Multi-Level Timing Simulation on GPUs

Eric Schneider, Michael A. Kochte, Hans-Joachim Wunderlich

Institute of Computer Architecture and
Computer Engineering (ITI)

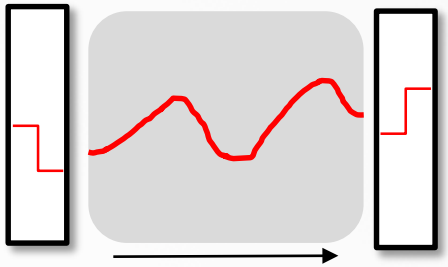
University of Stuttgart
Stuttgart, Germany

January 17, 2018

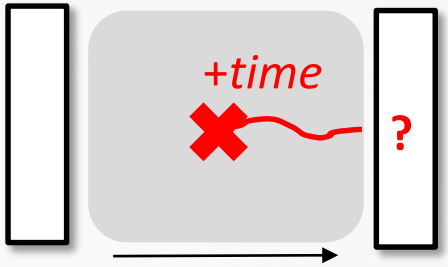
University of Stuttgart
Germany

Timing Simulation of Nano-electronic Circuits

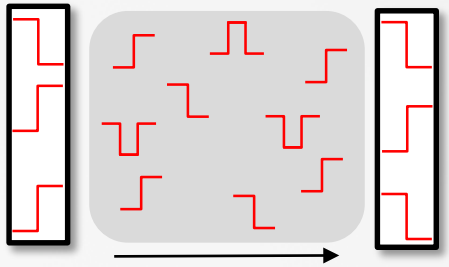
Design Validation (e.g. Variation Analysis)

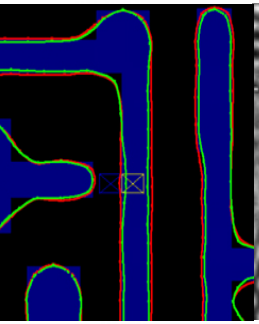


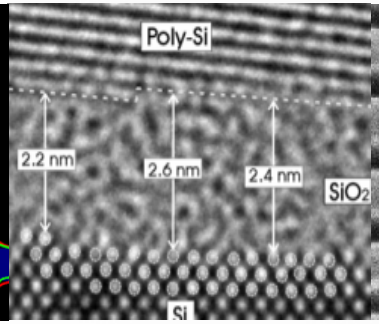
Delay Fault Simulation (e.g. Small Delays)

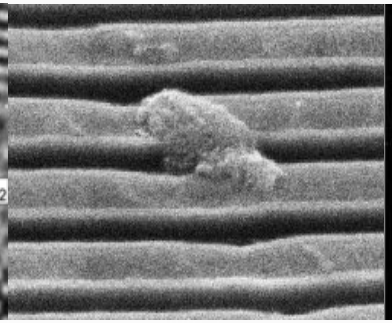


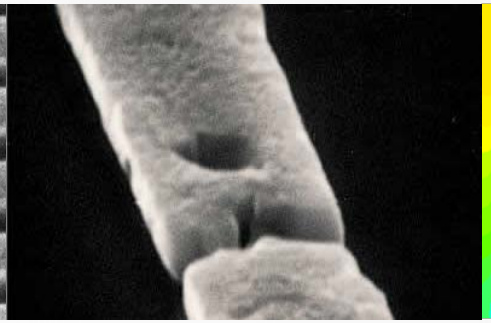
Non-Functional Properties (e.g. Power, IR-Drop)

