Pre and post-silicon techniques to deal with large-scale process variations

Jaeyong Chung, Ph.D.

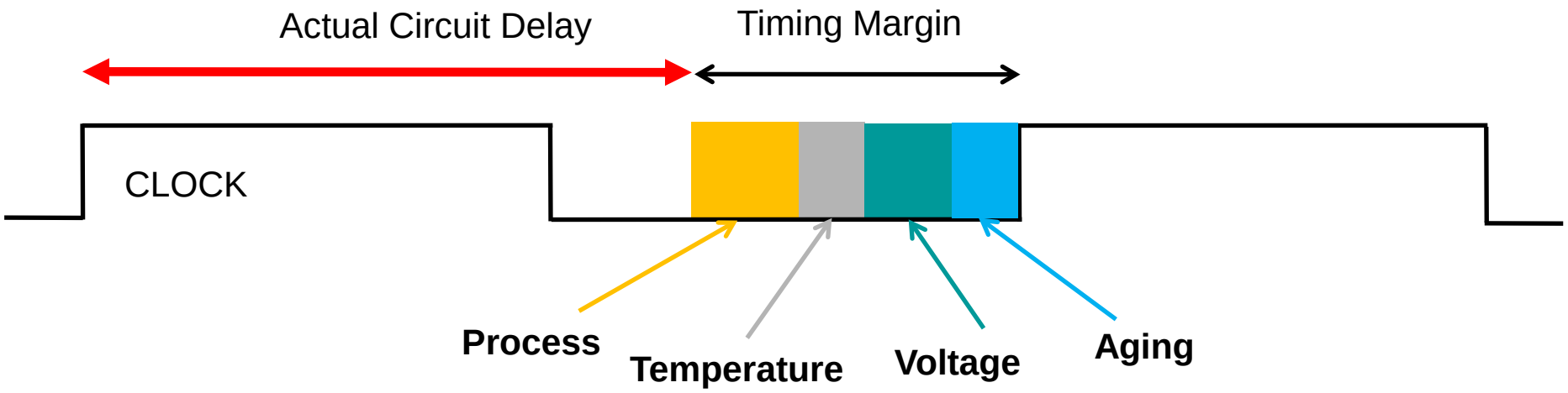
Department of Electronic Engineering
Incheon National University

Outline

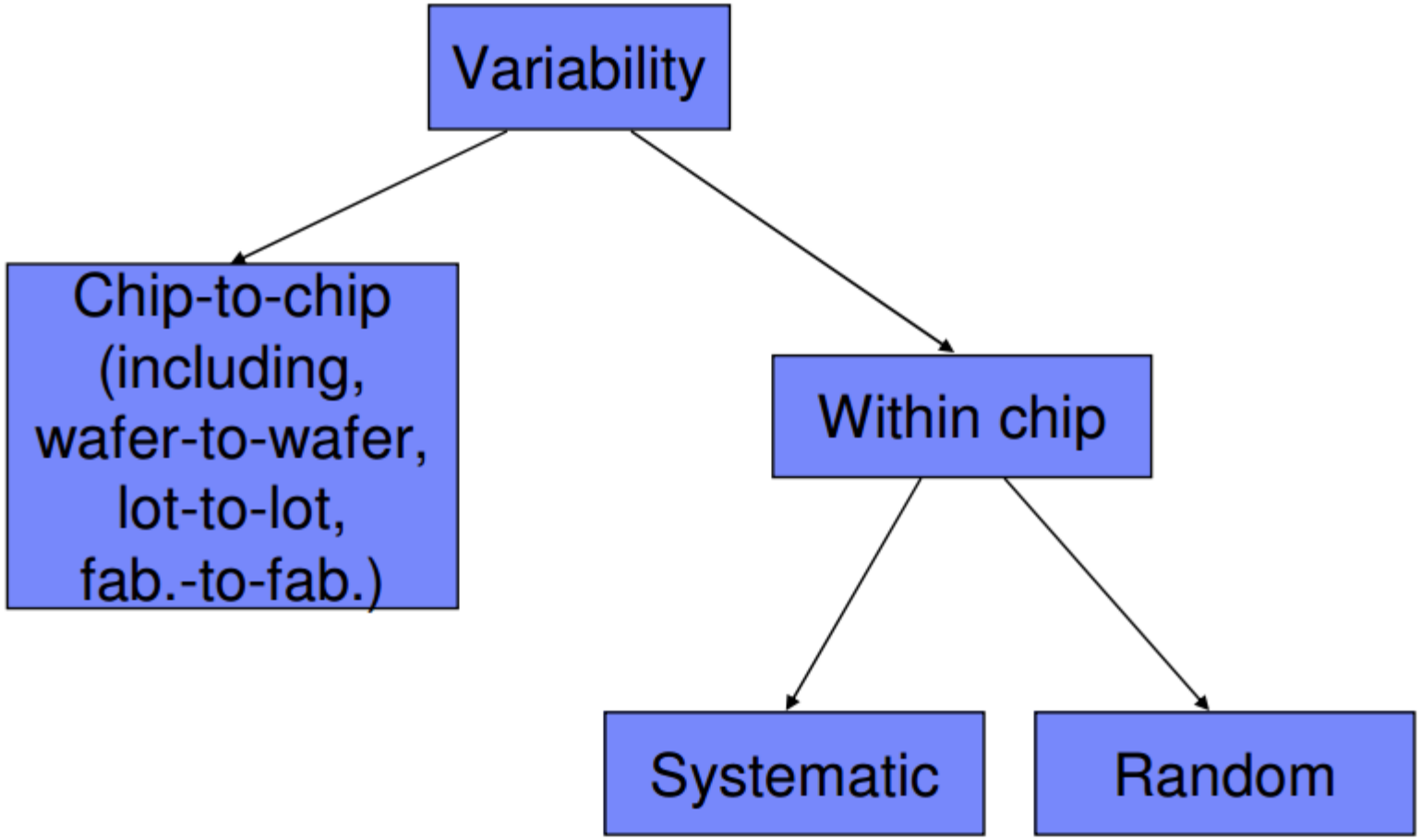
- Introduction to Variability
- Pre-silicon Techniques
  - Basics of traditional static timing
  - OCV
  - AOCV/LOCV
  - SSTA
  - POCV/SOCV
- Post-silicon Techniques
  - Compressed Sensing
  - Compressed Silicon Sensing (CSS)
  - Virtual Probe
  - Our Proposed Framework
  - Application of CSS

Timing Uncertainty

- Add Timing Margins For Delay Uncertainty
  - Process Variation
  - Voltage Variation
  - Temperature Variation
  - Aging Effects
- Associated Costs
  - Area, Power, Design Efforts/Time

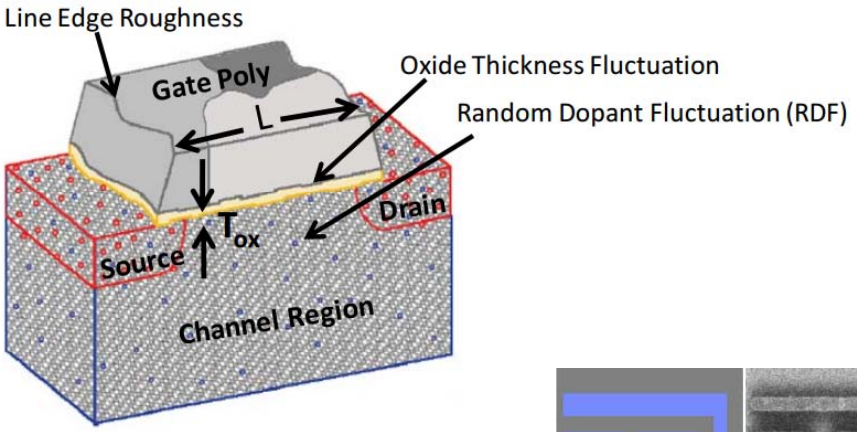


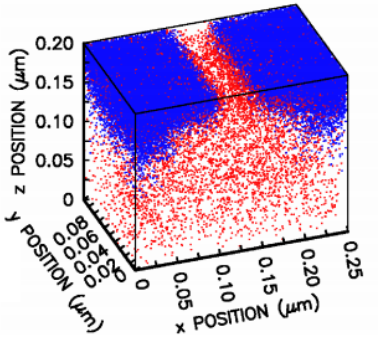
Classification of variability



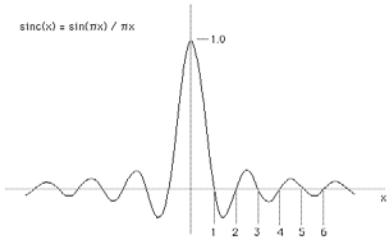
Sources of variation

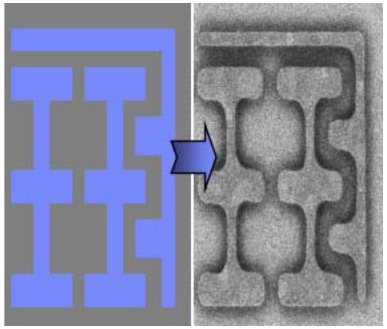
- FEOL (Front end of line) variation
- BEOL (Back end of line) variation



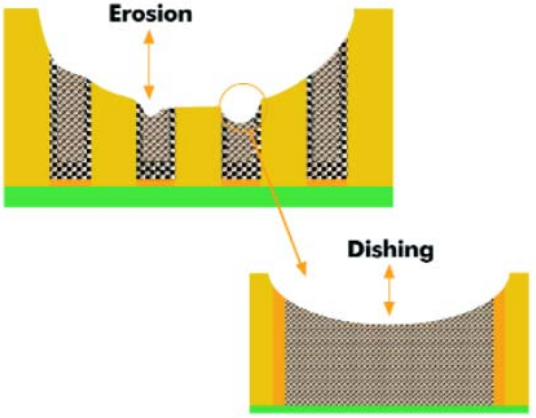


Random Dopant Fluctuation





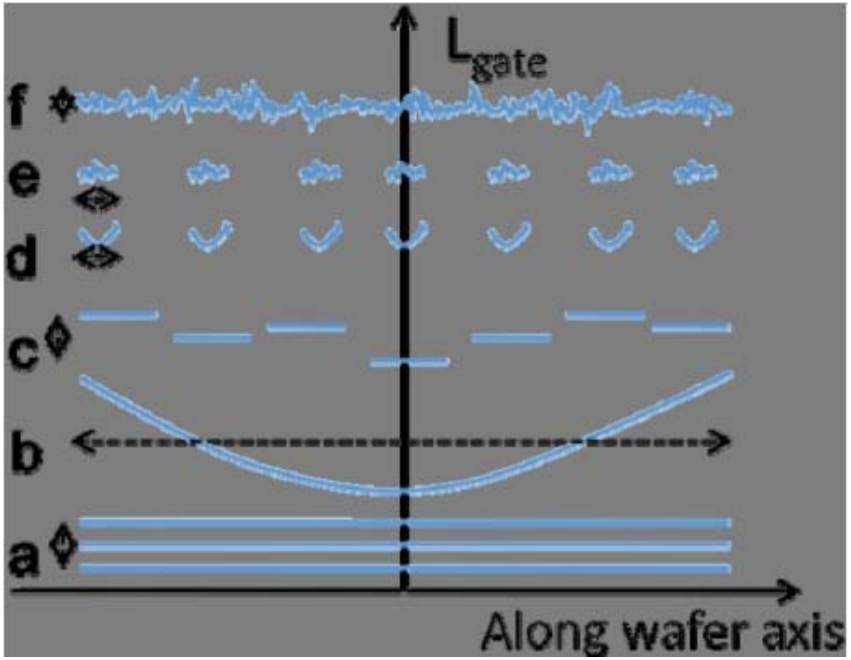
Lithography-induced variation /Proximity effects



Erosion and dishing in CMP process

CD (Lgate) Variation

- Critical dimension (a.k.a, Lgate, Leff,…)
  - The effective channel length of transistors
  - Affects delay and leakage substantially
  - Varies across-wafer and within-chip systemically
- A reduction of 1nm of the standard deviation of CD → $7.5/chip for a high end product

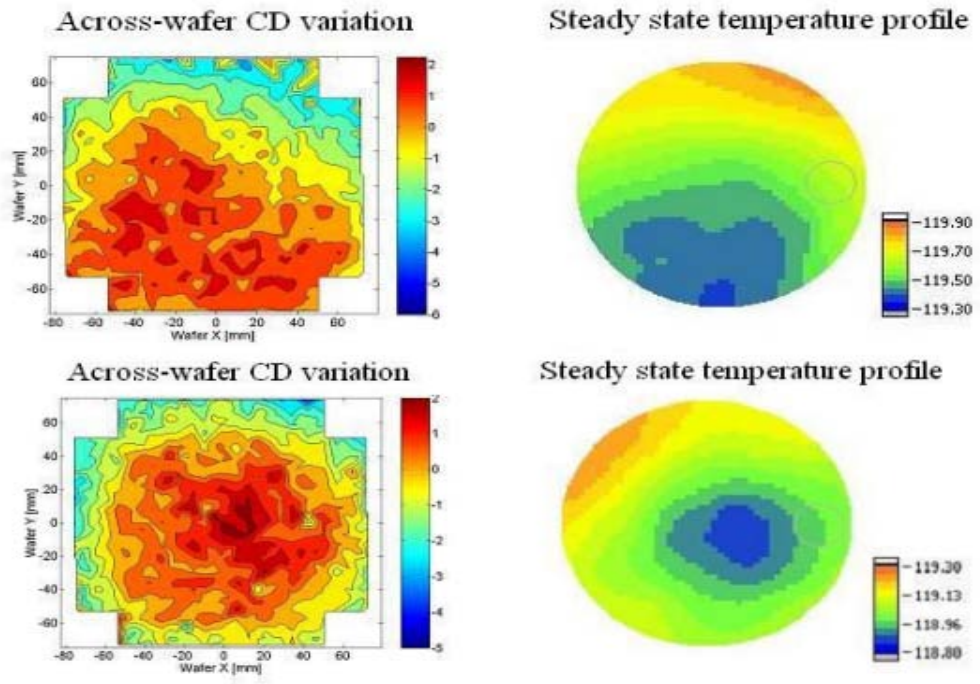


a) wafer-to-wafer
b) Across-wafer
c) Die-to-die
d) Across-die
e) Pattern dependent
f) Local random noise

http://www.eecs.berkeley.edu/~bora/Conferences/2009/SPIE09-Qian.pdf

CD (Lgate) Variation

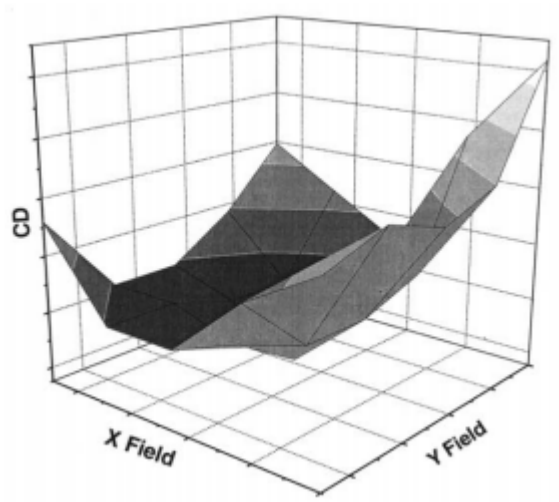
- Across-wafer CD variation
  - Post Exposure Bake (PEB) is the greatest variation culprit
  - In areas where the bake plate is relatively cool, CD is larger than average

