Tackling Test and Diagnosis Challenges Using GPU-Based High-Throughput Timing Simulation

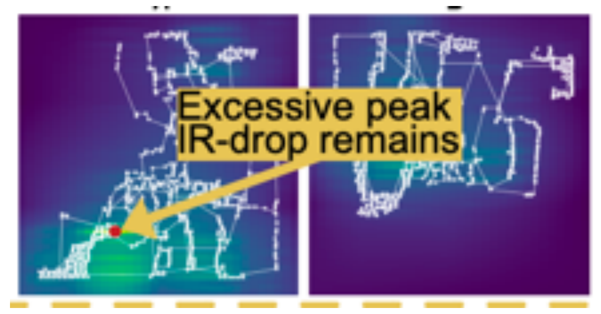
Stefan Holst

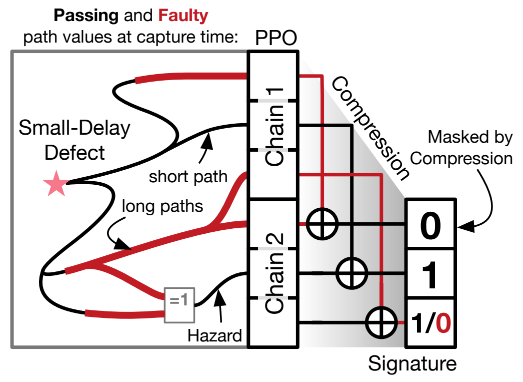
国立大学法人
九州工業大学

Motivation

Many Test/Diagnosis Tasks Require Timing Simulations

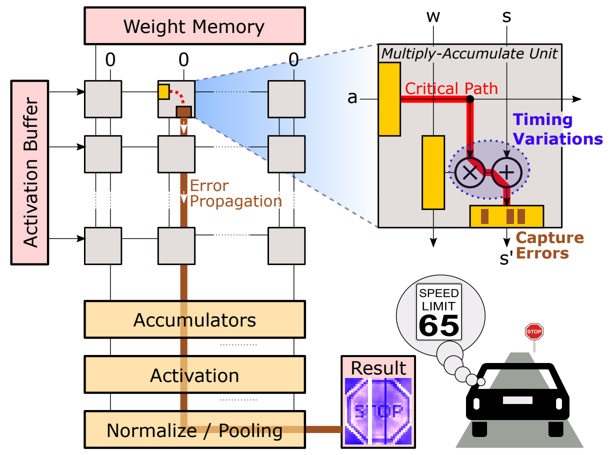
IR-Drop During Scan Test



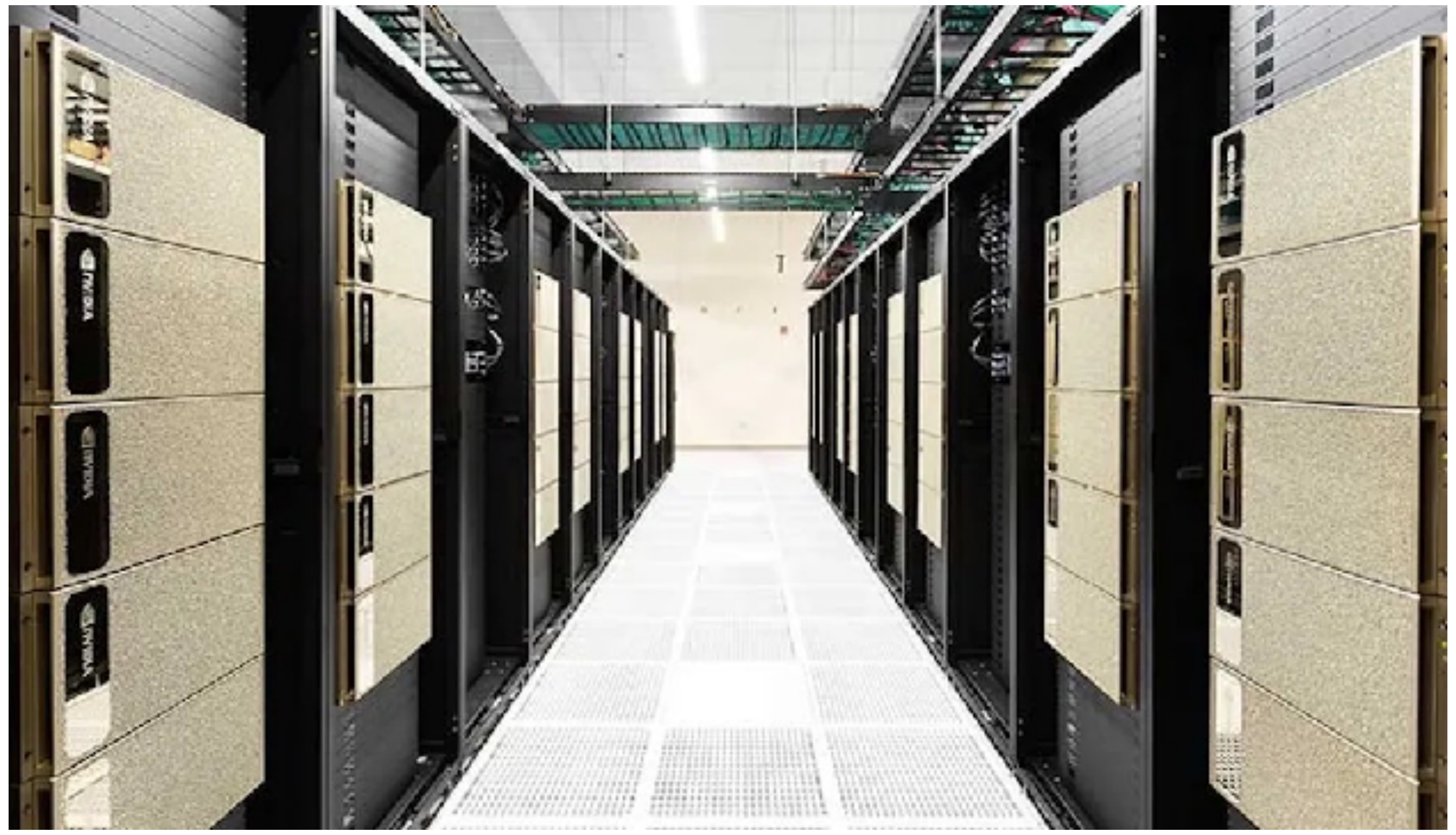


Small Delay Defect Diagnosis

Resilience Characterizations



ML Pushes New Data-Parallel Compute Architectures



[Nvidia]

Data-Parallel Architectures For High-Performance Timing Simulation

Agenda

GPU-Accelerated Timing Simulation

Scan-Test Power Analysis

Small Delay Fault Simulation and Diagnosis

AI Accelerator Resilience Analysis

Agenda

GPU-Accelerated Timing Simulation

Scan-Test Power Analysis

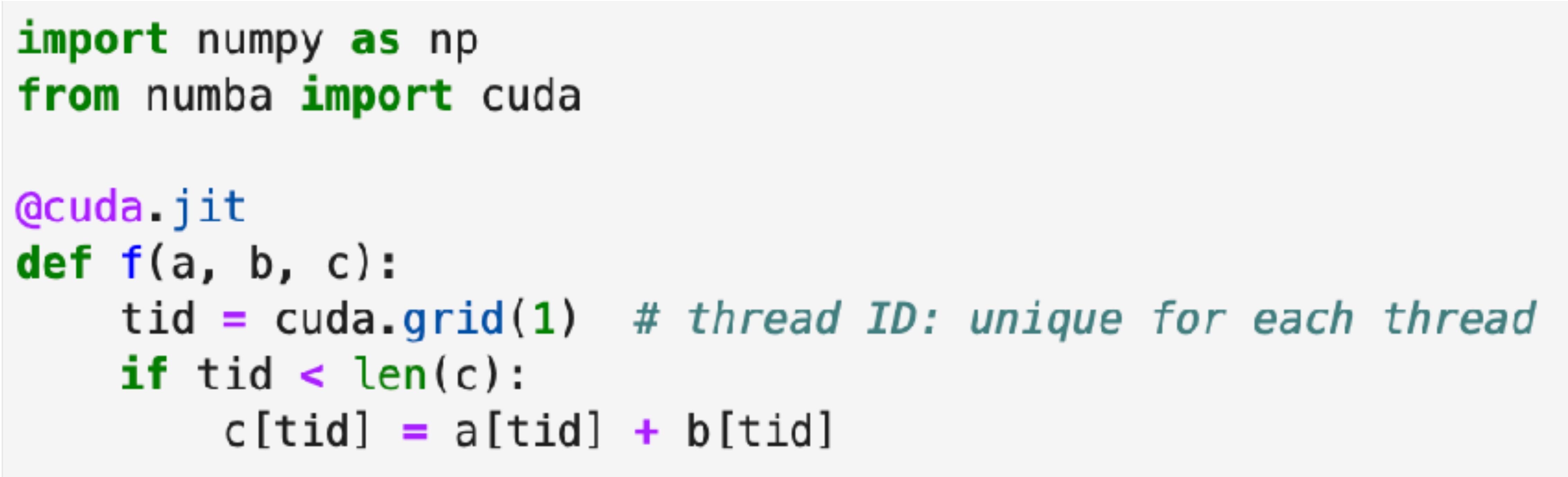
Small Delay Fault Simulation and Diagnosis

AI Accelerator Resilience Analysis

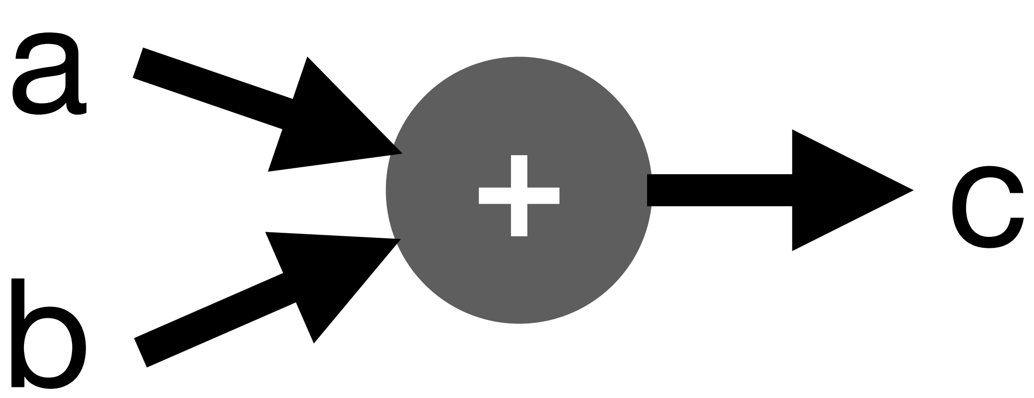
GPU Programming Principle

Single Instruction, Multiple Data (SIMD)

- **Kernel:** Short program that runs on GPU
- Addition on GPU:

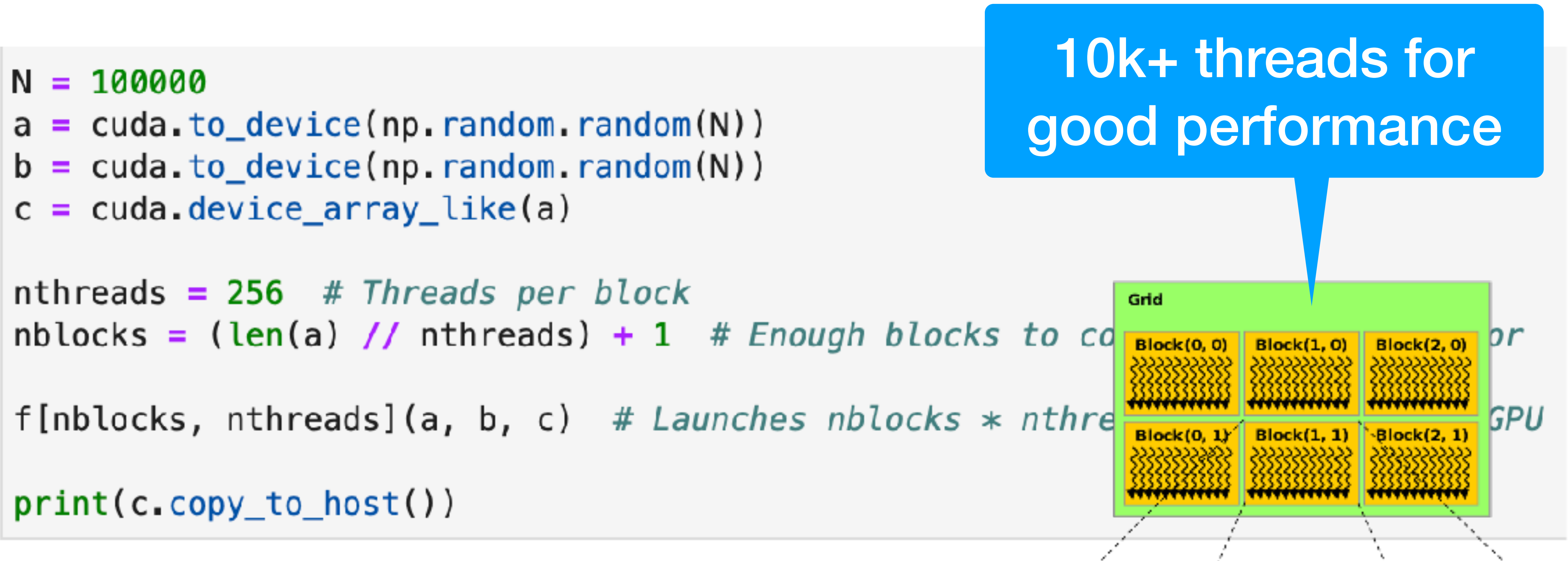


```
import numpy as np
from numba import cuda

@cuda.jit
def f(a, b, c):
    tid = cuda.grid(1)  # thread ID: unique for each thread
    if tid < len(c):
        c[tid] = a[tid] + b[tid]
```