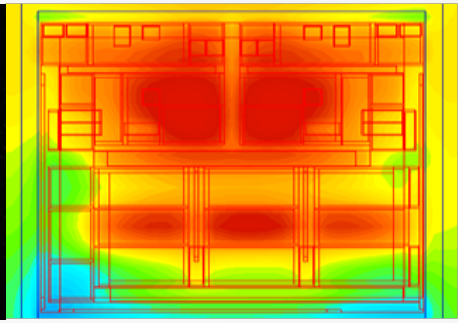




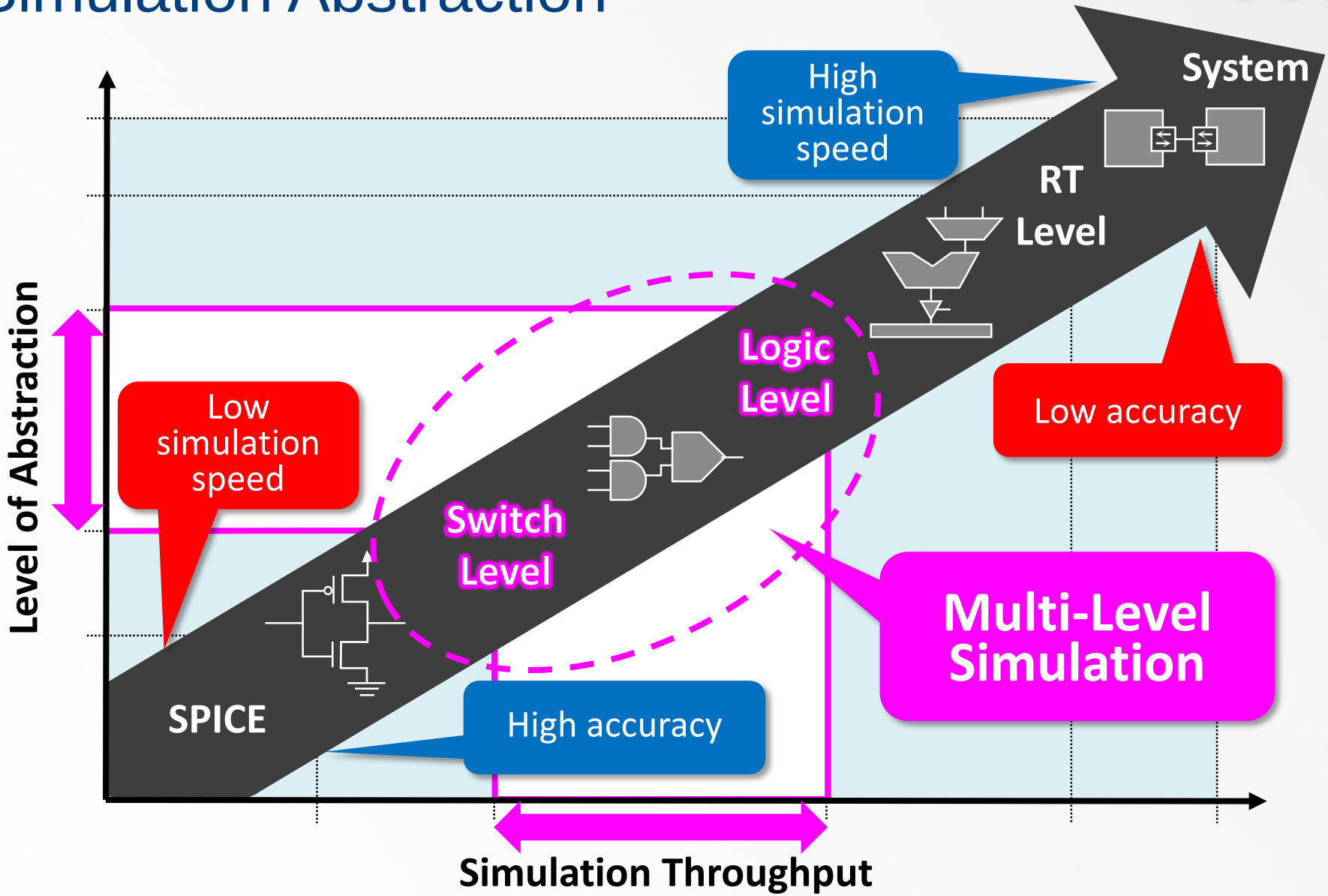






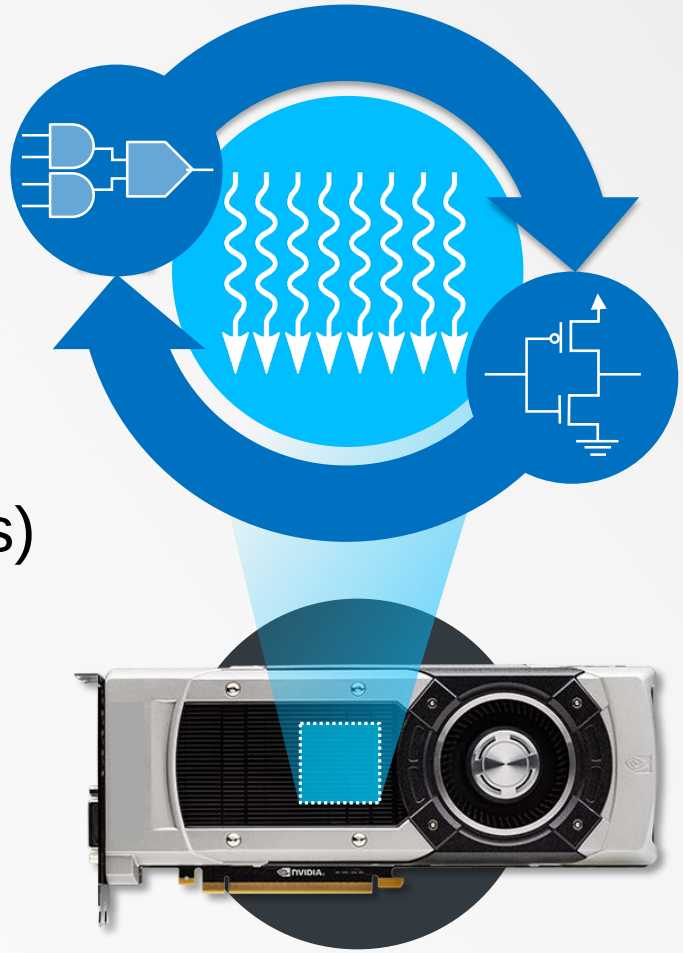


Simulation Abstraction



Efficient Multi-Level Timing Simulation

- **Flexible** timing-accurate simulation with mixed abstractions
  - **Logic** and **Switch** Level
- **High-throughput** parallelization on Graphics Processing Units (GPUs)
- Variable user-defined **Trade-off** in speed and accuracy

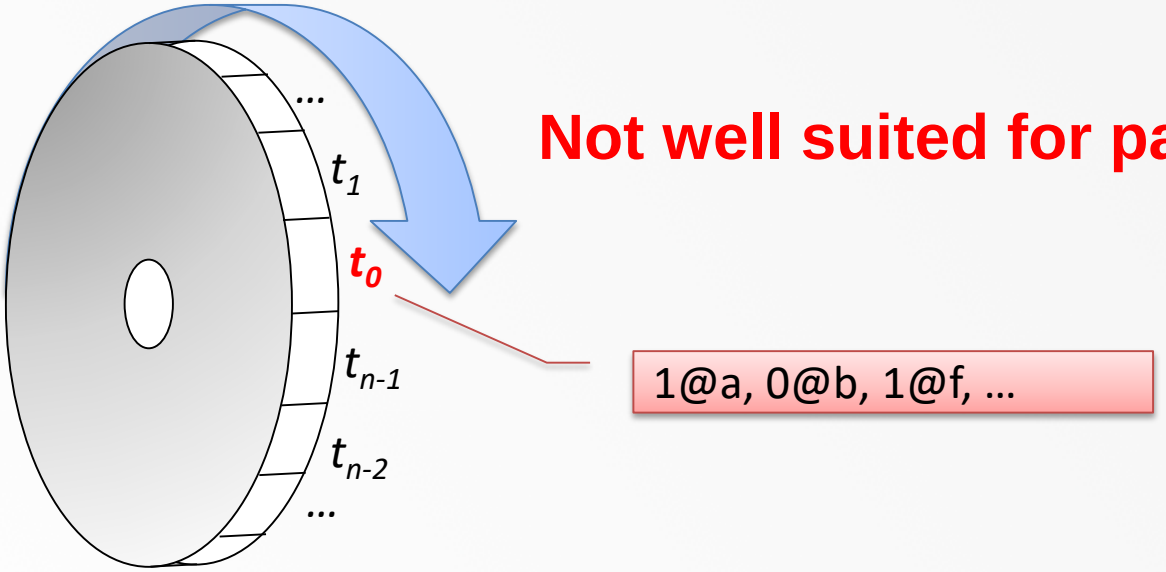


Agenda

- **GPU-accelerated Time Simulation**
- Transparent Multi-Level Time Simulation
- Experimental Results
- Conclusion

Conventional Time Simulation

- **Event-driven** time simulation for efficiency
  - **Time-wheel** implementation as cyclic list of event lists[1]
  - Requires dynamic memory management