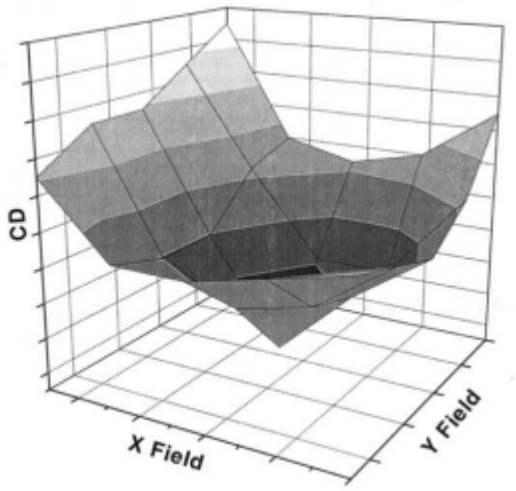


http://bcam.berkeley.edu/ARCHIVE/theses/Friedberg_PhD.pdf

CD (Lgate) Variation

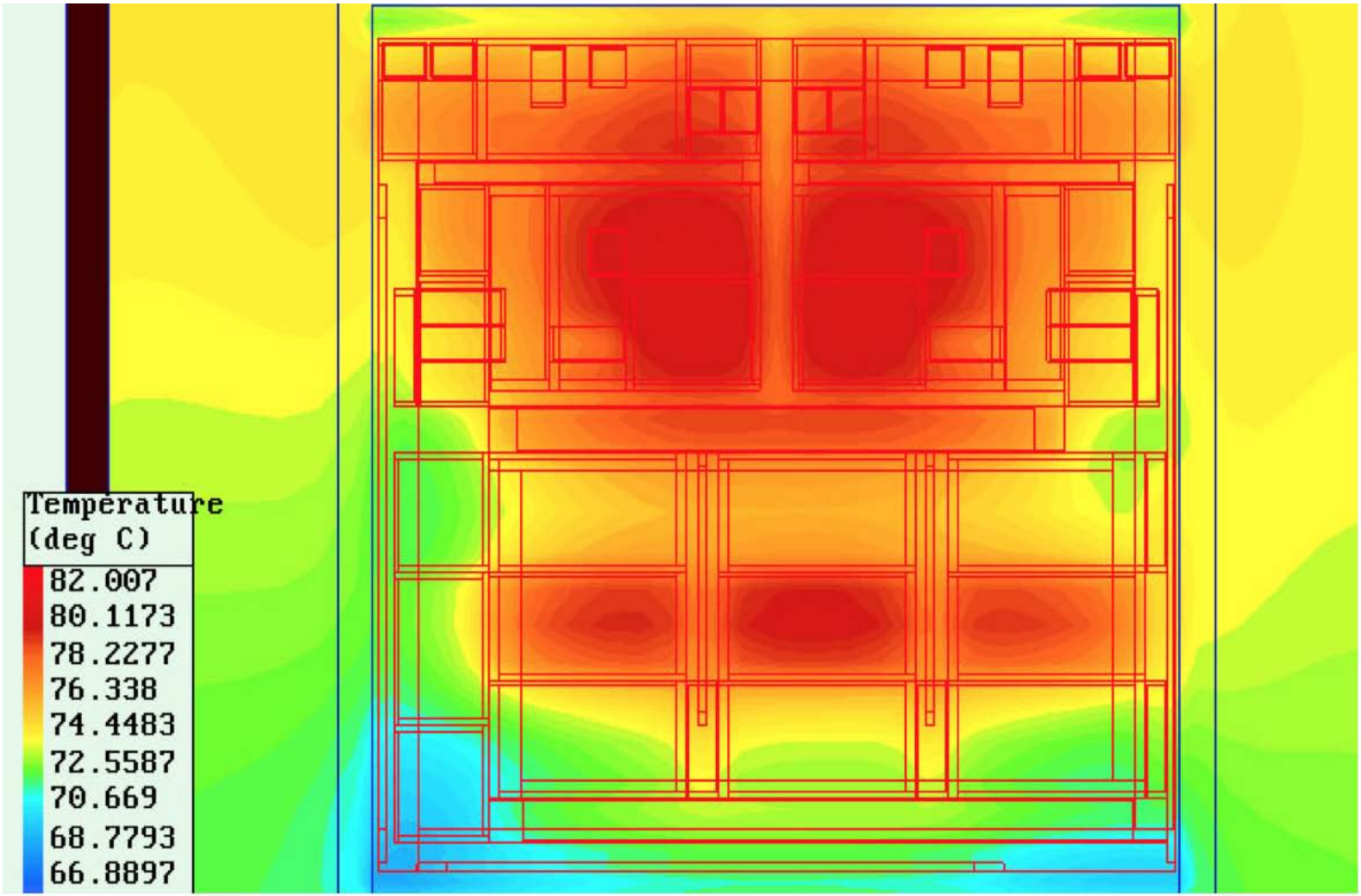
- Within-chip CD variation
  - Lens aberration induces spatially correlated variation
  - Different layout leads to different spatial patterns due to optical proximity effect





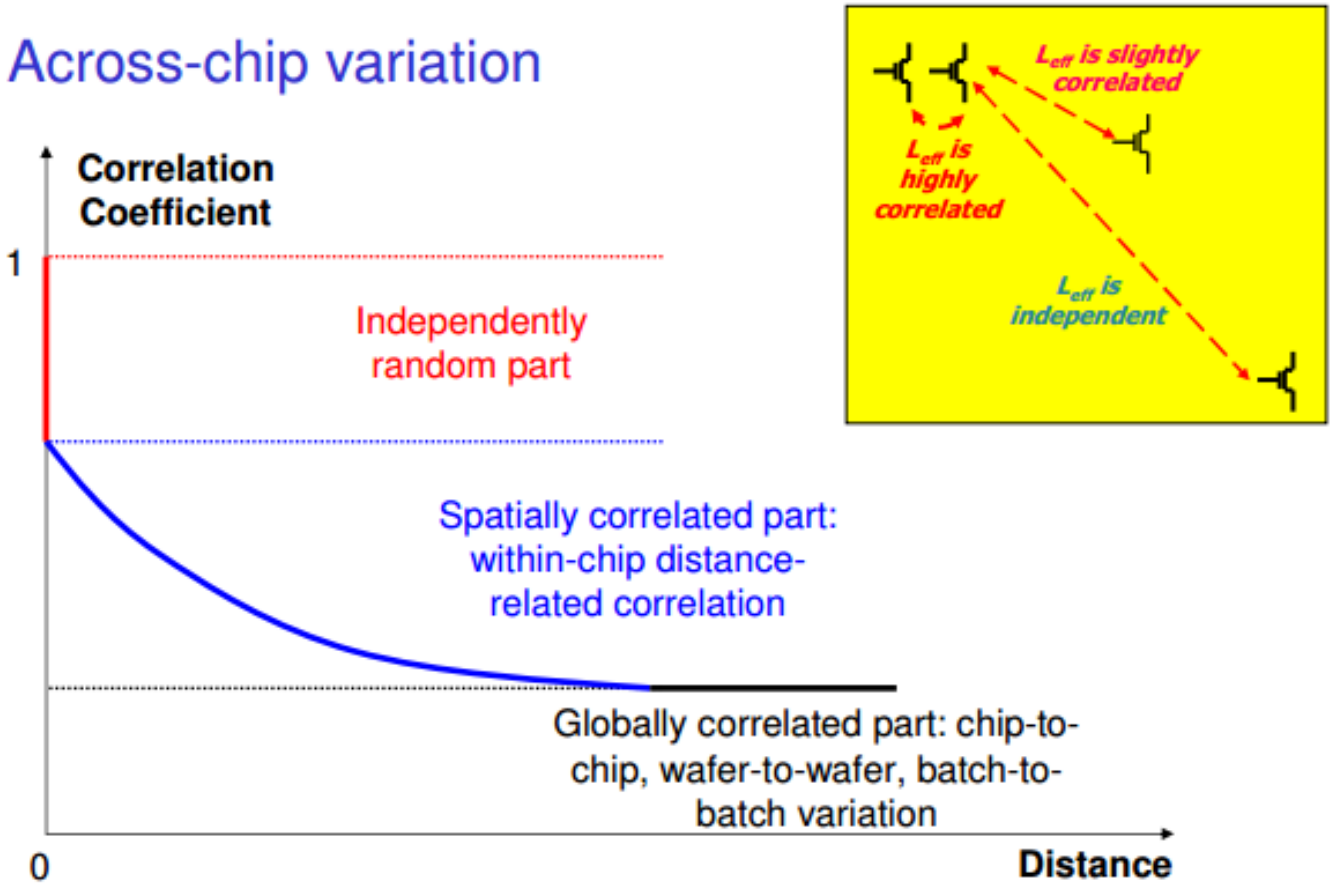
http://www.bioee.ee.columbia.edu/courses/upload/Bibliography/orshansky2004.pdf

Voltage, temperature, weather, …



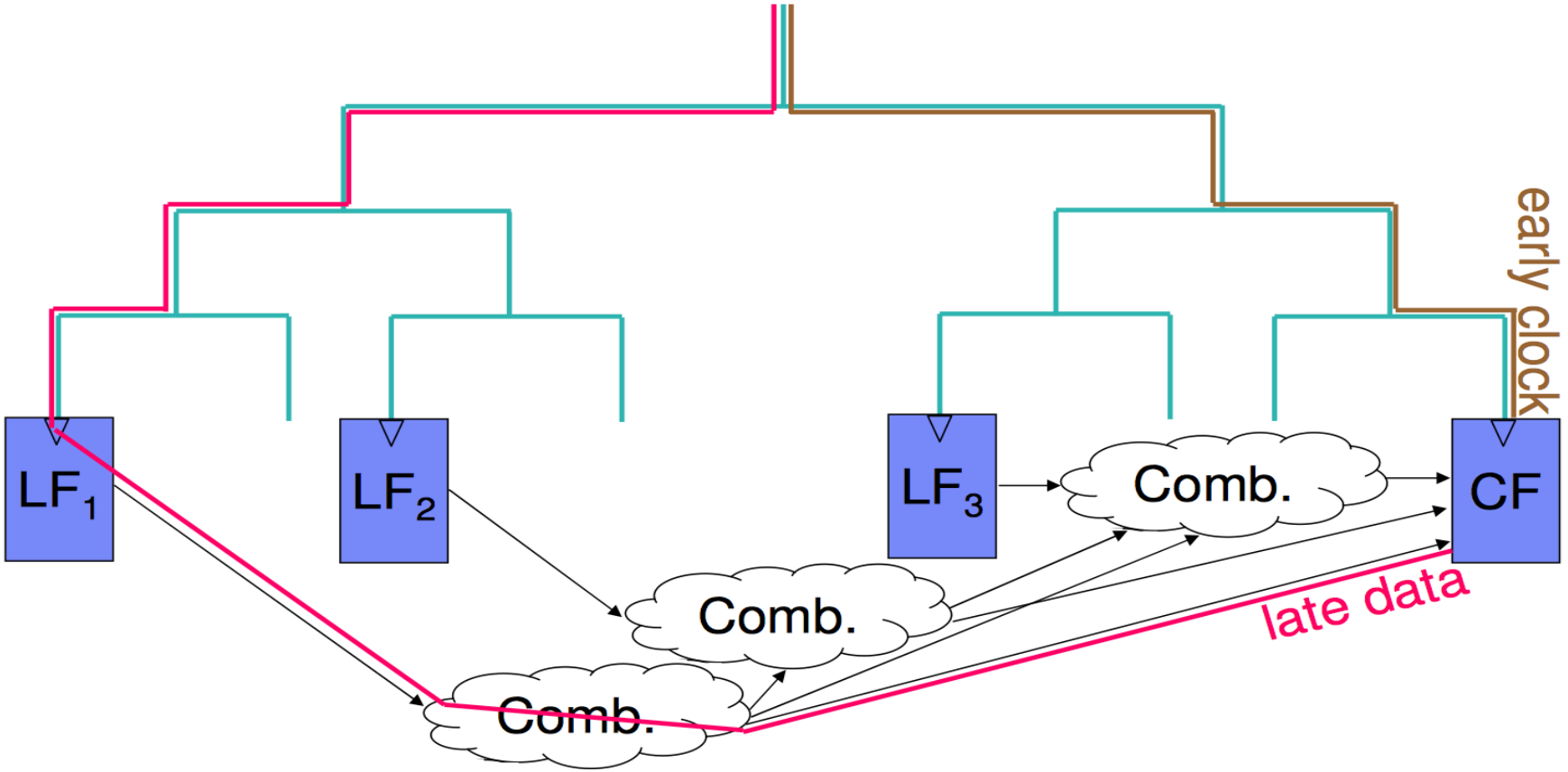
[IBM]

