EE382M
VLSI-II

Statistical Static Timing Analysis
Spring 2017

Mark McDermott
Jacob Abraham
Matthew J. Amatangelo

Agenda

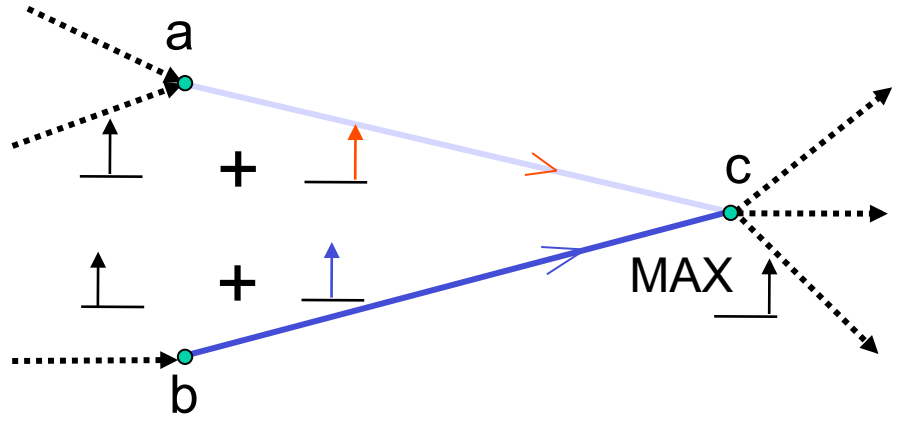
- SSTA
- Reporting
- Models
- Special considerations and limitations
- Implications for manufacturing test

Key points of Statistical STA (SSTA)

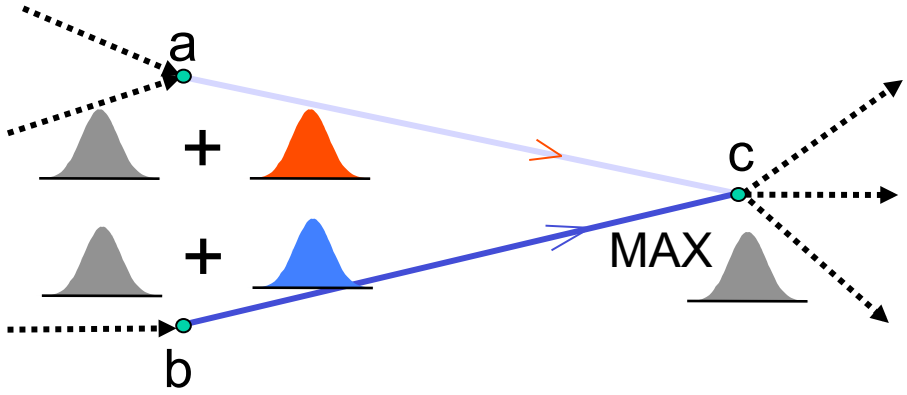
- Improvement to Static Timing Analysis by quantifying process variation as an effect on gate and wire delays
- Replaces STA guard-banding
- SSTA provides the probability that each path passes over the range of independent variables
- SSTA points to process parameters that need tweaked:
  - Given the circuit, it finds sensitivities of process parameter variance versus path delays
- Components of variance must be carefully considered
- Correlation is the absolute key in reducing over-pessimism
- SSTA relies on a great deal of process and circuit analysis
  - Most FABs don't go beyond standard STA cell libraries per corner
  - EDA industry generally doesn't have access to detailed process correlation data

STA vs. SSTA

- **Deterministic**

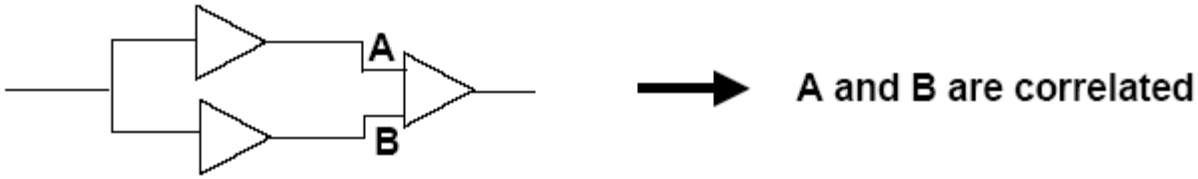


- **Statistical**

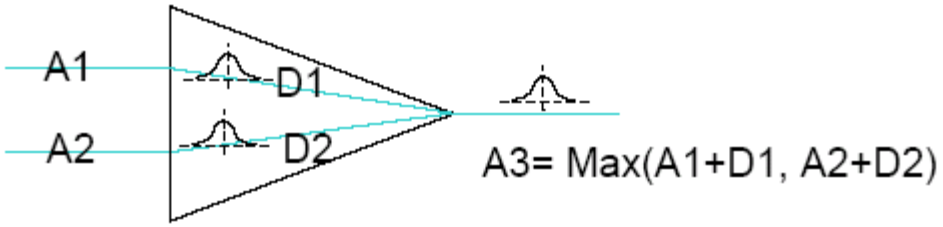


Statistical STA

- **Built on top of STA**
- **Handles correlations due to re-convergent paths**
  - **Assuming random process variation (R) is tracked**