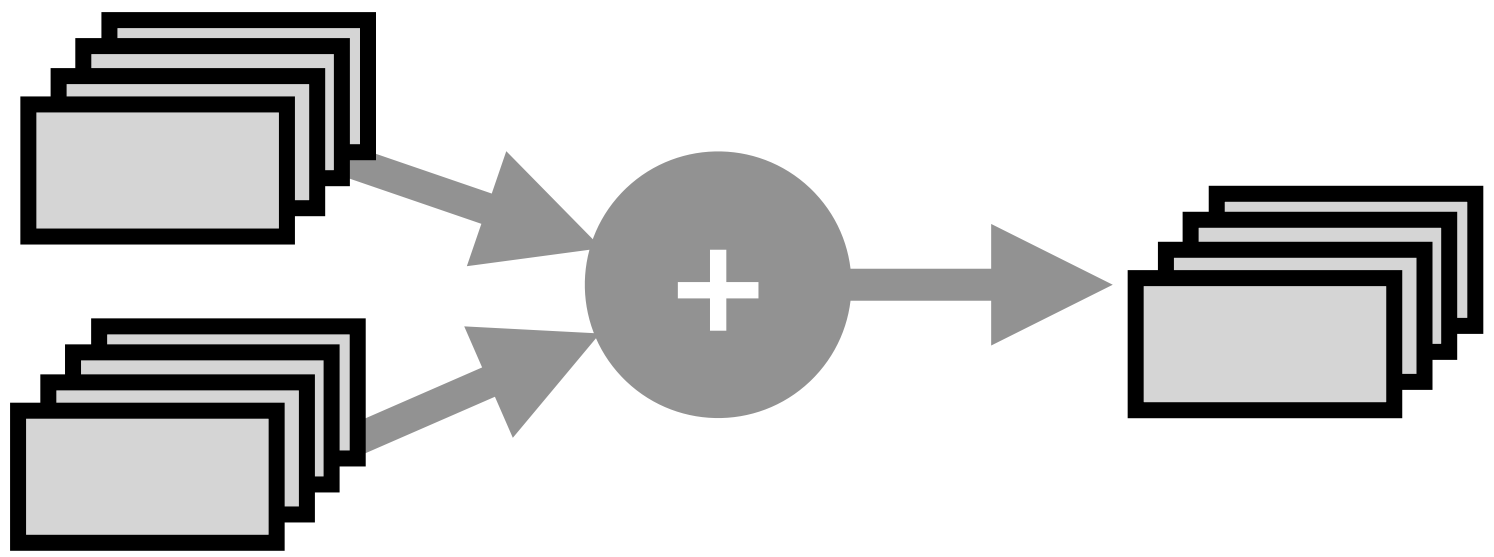


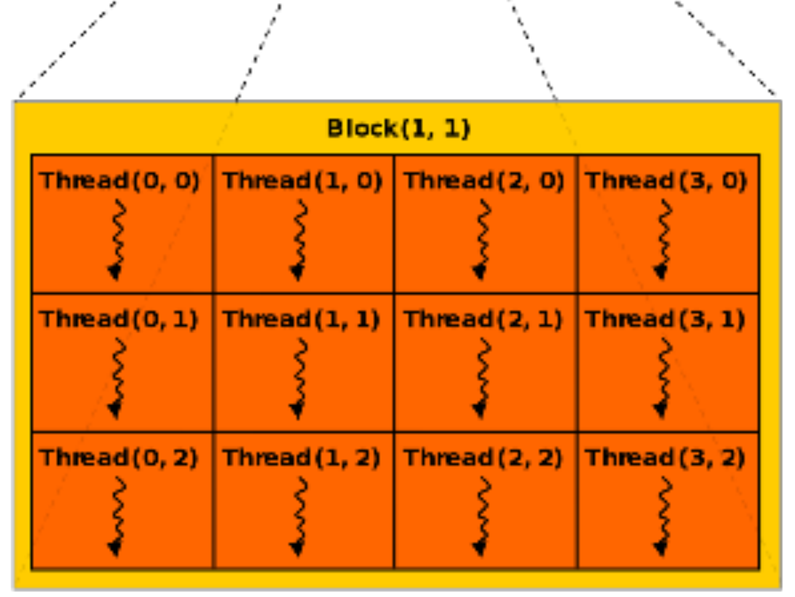
- **Thread:** Kernel running with unique ID
  Each thread can (should!) access different data
- Vector-Addition:

10k+ threads for good performance



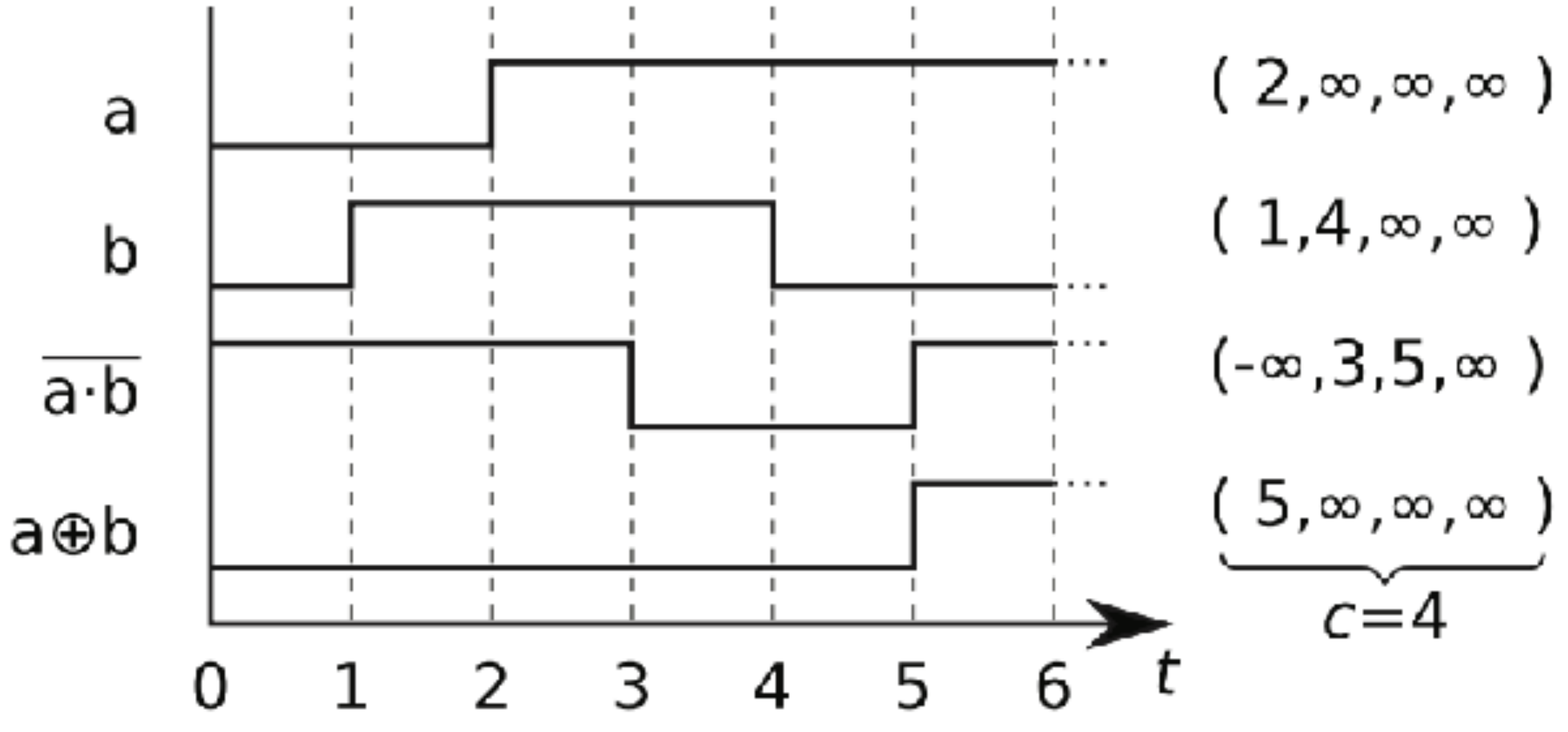
```
N = 100000
a = cuda.to_device(np.random.random(N))
b = cuda.to_device(np.random.random(N))
c = cuda.device_array_like(a)

nthreads = 256  # Threads per block
nblocks = (len(a) // nthreads) + 1  # Enough blocks to co or

f[nblocks, nthreads](a, b, c)  # Launches nblocks * nthre GPU

print(c.copy_to_host())
```

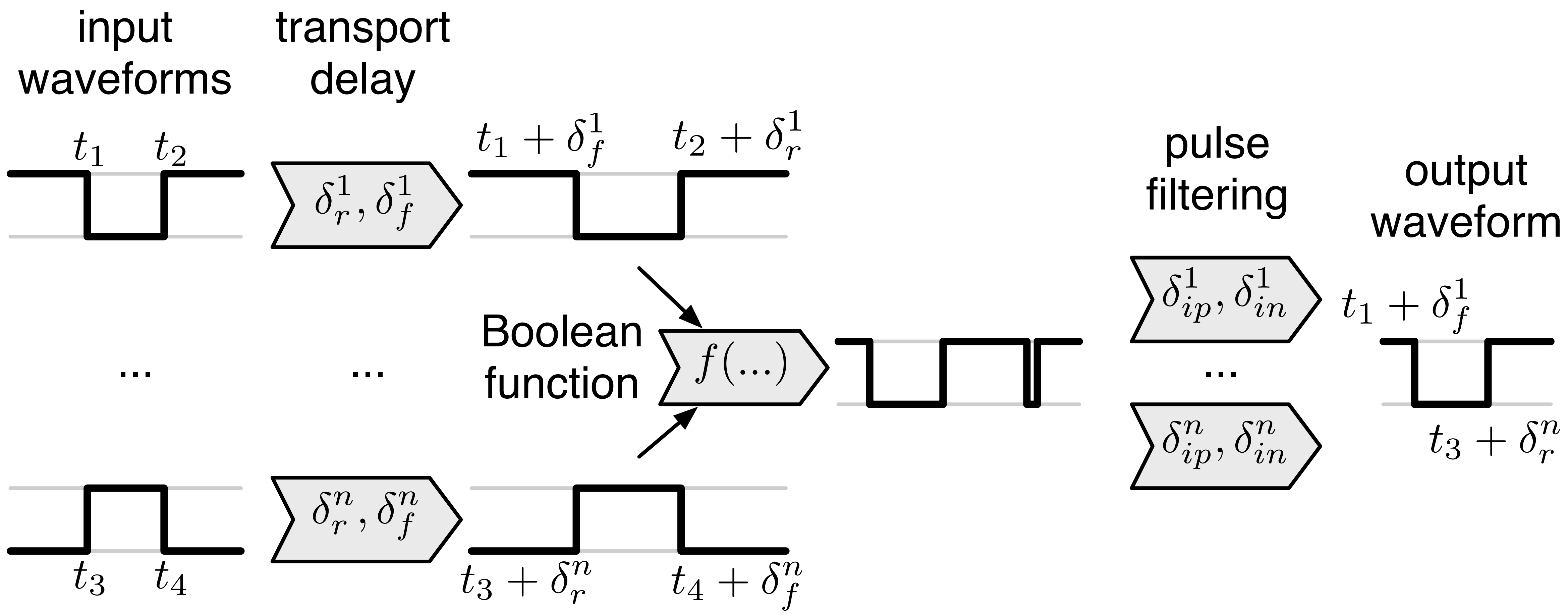




Kernel: Compute Output Waveform of One Cell

- **Waveform**: All transitions on a signal within one clock cycle
- Compute complete output waveform from complete input waveforms

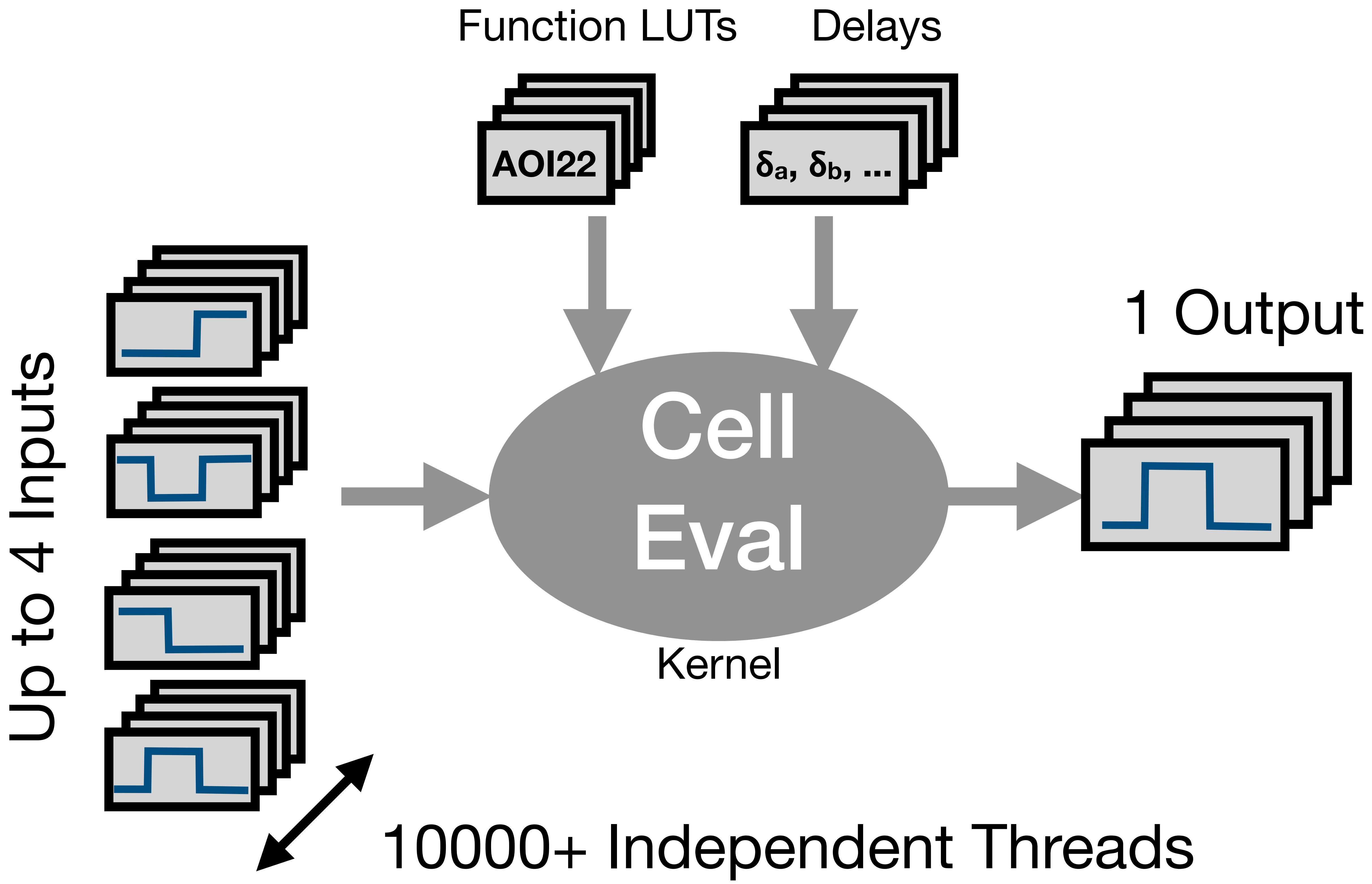




[Holst et al.: "High-Throughput Logic Timing Simulation on GPGPUs" ToDAES Vol. 20, No. 3, Article 37, June 2015]