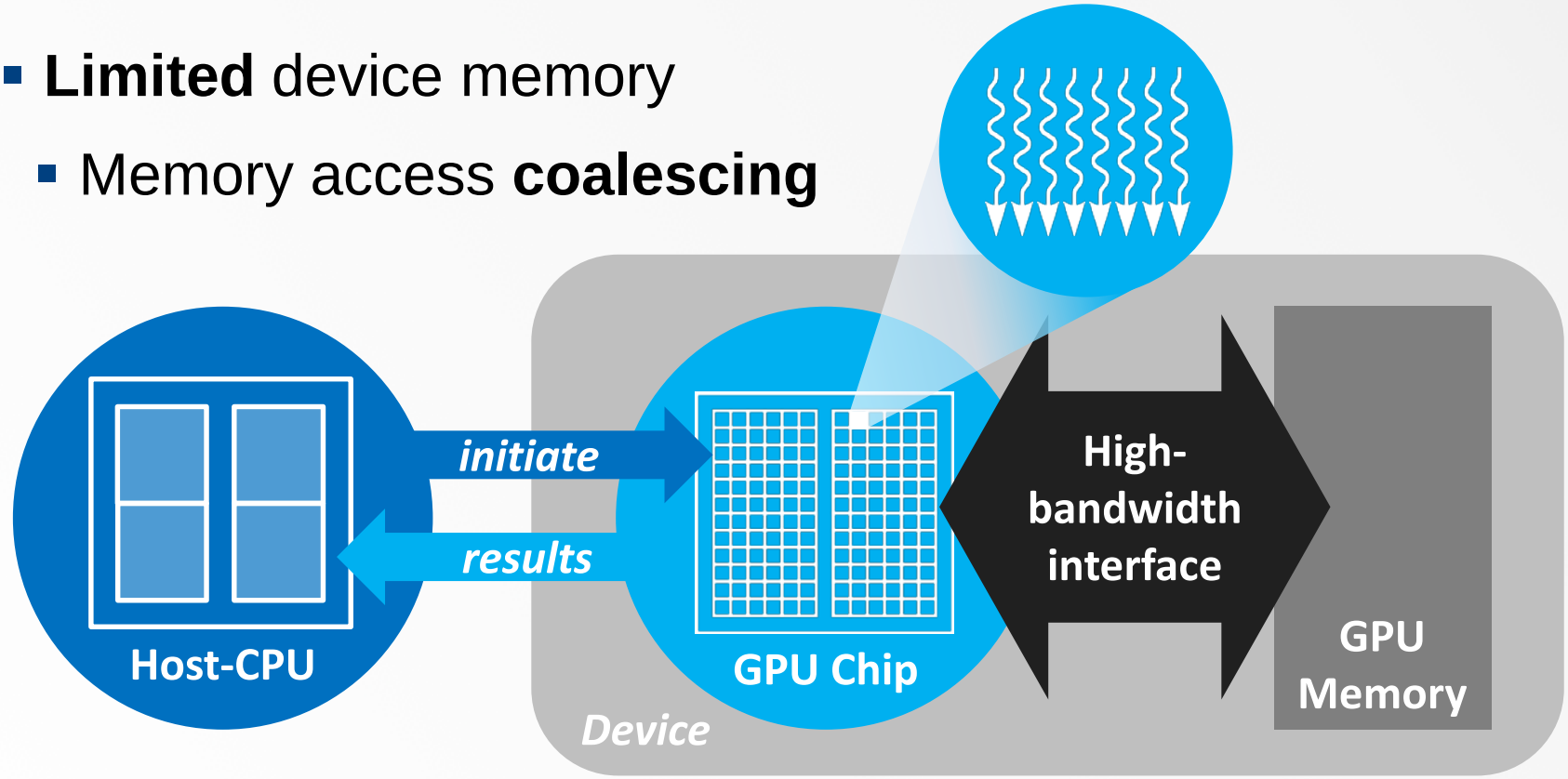


Not well suited for parallelization

[1] E. G. Ulrich. Exclusive Simulation of Activity in Digital Networks. Communications of the ACM, 12(2):102–110, Feb. 1969

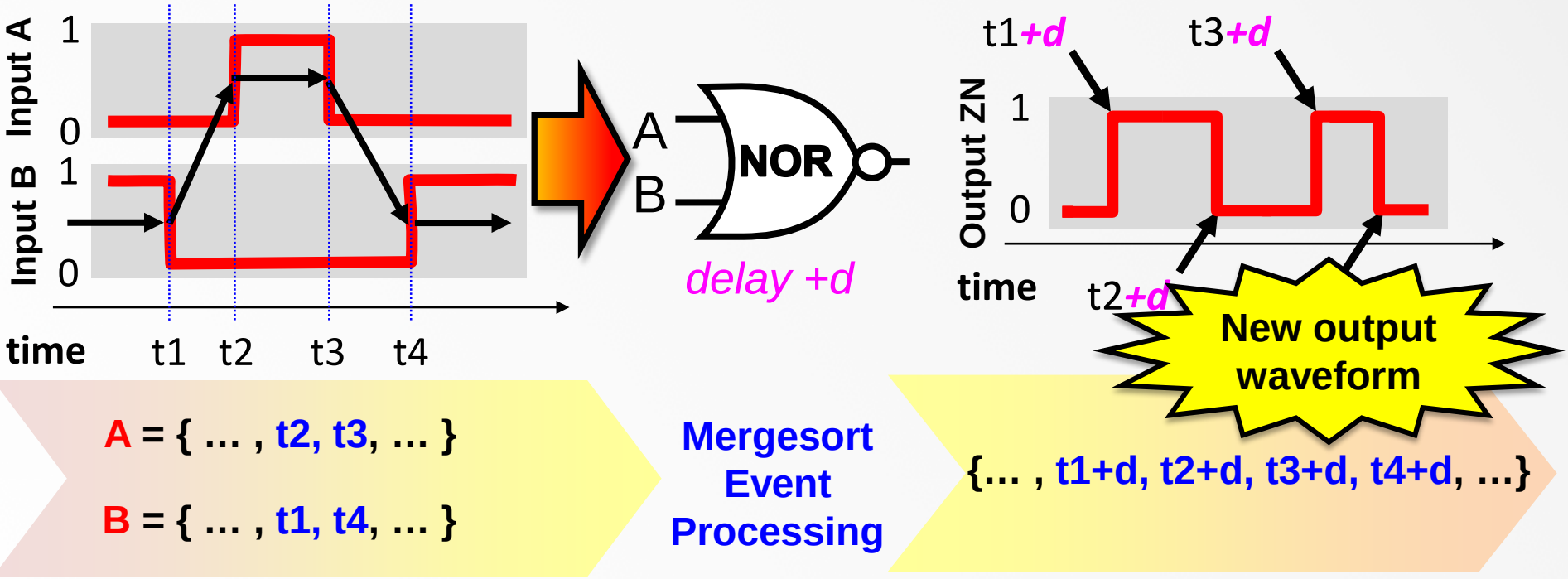
Graphics Processing Units (GPUs)

- Single-Instruction-Multiple-Data (SIMD) processing
  - **Uniform** control flow of threads
- **Limited** device memory
  - Memory access **coalescing**