- **Handles spatial correlations**
- **Statistical sum and max operations**



- **First-order parameterized sensitivity propagation**

$$A = A_0 + \sum A_i \Delta p_i$$

Where p includes both systematic and random variation

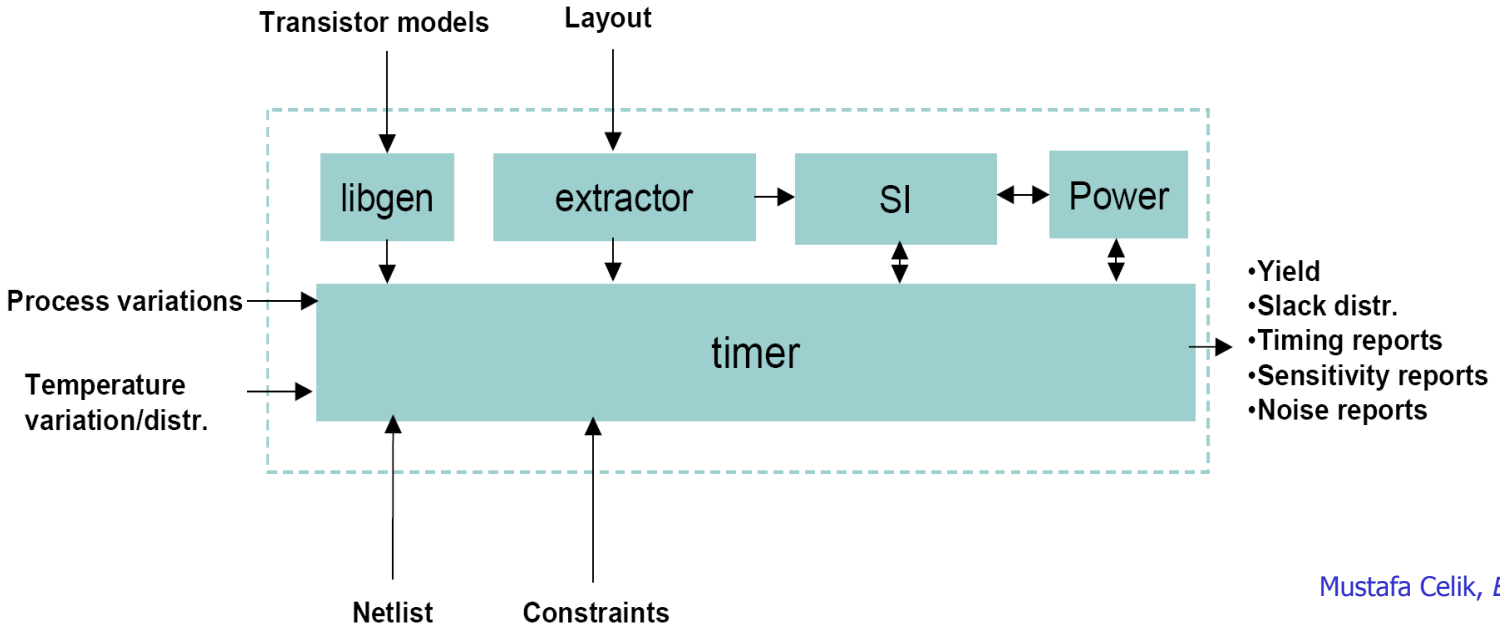
Mustafa Celik, *Extreme-DA*

Statistical STA

- **Statistical Static Timing Analysis is only as good as the quality of the input data**
  - Classical STA Timing models represent one point in Si process space
    - It does not consider variation of delay-critical parameters
    - Same goes for delay-critical parameters of each wiring layer
    - N parameters have 2**N corners
      - Exponential STA analysis points to cover process space
      - How to predict yield?
  - Proportion of inter- and intra-die process variation increases with decreasing feature sizes
    - Worst case STA models become too pessimistic
      - Some corners not reachable due to systematic correlations
    - Same goes for temperature and voltage deviations
  - Conclusion: Classical STA must improve by quantifying process variation as an effect on gate and wire delays

A Good Statistical Timer Provides:

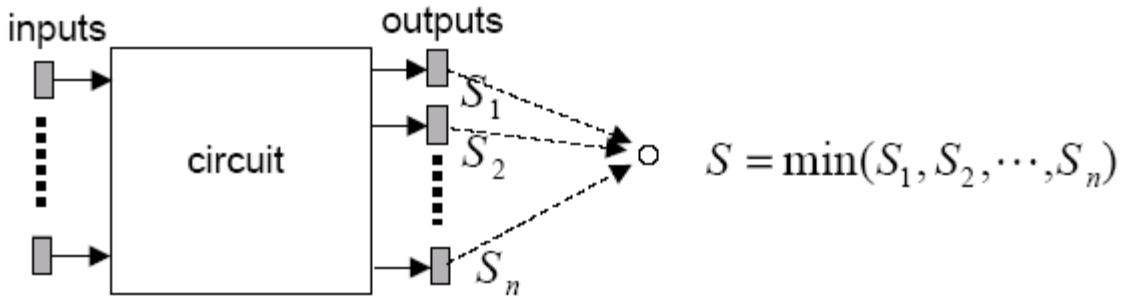
- Timing analysis for yield prediction
- Spatial and re-convergence correlations
- Analysis of on-chip and chip-to-chip variations
- Sensitivity analysis for parametric yield optimization
- Built-in variation-aware extractor
- Delay calculation and signal integrity analysis for silicon accuracy
- Monte-Carlo, SPICE, and 3D extraction capabilities
- Process variation modeling capabilities



