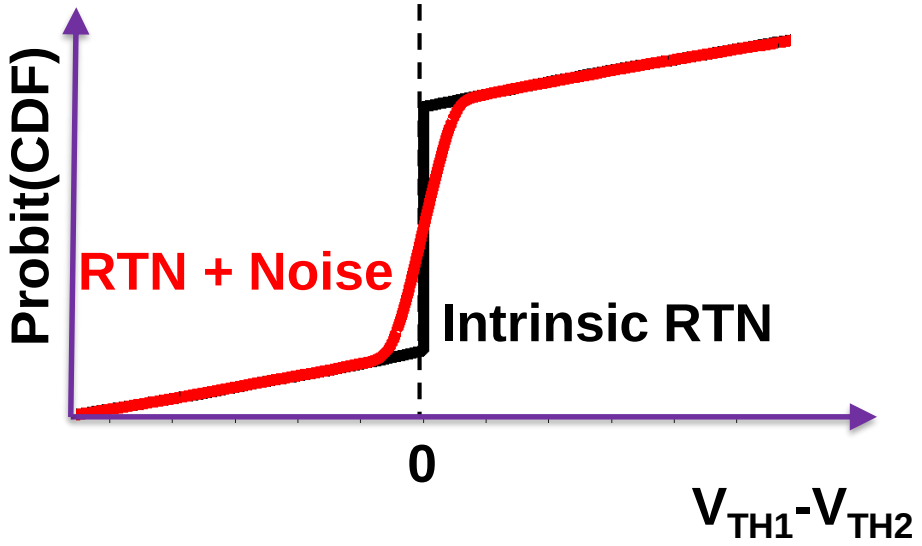


Temporal Noise extracted from two point current measurements with different integration time