Threads: Data-Parallel Cell Evaluations

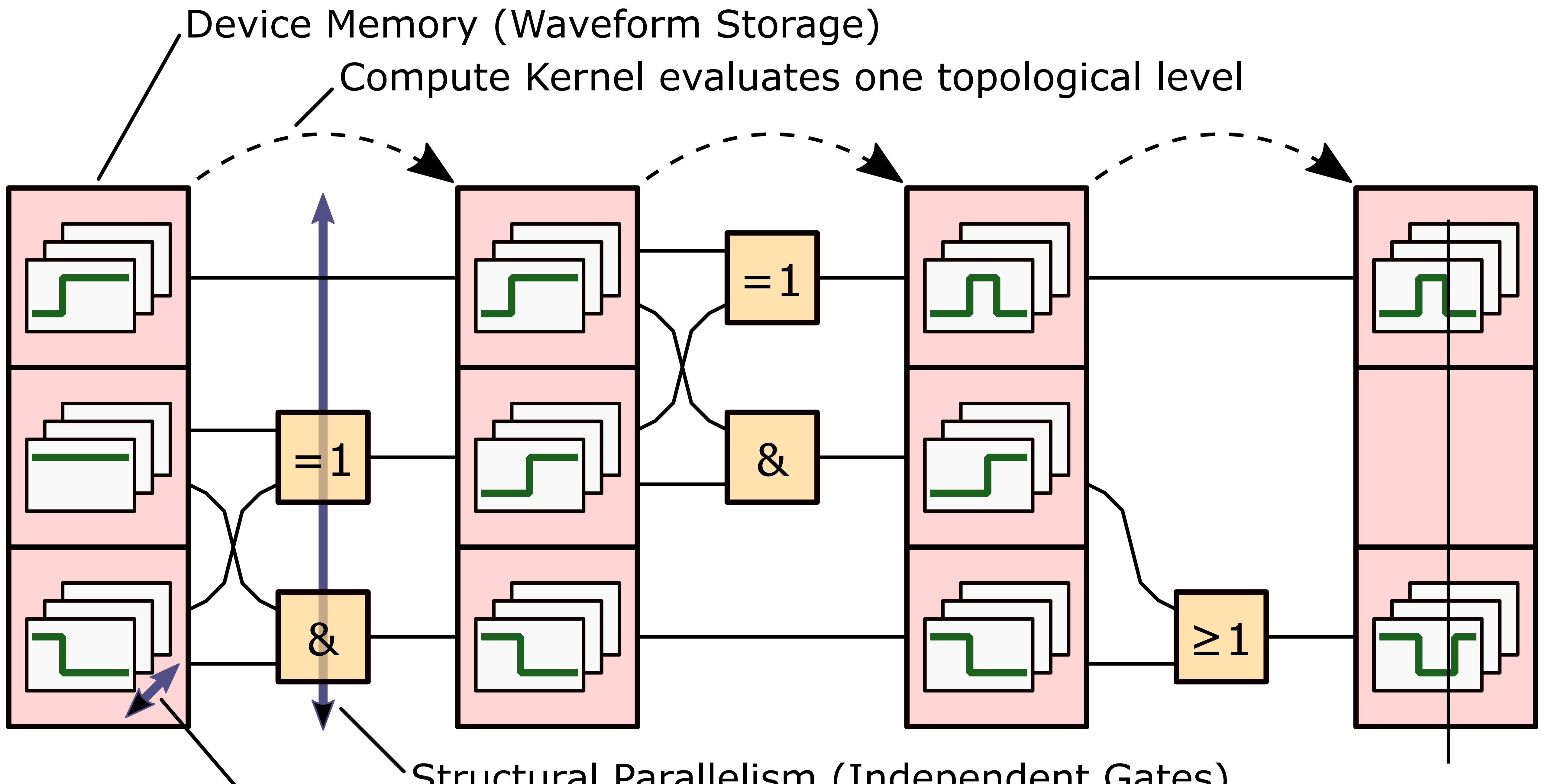
- Same Kernel Code, but Distinct ...
  - ... Input Waveforms
  - ... Cell Function (LUT)
  - ... Pin-to-Pin Delays
- One Kernel Launch for 10000+ Independent Cell Evaluations



[Holst et al.: "High-Throughput Logic Timing Simulation on GPGPUs" ToDAES Vol. 20, No. 3, Article 37, June 2015]

Gate-Level Timing Simulation on GPU

- Toposort the Gate-Level Combinational Logic
- One Kernel Launch per Level
- Maximizes Data-Parallelism
    - Parallel Evaluation of Independent Gates
    - Concurrent Sim of Many Independent Inputs


Device Memory (Waveform Storage)
Compute Kernel evaluates one topological level
=1
&
=1
&
≥1
Structural Parallelism (Independent Gates)
Data Parallelism (Independent Simulations)
capture

[Holst et al.: "High-Throughput Logic Timing Simulation on GPGPUs" ToDAES Vol. 20, No. 3, Article 37, June 2015]

Agenda

GPU-Accelerated Timing Simulation

Scan-Test Power Analysis

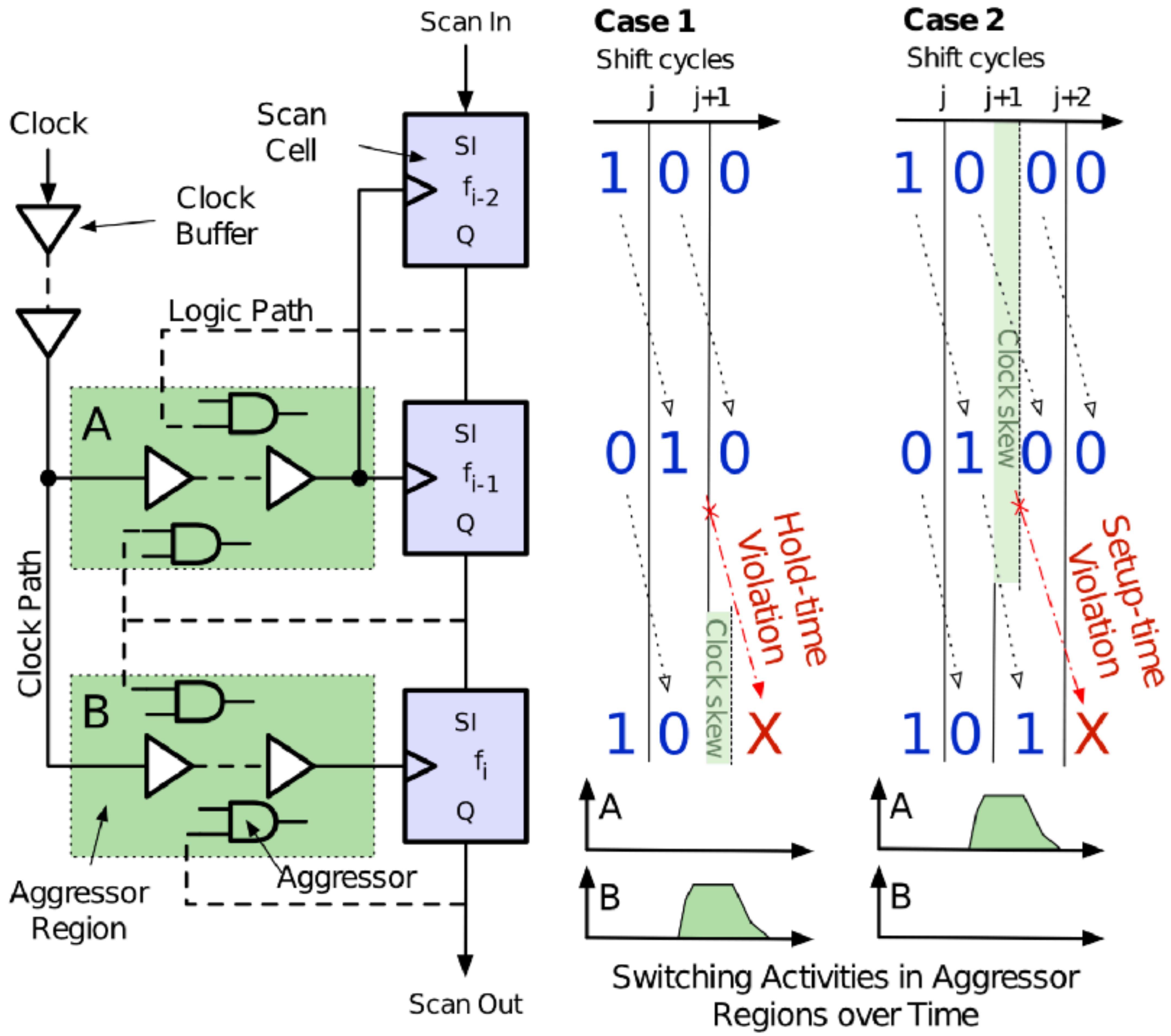
Small Delay Fault Simulation and Diagnosis

AI Accelerator Resilience Analysis

Scan Shift Clock Skew Problem

- Excessive IR-Drop During Shifting can Corrupt Test Data

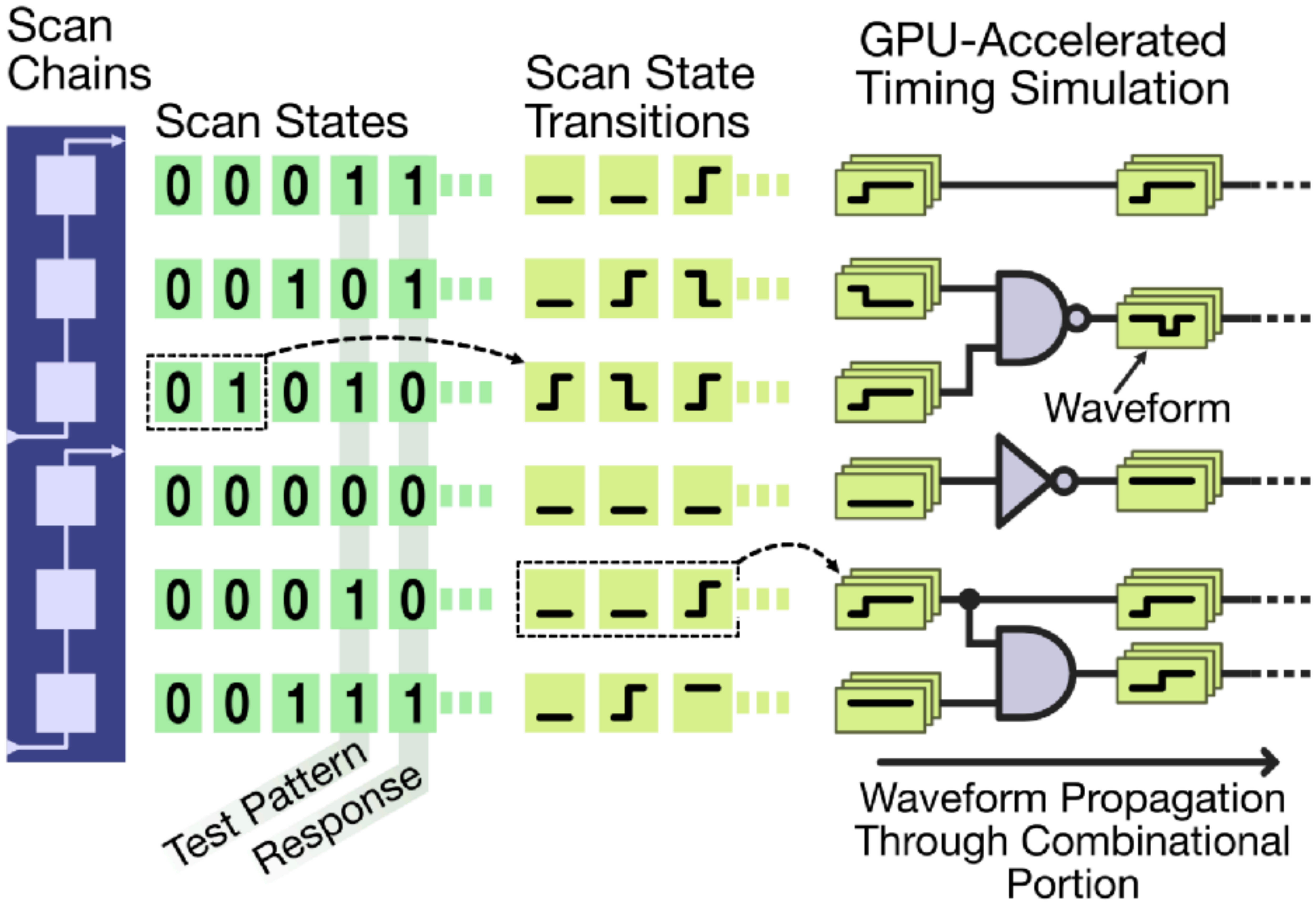
- Dynamic Power Simulation for Every Shift Cycle?



[Holst, Schneider, Kawagoe, Kochte, Miyase, Wunderlich, Kajihara, Wen: *Analysis and Mitigation of IR-Drop Induced Scan Shift-Errors* ITC 2017]

Shift Switching Activity Simulation

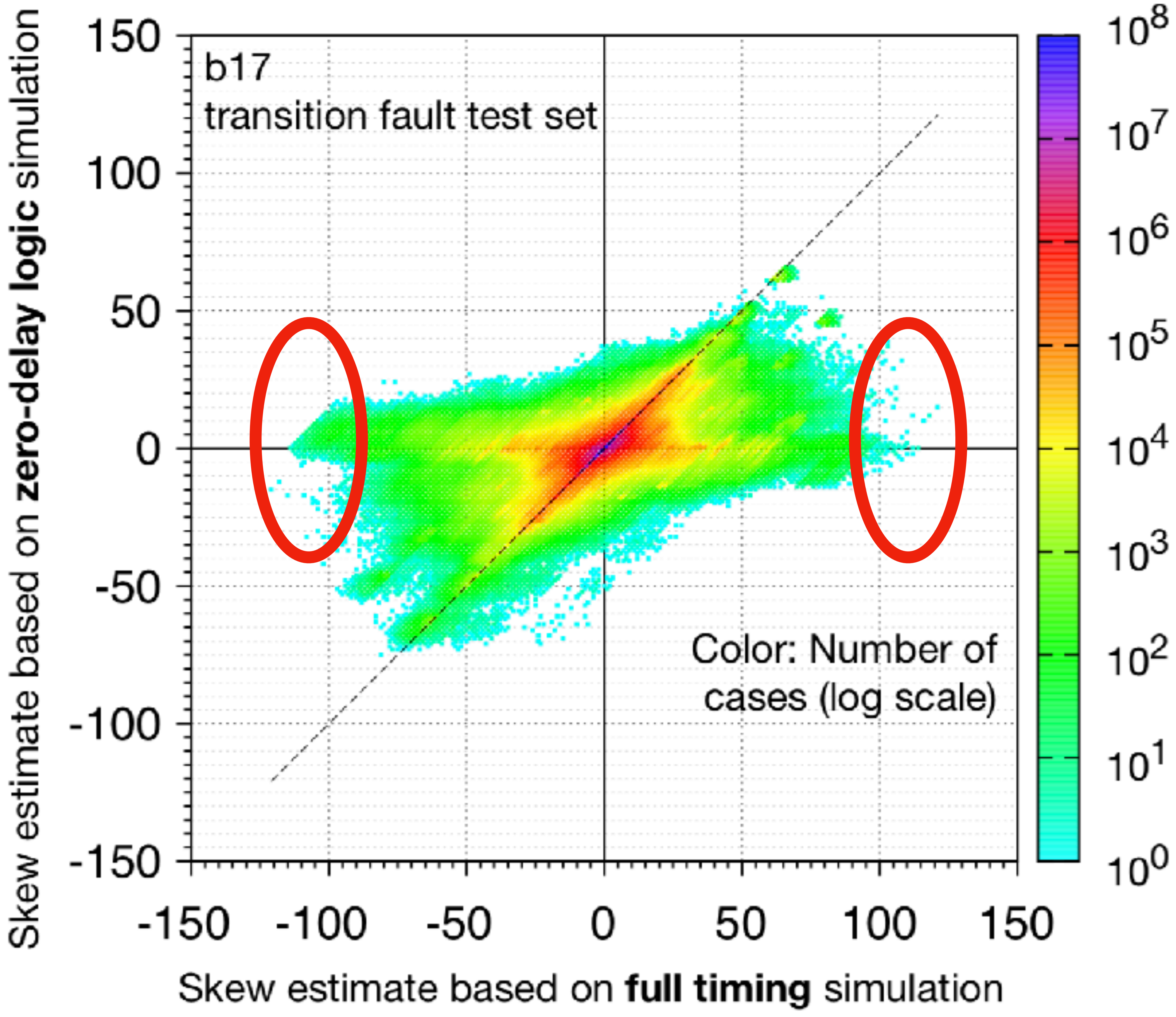
- Perfect Data-Parallel Workload
- All Scan-States are Known in Advance
- Simulate all PPI Transitions Data-Parallel