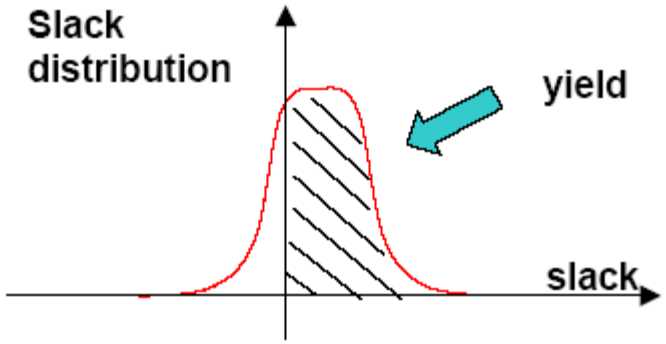
Mustafa Celik, *Extreme-DA*

Slack Distribution and Parametric Yield

- **Design slack is the minimum of all slacks**

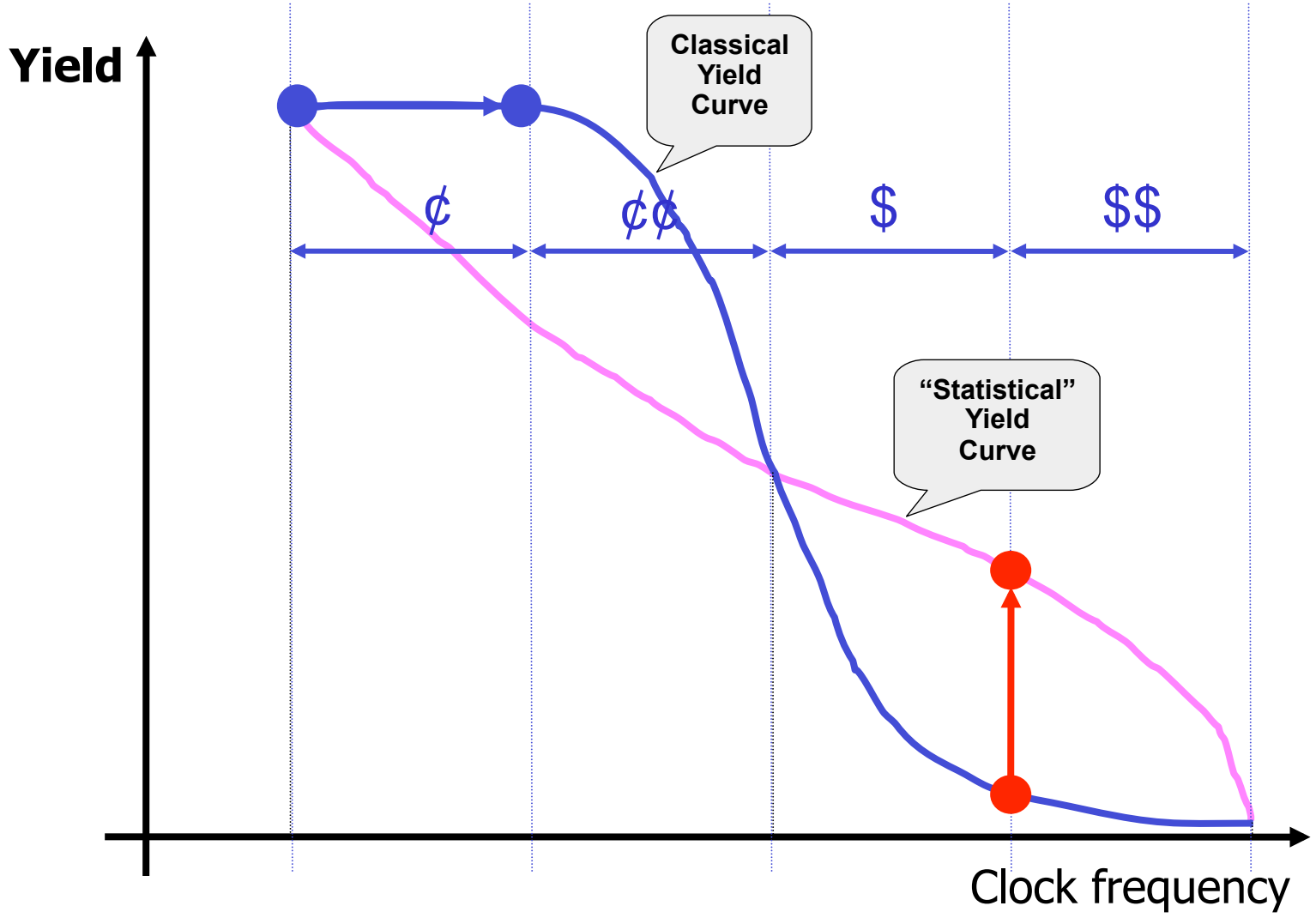


- **Parametric yield is obtained from the design slack distribution for a target clock frequency**



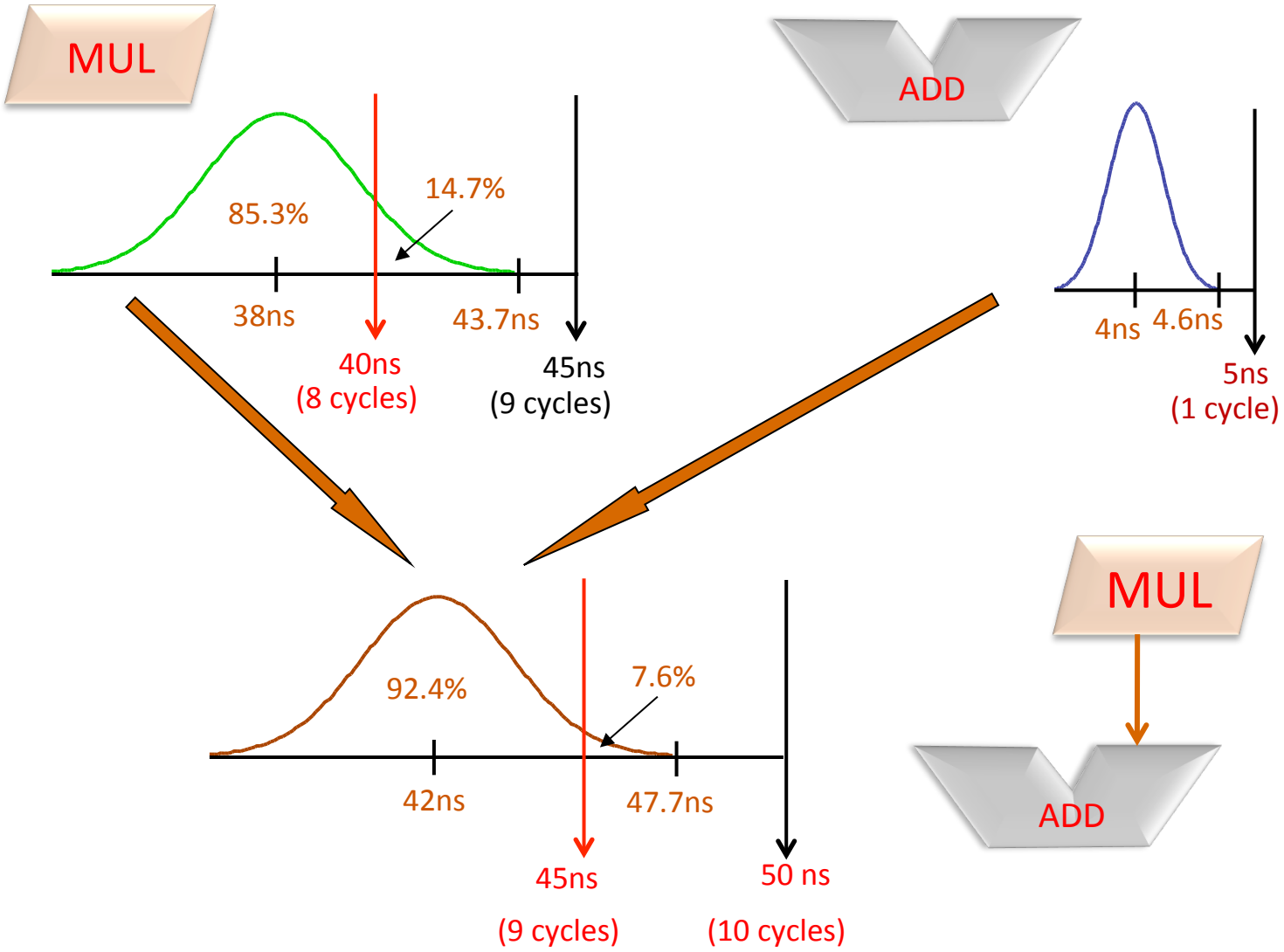
Mustafa Celik, *Extreme-DA*

Parametric yield curve



Chandu Visweswariah, *IBM Thomas J. Watson Research Center*

Trade-off between latency and parametric yield



Canonical variational delay model

- **Correlations are the problem**
  - In a circuit with 1M nodes and 2M edges and 12 timing values per node/ edge, we DO NOT want to store or manipulate a 36M x 36M covariance matrix!
  - Instead, parameterize all timing quantities by the sources of variation
  - Use a first-order canonical model:

$$a_0 + a_1 \Delta X_1 + a_2 \Delta X_2 + \cdots + a_n \Delta X_n + a_{n+1} \Delta R_a$$

Constant (nominal value)