Spatial Correlation



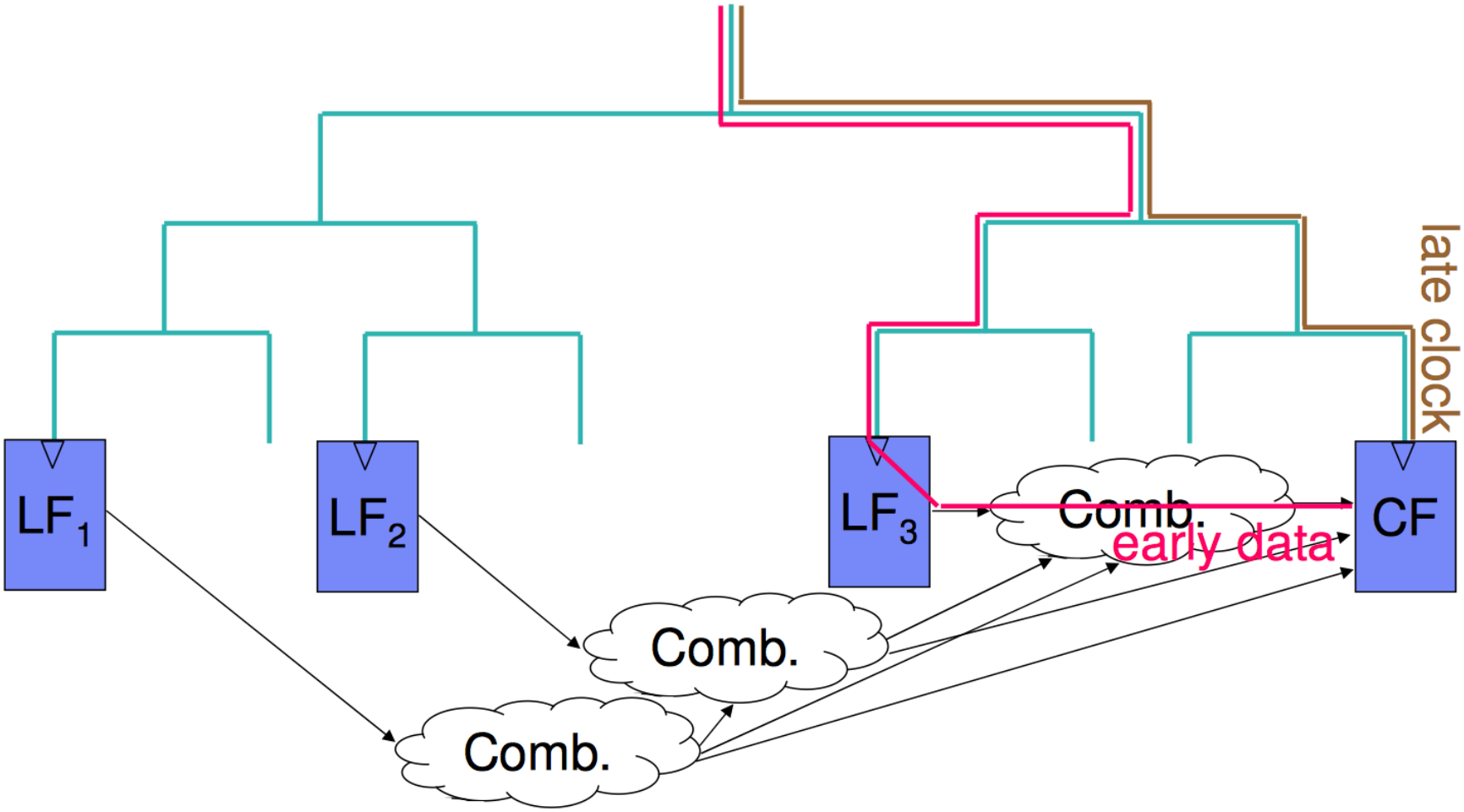
Traditional Static Timing

- Setup check



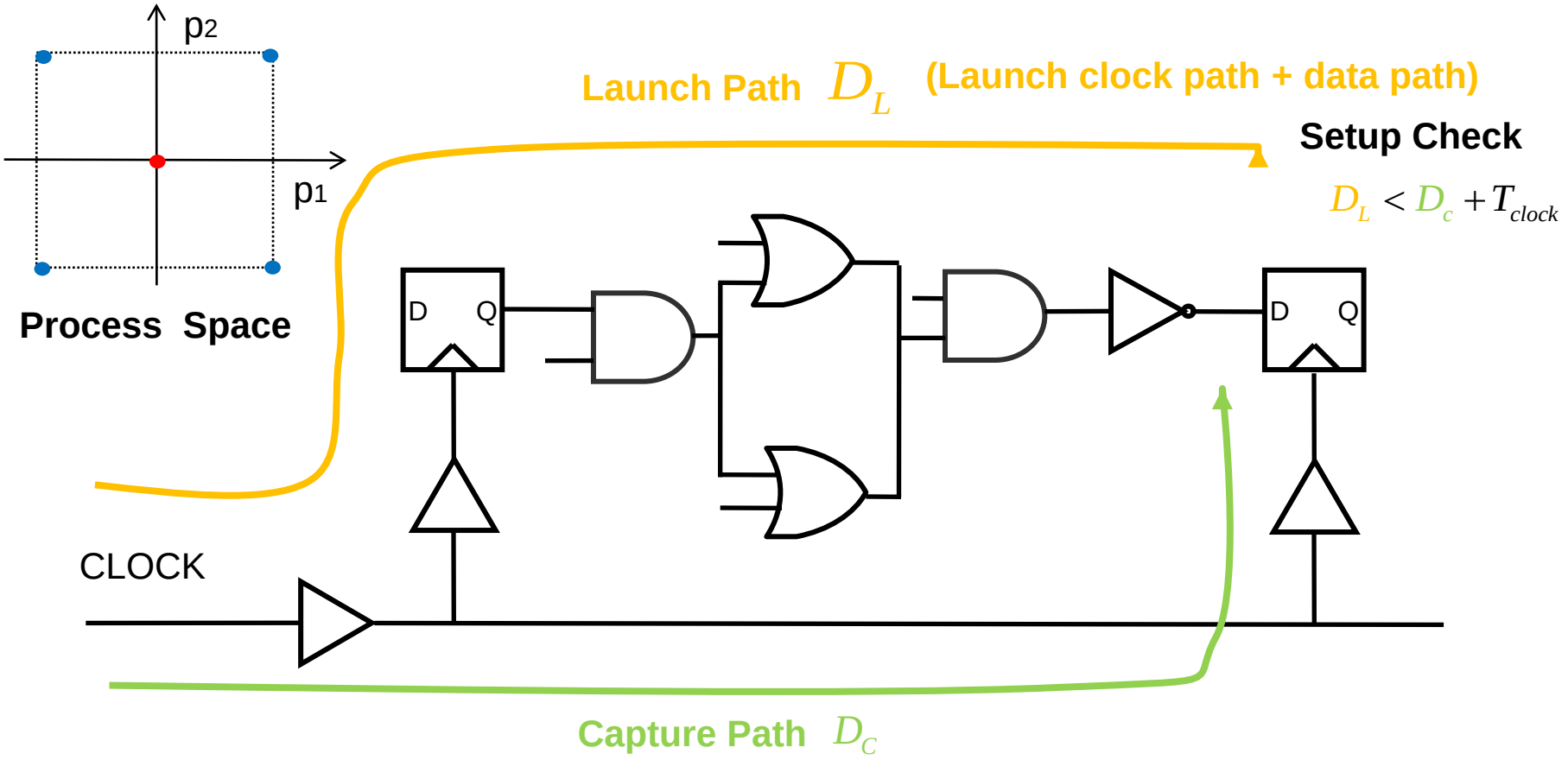
Traditional Static Timing

- Hold check



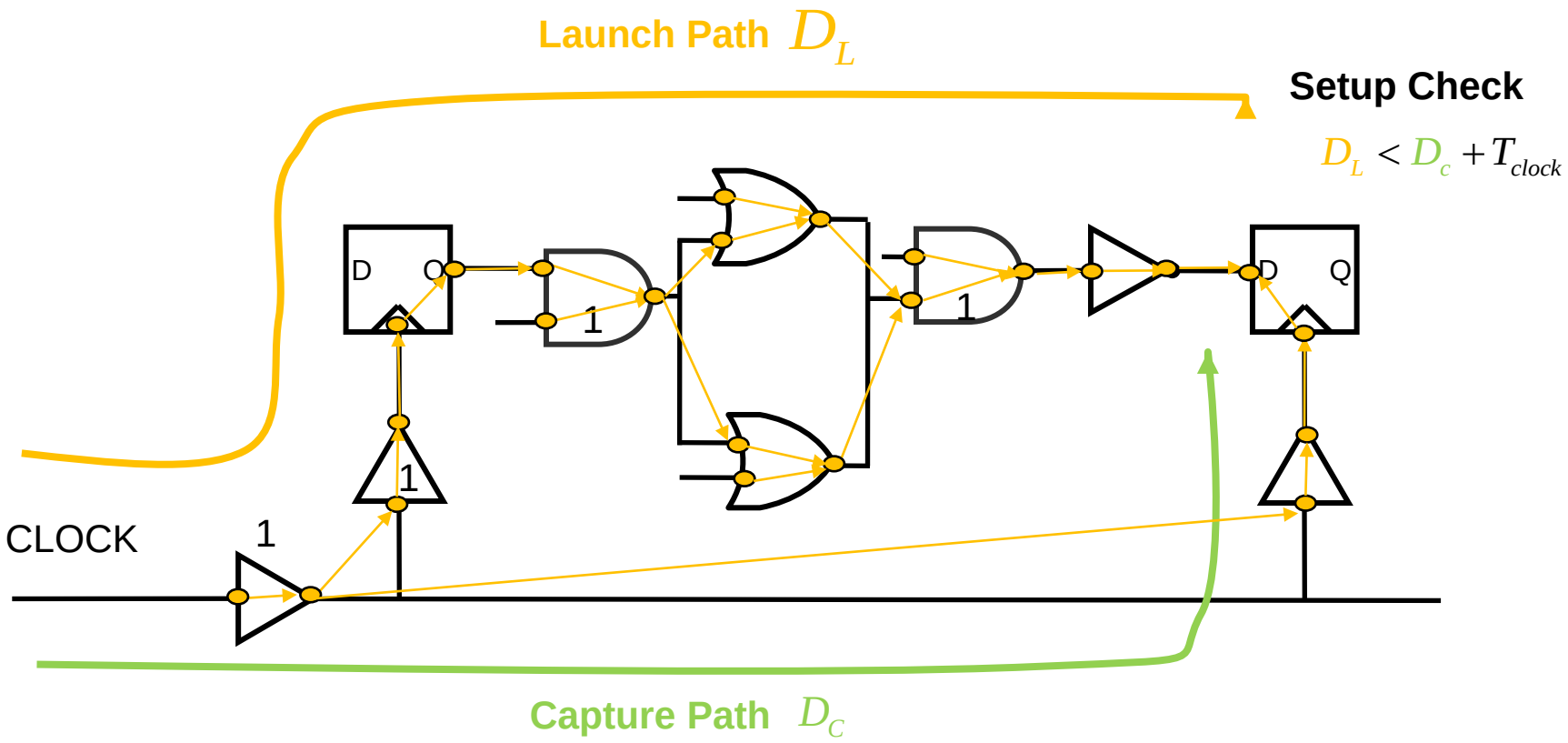
Traditional Static Timing

- Use worst/best corners for setup/hold checks



Traditional Static Timing

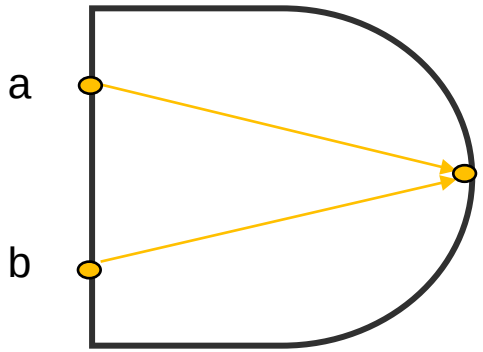
- GBA (Graph-Based Analysis) takes linear time in circuit size
- PBA (Path-Based Analysis) takes exponential time in circuit size



Traditional Static Timing

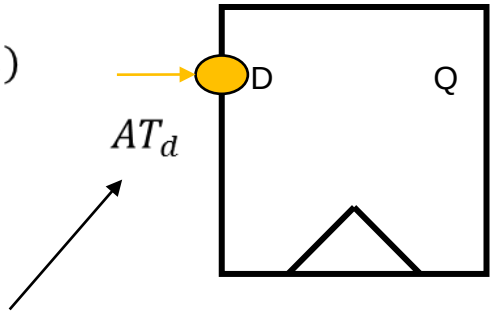
- GBA finds an upper bound of the worst path delay in linear time through the graph
- GBA is pessimistic than PBA

$$\text{Slew}_{ac} = F(Slew_a, Load_c)$$
$$D_{ac} = F(Slew_a, Load_c)$$



$$AT_c = \max(AT_a + D_{ac}, AT_b + D_{bc})$$
$$Slew_c = \max(Slew_{ac}, Slew_{bc})$$

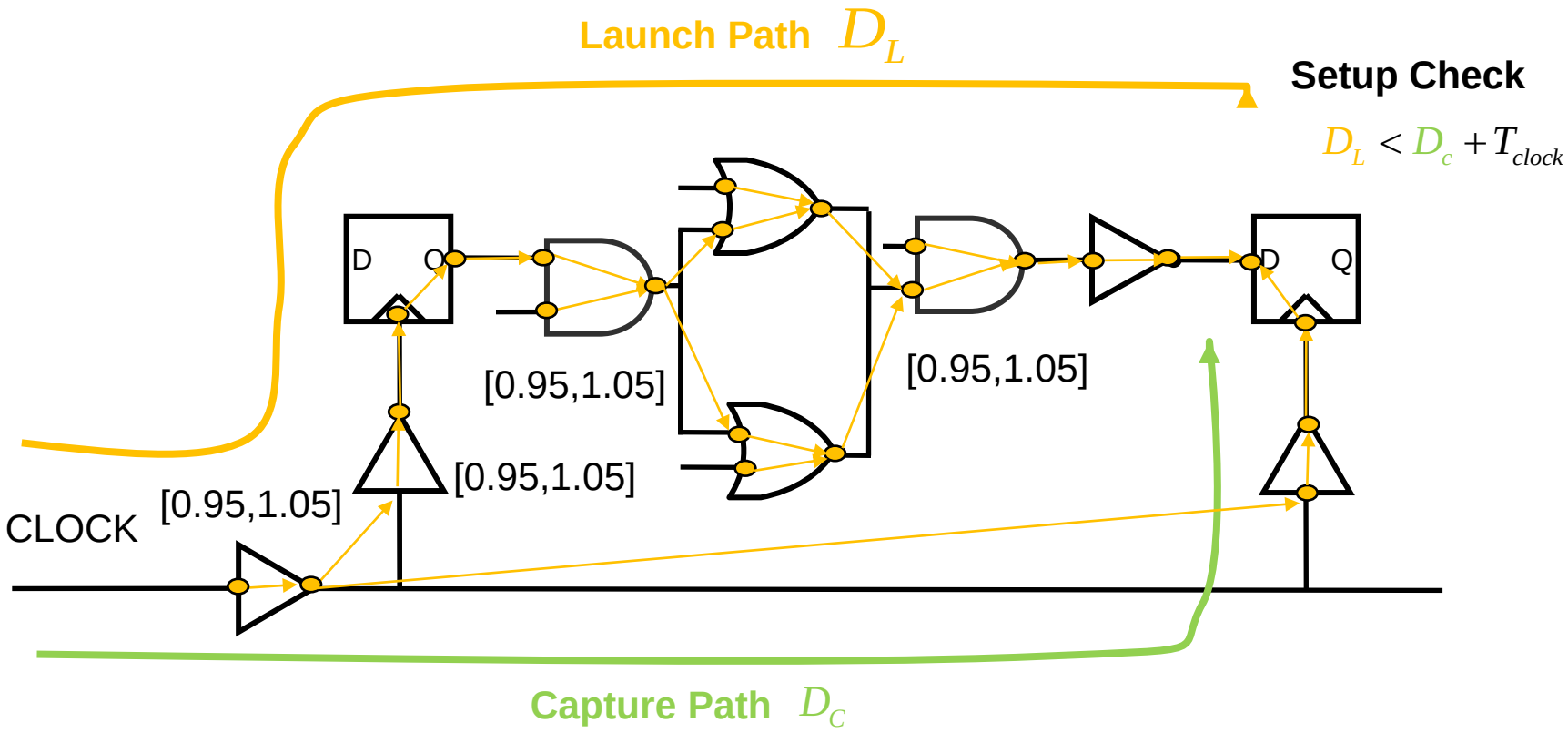
$$\text{Slew}_{bc} = F(Slew_b, Load_c)$$
$$D_{bc} = F(Slew_b, Load_c)$$



The upper bound of the worst path delay

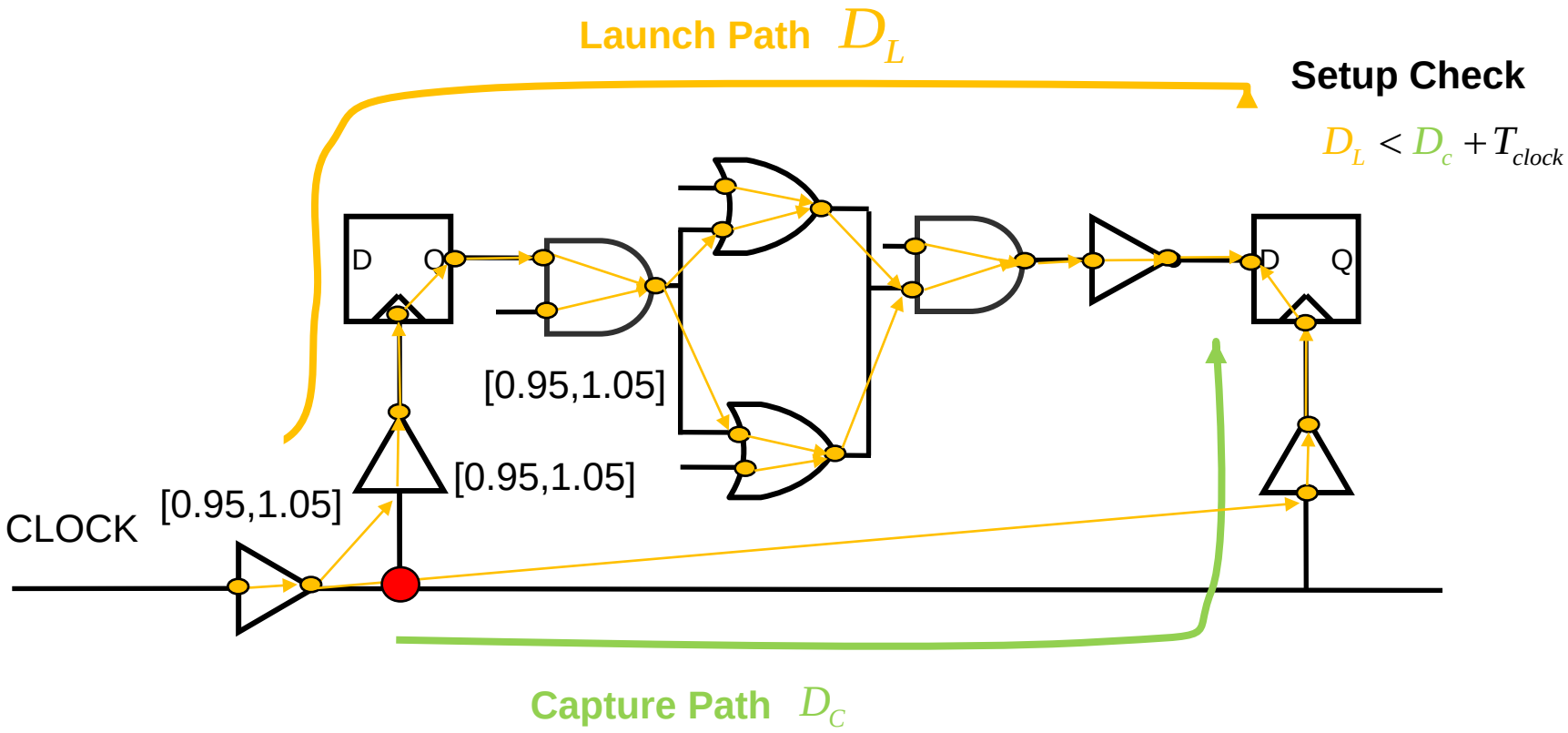
On-Chip Variation (OCV)