[Holst, Schneider, Kawagoe, Kochte, Miyase, Wunderlich, Kajihara, Wen: *Analysis and Mitigation of IR-Drop Induced Scan Shift-Errors* ITC 2017]

Skew Estimates: Zero-Delay vs. Full Timing

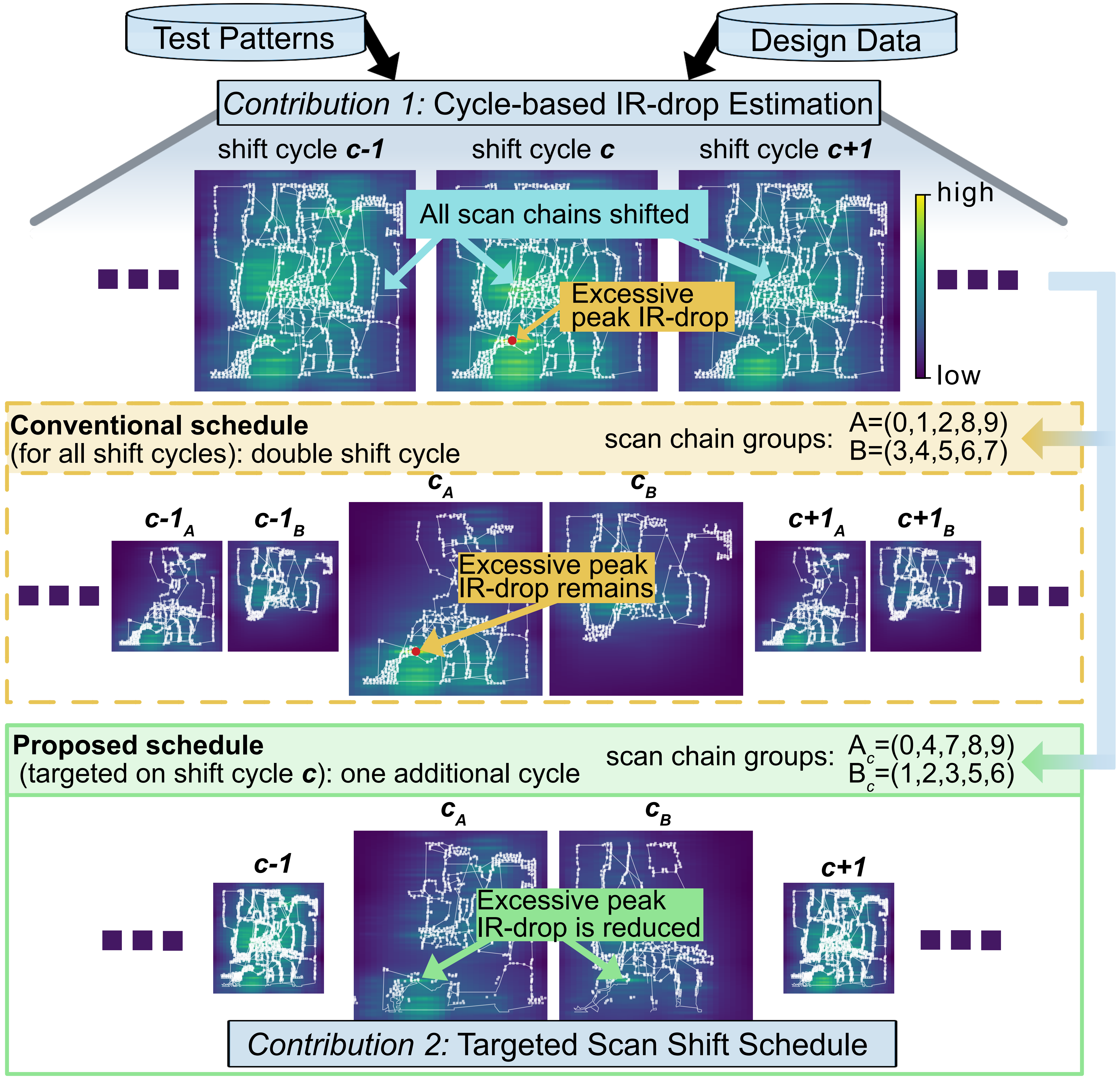
- Glitches have large impact in some shift cycles
  - It takes only one timing violation to corrupt the test
  - Need to simulate all shifts to find risky ones



[Holst, Schneider, Kawagoe, Kochte, Miyase, Wunderlich, Kajihara, Wen: *Analysis and Mitigation of IR-Drop Induced Scan Shift-Errors* ITC 2017]

Partial Shifting for IR-Drop Mitigation

- Estimate IR-Drop Map for Every Shift Cycle
- Identify Risky Cycles with IR-Drop Hotspots
- Assign Chains into Shift-Groups to Balance Out Power Demand
  - Additional Simulations to Find the Sweet-Spot



[S. Holst, S. Shi, and X. Wen, "Targeted Partial-Shift for Mitigating Shift Switching Activity Hot-Spots During Scan Test," PRDC 2019]

Agenda

GPU-Accelerated Timing Simulation

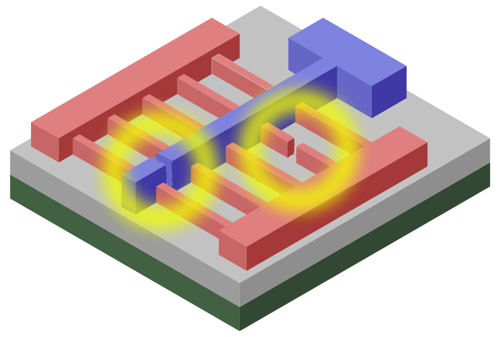
Scan-Test Power Analysis

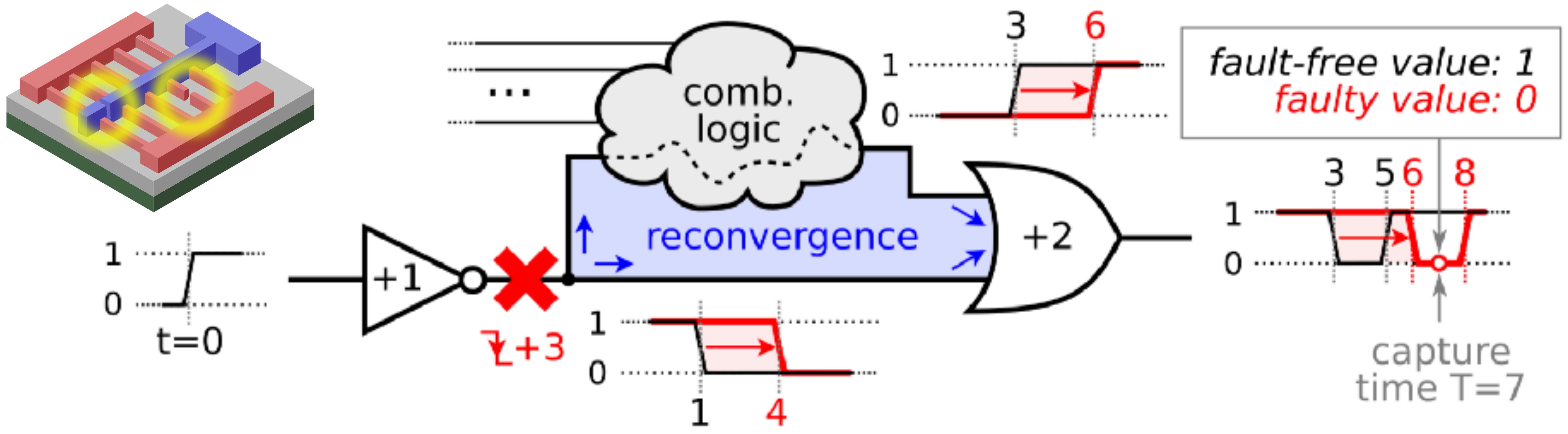
Small Delay Fault Simulation and Diagnosis

AI Accelerator Resilience Analysis

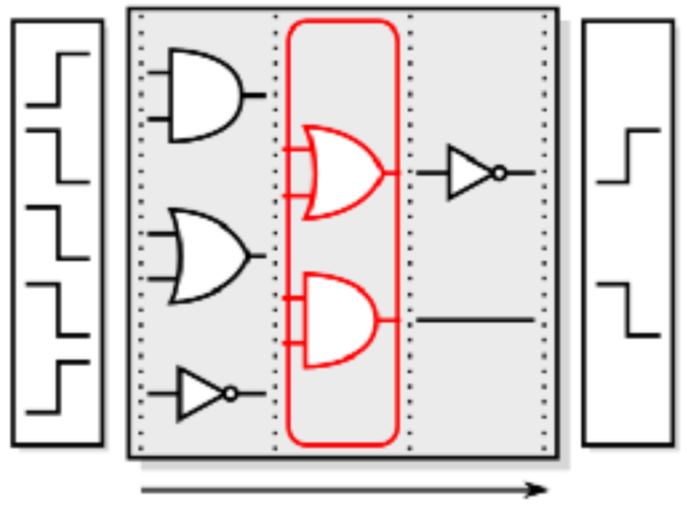
Small-Delay Fault Simulation

- Fin-FET: More Small-Delay Faults
- Complex Timing Behavior
- 4 Dimensions of Parallelism:

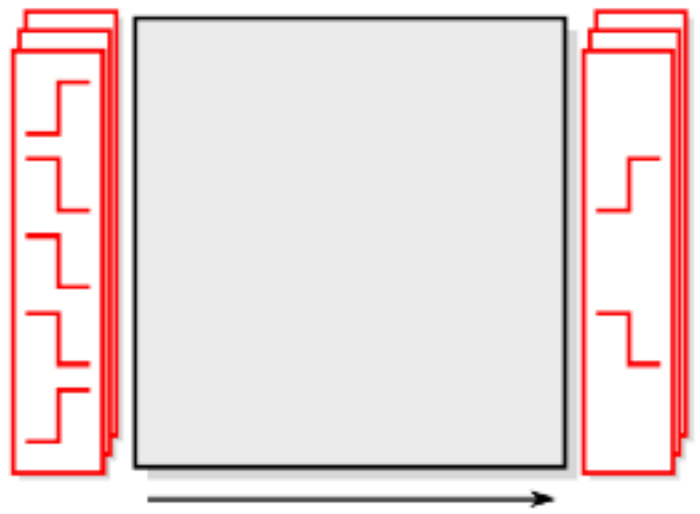




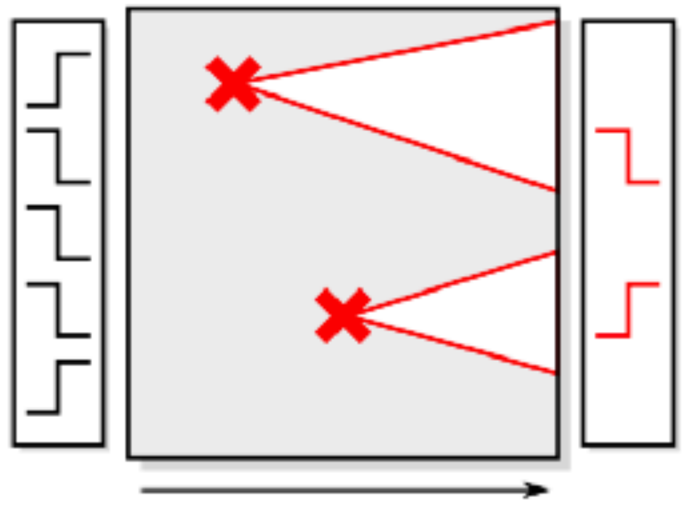
Structural Parallelism



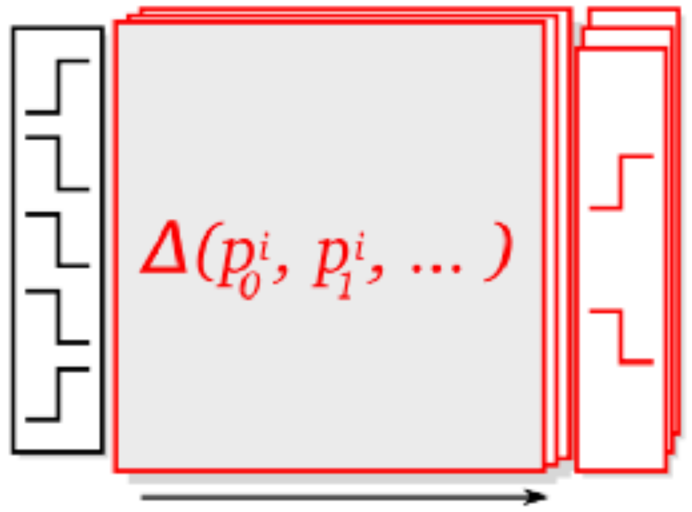
Stimuli Parallelism



Fault Parallelism



Variation Instance Parallelism

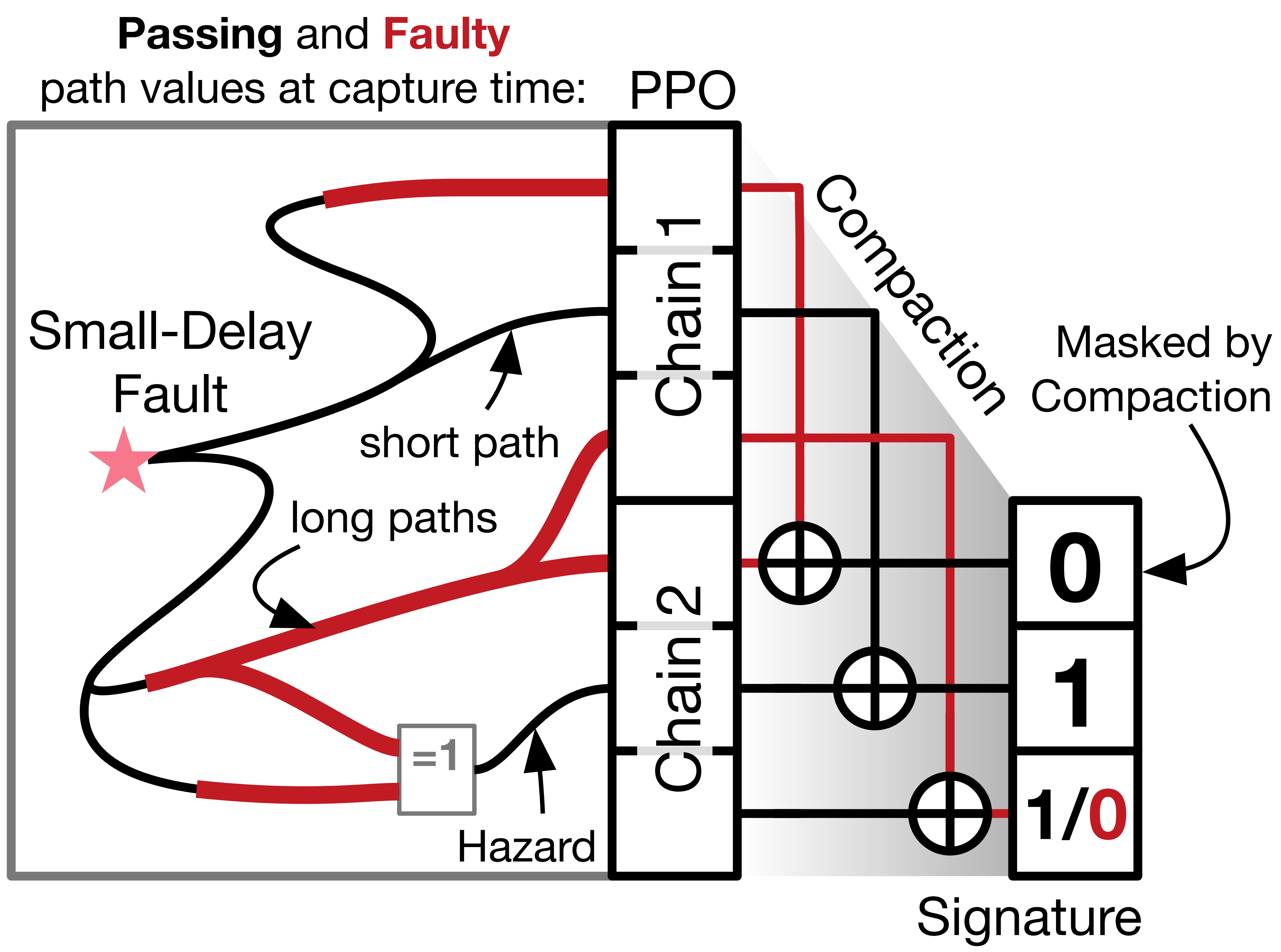


[Schneider, Holst, Kochte, Wen, Wunderlich: *GPU-Accelerated Small Delay Fault Simulation* DATE 2015]

[Schneider, Kochte, Holst, Wen, Wunderlich: *GPU-Accelerated Simulation of Small Delay Faults* TCAD 2017]

Small-Delay Fault Diagnosis for Yield Learning

- **Challenges:**
  - Complex Behaviour of SDFs
  - Test Response Compression
  - Process Variation

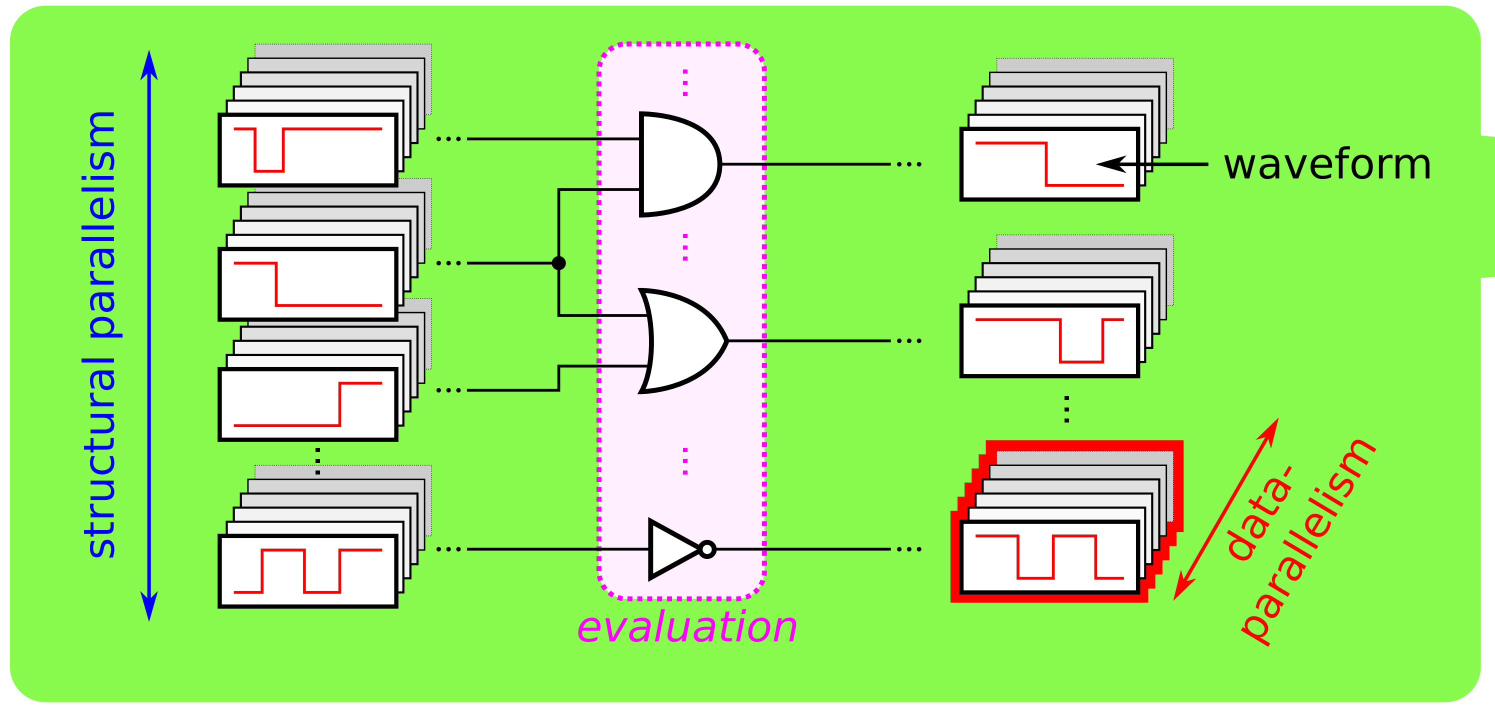


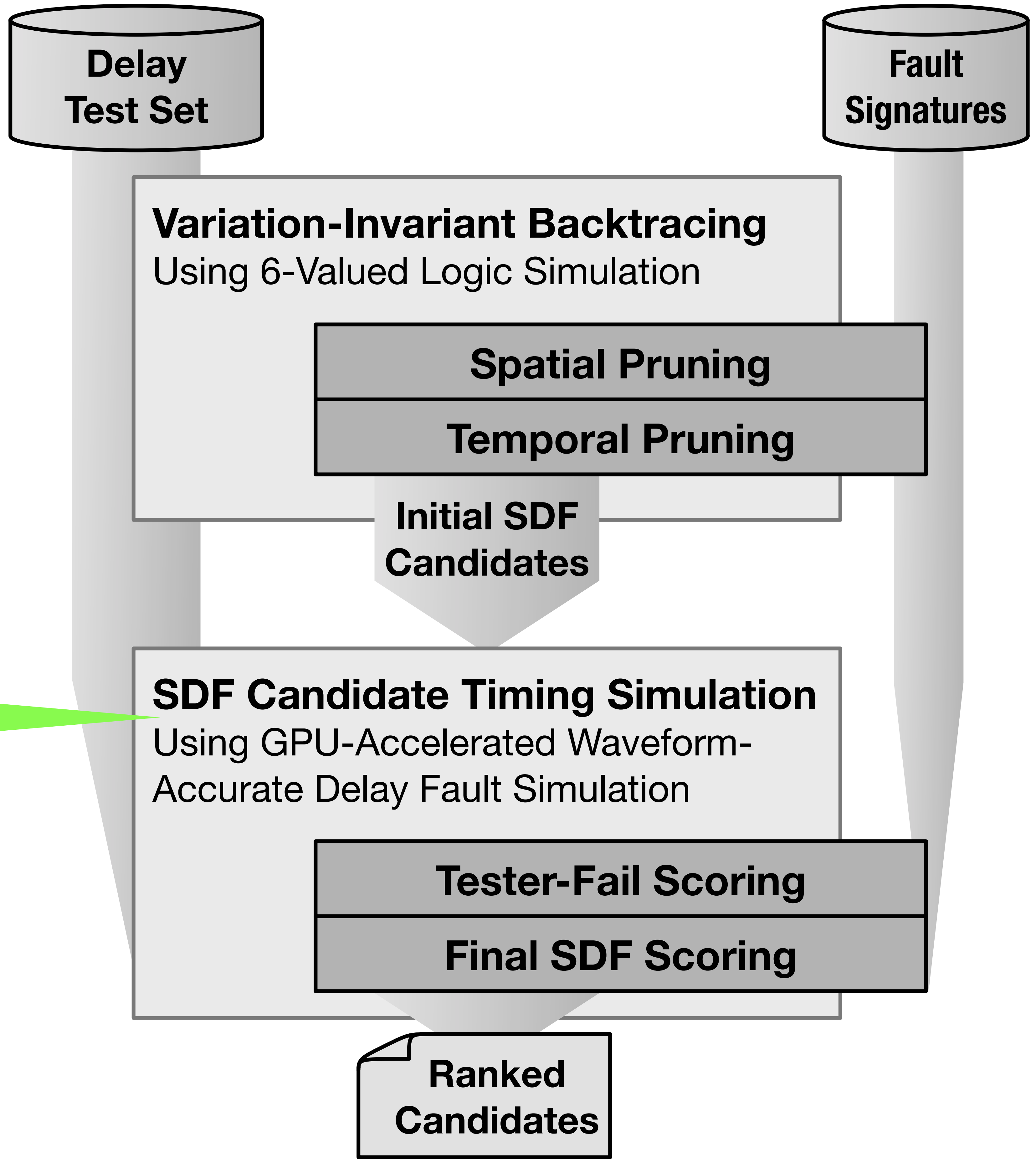
[Holst, Schneider, Kochte, Wen, Wunderlich: *Variation-Aware Small Delay Fault Diagnosis on Compressed Test Responses* ITC 2019]

Small-Delay Fault Diagnosis Flow

- Scoring SDF Candidates for all Patterns:

  Many Data-Parallel Timing Simulations





Delay Test Set

Fault Signatures

Variation-Invariant Backtracing
Using 6-Valued Logic Simulation

Spatial Pruning

Temporal Pruning

Initial SDF Candidates

SDF Candidate Timing Simulation
Using GPU-Accelerated Waveform-Accurate Delay Fault Simulation

Tester-Fail Scoring

Final SDF Scoring

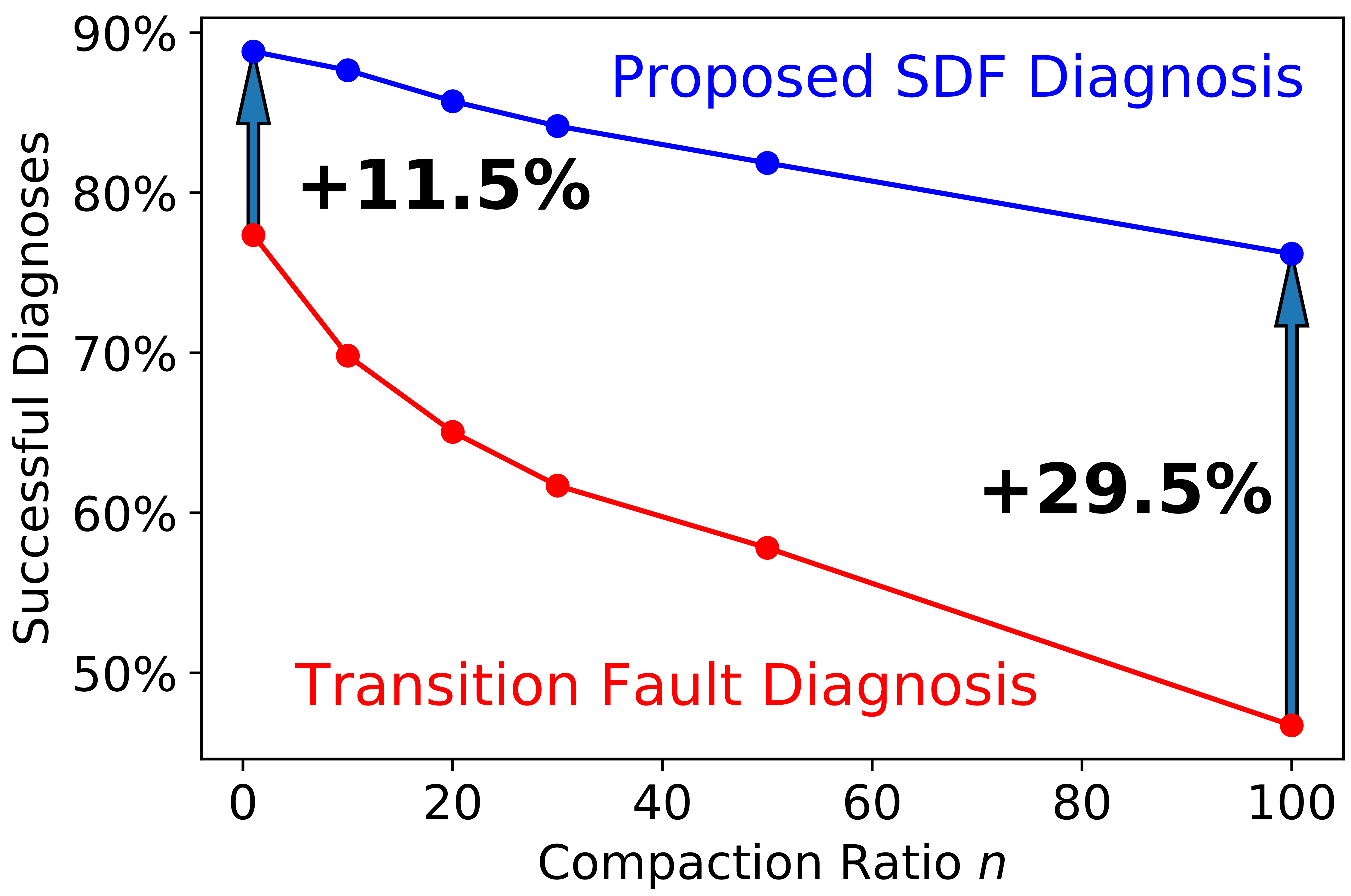
Ranked Candidates

[Holst, Schneider, Kochte, Wen, Wunderlich: *Variation-Aware Small Delay Fault Diagnosis on Compressed Test Responses* ITC 2019]

Impact of Compaction on Diagnosis

- Response Compaction has Huge Impact on Traditional Diagnosis

- GPU Provides the Necessary Simulation Performance to Restore Diagnostic Resolution