Weckx et al., IRPS 2015

Noise Due to (Dis)Charging During Measurement



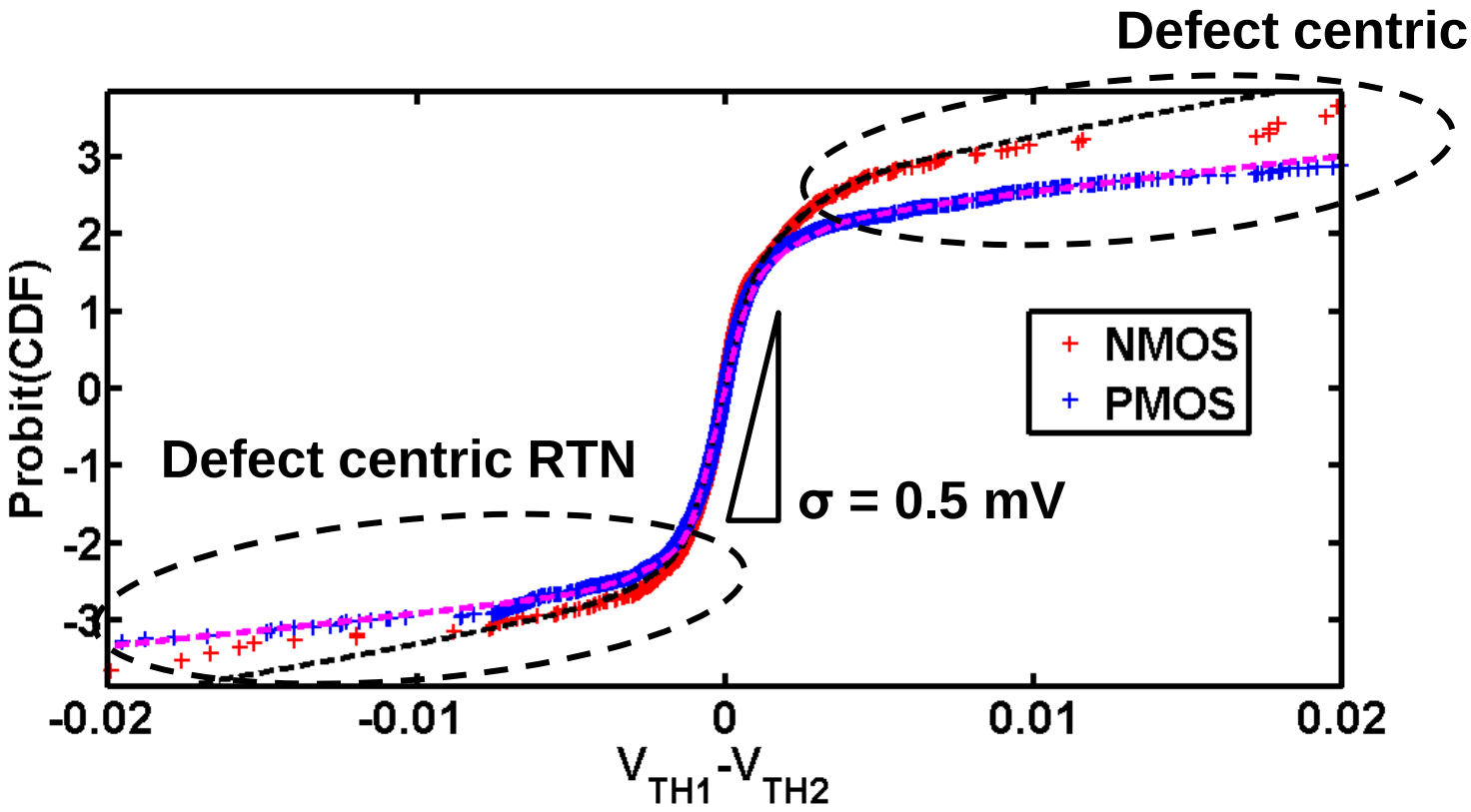


- **Measured RTN incorporates noise to (dis)charging during measurement**

$$\Delta V_{TH,RTN\_meas} = \Delta V_{TH,t1} - \Delta V_{TH,t2} + Noise$$

Weckx et al., IRPS 2015