- Accounts for within-chip variation
- Global derating (e.g., ±5%) → early/late split

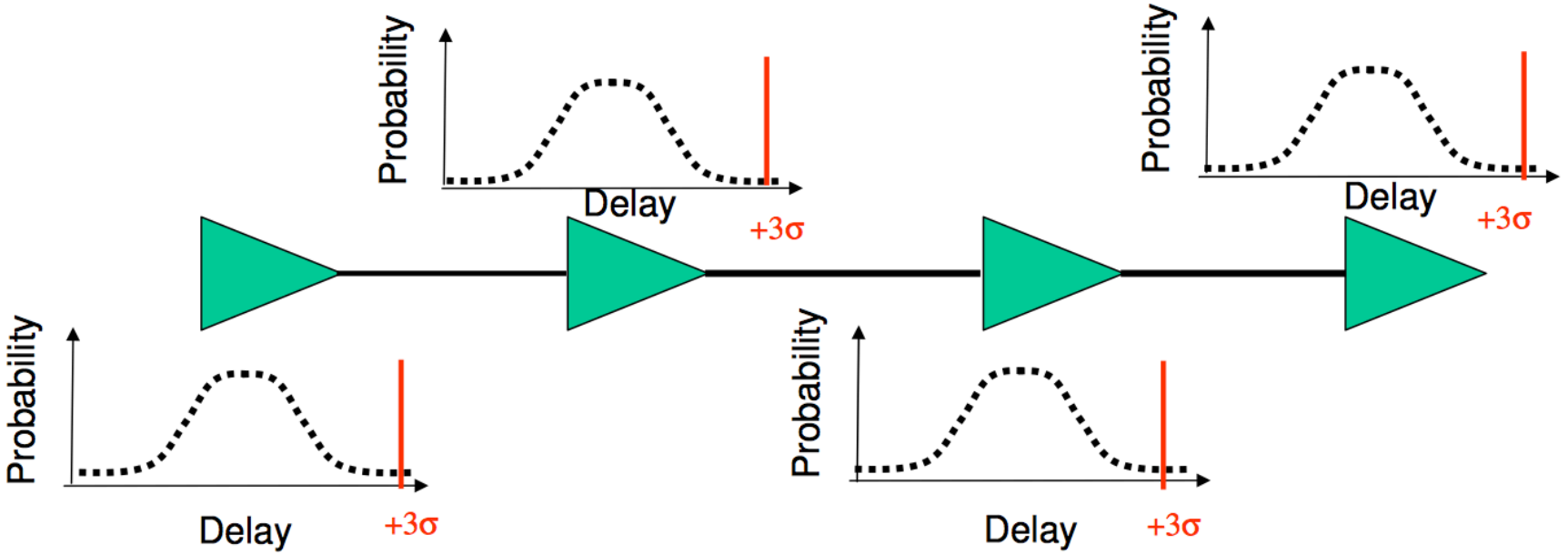


On-Chip Variation (OCV)

- Common Path Reconvergence Pessimism Removal (CRPR) (a.k.a, CPPR)



Statistical Cancellation



- **If each gate has a delay of 50 with a standard deviation of 2**
  - Method 1: Set each gate to its 3σ limit, total delay = 4*(50+6) = 224 (OCV)
  - Method 2: Compute the 3σ value of the sum of 4 random variables
    = $4*50 + 3*(2^2 + 2^2 + 2^2 + 2^2)^{1/2} = 212$
- **The difference between these two is called RSS credit**
  - RSS = Root of the Sum of the Squares (reflects statistical cancellation)
  - In this case, the credit is 12

Statistical Cancellation

- a.k.a., RSS Credit (Root of the Sum of the Squares)

$$E[X+Y] = E[X] + X[Y]$$

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

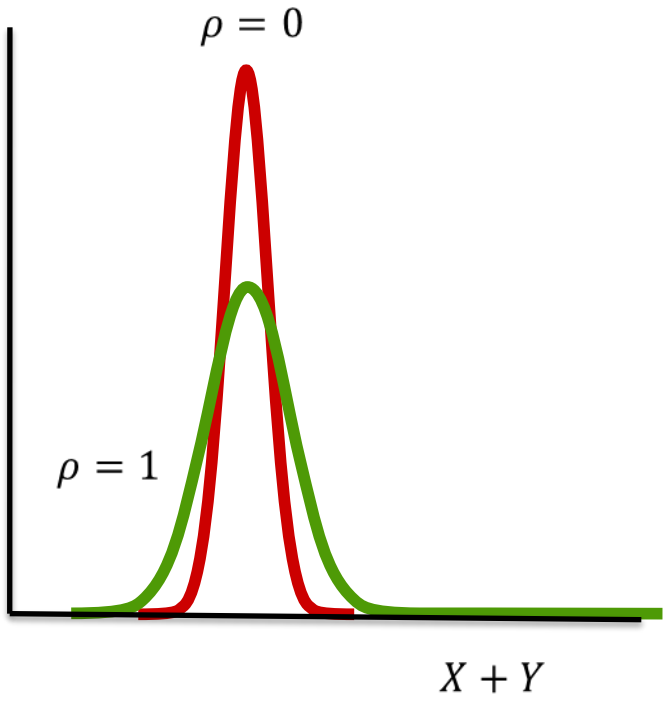
$$E[X+Y] = 2\mu$$

$$Var(X+Y) = 2\sigma^2 + 2\sigma^2\rho = \begin{cases} 4\sigma^2 & (\rho = 1) \\ 2\sigma^2 & (\rho = 0) \end{cases}$$

$$\frac{3\sigma}{\mu}$$

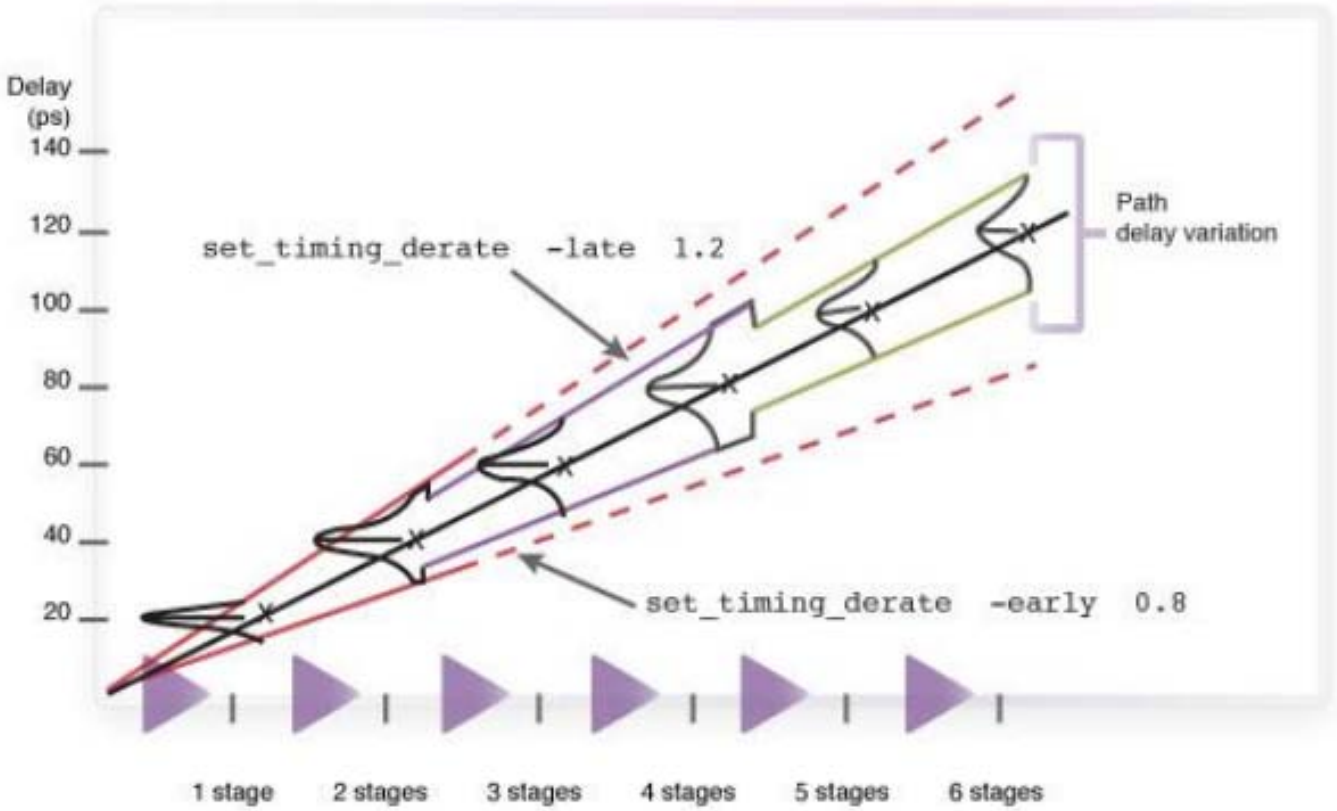
$\rho = 1$ $\qquad 3 \times \frac{2\sigma}{2\mu} = \frac{3\sigma}{\mu}$

$\rho = 0$ $\qquad 3 \times \frac{\sqrt{2}\sigma}{2\mu} = \frac{3\sigma}{\sqrt{2}\mu}$



For *N* indep. variables, variation is reduced by a factor of $1/\sqrt{N}$

Advanced OCV (AOCV) (aka LOCV)

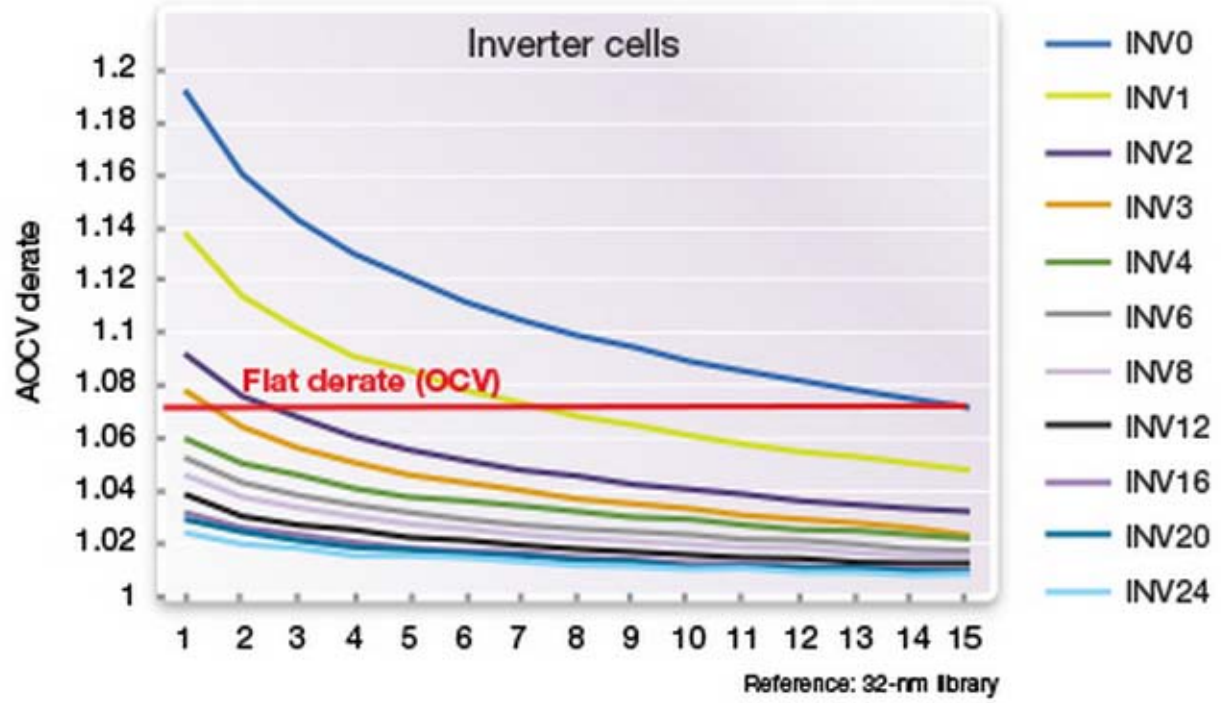


Depth	1	2	3	4	5	6
Derate	1.2	1.2	1.15	1.15	1.08	1.08

Stage-based AOCV

[Synopsys Whitepaper]

Advanced OCV (AOCV) (aka LOCV)



Depth	1	2	3	4	5	15
Derate	1.2	1.16	1.14	1.13	1.12	1.08

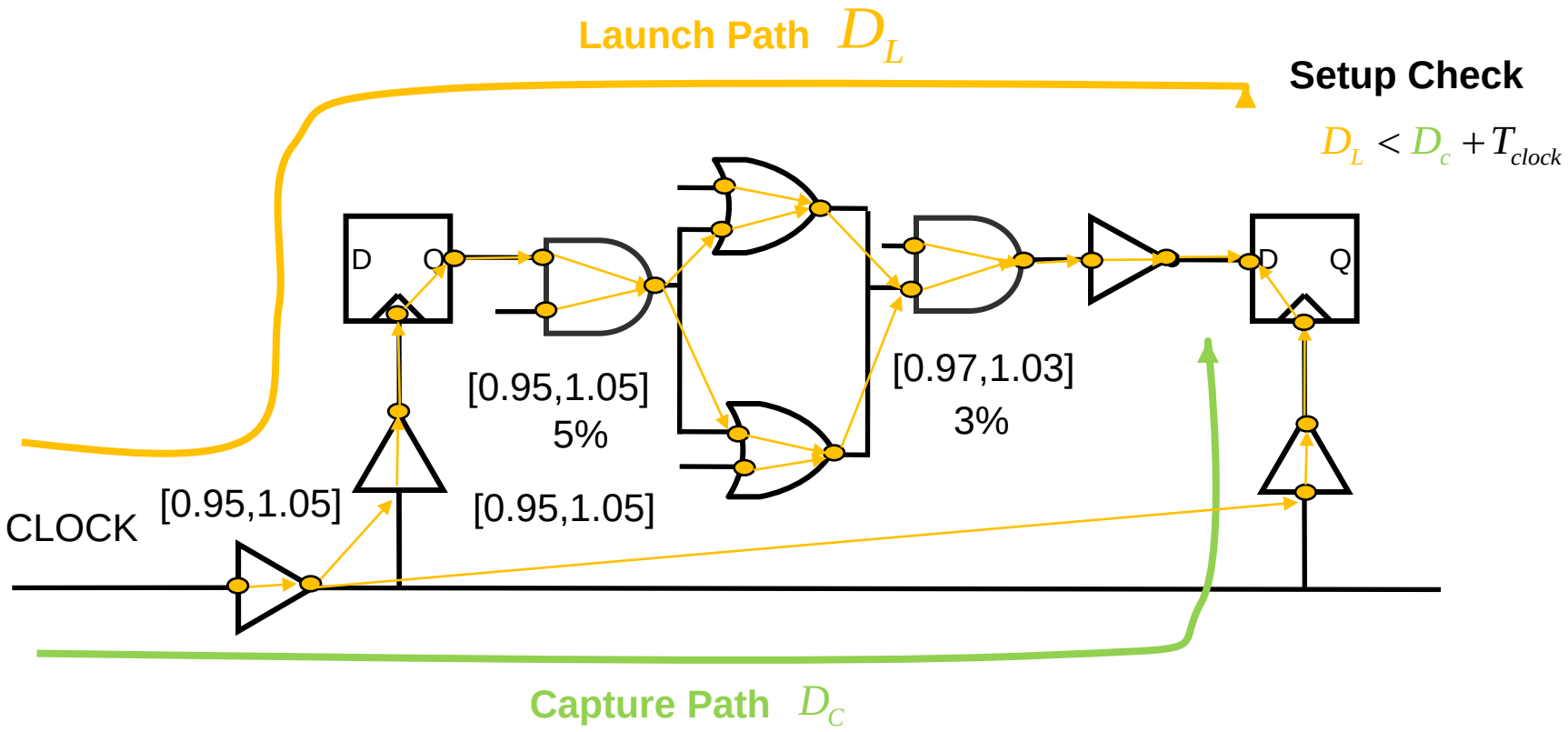
On chip variation (OCV)
- Fixed derate
- Limited accuracy