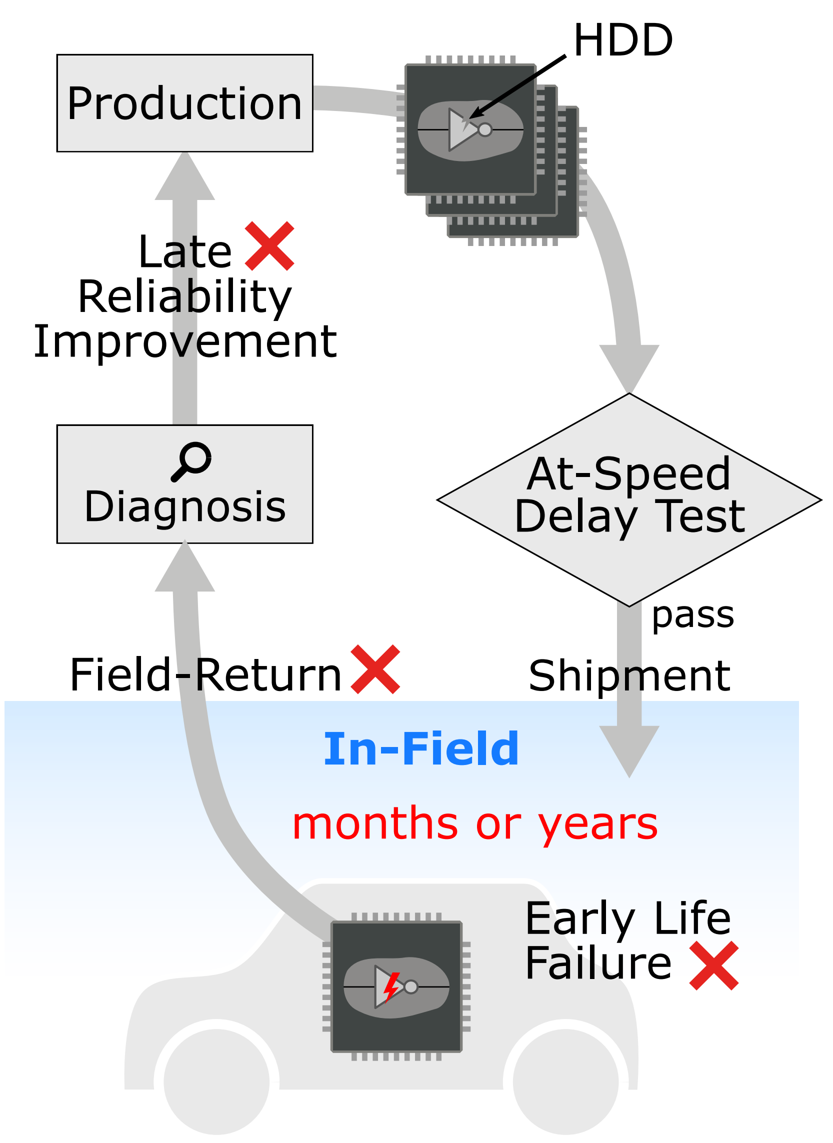
[Holst, Schneider, Kochte, Wen, Wunderlich: *Variation-Aware Small Delay Fault Diagnosis on Compressed Test Responses* ITC 2019]

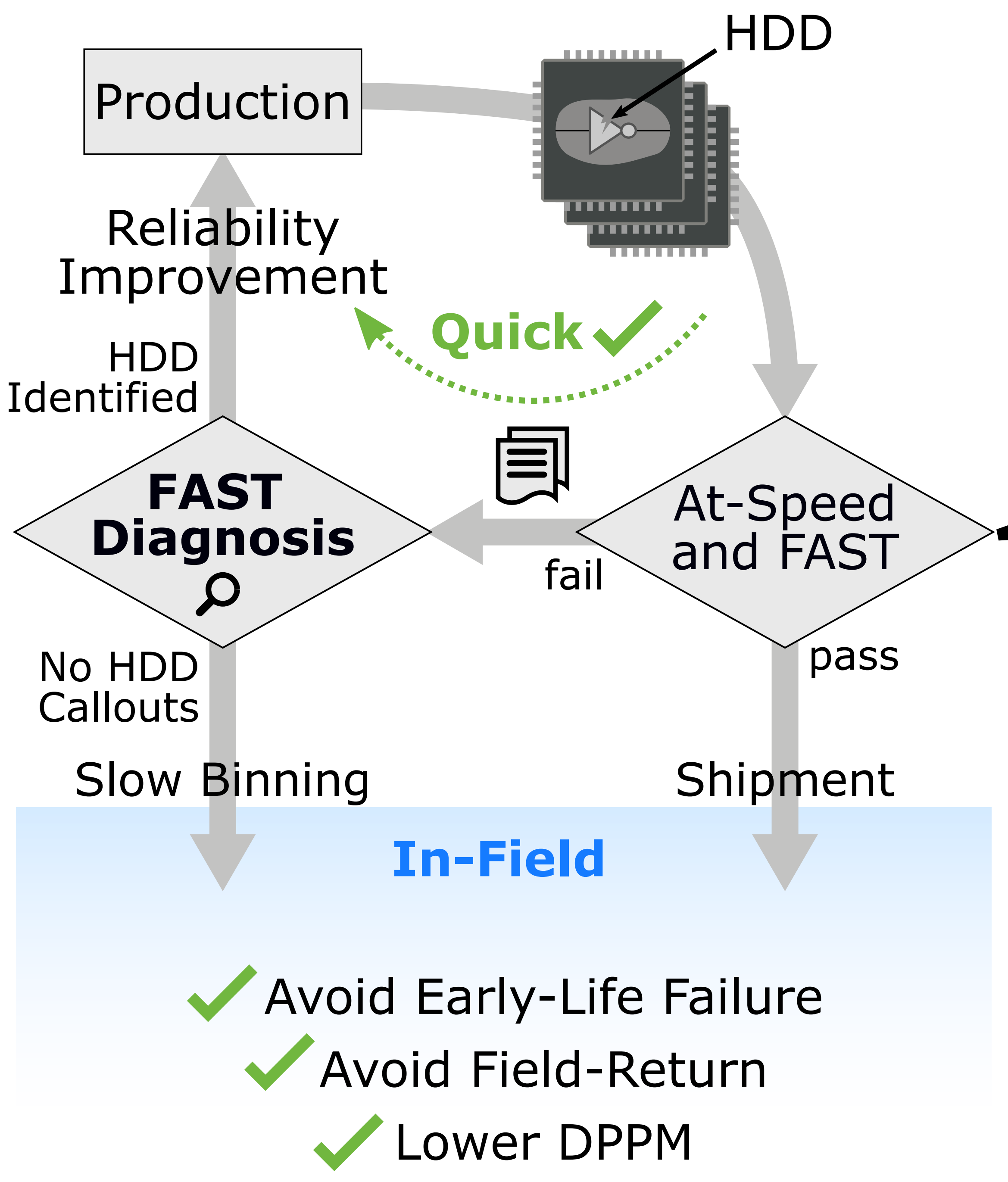
TF Diagnosis: [Holst, Wunderlich: *A Diagnosis Algorithm for Extreme Space Compaction*, DATE 2009]

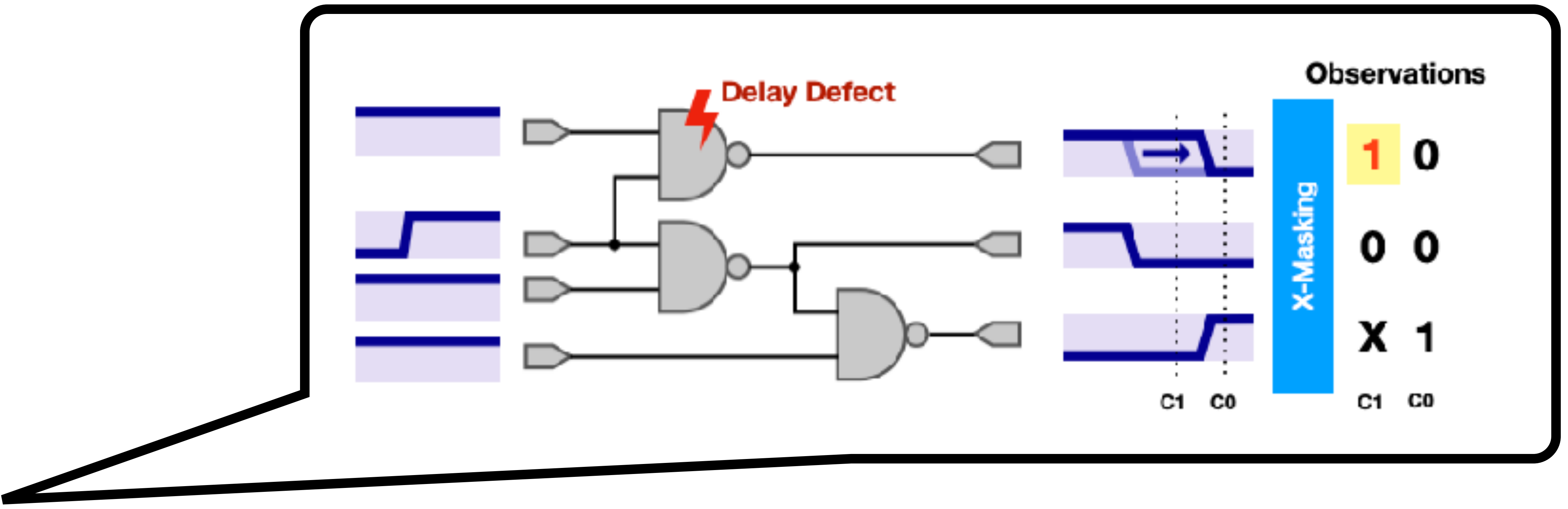
Diagnosis For Reliability Improvement

Diagnose Faster-Than-At-Speed Test Signatures

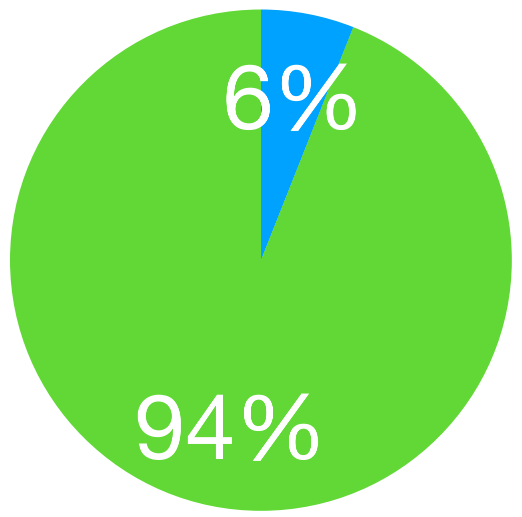
At-Speed Test and Conventional Diagnosis:







Diagnose Hidden Delay Defects (HDDs)



In 89-98% of all cases the real Hidden Delay Defect is included in the final ranking

[S. Holst, M. Kampmann, A. Sprenger, J. D. Reimer, S. Hellebrand, H.-J. Wunderlich, and X. Wen, "Logic Fault Diagnosis of Hidden Delay Defects," ITC 2020]