Defect-Centric Model Fit to RTN

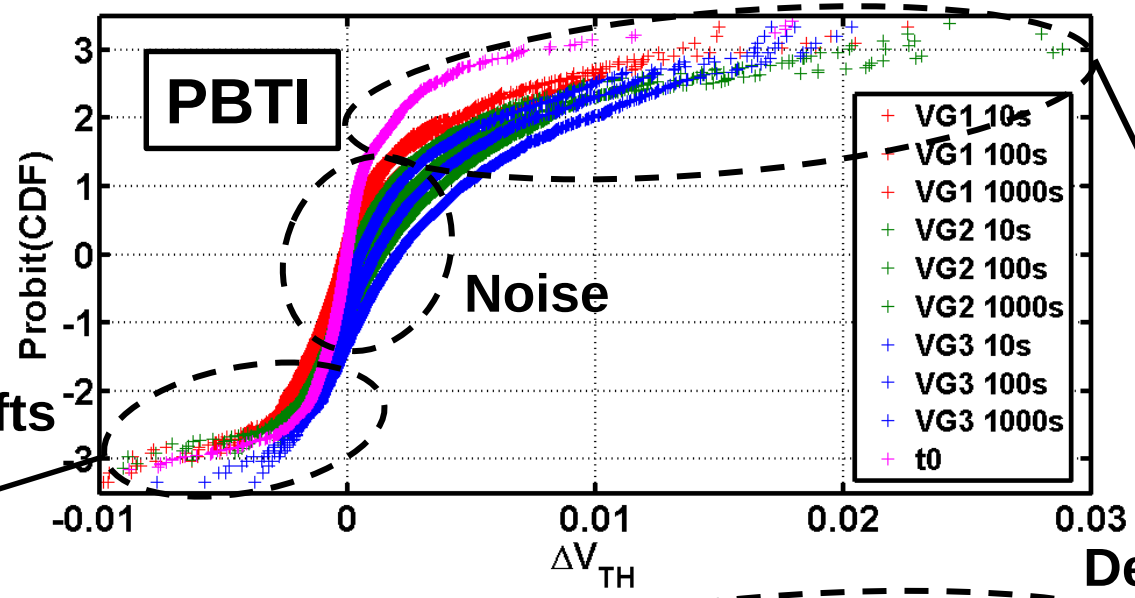


$$\Delta V_{TH,RTN\_meas} = \Delta V_{TH,t1} - \Delta V_{TH,t2} + Noise$$

Weckx et al., IRPS 2015

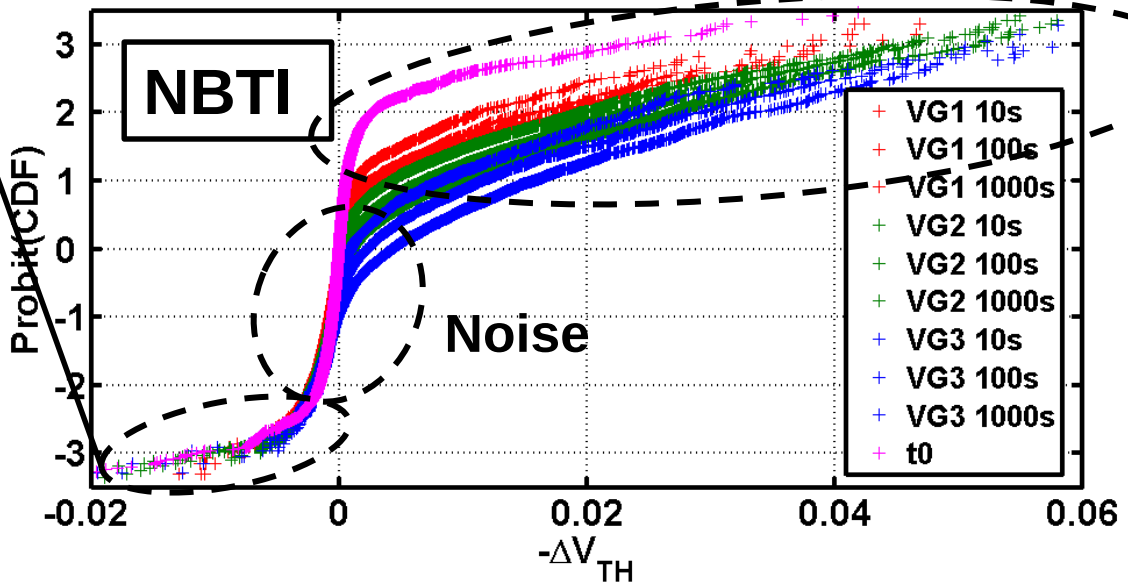
$$f(\Delta V_{TH}) = f_{RTN}(\Delta V_{TH}) * f_{Gaussian}(\Delta V_{TH})$$