Advanced OCV
- Derate based on cell type and path depth
- Improved accuracy

[Synopsys Whitepaper]

Advanced OCV (AOCV) (aka LOCV)

- Cells gets different derate depending on the logic depth and the cell type
- (Design-specific OCV, CLK DA) it also depends on the loads and slews



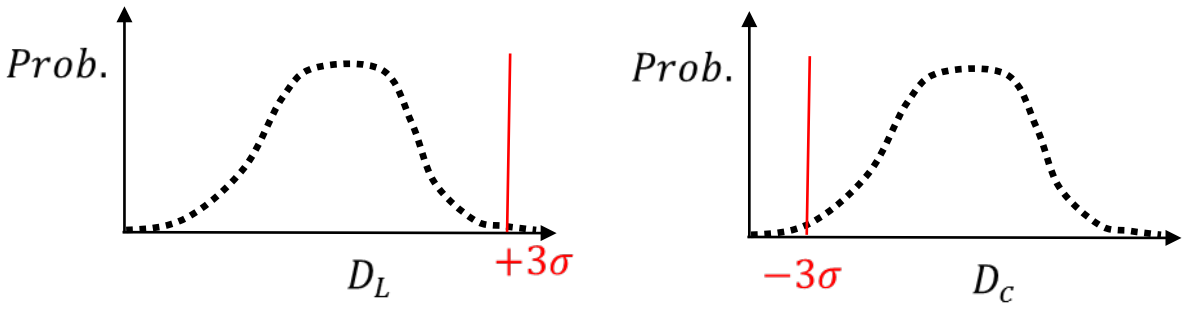
RSS credit in setup/hold check

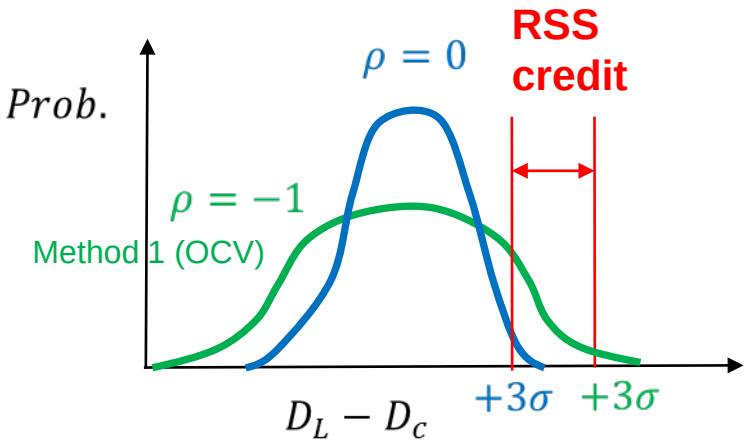
Setup Check

$$D_L < D_c + T_{clock}$$

$$\Leftrightarrow D_L - D_c < T_{clock}$$

Method 1 (OCV)





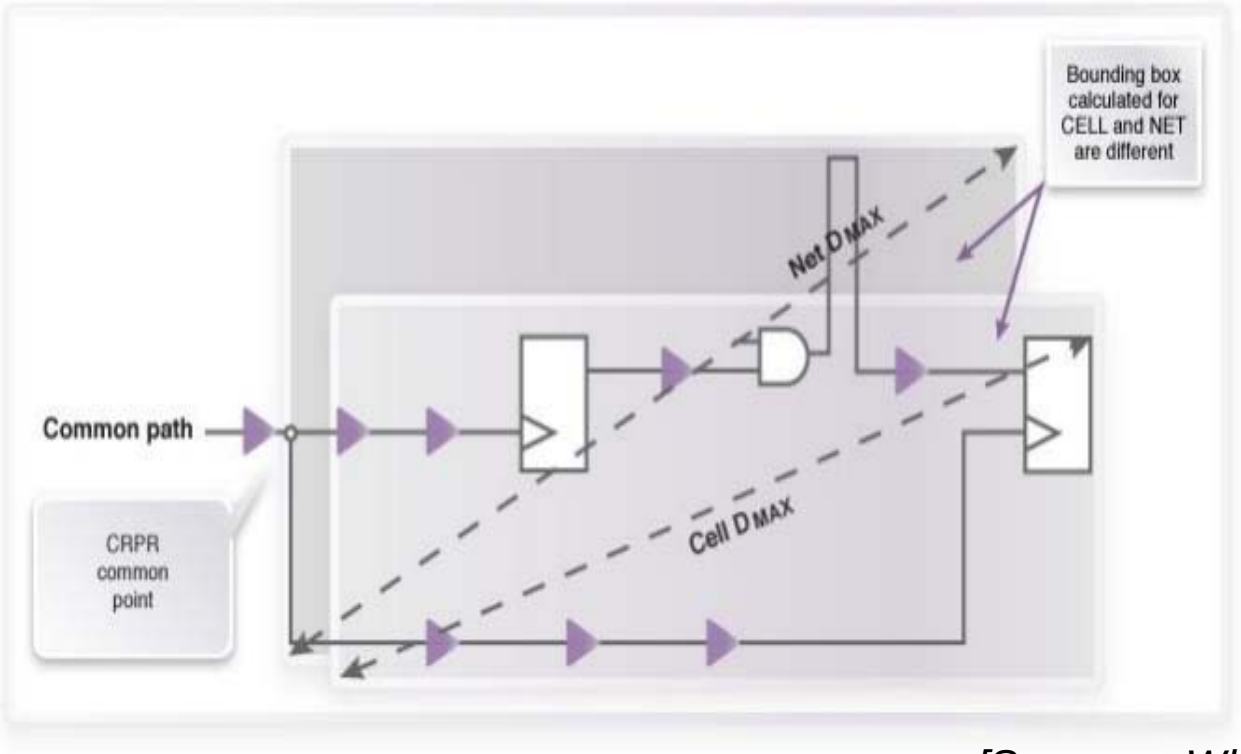
$$Var(X - Y) = 2\sigma^2 - 2\sigma^2\rho$$

$$= \begin{cases} 4\sigma^2 & (\rho = -1) \\ 2\sigma^2 & (\rho = 0) \\ 0 & (\rho = 1) \end{cases}$$

Method 1 (OCV)

If perfectly correlated, variation will be canceled

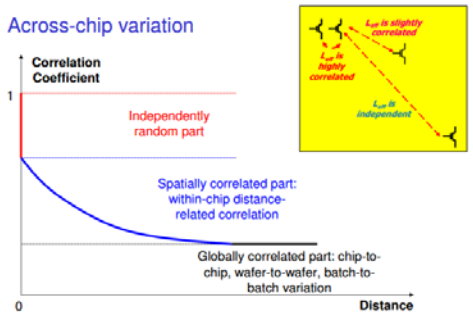
Advanced OCV (AOCV) (aka LOCV)



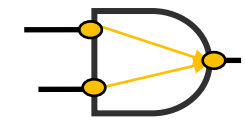
[Synopsys Whitepaper]

Distance-based AOCV

$$D_L < D_c + T_{clock}$$

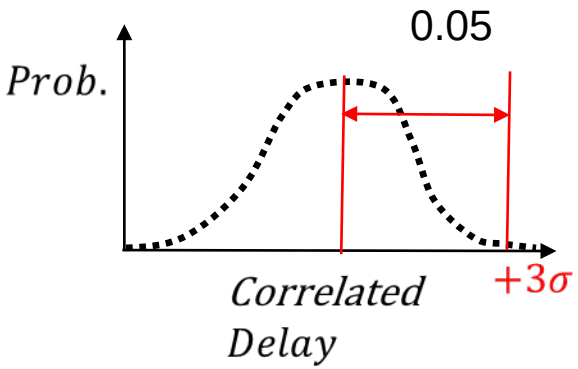


Advanced OCV (AOCV) (aka LOCV)

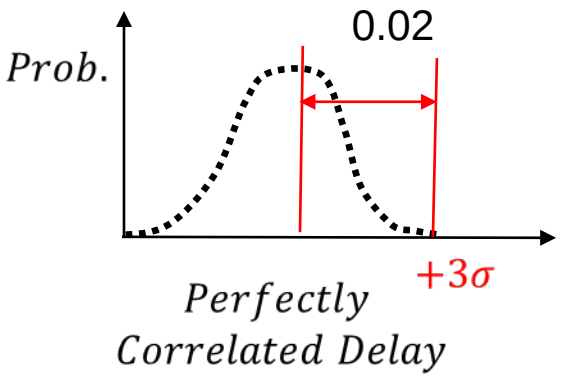


[0.95,1.05]

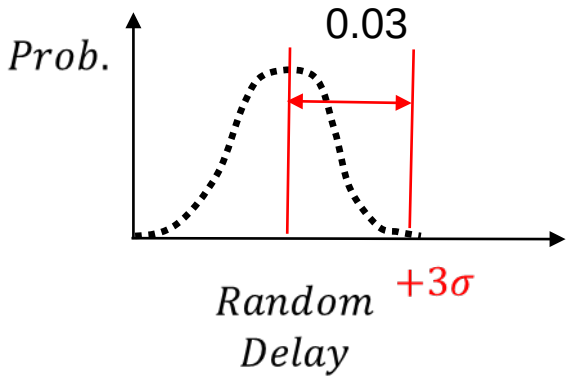
5% → 3%



=



+



Advanced OCV (AOCV) (aka LOCV)

Distance \ Depth	1	2	3	4	5	10	50	100
1000	1.130	1.099	1.085	1.078	1.074	1.063	1.055	1.053
2000	1.130	1.099	1.086	1.079	1.074	1.063	1.055	1.054
3000	1.131	1.100	1.186	1.079	1.075	1.064	1.056	1.054
4000	1.131	1.100	1.187	1.080	1.076	1.065	1.057	1.056
5000	1.133	1.102	1.189	1.082	1.078	1.068	1.061	1.059
6000	1.135	1.105	1.192	1.086	1.082	1.072	1.065	1.063
8000	1.137	1.108	1.195	1.089	1.085	1.076	1.070	1.068
10000	1.140	1.112	1.100	1.094	1.090	1.082	1.075	1.074
15000	1.147	1.120	1.110	1.104	1.101	1.093	1.088	1.087

Derating Table for each cell type

[Synopsys Whitepaper]

Advanced OCV (AOCV) (aka LOCV)

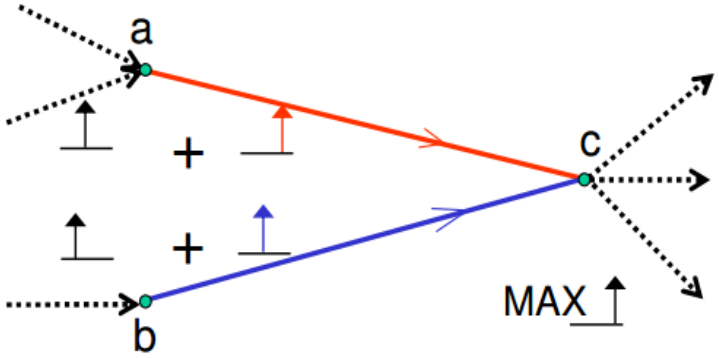
- AOCV requires a lot of library characterization efforts
  - Derating values for each cell type, each depth, each location, each slew, each load
  - Worst-case derating is selected across each load and each slew
    - A source of pessimism
- AOCV tables doesn't have much information
  - Can be predicted by the simple analytic model

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

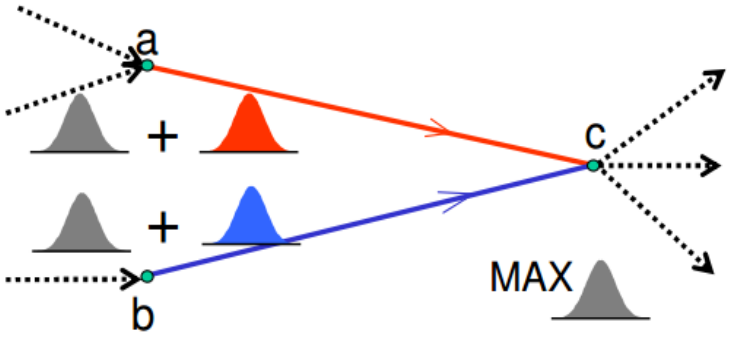
- Path credit is mapped into the credit of segment delays
  - Paths do not consist of a single type of gates
  - Not graph-based analysis (GBA) friendly

Statistical Static Timing Analysis (SSTA)

- Deterministic



- Statistical



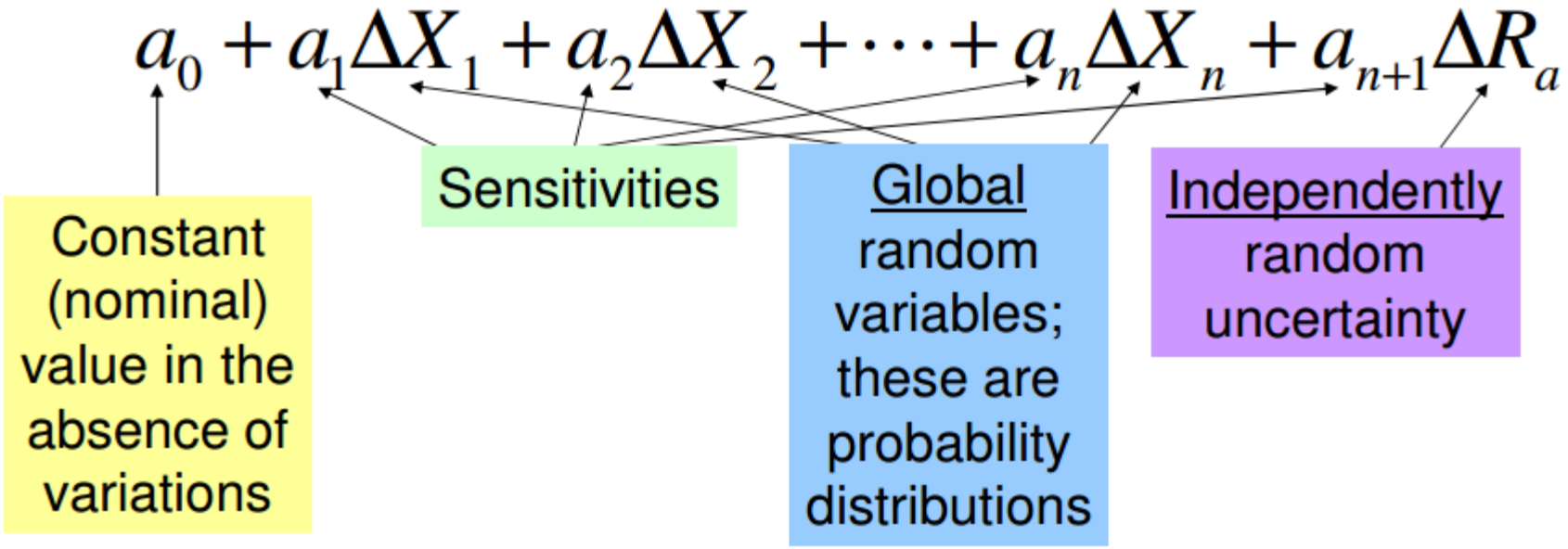
- Correlations
  - global
  - spatial
  - none!