GPU-accelerated Logic Time Simulation

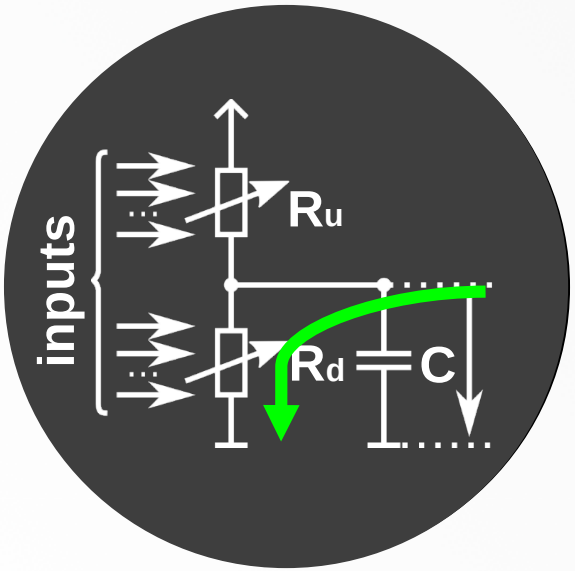
- Simple and compact **data structures**
  - Full signal histories (**waveforms**) stored as **event lists**

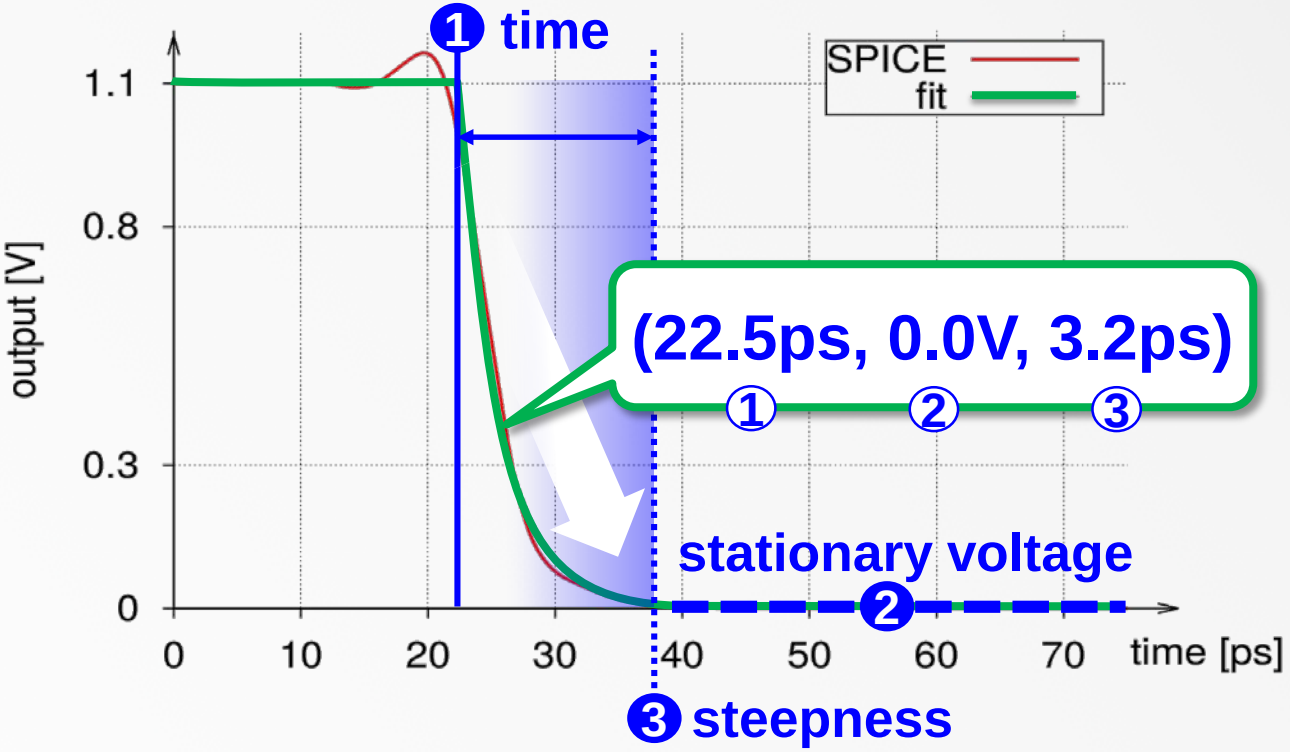


[1] Holst, Imhof, Wunderlich, *High-Throughput Logic Timing Simulation on GPGPUs*, TODAES, 2015.

GPU-accelerated Switch Level Simulation

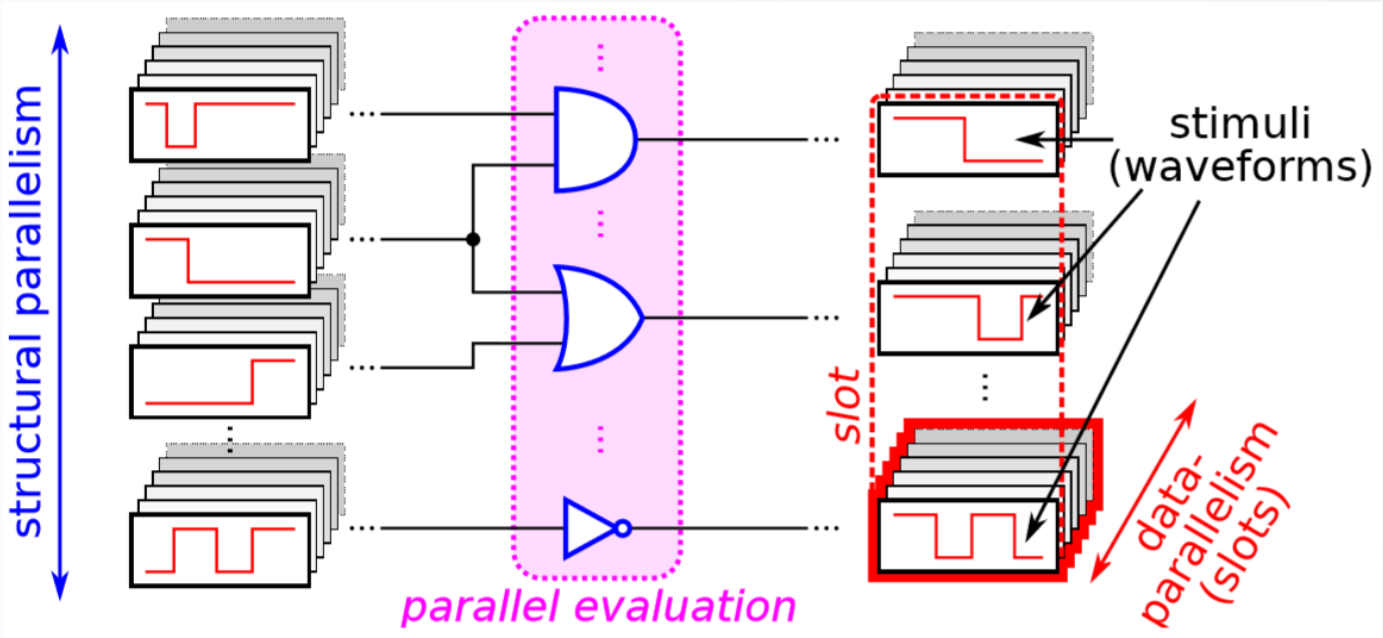
- Cell behavior modeled using **first-order** electrical parameters
  - Switching events expressed by **exponential curves**





Schneider, Holst, Wen and Wunderlich, *Data-Parallel Simulation for Fast and Accurate Timing Validation of CMOS Circuits*, Proc. ICCAD, 2014.