Sensitivities

Deviation of global sources of variation from their nominal values

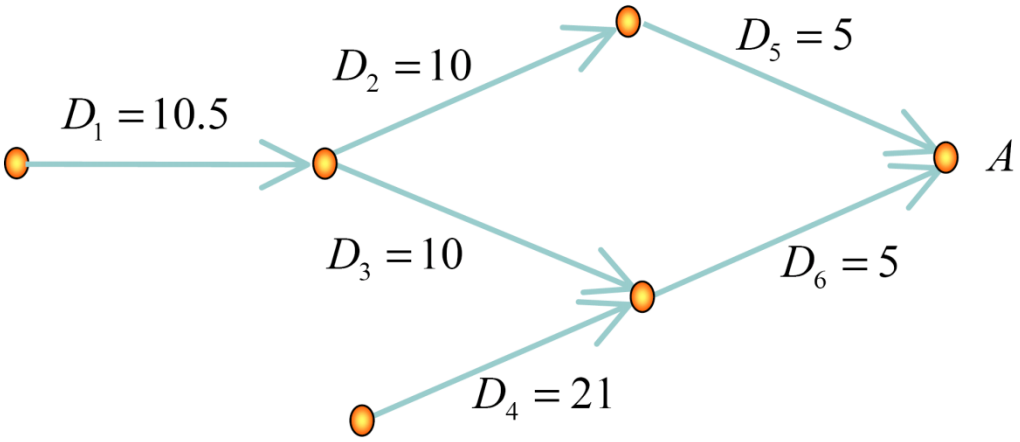
Random uncertainty (deviation from nominal value)

Chandu Visweswariah, *IBM Thomas J. Watson Research Center*

Sensitivity Analysis

- Define the sensitivity of output WRT a delay $\frac{\Delta A}{\Delta D_i}$

- Without delay variations, sensitivities w.r.t. delay arcs in the critical path will be one, and others will be zero
- Only these delay elements will be optimized



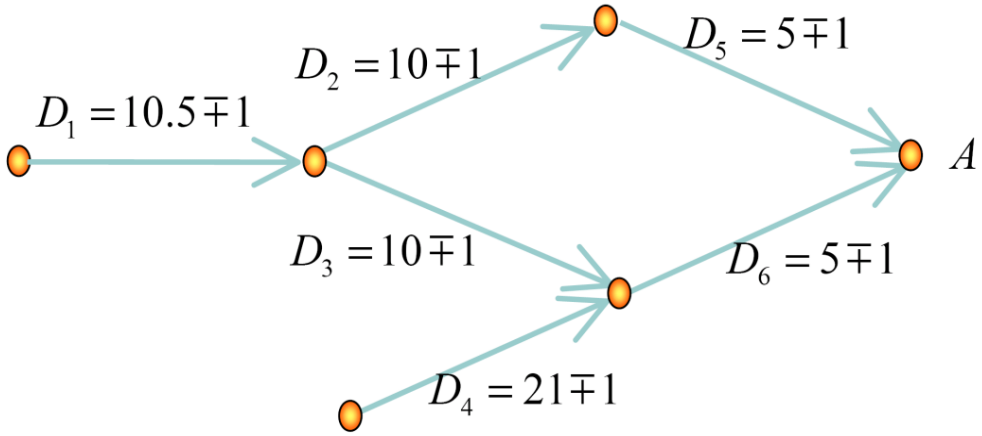
$$\frac{\partial A}{\partial D_1} = 0 \quad \frac{\partial A}{\partial D_2} = 0$$
$$\frac{\partial A}{\partial D_3} = 0 \quad \frac{\partial A}{\partial D_4} = 1$$
$$\frac{\partial A}{\partial D_5} = 0 \quad \frac{\partial A}{\partial D_6} = 1$$

→ Optimize D4 and D6

Mustafa Celik, *Extreme-DA*

Sensitivity Analysis

- **With process variations**
  - Critical path is not well defined
  - Every path can be "critical" under certain probability
- **Now, all the sensitivities have nonzero values and their criticalities are different**



$$\frac{\partial A}{\partial D_1} = 0.56 \quad \frac{\partial A}{\partial D_2} = 0.33$$

$$\frac{\partial A}{\partial D_3} = 0.23 \quad \frac{\partial A}{\partial D_4} = 0.44$$

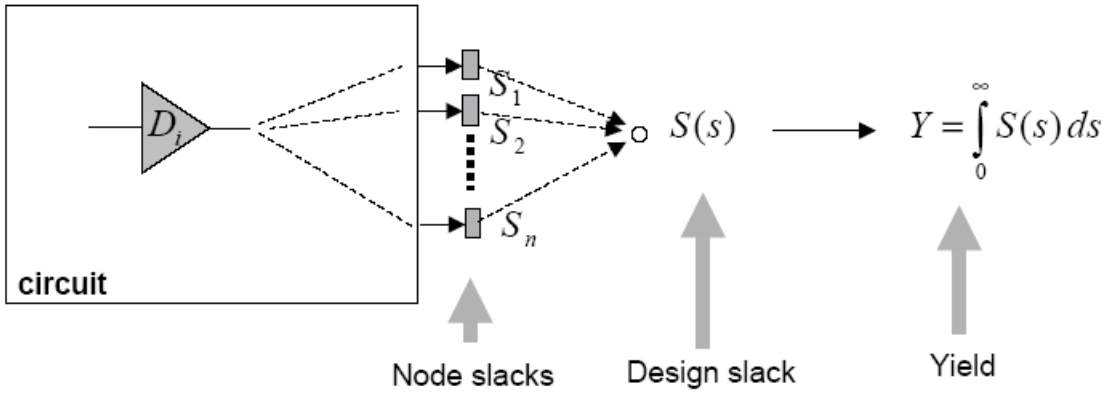
$$\frac{\partial A}{\partial D_5} = 0.33 \quad \frac{\partial A}{\partial D_6} = 0.67$$

→ Optimize D1 and D6

Mustafa Celik, *Extreme-DA*

Sensitivity Analysis

- Yield sensitivity w.r.t. every instance delay $\frac{\partial Y}{\partial D_i}$



- Statistical sensitivities provide much more useful information than critical path analysis
- Calculates the sensitivities of delays, arrivals, slacks, design slack, and parametric-yield w.r.t.
  - design parameters (cell size, wire size, wire spacing)
  - system parameters (vdd, temperature)
  - process parameters

Mustafa Celik, *Extreme-DA*

SSTA Summary

- **Motivation**
  - Variability is proportionately increasing
    - As feature sizes decrease, variances increase
- **Better predicts performance and yield than STA**
  - e.g., P(ATdata < ATclk)
  - Correlations matter
    - 2 sources
      - Design: reconvergent paths
      - Process: systematic
  - Variation sensitivities matter
  - Robustness is an important metric
    - Higher order models (e.g., quadratic)

- **As with STA, SSTA lacks accuracy compared to dynamic simulation considering simultaneous switching, parallel drivers**

Agenda

- SSTA
- **Reporting**
- Models
- Special considerations and limitations
- Implications for manufacturing test

Reporting

- **Variety of report types**
  - Instance reports
    - Provides ATs, RATs, corresponding phases, slews, and other node information for every port of instances
    - Enables discovery of unwanted ATs and missing data
  - Path slack report
    - Shows progression of signal through each node of the graph between its launch and capture points including these node attributes: cumulative delay, clock phase, net name and edge direction, slew, capacitance
    - Normally organized in order of largest violations
    - Usually provides downstream path violations that are independent of upstream problems – "Re-Launching"
    - STA adjusts arrival times to be modulo the clock cycle -- "Cycle Accounting"
  - Slack histograms
    - Shows slack distribution
  - STA run errors and warnings