P/NBTI Time-Dependent Variability Data



Opposite shifts due to RTN

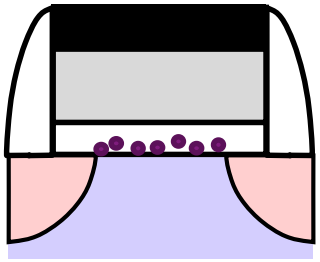
Defect centric BTI

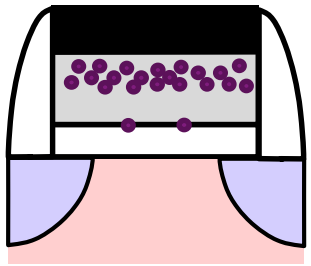


Weckx et al., IRPS 2015

Unimodal vs Bimodal Defect-Centric Distribution

PMOS NBTI





NMOS PBTI

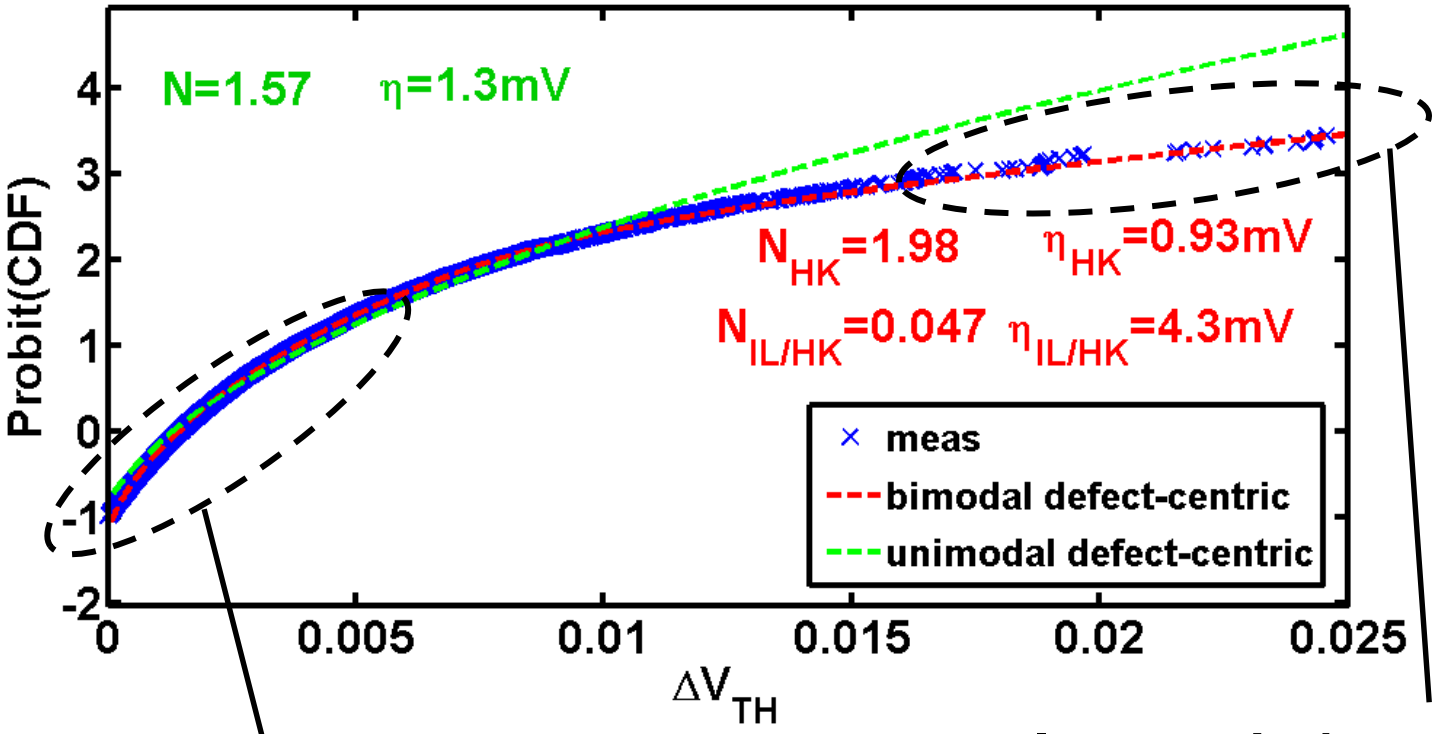
Single defect distribution	$\Delta V_{TH} \sim Exp(\eta)$	$\Delta V_{TH,i} \sim Exp(\eta_i)$
Given #traps CDF	$F_{n,\eta}(\Delta V_{TH}) = 1 - \frac{n}{n!}\Gamma(n, \Delta V_{TH}/\eta)$	$F_{n_1,n_2,\eta_1,\eta_2}(\Delta V_{TH}) = 1 - \alpha\, exp(S\Delta V_{TH})\, 1$
Distributed #traps CDF	$F_{N,\eta}(\Delta V_{TH}) = \sum_{n=0}^{\infty} \frac{e^{-N} N^n}{n!} F_{n,\eta}(\Delta V_{TH})$	$F_{N_1,N_2,\eta_1,\eta_2}(\Delta V_{TH}) = \sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty} \frac{e^{-N_1} {N_1}^{n_1}}{n_1!}\frac{e^{-N_2} {N_2}^{n_2}}{n_2!} F_{n_1,n_2,\eta_1,\eta_2}(\Delta V_{TH})$

Weckx et al., IRPS 2015

***N*: average number of charged defects per device (Poisson distribution)**

η: average V_{TH} impact per charged defect (Exponential distribution)