$$E[X+Y] = E[X] + X[Y]$$

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

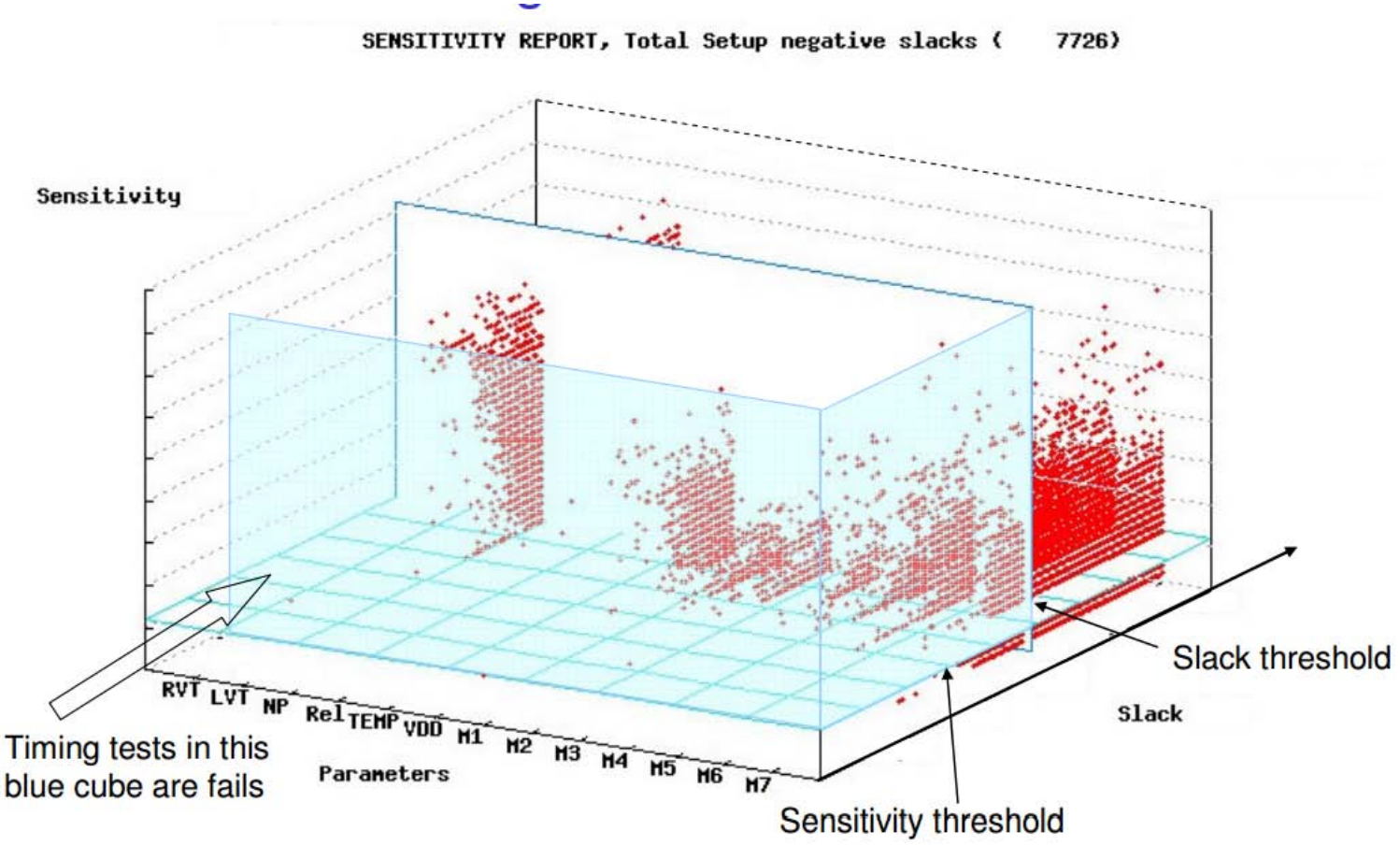
Statistical Static Timing Analysis (SSTA)

Canonical Form



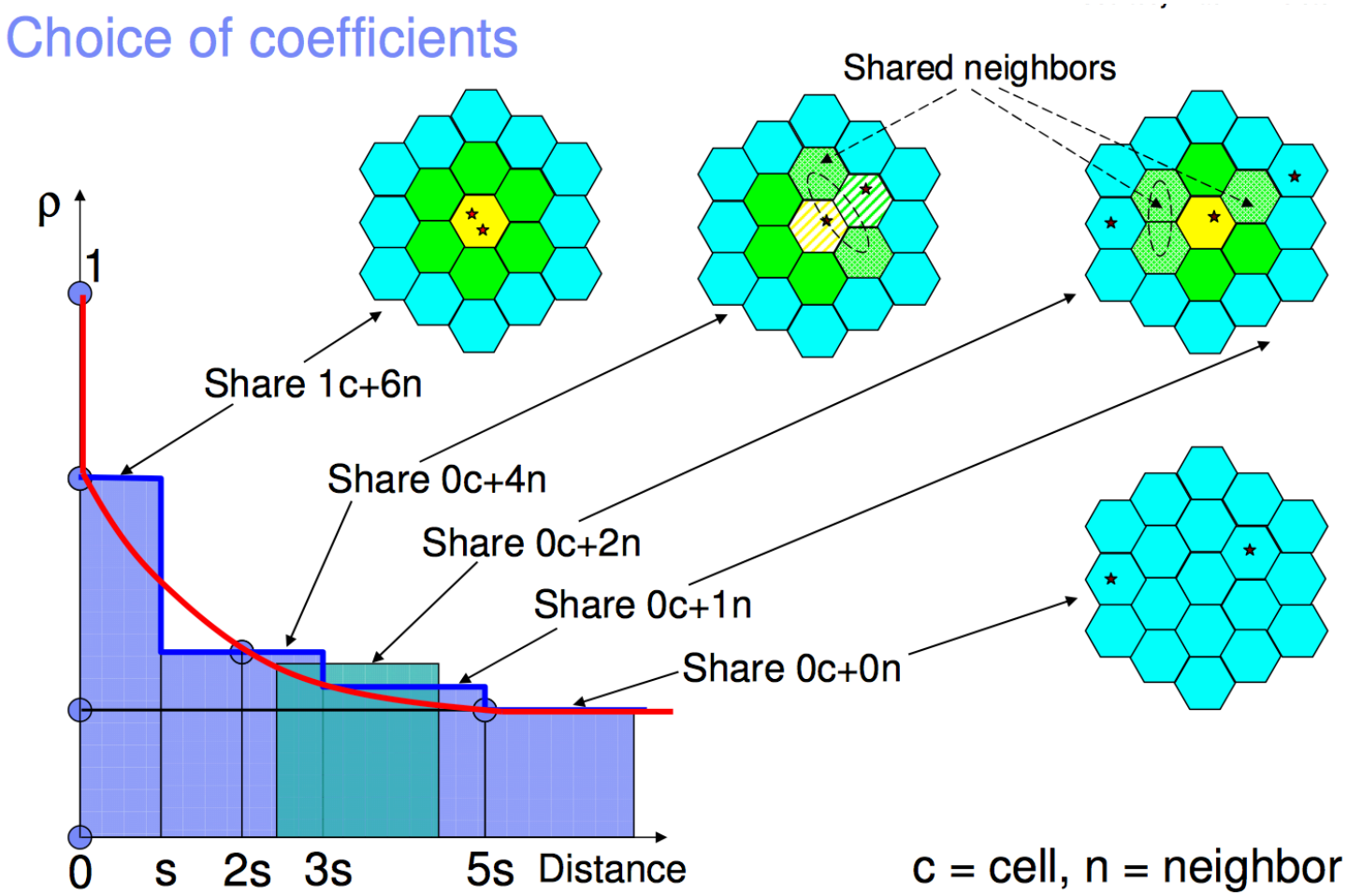
- All timing quantities computed and propagated in a parameterized form
  - ATs, RATs, slacks, slews, delays, etc
- Need to characterize sensitivities for each cell type, each delay, each slew

Statistical Static Timing Analysis (SSTA)



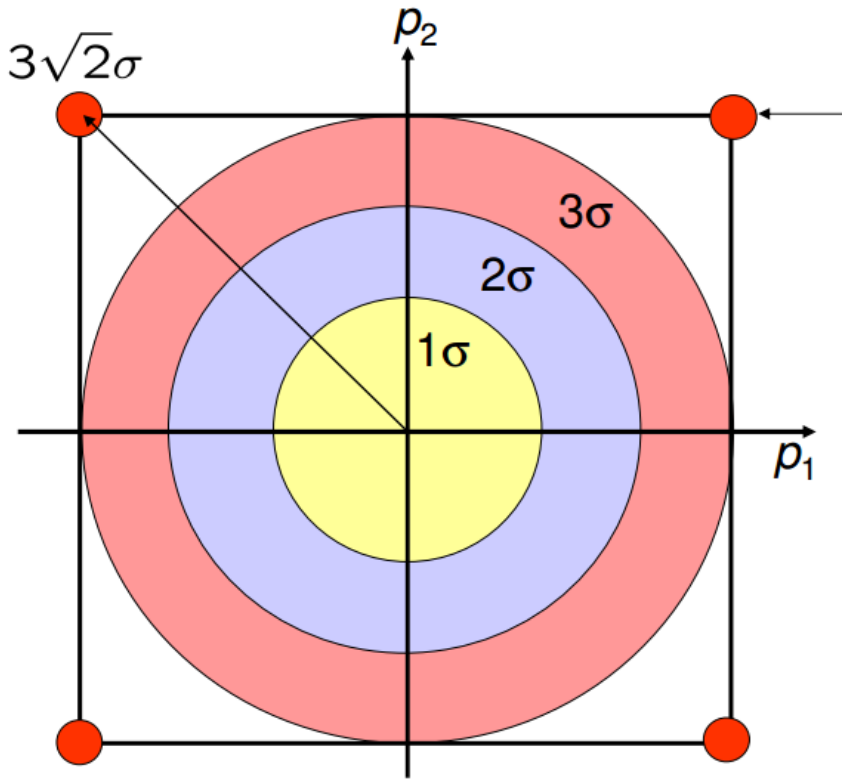
Statistical Timing: Where's the tofu? ICCAD 2009, IBM

Statistical Static Timing Analysis (SSTA)



Use 7 variables in the Canonical Form

Statistical Static Timing Analysis (SSTA)



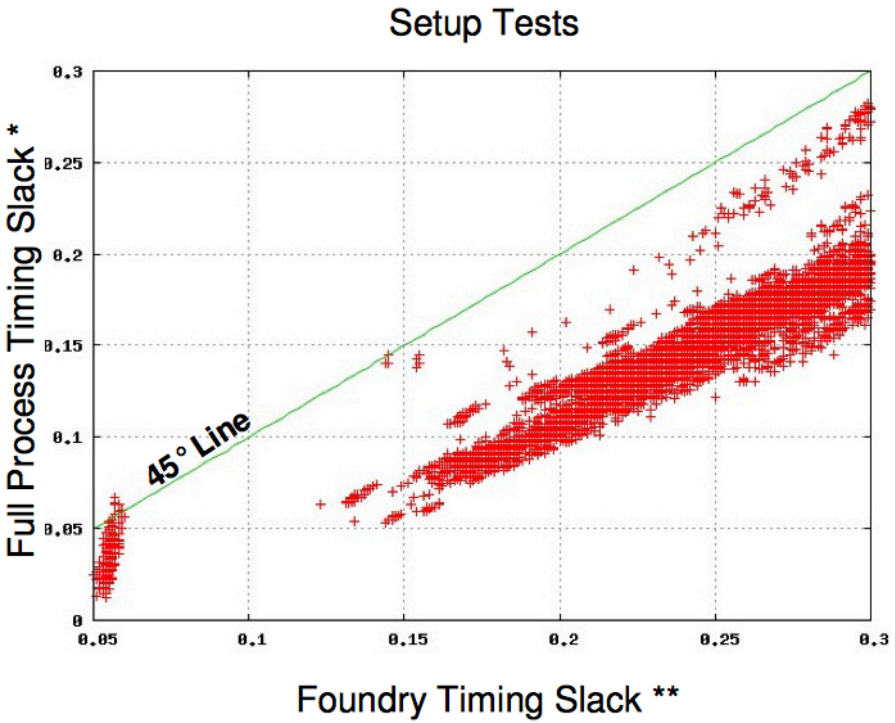
- SSTA benefits
  - Chip-to-chip variation
    - No corners
    - Safe
    - RSS credit
  - Within-chip variation
    - RSS credit down a path
    - RSS credit in setup/hold check

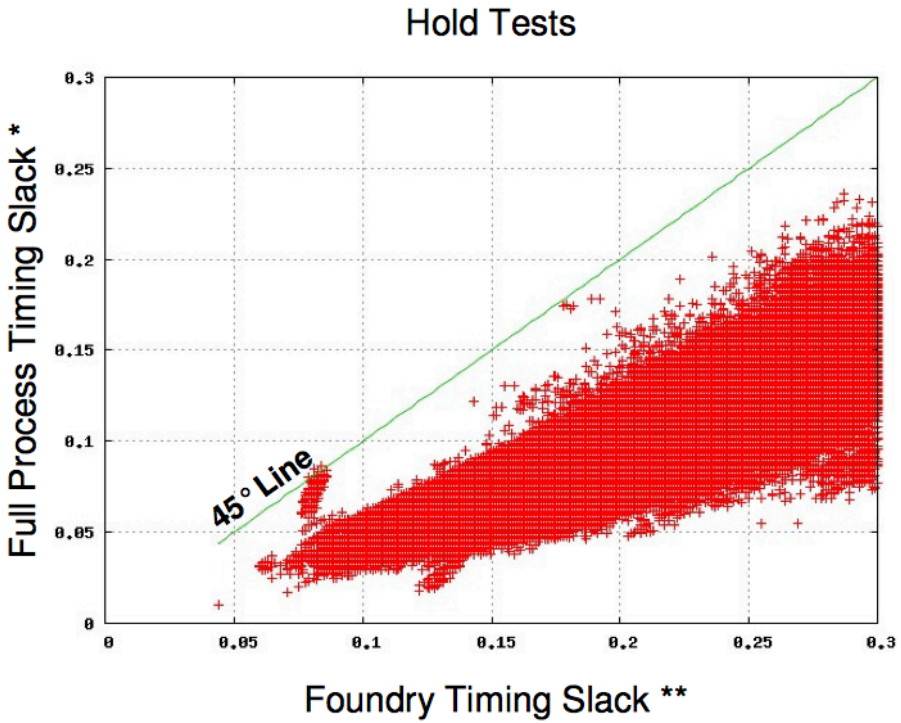
$$E[X+Y] = E[X] + E[Y]$$

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

Statistical Static Timing Analysis (SSTA)