High-Throughput Parallelization



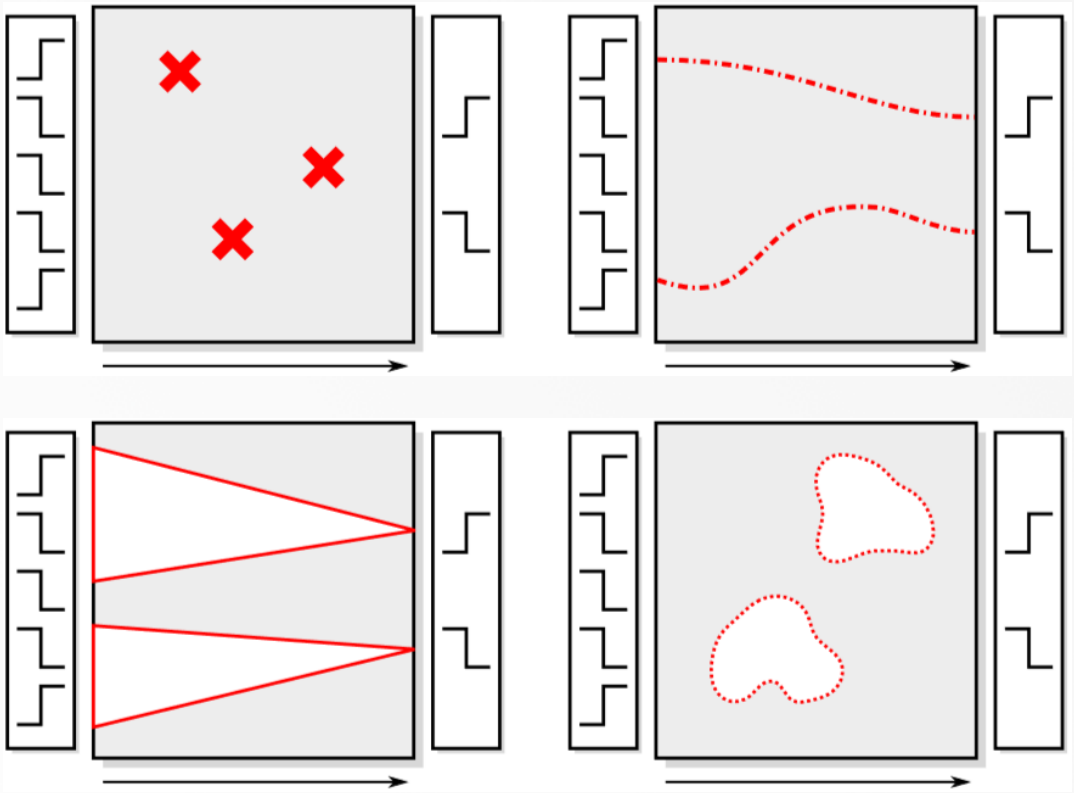
- **Multi-dimensional** kernels
  - Gate-parallelism and waveform-parallelism

Agenda

- GPU-accelerated Time Simulation
- **Transparent Multi-Level Time Simulation**
- Experimental Results
- Conclusion

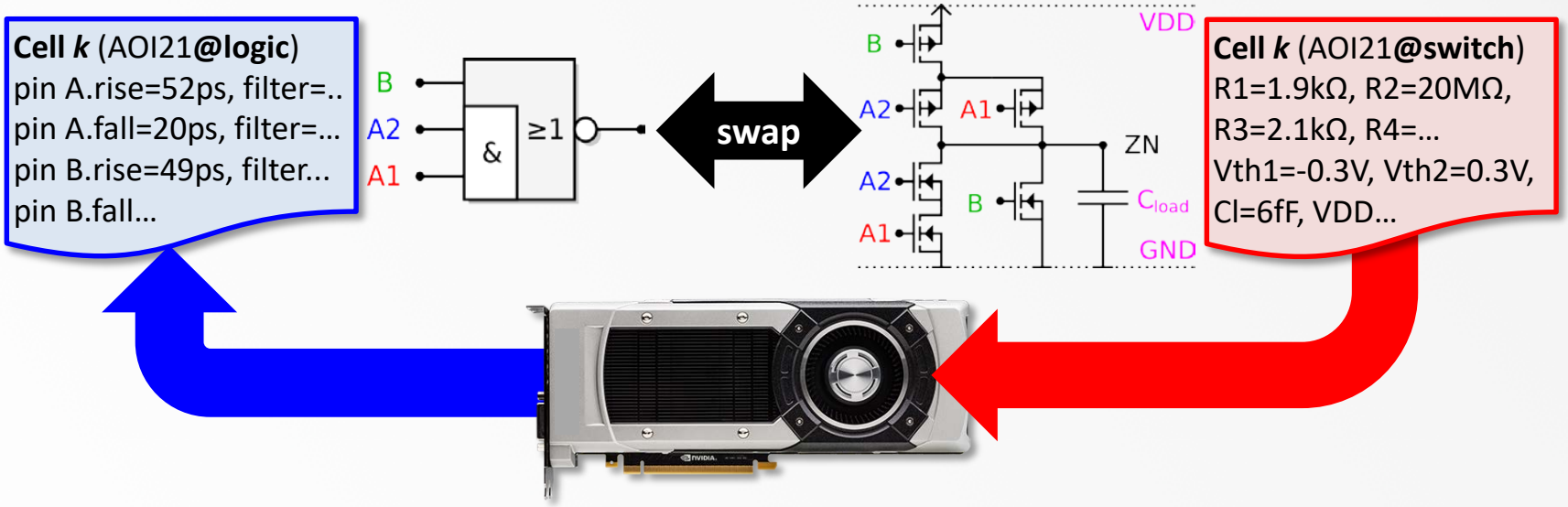
Regions of Interest (ROI)