Agenda

GPU-Accelerated Timing Simulation

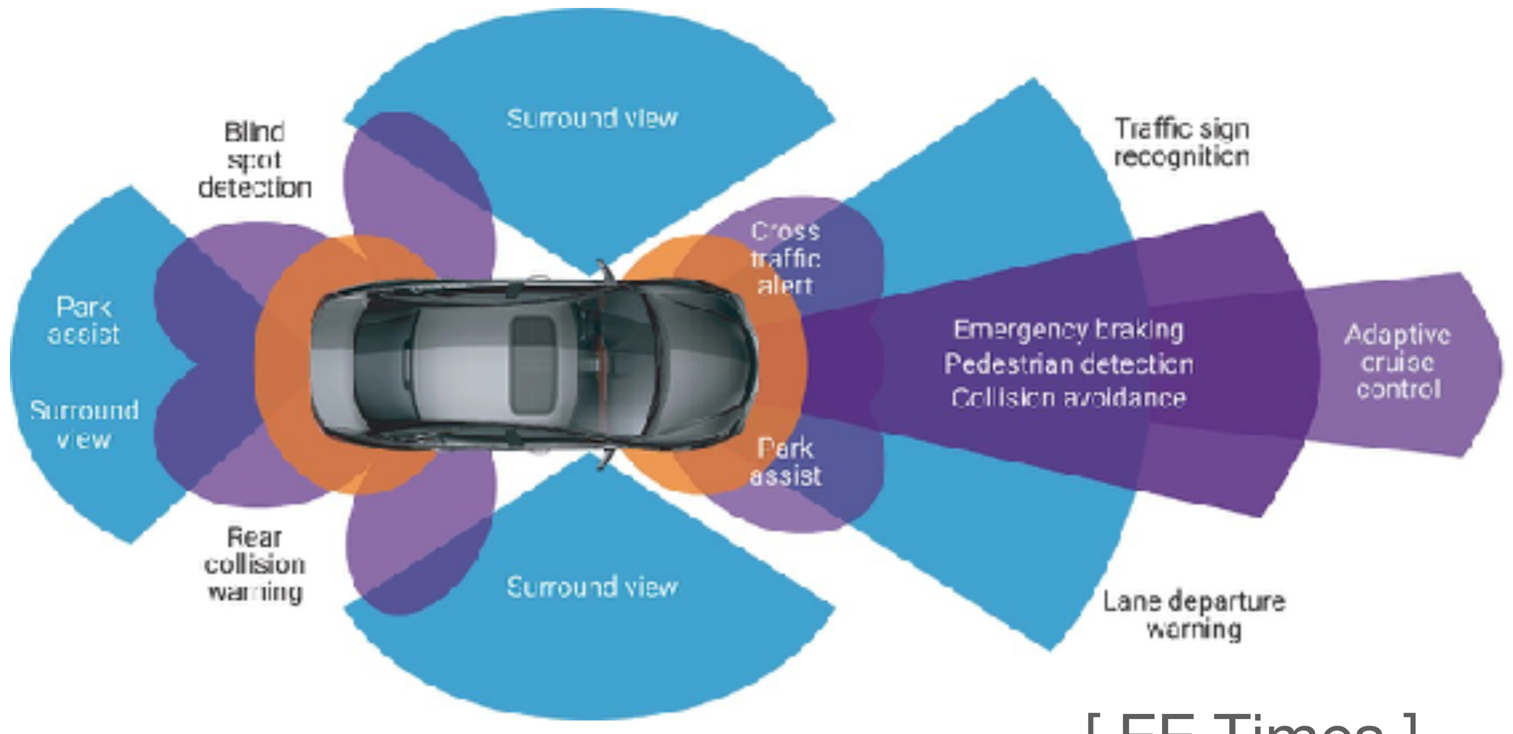
Scan-Test Power Analysis

Small Delay Fault Simulation and Diagnosis

AI Accelerator Resilience Analysis

Cutting-Edge VLSI Meets Safety-Critical Systems

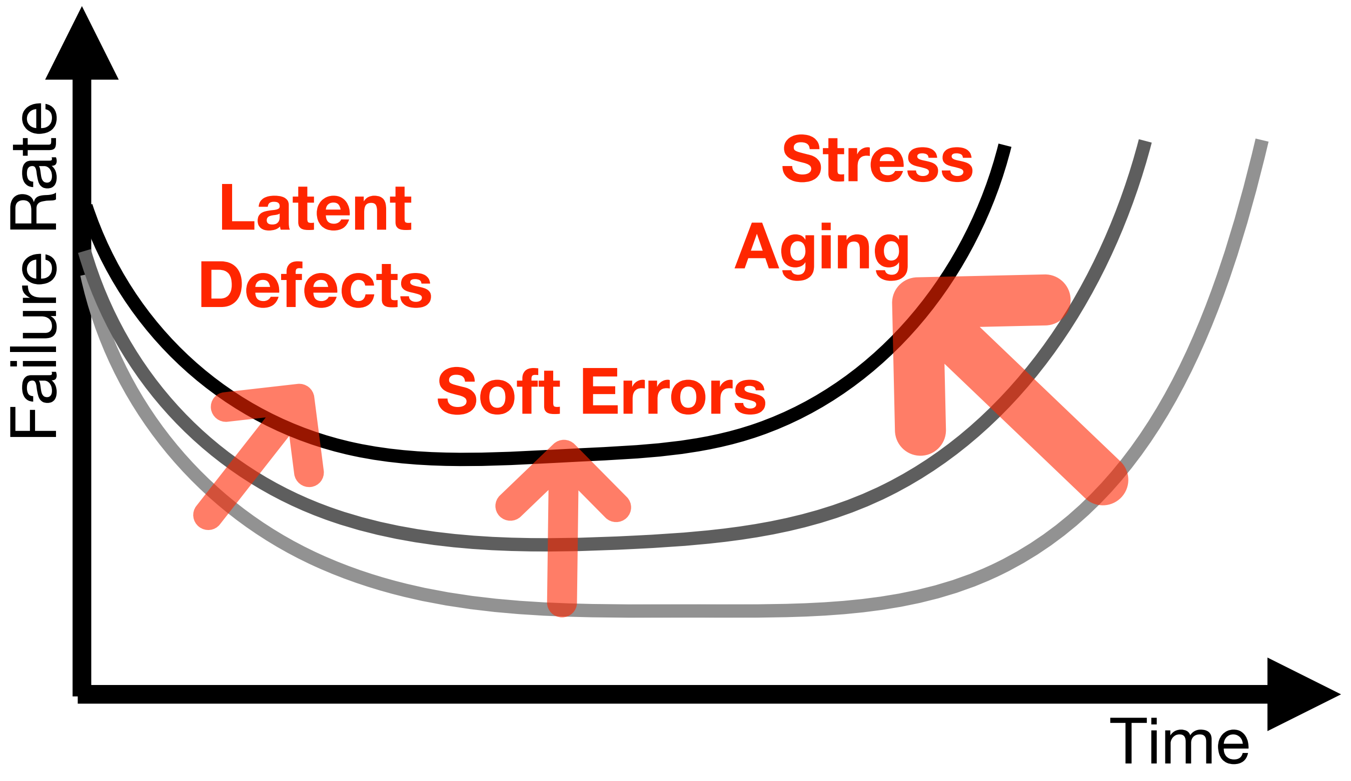
Compute Performance Needs ↗



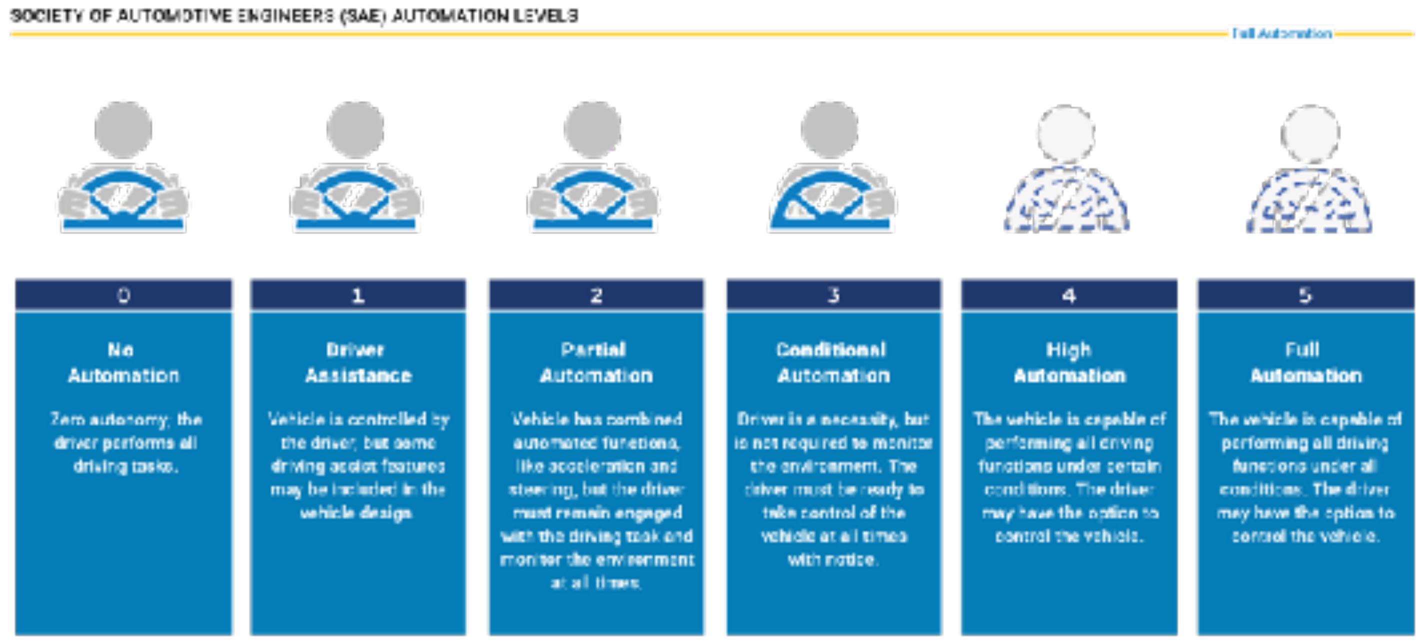
AI Inference

[EE Times]

Cutting-Edge Process: Reliability ?



Functional Safety Requirements ↗


SOCIETY OF AUTOMOTIVE ENGINEERS (SAE) AUTOMATION LEVELS

Full Automation

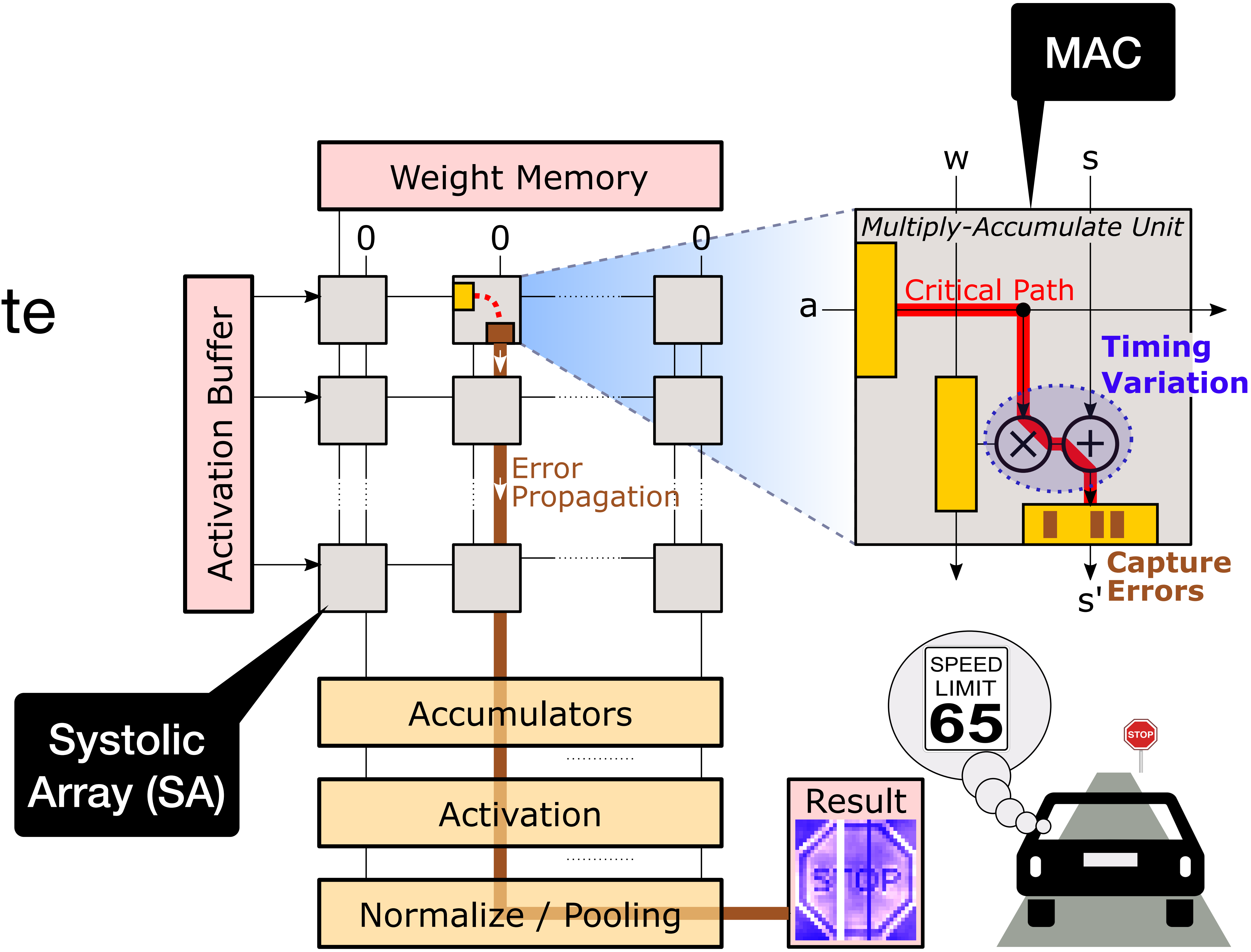
0	1	2	3	4	5
No Automation	Driver Assistance	Partial Automation	Conditional Automation	High Automation	Full Automation
Zero autonomy; the driver performs all driving tasks.	Vehicle is controlled by the driver, but some driving assist features may be included in the vehicle design.	Vehicle has combined automated functions, like acceleration and steering, but the driver must remain engaged with the driving task and monitor the environment at all times.	Driver is a necessity, but is not required to monitor the environment. The driver must be ready to take control of the vehicle at all times with notice.	The vehicle is capable of performing all driving functions under certain conditions. The driver may have the option to control the vehicle.	The vehicle is capable of performing all driving functions under all conditions. The driver may have the option to control the vehicle.

[Accolade Technology]

- Need to Understand Impact of Hardware-Related Errors

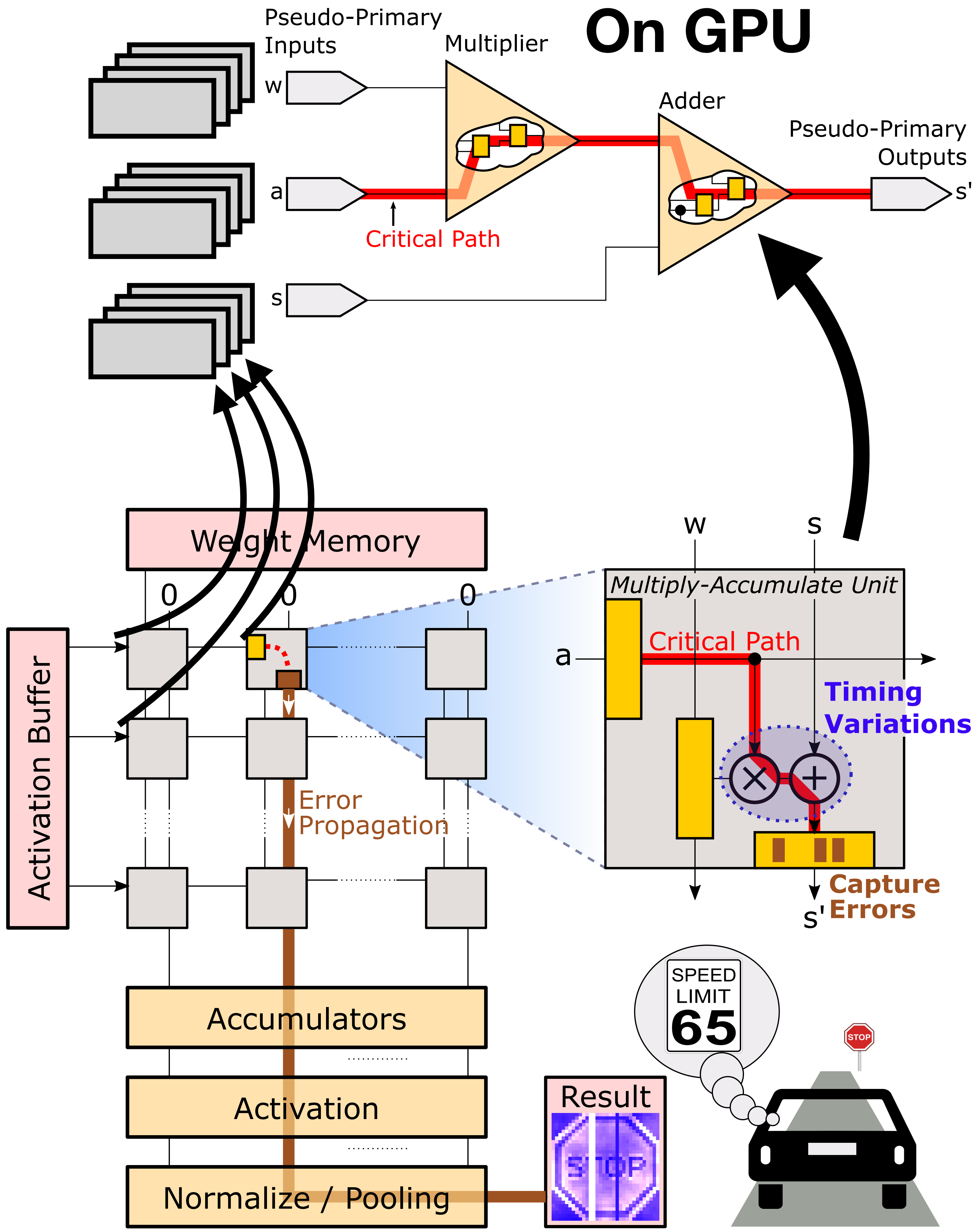
Typical Neural Processing Units

- NN Inference = Many Dot-Products
- Large 2D-Grid of Multiply-Accumulate (MAC) Units
  - 96 x 96 for a Tesla NPU
  - 256 x 256 for Google TPUv1 (Systolic Array)
- Hardware errors can lead to **Silent Data Corruption.**



Simulating Systolic Arrays

- Paths of interest are within the MAC units
  - Load MAC Gate-Level Netlist on GPU
- Parallelism of the SA directly translates to data parallelism on GPU
  - Can simulate all 256 x 256 MAC units in parallel
  - Each MAC can have distinct simulation parameters: timing variation, fault conditions, ...