Derating Factors of ±0% (Chip 2)

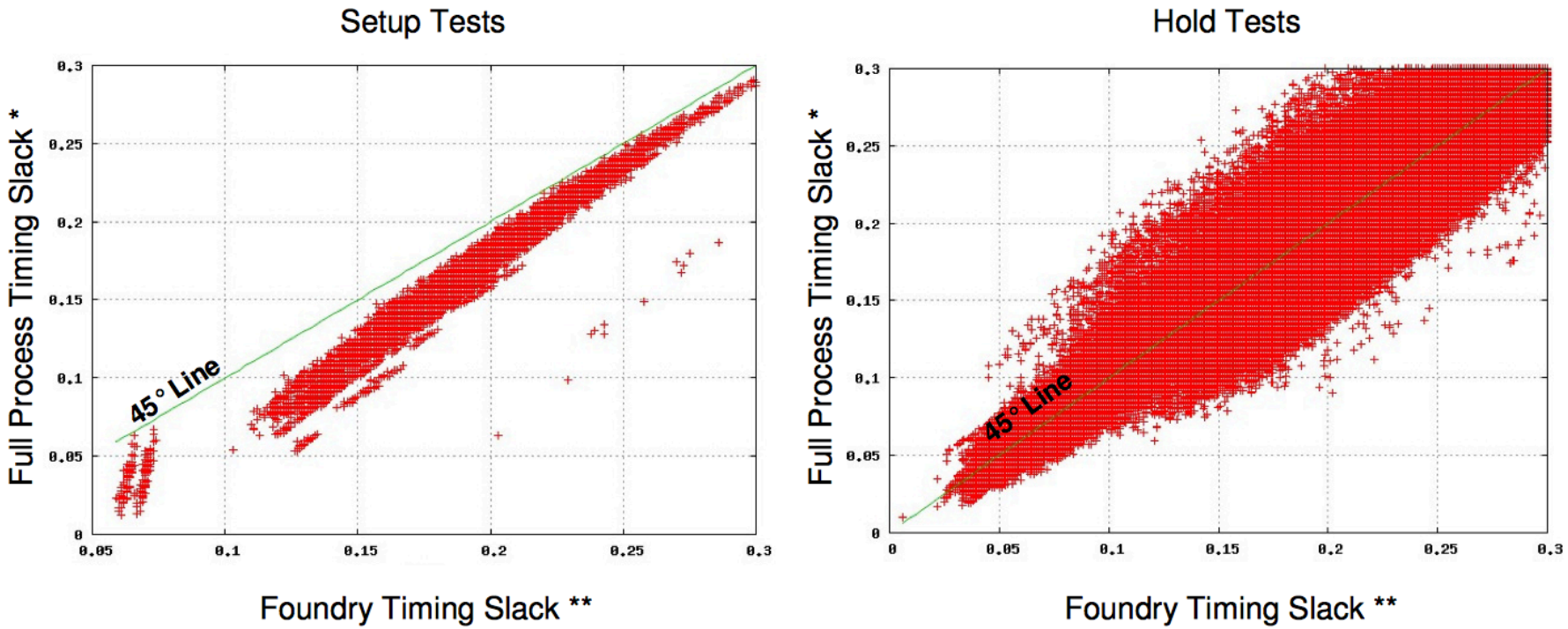




Statistical Timing: Where's the tofu? ICCAD 2009, IBM

Statistical Static Timing Analysis (SSTA)

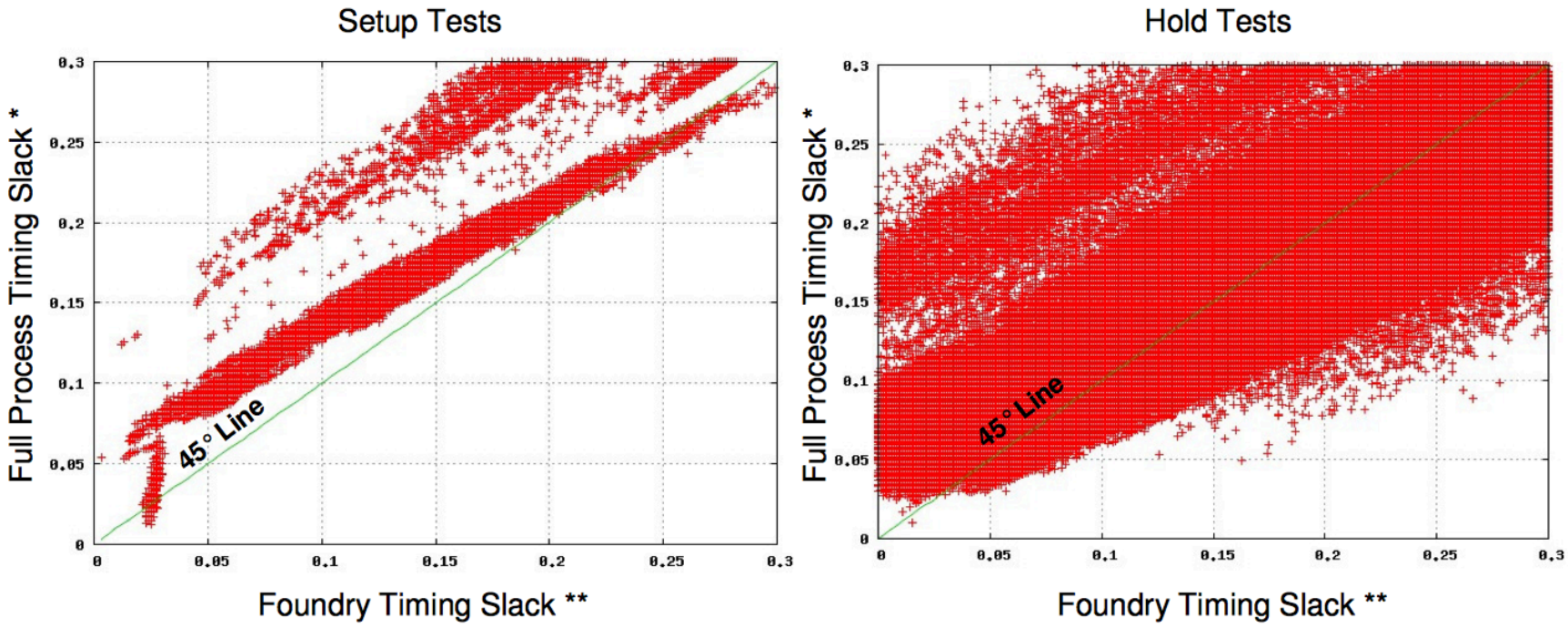
Derating Factors of ±5% (Chip 2)



Statistical Timing: Where's the tofu? ICCAD 2009, IBM

Statistical Static Timing Analysis (SSTA)

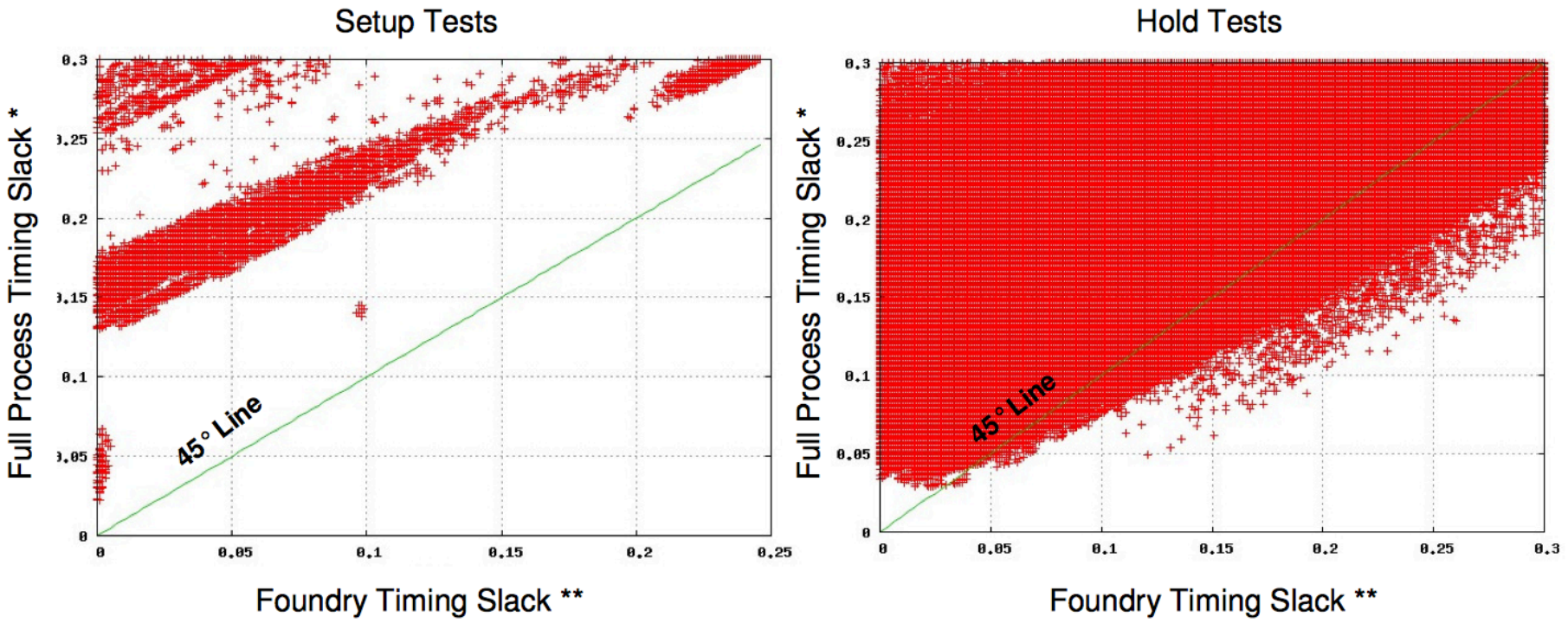
Derating Factors of ±13% (Chip 2)



Statistical Timing: Where's the tofu? ICCAD 2009, IBM

Statistical Static Timing Analysis (SSTA)

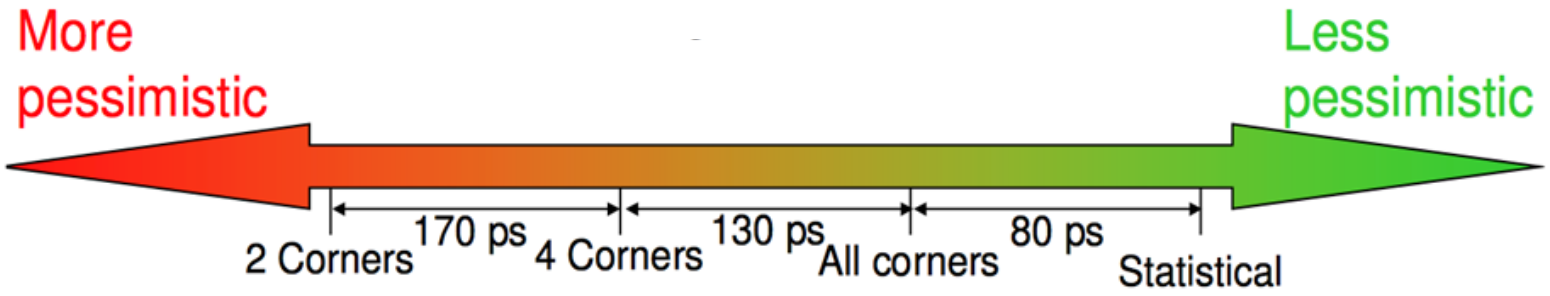
Derating Factors of ±25% (Chip 2)



Statistical Timing: Where's the tofu? ICCAD 2009, IBM

Statistical Static Timing Analysis (SSTA)

- Apples-to-apples comparison of statistical flow to:
  - 2 corner foundry-like timing with derating
  - ‘n’ corner industry-standard flow
  - Exhaustive corner timing



- 380ps total
  - 200ps from RSS credit in chip-to-chip variation
  - 80ps from RSS credit in on-chip variation

Statistical Timing: Where’s the tofu? ICCAD 2009, IBM

Parametric OCV (POCV) (aka SOCV)

- Use SSTA for within-chip variation only
- Eliminate a lot of characterization burden from SSTA, giving up the benefits in chip-to-chip variation
- Use a few variables only in the canonical form

$$delay = d_0 + \Delta d + \frac{\partial d}{\partial r}\Delta r + \frac{\partial d}{\partial c}\Delta c + \frac{\partial d}{\partial c_L}\Delta c_L$$

- Statistical OCV (SOCV) is a similar technique
- In theory, POCV/SOCV is clearly a better engineering than AOCV
  - Better accuracy and less characterization effort

"A parametric approach for handling local variation effects in timing analysis", DAC 2009, Mutlu. A (Extreme DA)

Remaining Pessimism in SSTA/POCV

- Refactoring - CRPR for Combinational Networks

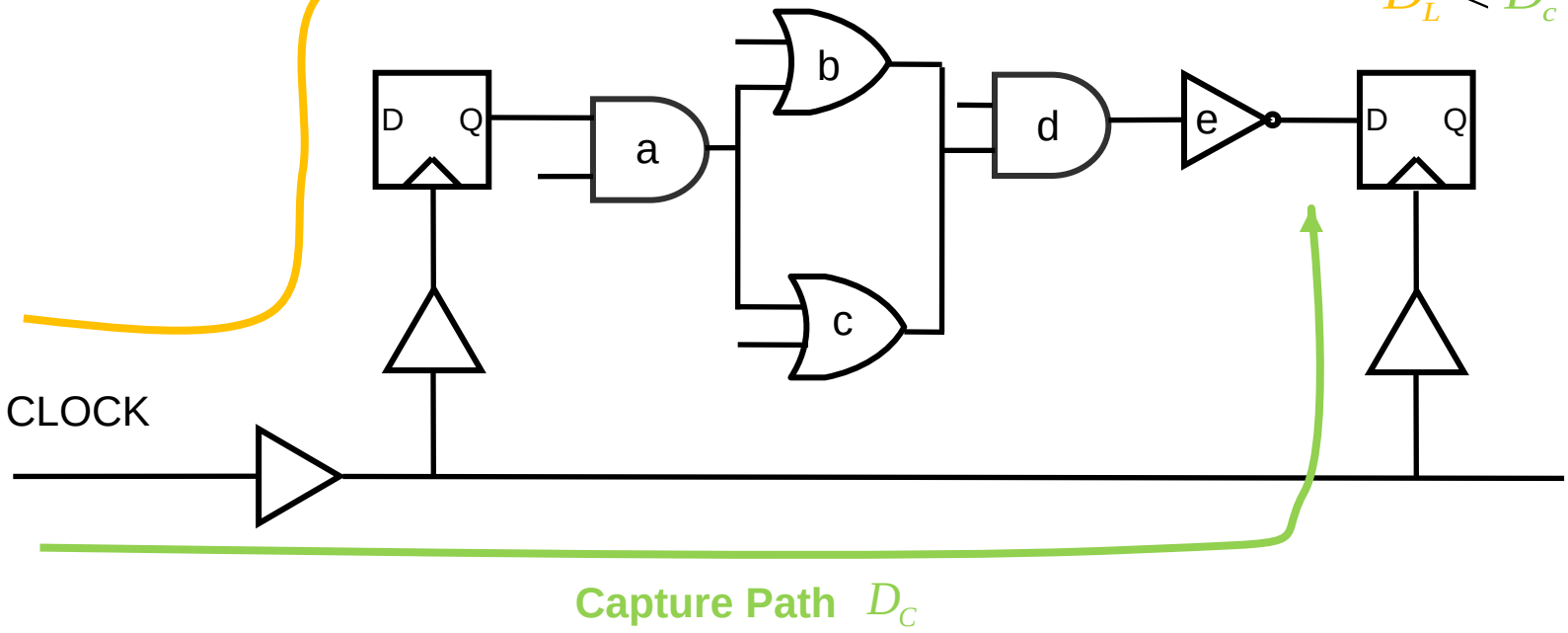
Launch Path $D_L = D_{cd} + \max(a+b, a+c) + d + e$