- Demand for areas with **higher simulation** accuracy



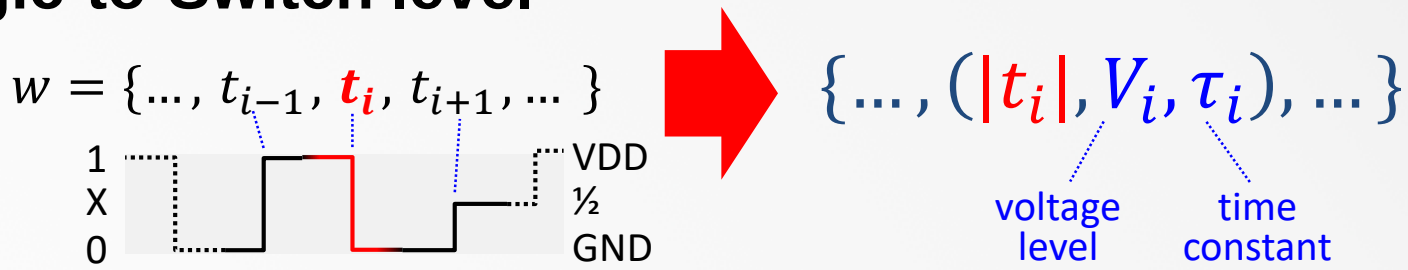
Abstraction Switching

- **Logic** and **switch** level descriptions for nodes
  - During simulation only **one active** abstraction per node

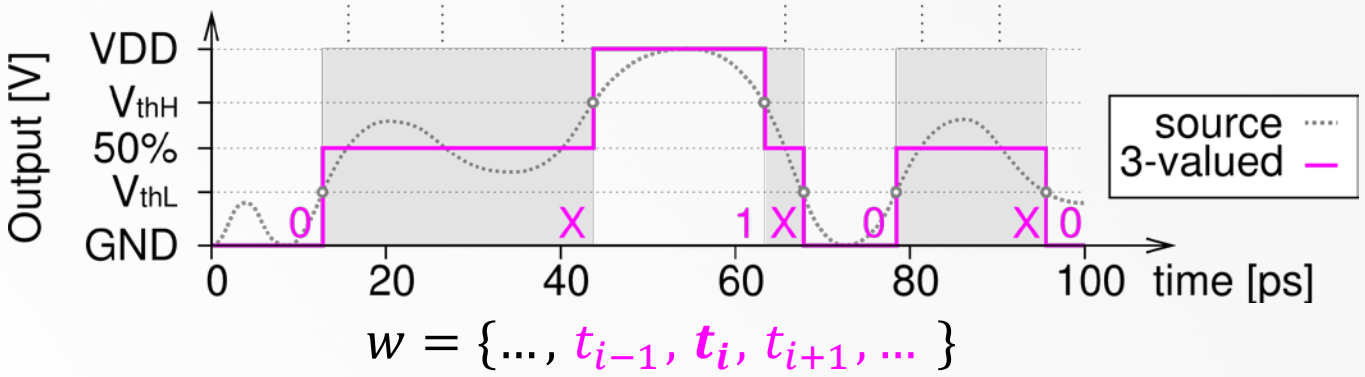


Waveform Event Transformation

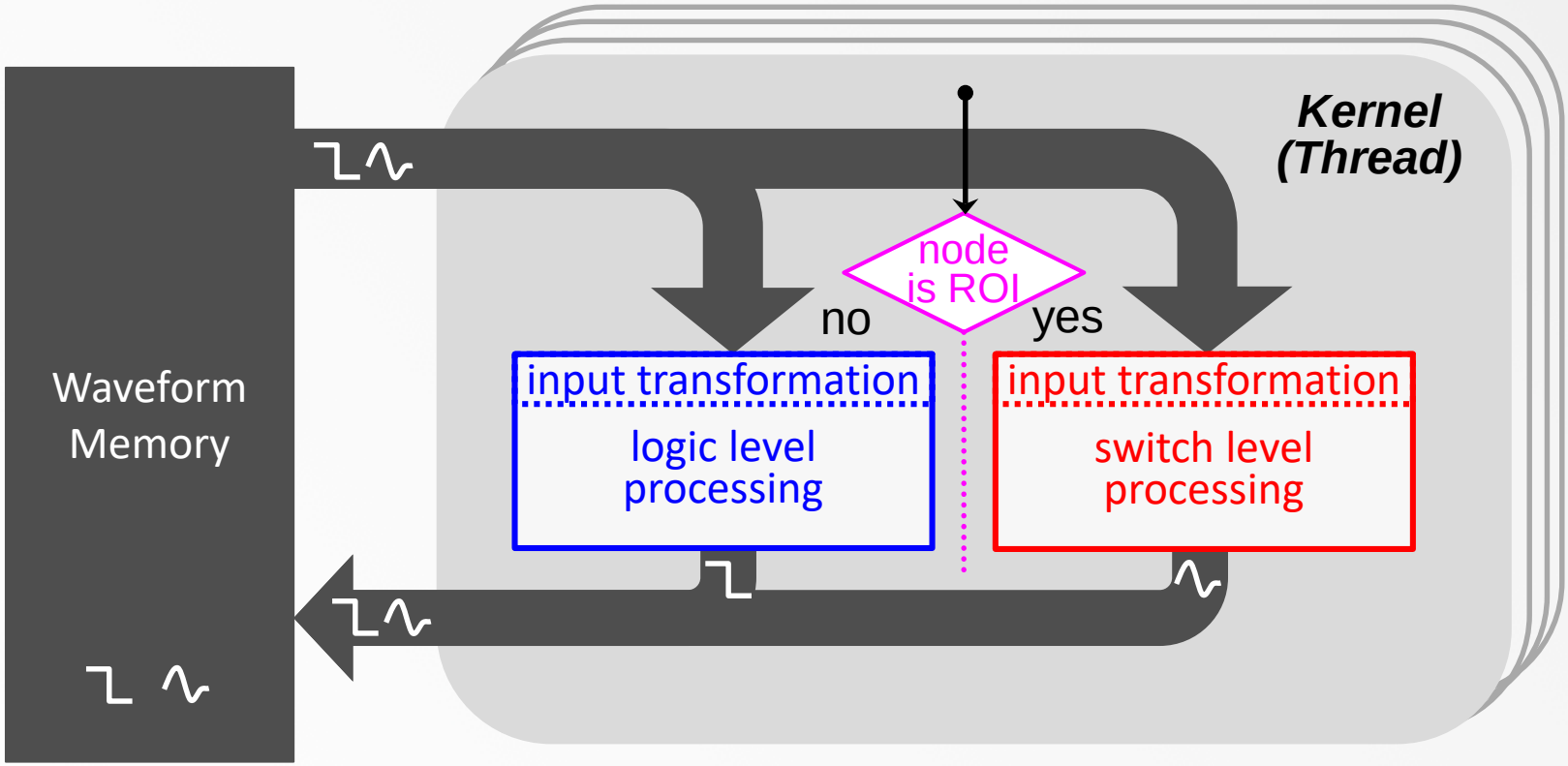
- **Logic-to-Switch level**



- **Switch-to-Logic level**
  - Threshold-based mapping to (**ternary**) **logic** symbols