PBTI Bimodal Defect-Centric Model

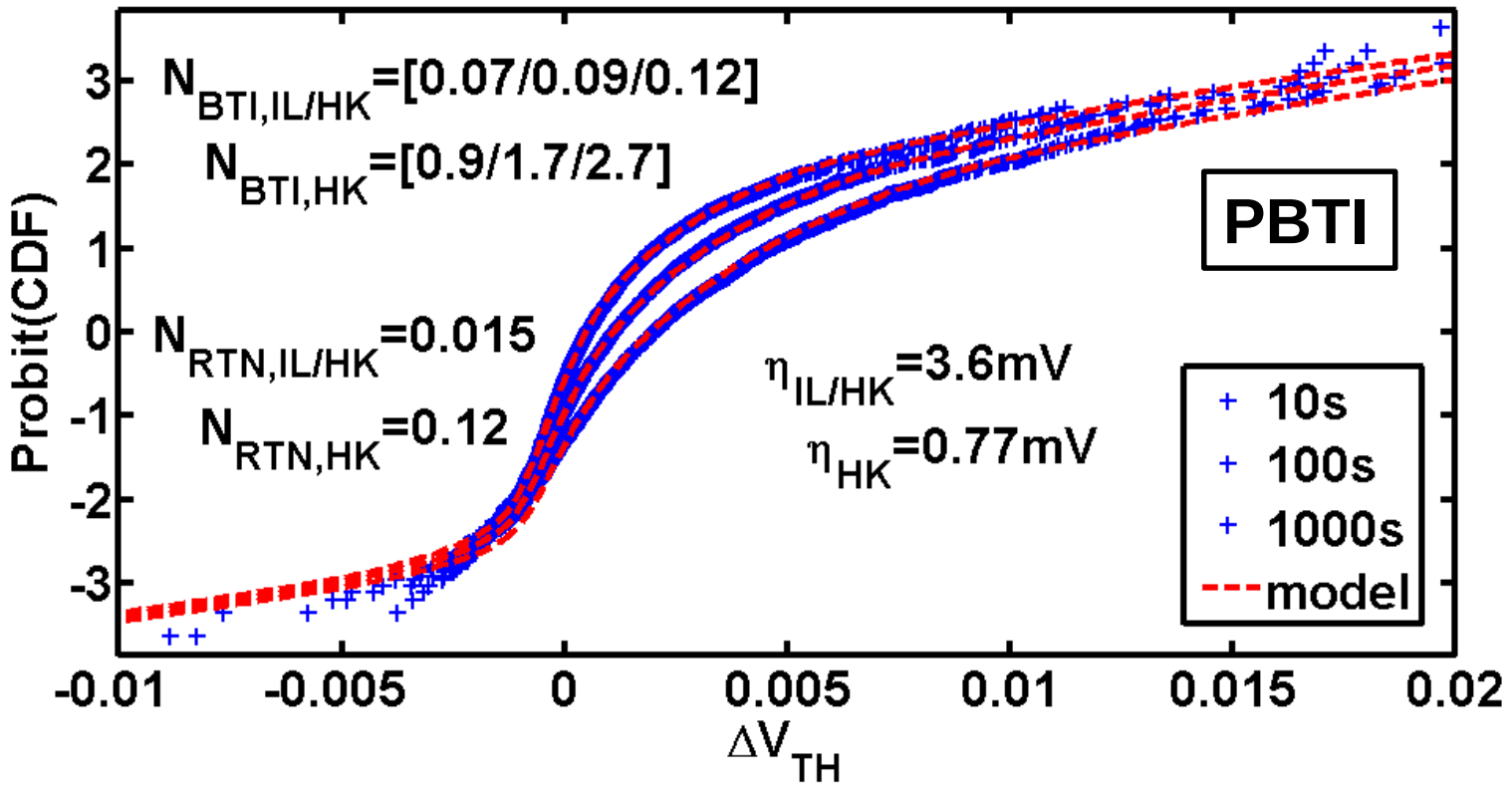


Low *η* (HK traps) gives good agreement in the bulk

large *η* (IL/HK traps) gives a good agreement in the tail

Weckx et al., IRPS 2015

Multivariate Model for RTN and BTI

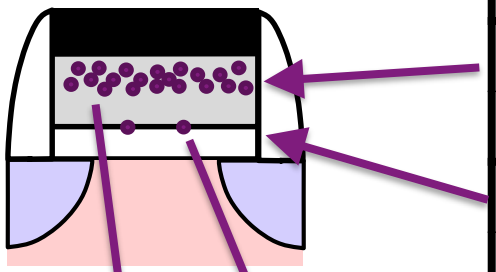


One set of η values (trapping HK and IL/HK interface) accurately models combined RTN and PBTI for various stress times