Reporting (cont)

- **SSTA provides**
  - Probability that each path passes over the pvt range
    - System reliability is as good as the worst path
  - Points to process parameters that need tweaked
    - Find sensitivities of parameter variance vs path delays
    - Large sensitivity coefficients means greater impact on delay
  - Most critical gates/interconnect

Agenda

- SSTA
- Reporting
- **Models**
- Special considerations and limitations
- Implications for manufacturing test

Timing Models

- **What they are**
  - Models represent the timing for a particular block of logic
    - Reduced data set specifying only what the higher level of analysis needs to experience from this logic block
    - Normally separate model exists per logic block for both max and min modes
    - Programming languages exist to describe complex timing behavior of almost any circuit (e.g., TLF, DCL = Delay Calculator Language, an IEEE Standard).
  - They define the delay arcs between nodes
  - They define the tests at nodes (including the margins) and the reference node and required setup or hold times
    - Should delineate the operating frequency of the logic (pulse widths on clocks)
  - Model types
    - Black box
      - Appropriate for flushless designs – test points at model boundary
      - Model only combinatorial paths and paths to and from hard timing boundaries (flip flops and clock gates)
      - Unfortunately internal reg to reg paths are not represented normally
    - Gray box
      - Appropriate for latch and domino designs
      - Includes internal test points

Timing Models (cont.)

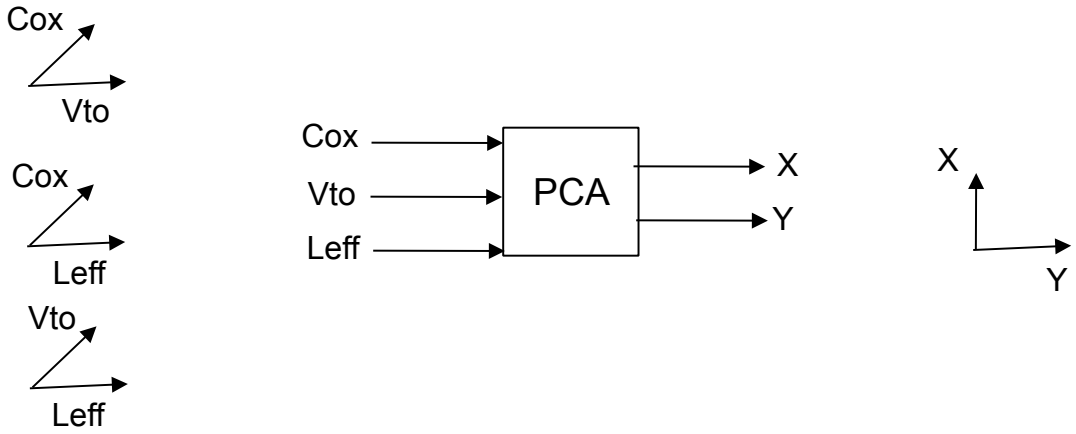
- **Where the numbers come from**
  - Early design phase: estimates, scaled delays
  - cell based: models created from simulation
    - delays and slews are modulated by input slews and output loads
    - accuracy degraded because any path is comprised of multiple cells, each having an accuracy tolerance
    - the broader the range of electrical tolerances, the greater the error
  - Custom or true path analysis: models created by transistor timing tool
    - accuracy greater than cell based approach because each channel connected component (CCC) is simulated to the actual in-circuit electrical conditions
      - Except for the pins exposed to the block's boundary
      - Often represented with tables
      - Can include simultaneous switching routinely
    - can be extremely time consuming to create models for blocks
- **Delay, Slews, and Required Setup/Hold Times (elements where clock meets data) are appropriate for a particular PVT (process, voltage, temperature) corner**
  - SSTA would alleviate need for multiple model corners

Timing Models (cont.)

- **Models are characterized for a range of circuit conditions**
  - Input slews, output loads
  - Never use the model outside of these ranges (STA warning)
- **Custom Logic Macros**
  - Transistor level timers perform STA and timing rule generation.
    - Black & Gray Box models.
- **Synthesized logic**
  - Standard Cell Libraries typically come with timing rule libraries in a variety of industry standard formats.
- **Embedded Arrays (SRAMs)**
  - Spice analysis of hand generated array cross sections is still common.
  - Transistor level timers such as PathMill and Dynacore have abstraction and modelling capabilities that enable them to do timing analysis and timing rule generation on a wide assortment of embedded array macros for Gray Box modelling.

Timing Models (cont.)

- **SSTA cell models**
  - Add parameter fluctuations to independent variables of combinational logic and input slew rate
    - Parameters could include Vt, Leff, etc, or be based upon principal component analysis (derived to cover variance)
      - PCA makes all variables orthogonal to each other in order to best represent variance



  - Outputs both delay and slew with mean and variance
    - PDF options include linear relationship with parameters (assumes pdf always Gaussian) or quadratic (etc) form to achieve better approximations to cover when inputs have vastly differing distributions

PCA – Principle Component Analysis

Agenda

- SSTA
- Reporting
- Models
- **Special considerations and limitations**
- Implications for manufacturing test

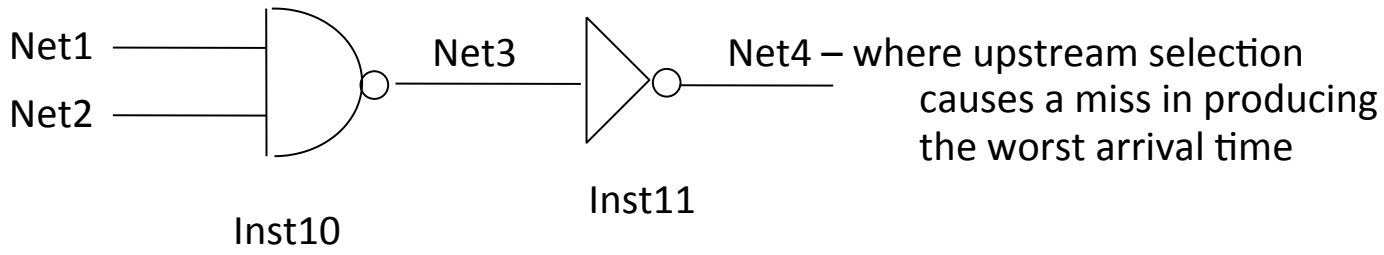
Special Considerations

- **Path considerations (See Extras)**
  - Tests, end points, and path enumeration
  - Loops must be broken – what happens
- **Problems with static timing**
  - What can happen with arrival time propagation
  - Worst case delays (min and max) -- simultaneous switching and contention
  - Parallel drivers
- **Wire delay**
- **Special path control**
- **Skew reduction**