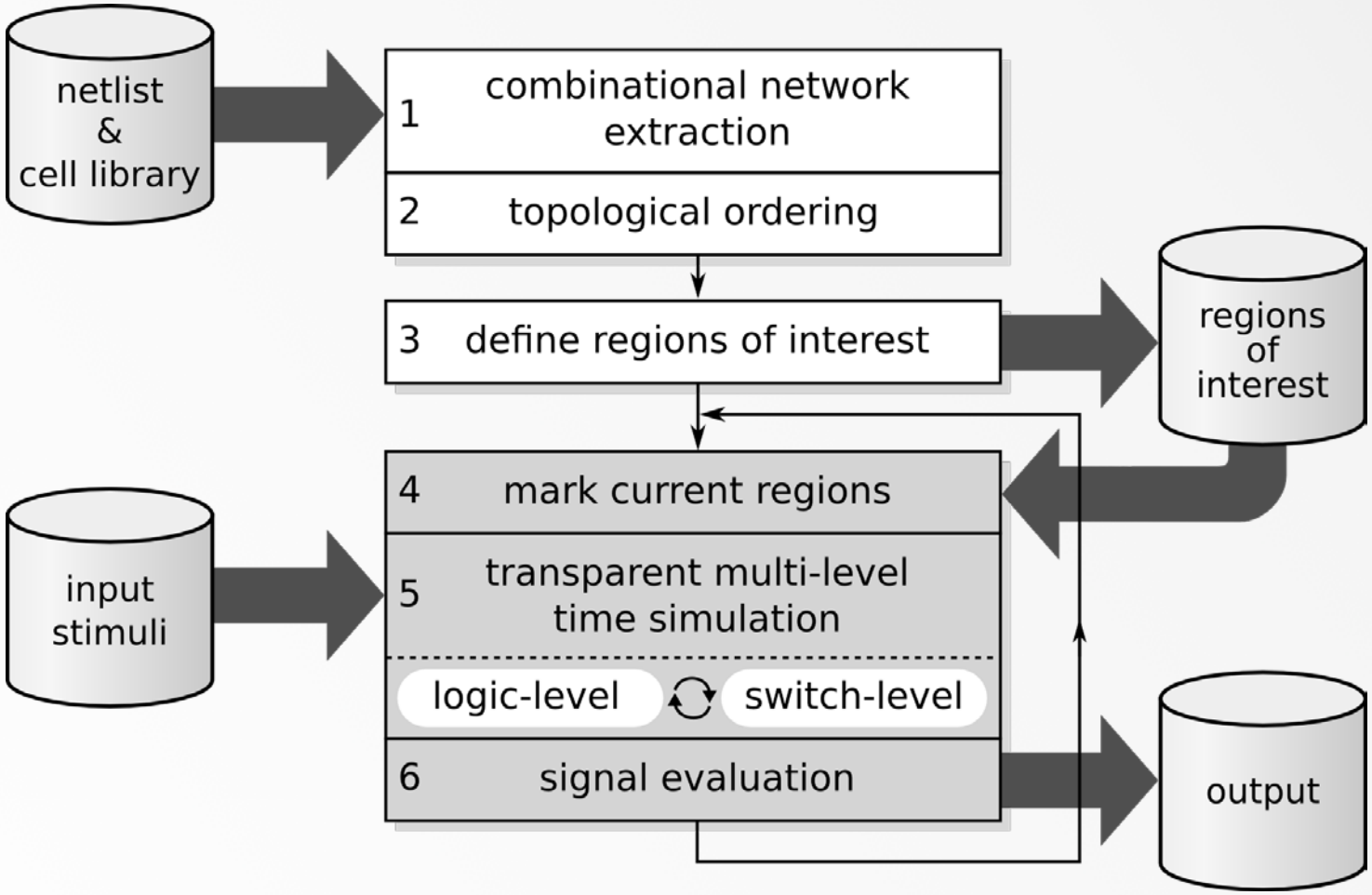


Transparent Multi-Level Simulation



- Input waveform events are mapped to **target abstraction** level of node during processing