$$f(\Delta V_{TH}) = f_{BTI}(\Delta V_{TH}) * f_{RTN}(\Delta V_{TH}) * f_{Gaussian}(\Delta V_{TH})$$

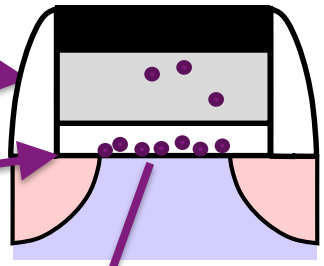
Weckx et al., IRPS 2015

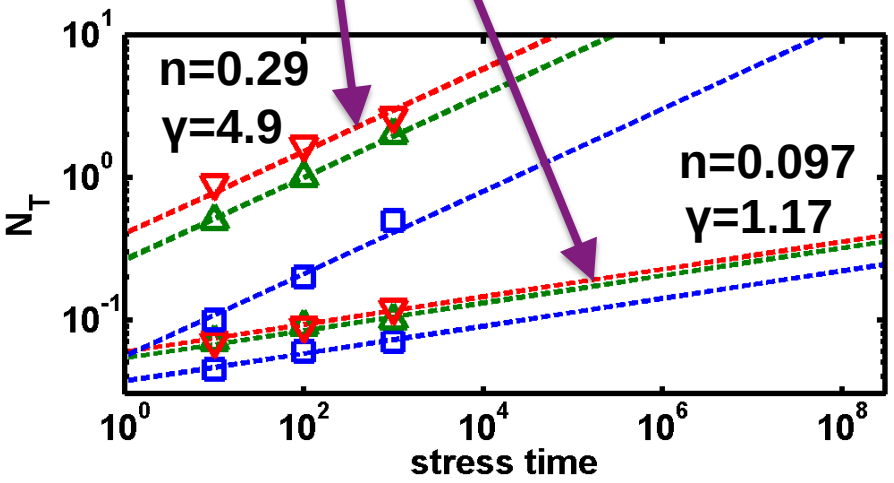
Time and Voltage Acceleration

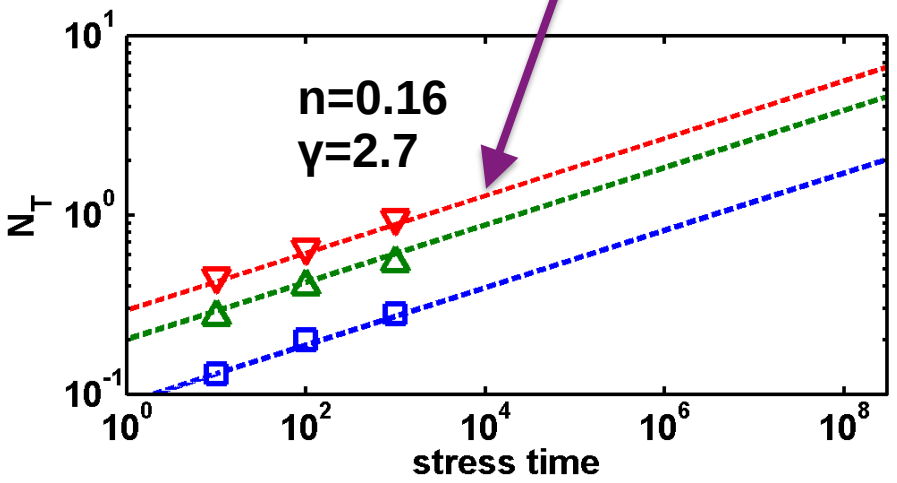


NMOS	PMOS
η=0.77mV	η=0.9mV
N_T=3.6	N_T=0.21
η=3.6mV	η=7mV
N_T=0.12	N_T=0.88

?