What Can Happen with AT Propagation

- **STA keeps min and max AT for each edge/phase.**
  - **However, the output slew usually is that corresponding to the particular AT**
    - **Could miss the true worst case path, particularly in max mode**
    - **Breadth-first: prunes away all non-extreme arrival times at the output node**
    - **It's possible that one of the non-extreme ATs had a much worse slew than the extreme AT as in Net1 -> Net3 below. This gets omitted because from the subsequent delay calculation (for Inst11)**
  - **Some timers allow propagating worst slew or some composite of worst AT and "nearby" worse slew**

Node	AT/slew	Net3 result	Net4 (STA)	Net4 (depth-first)
Net1	150/300	180/220	---------	210/150
Net2	160/130	182/135	195/100	195/100



Simultaneously Switching Inputs

- For multi-input gates, if 2 or more data signals switch together the delay will differ from that calculated when switching is monotonic
  - Series FETs will have slower response
  - Parallel FETs will have faster response
- Problem: Timing models only store delay info to calculate AT from an input despite what other inputs do
  - If the model is built under simultaneously switching (ss) inputs, then it's over-pessimistic for monotonic (usual) cases
  - ..or it is optimistic if built without ss for the unusual case
- Similar problem exists when contention occurs between two signals – the actual delay is slower (max mode) than what STA predicts

Parallel Drivers

- Problem: STA path analysis not suitable for calculating correct timing when blocks are shorted
- Some timers have provision to address this, though not very well
  - Input AT differences ignored
  - Current capacity per driver may be estimated
- Be careful. Simulate such situations in spice.
  - It may be better to add user AT constraint at common outputs

Wire Delay Consideration

- **Dilemma: avoid timing escape because of a gracious wire resistance**
  - Min mode concern
  - At chip timing level, up the capture path resistance (or reduce launch path resistance) by some proportion
    - Select proportion based on min time path accuracy versus spice
    - Often achieved by 'scaling' the max mode arrivals to be later when checking hold times
  - Use biggest wire delays for clk and data in max mode

- **SSTA, done properly, would remove the need for special biasing of wires to cover corner cases**
  - Parameter variation (e.g., usually covers the 'lumped' fluctuations of metal width/thickness/spacing and dielectric) would be used to determine sensitivities against wiring artifacts (e.g., resistance) or, more likely, against the delay: $S = \Delta X / \Delta R$

    $S = \Delta X / \Delta D$

Parasitic Considerations

- **Parasitic data is annotated on top of the netlist**
  - Load and driver impedance are analyzed
  - timer's delay calculator reduces to simple PI model with Cnear, Cfar, Rwire
    - algorithms create entries for driver and wire delays in path reports
  - series inductive effects must be somehow accounted for
    - eventually wave reflections for mismatched impedances may be important
- **Noise effect on delay considered via coupling capacitances**
  - eventually inductive coupling will be important

Special Path Control

- **False path control**
  - Stop propagation that can't logically happen (but the timer finds a path anyhow)
    - Often used to hide asynchronous paths
- **Multicycle paths**
  - Paths that require multiple clock cycles to complete should be specified
- **Multifrequency paths**
  - Unless defined as asynchronous, all defined clock pairs are assumed to be synchronous
  - GCD determines closest launch/capture edge interval ω for setup
    - Defines pseudo base frequency
    - If pseudo base cycle phase shift exists between any clock pair:
      $\omega = \min \{\phi \,\%\, GCD(T_1,T_2)\ ,\ GCD(T_1,T_2) \,\%\, \phi\}$; modulo operation
    - Assumption for hold: capturing (master) edge comes as close to next cycle launching (master) edge without being later
- **Case analysis**
  - Used to analyze mutually exclusive modes of operation
- **Similar provisions in both STA and SSTA**

Skew Reduction

- Reduce skew penalty in min mode if launch and capture path have a common event (one edge) and common node closer to the sequential elements than is the master clock
  - Static timers can't distinguish that an event on a particular node can only occur once so it normally applies a skew corresponding with the entire clock path
  - Common Path Pessimism Removal (CPPR) reduces the skew by removing that uncertainty normally attributed to the common driver members of both launch and capture paths
- Min skew relief sometimes sought when skew can be controlled for certain circuits
  - distance-based skew is attempted to reduce skews when launch and capture clock grid origination points are significantly closer than the distance used for foundation of the skew
  - OR sometimes fast SPICE analysis of clock generation circuitry
- SSTA removes over-pessimism and alleviates the need to perform such alternative analysis (correlations)
  - Clk uncertainty (u) proportion of cycle time (T) generally increases as the critical dimension decreases

Call for Statistical Static Timing Analysis

- **Clock network complexity yielded STA skew**
  - Spec from clock team; avoid STA of clock network
  - The simpler the spec, the greater the guard-banding and more difficult to meet both min and max mode timing
    - Specs are now given per cluster or even per block
    - Further reductions provided by distance-based skew
    - But manufacturing physical deviations on chip and chip-to-chip present greater entities to consider, tending to larger relative skews
  - Race ratio scaling is used to assure races are won (hold margins) which further exacerbates over-pessimism
- **SSTA Goal: more directly apply fab data to timing**
  - Avoid allowances to cover corner cases when assessing and accounting components of skew (e.g, vdd deviation, temperature variation, etc.)
  - Reduce skew guard-banding and hold margins since spatially local circuits have high PVT correlation

Agenda

- SSTA
- Reporting
- Models
- Special considerations and limitations
- **Implications for manufacturing test**