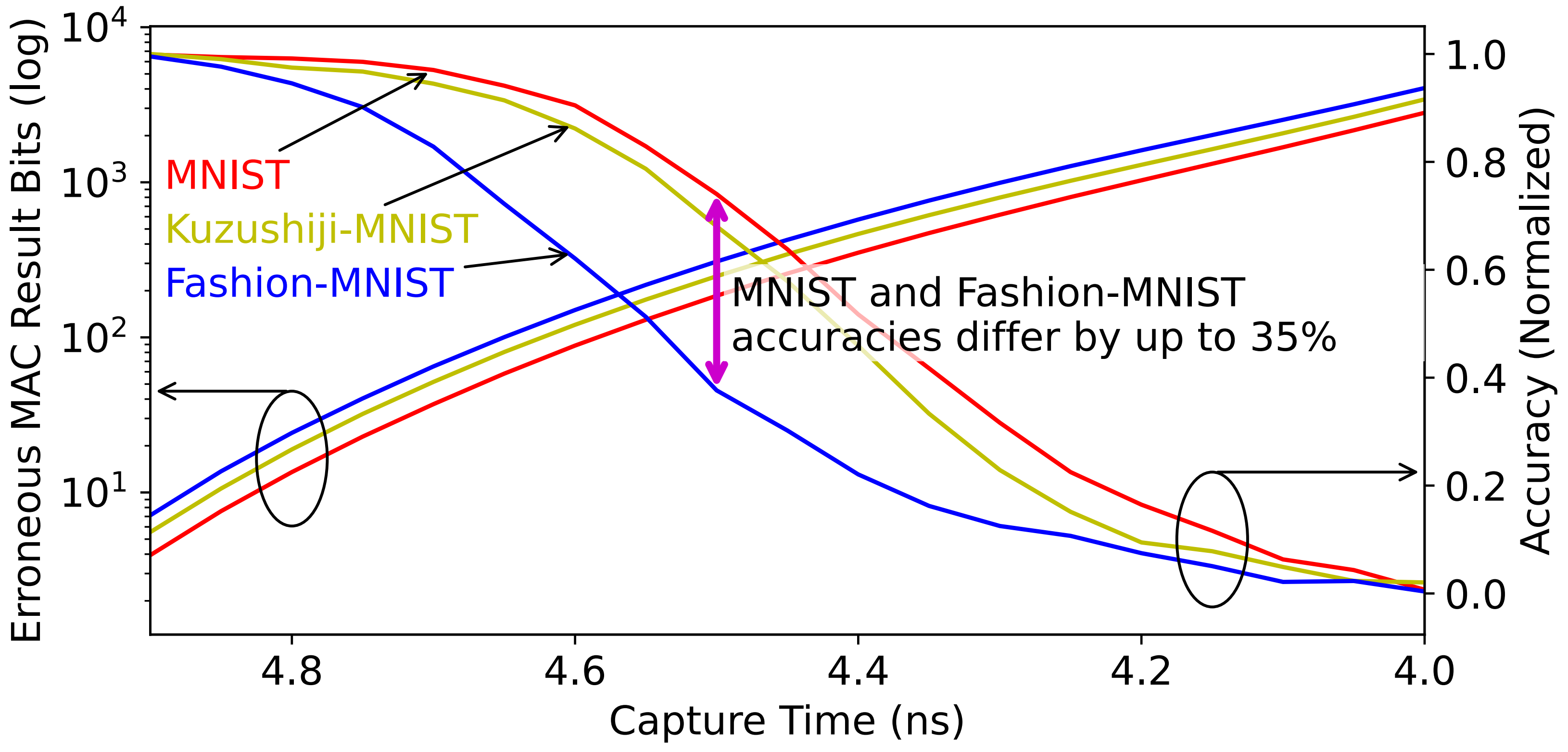
Holst et al.: "GPU-Accelerated Timing Simulation of Systolic-Array-Based AI Accelerators"
ATS 2021 *Best Paper*

Complete Timing Simulation of NN Inferences

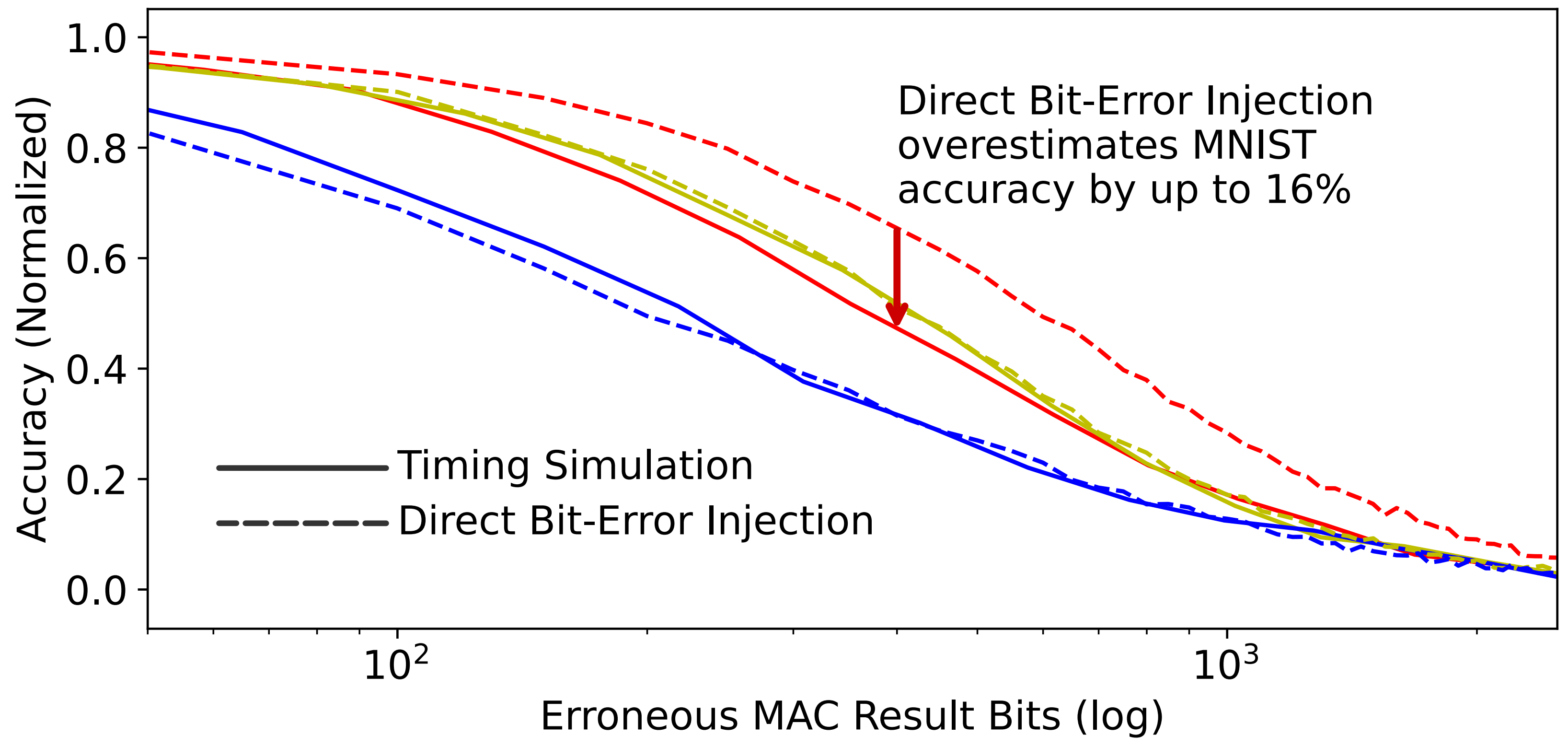
INT8 MAC Unit

- GPU: Nvidia RTX 3090 with 24 GB Memory
- LeNet-5: 417k MAC operations per inference
- 128 images = 53M MAC operations = 8 min sim time

Errors and Accuracy Impact of Overclocking



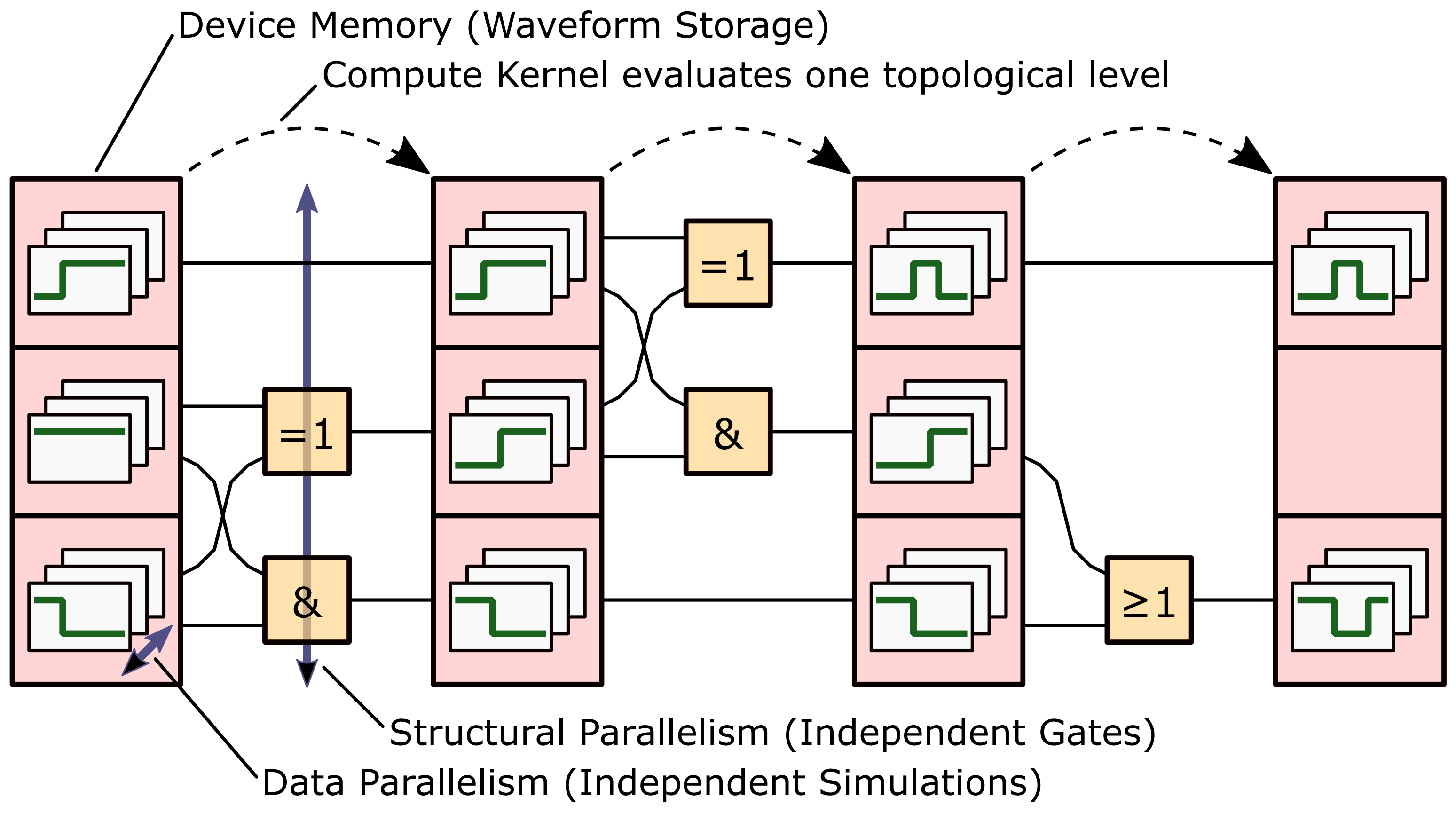
Better Results Than Bit-Error Injection



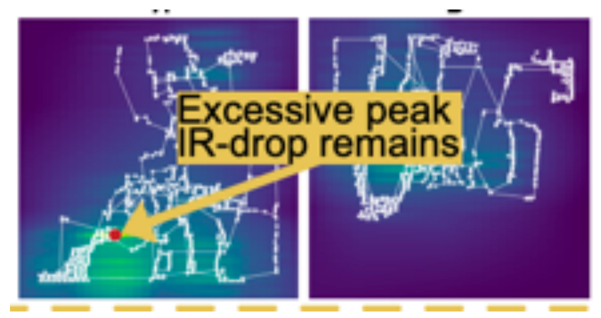
[Holst et al.: "The Impact of Timing Errors in Systolic-Array-Based AI Accelerators" AI-TREATS 2023]

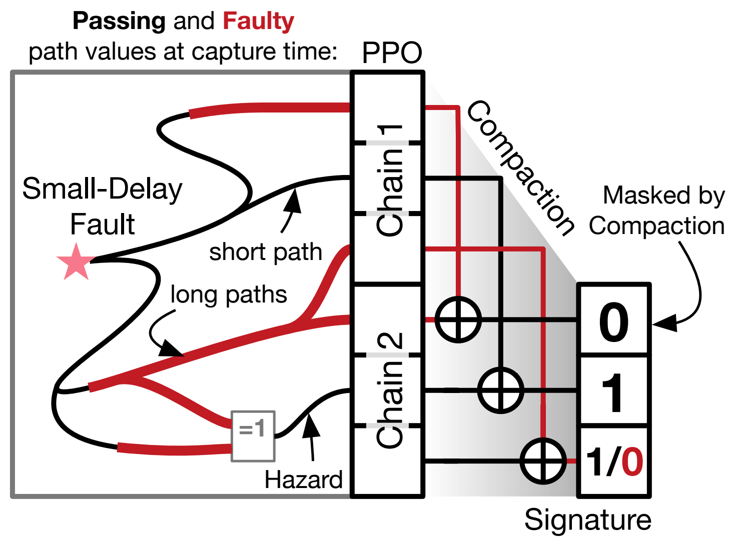
Summary

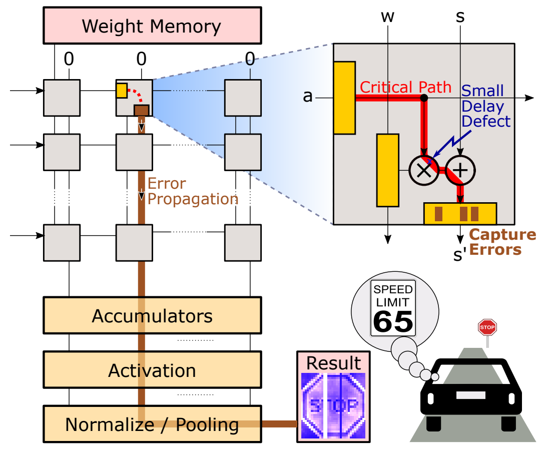
- GPU-based High-Throughput Timing Simulation
  → Performance by Data-Parallelism
- 10000+ Independent Simulations
  "Feeding the Monster"
  - Scan-Test Power Analysis
  - Small Delay Fault Simulation and Diagnosis
  - AI Accelerator Resilience Analysis
- Code: `https://github.com/s-holst/kyupy`



Device Memory (Waveform Storage)
Compute Kernel evaluates one topological level
=1
&
=1
&
≥1
Structural Parallelism (Independent Gates)
Data Parallelism (Independent Simulations)







Thank You!