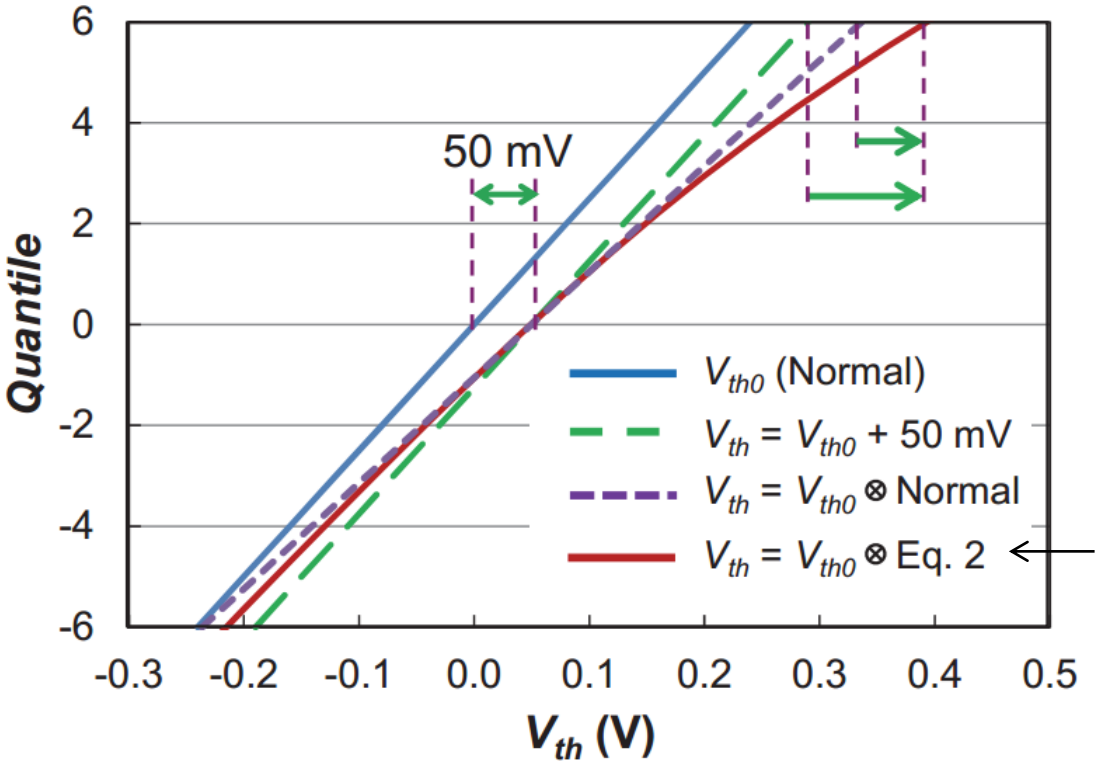




- NMOS: different components due to HK and IL/HK defects
- PMOS: high N_T/high η component which is a signature for SiO_2 NBTI, unresolved second component
- A simplified power-law model fits $\boldsymbol{N_T = AV_{OV}{}^{\gamma} t^{n}}$

Weckx et al., IRPS 2015

Combining Time-Zero and Time-Dependent Variability



← Defect-centric model

Kaczer et al., ESSDERC 2015

- Total variability distribution is convolution of time-zero and time-dependent distributions
- Defect-centric time-dependent variability (matching data) results in longer tail than assumption of normally distributed variability

Summary

- Device local variation increases with scaling
- Transistor array enables efficient characterization of variability to high-sigma values
- Defect-centric model can explain time-dependent variability of both RTN and NBTI/PBTI
- Distribution of time-dependent variability is non-normal
- Time-zero and time-dependent variability must be accounted for to project circuit performance and yield over lifetime of product

Acknowledgements

- Altera:
  - Chris Chen, Liping Li, Queennie Lim, Hong Hai Teh, Noor Fadillah Binti Omar, Chun Lee Ler, Kaushik Chanda, Sue Chen
- IMEC:
  - Pieter Weckx, Ben Kaczer, Jacopo Franco