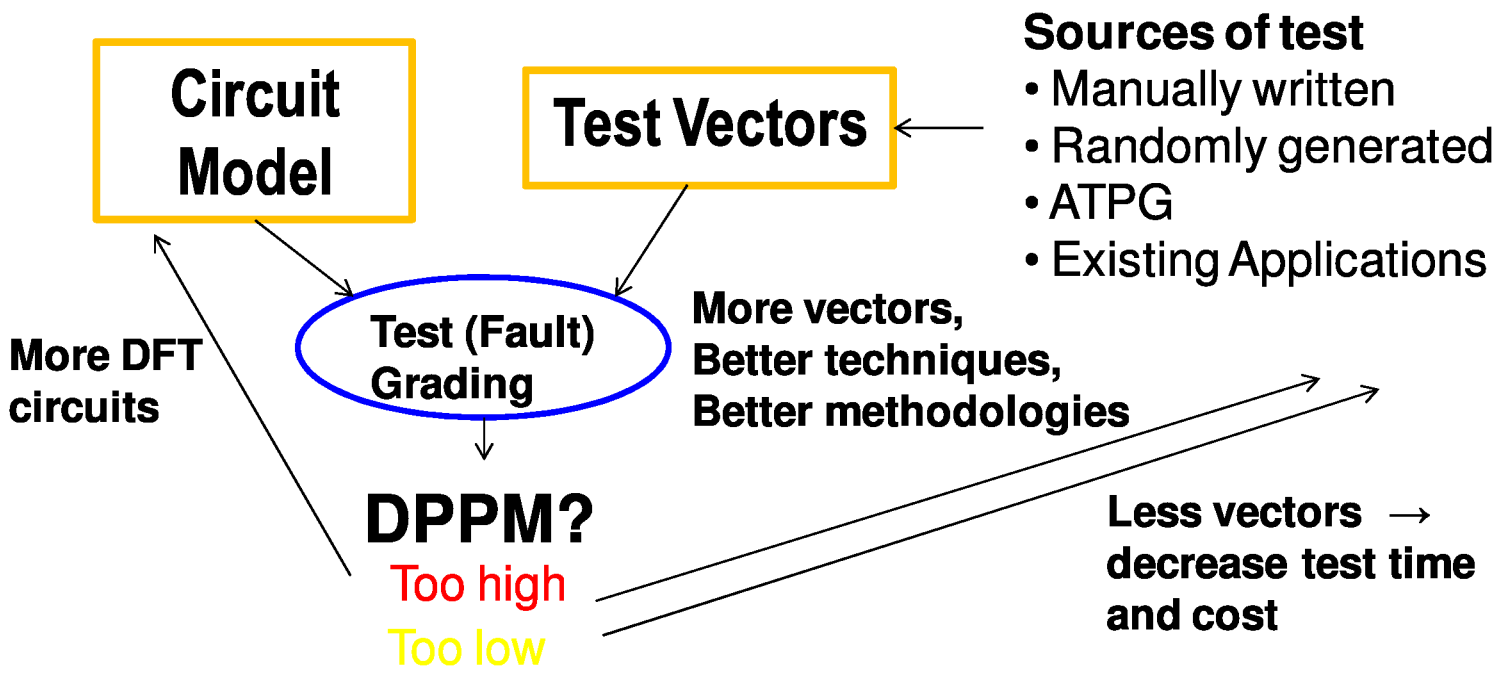
Testing Manufactured Chip for Delay Faults

Delay Test

	Silicon Debug	Manufacturing Test
Objective	Find speed-limiting Vectors	Check if DUT can run at a desired freq.
Where it is done	Microprocessor design	ASIC design/Micro. design
Need to hit Fmax	Necessary	Good correlation is ok
Effects of interest	Unmodelled effects, model/silicon mismatch (e.g., MIS, crosstalk, power droop)	Effects that differentiate each chips (e.g., process variation, delay defects)
Important step	Test vector generation	Path selection
Test type	Transition tests may work better than path tests	Path delay tests are appropriate

1

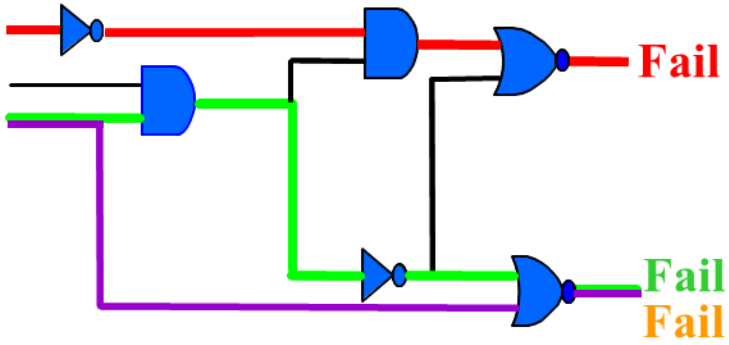
Delay Testing in Nanoscale Technologies

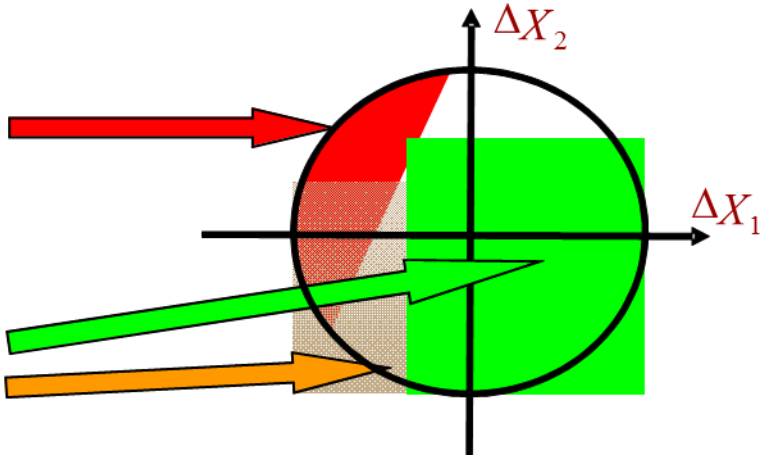


- For catastrophic faults, stuck-at and transition fault coverages are correlated well with DPPM
- For parametric delay faults, there is no such test metric
- SSTA can directly measure DPPM due to parametric delay faults

Critical Path Not Unique in Presence of Variations

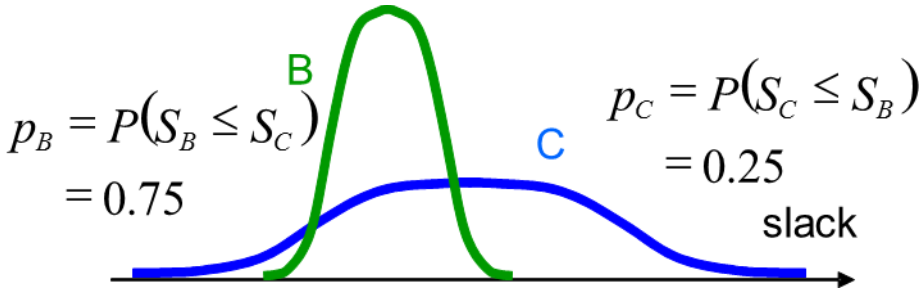
- **Different paths will become frequency-limited at different points in the process subspace**





- **"Criticality Probability": among all manufactured chips, the probability of the path (edge/node) being critical**
  - **Natural extension of the useful critical path concept**
  - **First proposed to the EDA community by Visweswariah (DAC 2004)**