$= D_{cd} + a + \max(b,c) + d + e$

Using Distributivity Of + over *max*

Setup Check

$D_L < D_c + T_{clock}$

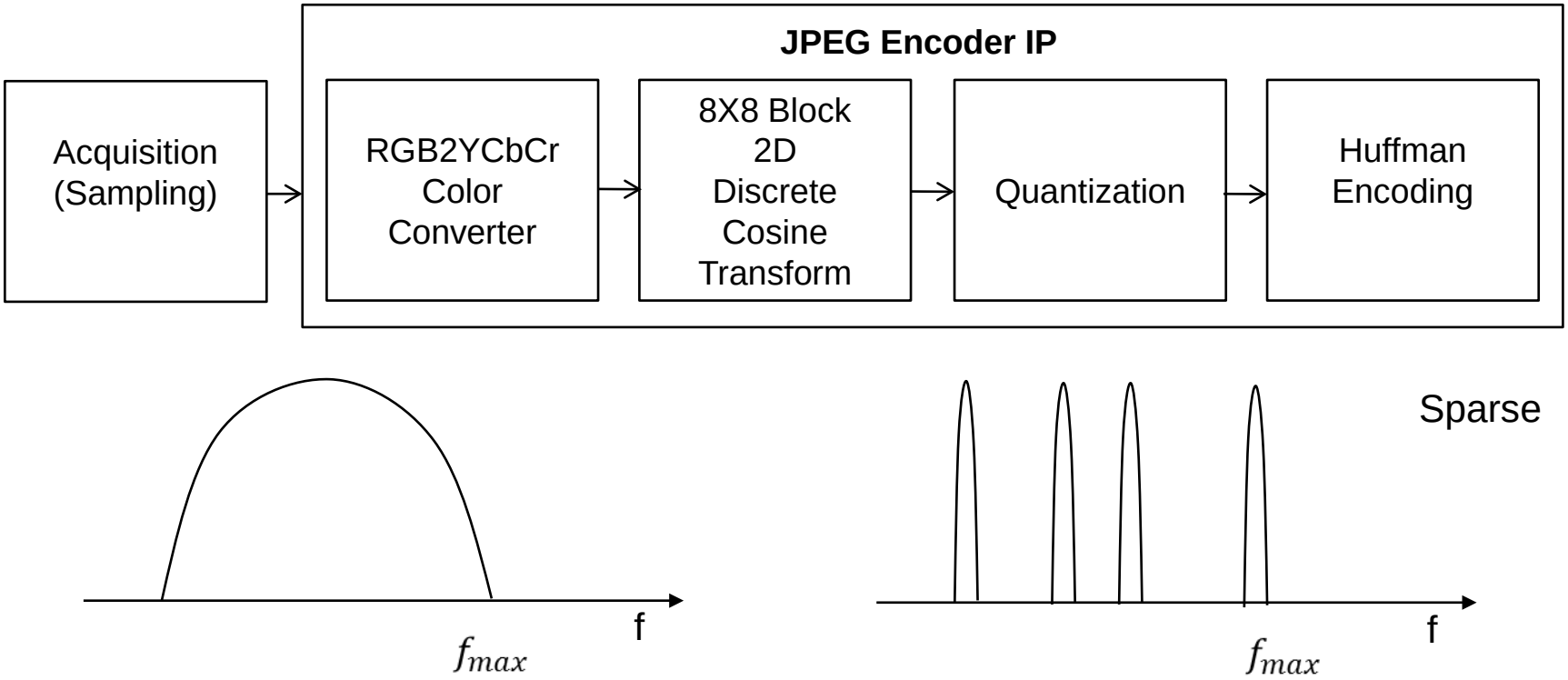


Capture Path D_C

[Chung and Abraham, ICCAD 2009] (Best Paper Award Nomination)
[Chung and Abraham, TCAD 2012]

Compressed Sensing

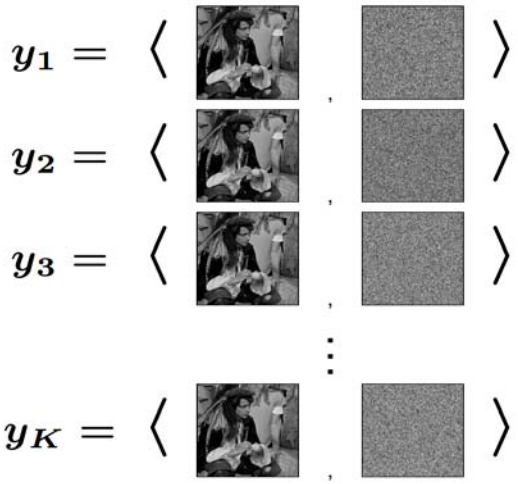
- Well-known that natural signals are *compressible*
- Traditional DSP Systems



Compressed Sensing

- Tremendous impact on signal processing, machine learning, statistics,..
- The original groundbreaking paper [Donoho 2004] has been cited 8769 times (200+ papers in the last 3 years.)

Linear measurements



or

Non-uniform sampling

Decoding or Recovery

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_k \end{bmatrix} = \begin{bmatrix} T_{1,1} & \cdots & T_{1,n} \\ T_{2,1} & \cdots & T_{2,n} \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ T_{n,1} & \cdots & T_{n,n} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ \theta_1 \\ 0 \\ \vdots \\ 0 \\ \theta_m \\ 0 \end{bmatrix}$$

random sampling matrix (k x n)

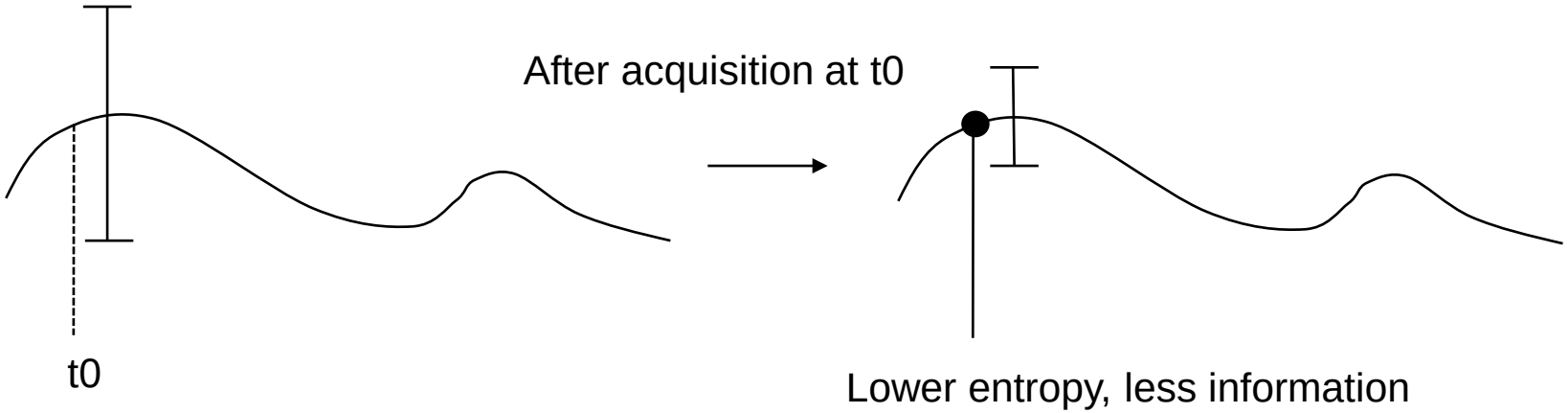
signal representation in terms of m << n coefficients

- Classical answer:
  - Underdetermined → cannot solve
- We have k equations and 2m unknowns,
  - If k>2m, we may have a unique solution
- New answer: Information on 2m unknowns

are encoded into k measurements, and
we can recover it *perfectly and efficiently*
(*In practice, around 4m are needed*)

Compressed Sensing

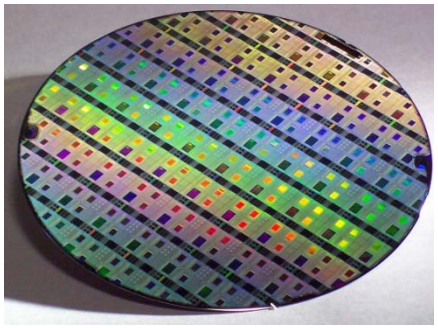
- Images and sounds have continuation
- Samples adjacent in time or space are highly correlated (high energy at low frequencies)
  - Conventional measurements are not efficient



- CS recovers/*predicts* unobserved quantities from a few observations

Compressed Silicon Sensing

- In IC manufacturing, measurements are expensive
  - IC cost = die cost + test cost + package cost
- Could be applicable to pre-silicon as well (where some simulations are expensive or interpolation is used)

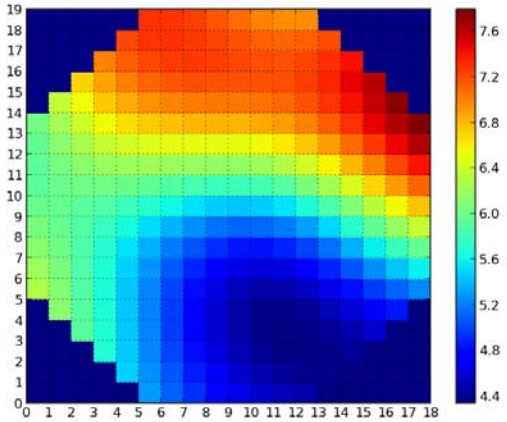


Wafer

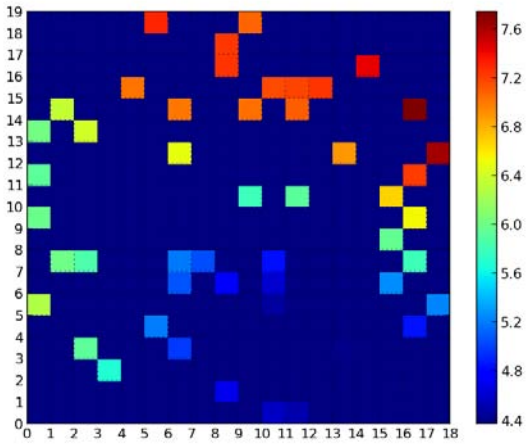
Automatic Test Equipment (ATE)

Virtual Probe

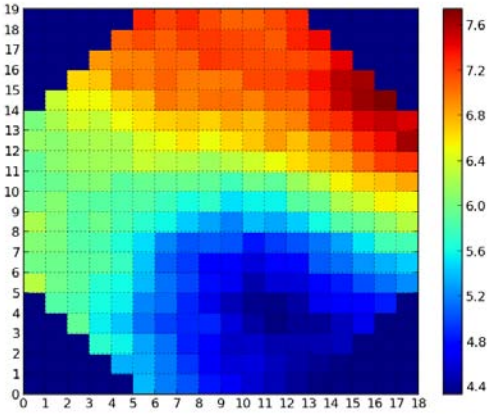
- Framework for wafer characterization
- Many wafer test results are spatially correlated across wafer



Spatially correlated data (282 measurements)



Random 50 measurements



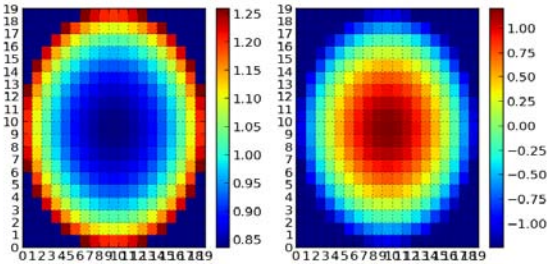
Predicted from 50 samples

1.8% Error

Our CSS Framework

- Test-items are also correlated strongly
  - VP recover results of each test-item independently
  - Our approach does it simultaneously

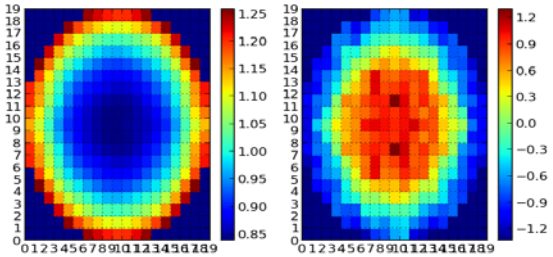
Synthetic wafer



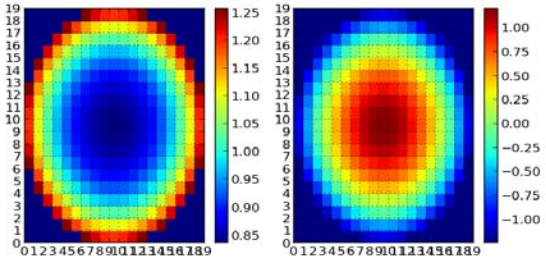
Normalized Flushed delay

Normalized Log(IDDQ)

50 samples/item



VP's Prediction (12% Error)

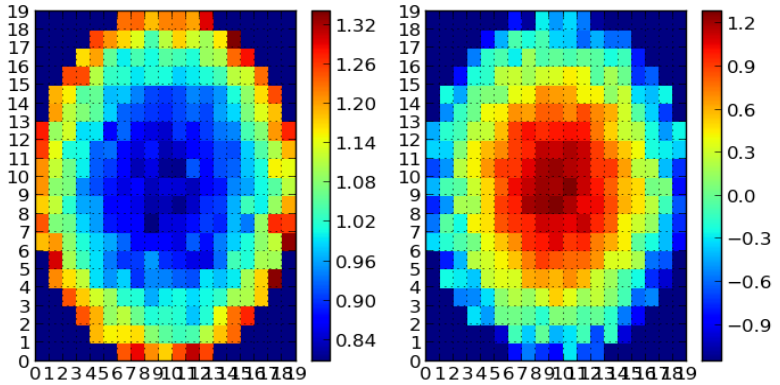


Our Prediction (0.4% Error)

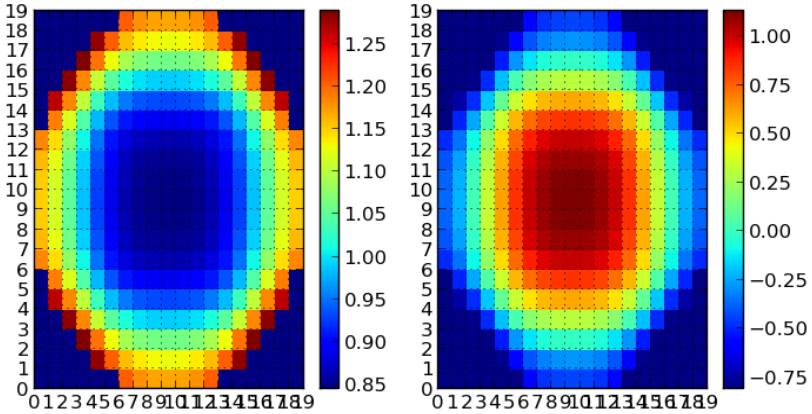
Our CSS Framework

- Can decompose it into correlated variation and random variation

Synthetic wafer



50 samples/item



Applications of CSS

- What can we do if we have a very good predictor?
- At wafer characterization step, sampled measurements are common

Conclusions

- Robustness is the key to success in nanometer technologies
  - Margins are the easiest way to obtain robustness
  - Margins eat up competitiveness
  - Needs sophisticated engineering for margining (OCV, AOCV, POCV,…)
- Post-silicon engineering (silicon debug, characterization, etc) is very important under large-scale process variations
  - Compressed Silicon Sensing
    - CS is a revolutionary theory
    - Let’s take advantage of it at IC design and manufacturing!

Q/A

Thank you!