$$p_i = P\left(S_i \le \min_{j \ne i}(S_j)\right)$$

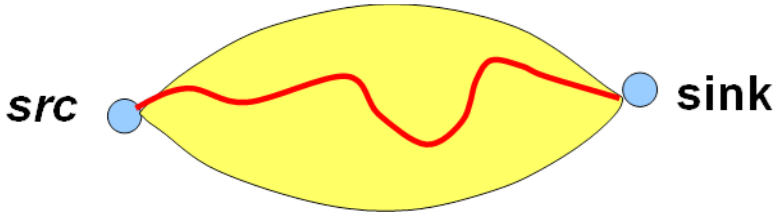


Criticality Computation and Application to Testing

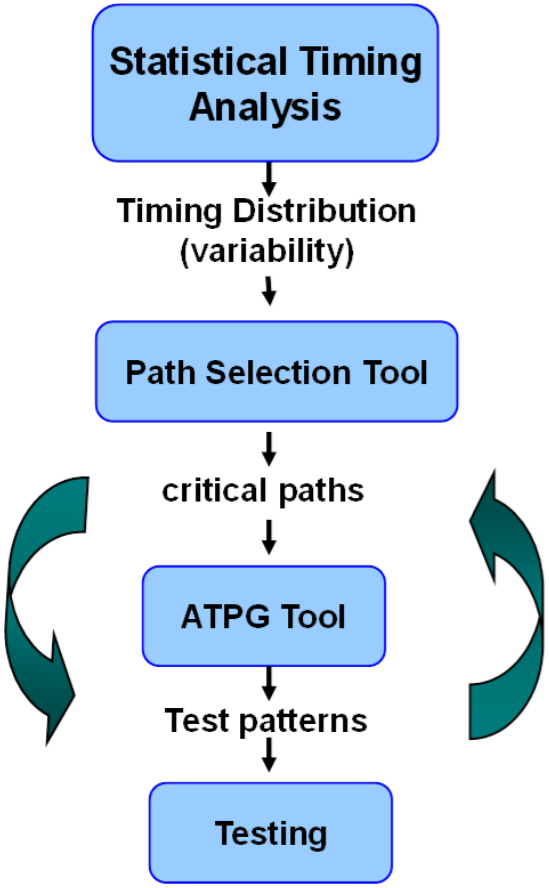
- **Definitions**
  - Chip slack: min. slack of all paths in design
  - Complement path slack: min. slack of all paths except the given path of interest

  Chip slack: $S{\downarrow}chip = min(S{\downarrow}path, S{\downarrow}complement)$

- **For a given path, its criticality can be computed in constant time (Xiong, DATE 2008)**



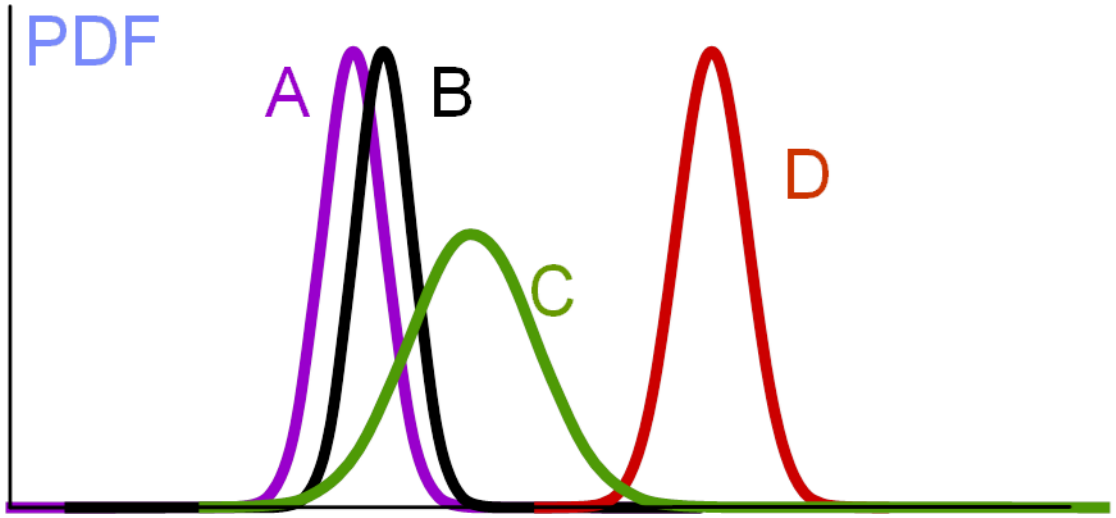
$$P{\downarrow}path = P(S{\downarrow}path \leq S{\downarrow}complement)$$



Two-step approach

Critical Path Selection

Important for Optimization, Testing, ….



- **Assuming A and B perfectly correlated, but A is untestable**
- **Which top 2 paths to choose?**
  - **Ignoring testability of paths**
    - **Deterministic timing: A and B**
    - **Statistical timing: A and C**
  - **Considering testability**
    - **Statistical timing: B and C**

GLOSSARY

Glossary

arc – a path between pins or nodes of a timing graph that illustrates a signal can pass arrival time and slew from the input pin/node to the output (considering polarity); represents delay/slew of logic blocks or wires between pins of logic blocks

AT – arrival time; the time a signal arrives at a node

check – a test between two signals to guarantee event order; setup checks test that the signal arrives before the reference does, hold checks test that the signal arrives after the reference

clipping – an arrival time adjustment by an amount that would barely allow the failing upstream check to pass in order to judge the arrival times downstream

clock – signal which defines synchronous behavior; requirement that defines cycle time and required arrival times at timing elements

clock domain – see "phases"

clock gate – non-transparent timing element which imposes a half cycle path; setup tested against asserting clock edge, hold against de-asserting edge

clock tracing – operation of the static timer to identify all block pins that are clocks

controlled arc – a timing arc that will not propagate data unless an enable signal is in the proper state

Glossary (cont.)

cycle accounting – adjusting the cumulative arrival time to maintain events within the cycle modulus

cycle adjusts – cycle modulus changes to the required arrival time in order to relate the correct data to the clock event that it must be checked against

domino gate – transparent timing element

data – a signal that is not a clock; a signal that changes once per cycle unless it is a dynamic signal (e.g., domino) which is affected by both edges of the clock

early mode – timing mode to check that shortest paths meet proper registration/don't arrive too early; use earliest arriving data and compare to latest required arrival time

flip flop – edge triggered (non-transparent) timing element

hold time – timing check to assure new data didn't change until after reference closed; sometimes refers to the margin associated with particular timing elements to assure the slack calculation takes all necessary circuit issues into consideration

latch – transparent timing element

Glossary (cont.)

late mode – mode to check that longest paths meet timing goals; use latest arriving data and compare to earliest required arrival time

near-domino gate – transparent timing element

phases – clock and data signal attribute which indicates the relationship of the signal compare to the master clock (signal from which its timing context is derived); sometimes referred to as signal's clock domain; phases are used in order to provide the static timer the ability to perform cycle adjusts

RAT – required arrival time; defines the acceptable boundaries of a signal's min and max arrival time and considers the reference events, circuit mechanics, and system tolerances

setup time – timing check to assure data arrives before reference closes; sometimes refers to the margin associated with particular timing elements to assure the slack calculation takes all necessary circuit issues into consideration

skew – difference in same clock event arrival time at different pins throughout the design; usually not analyzed by static timer

slack – residual time of the difference between when an event actually occurs versus when it is required to occur; a negative value means that the order of the two events was wrong.

SSTA – Statistical static timing analysis

Glossary (cont.)

static timing – exhaustive method of measuring a design's timing robustness by building a timing graph of the design, providing signal arrival times, propagating these and identifying critical paths

timing elements – logical context defining arcs and checks among three points in a timing graph; examples: latch, flip flop, clock gate

timing graph – a collection of arcs and checks which represents the timing behavior of a logic design

transparent – data propagation through a controlled arc when the controlling signal is enabled

uncertainty – safety margin included in the slack calculation to account for clock skew and other variables affecting clock edge events