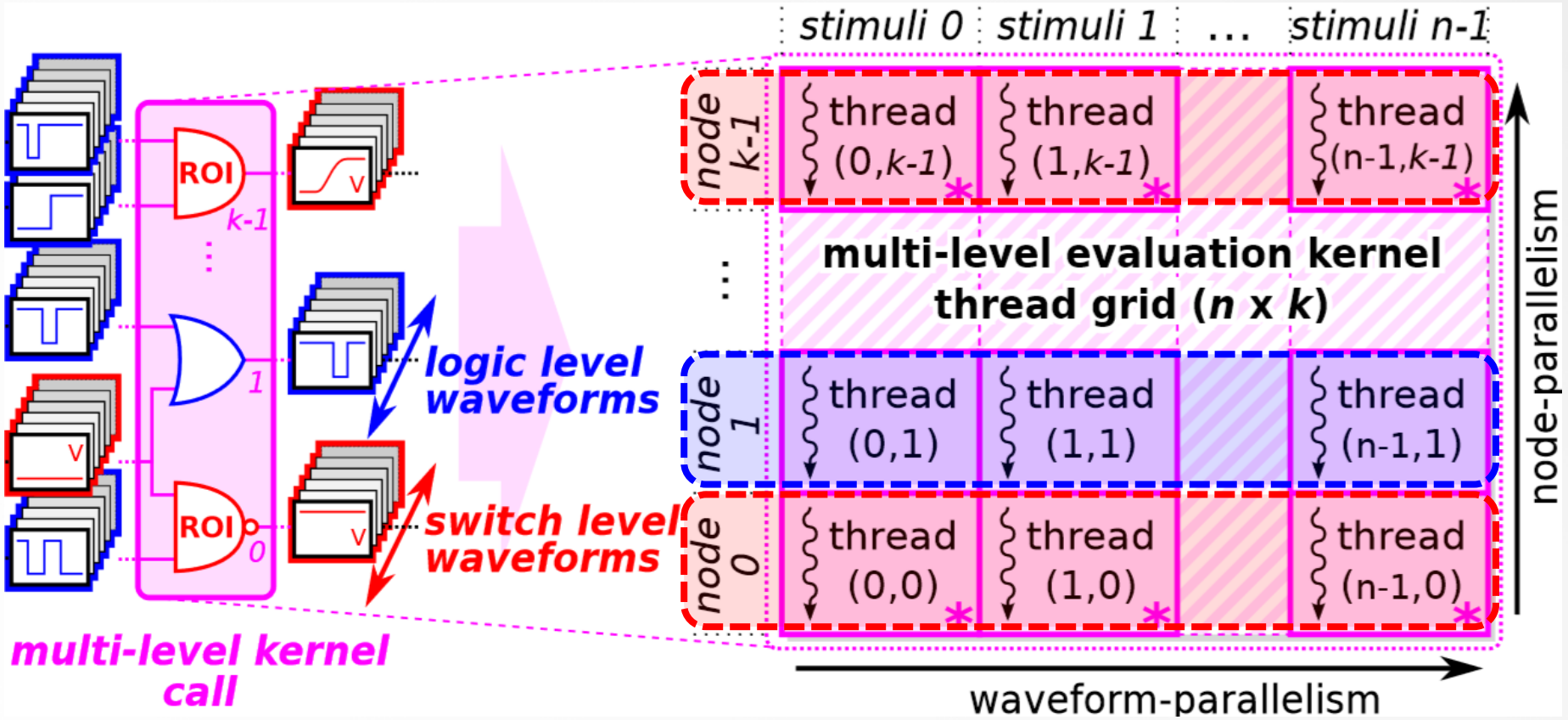
Simulation Flow



[1] Holst, Imhof, Wunderlich, *High-Throughput Logic Timing Simulation on GPGPUs*, TODAES, 2015.

[2] Schneider. et. al. *Data-Parallel Simulation for Fast and Accurate Timing Validation of CMOS Circuits*, ICCAD, 2014.

2D-Thread Grid Organization



- Threads of a SIMD group compute the **same** abstraction

Agenda

- GPU-accelerated Time Simulation
- Transparent Multi-Level Time Simulation
- **Experimental Results**
- Conclusion

Experimental Setup

- **ISCAS'89**, **ITC'99** and industrial designs from **NXP**
- Simulation of random ROI scenarios
  - Varying ROI count and distribution
  - 10-detect transition delay fault test set by ATPG

- Intel® Xeon® @3.0GHz, 8 cores, 128GB RAM
- NVIDIA® Tesla® K80 @875MHz, 2x2496 cores, 2x12GB

Simulation Runtimes

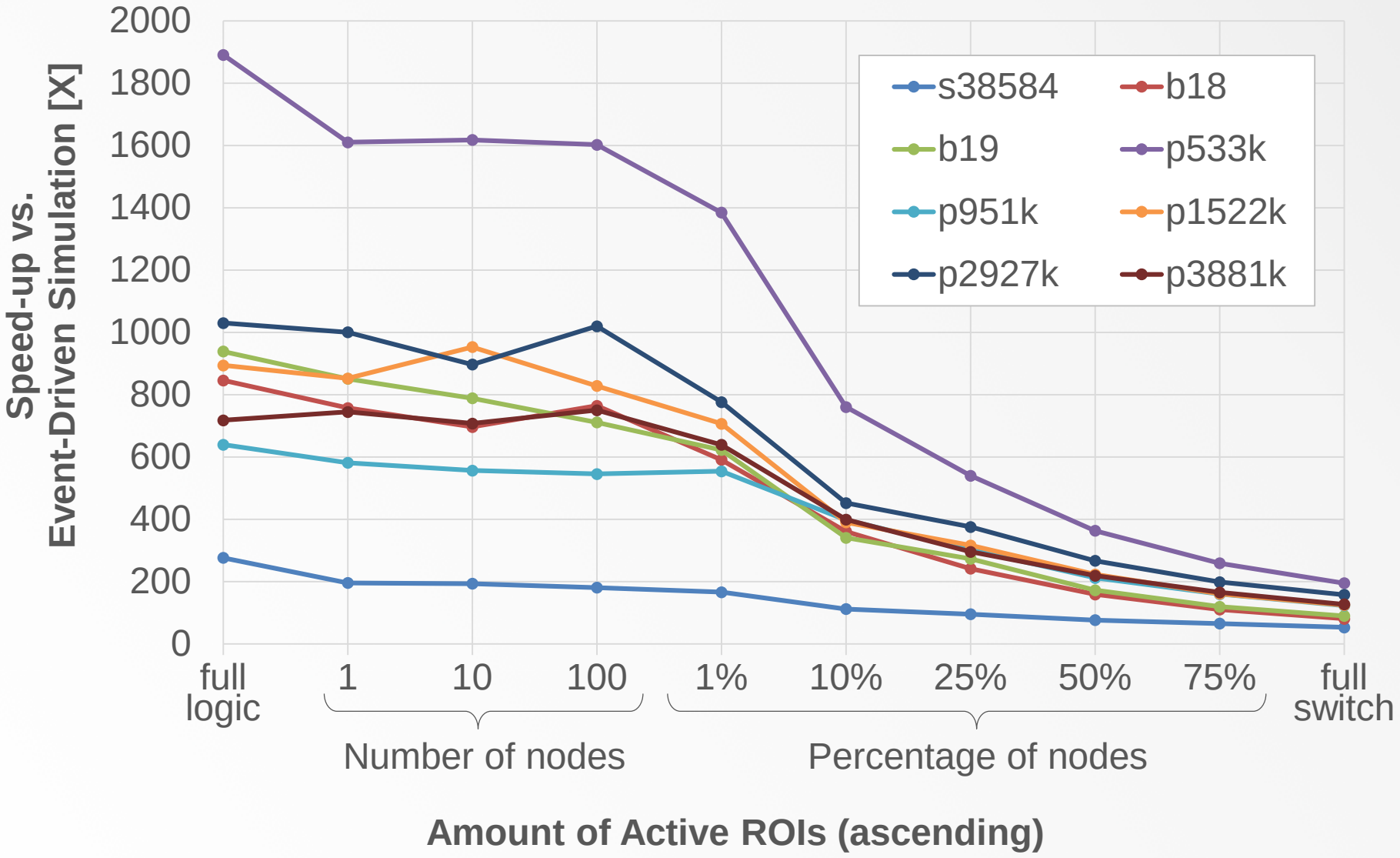
			Commercial Event-Driven	Proposed GPU-accelerated			
Circuit	Nodes*	Pattern-Pairs	T_{logic}	T_{logic}	$MEPS_{log}$	T_{switch}	$MEPS_{swi}$
s38584	23053	563	5.49s	51ms	254	266ms	49
b18	125305	3174	0:13h	922ms	431	9.69s	41
b19	250232	4651	0:41h	2.62s	444	27.36s	43
p533k	676611	3417	2:28h	5.33s	434	51.73s	45
p951k	1090419	7063	4:18h	24.20s	318	2:06m	61
p1522k	1088421	17980	12:34h	50.60s	387	6:03m	54
p2927k	1672479	22107	28:54h	1:41m	366	0:11h	56
p3881k	3691849	12092	27:31h	2:18m	323	0:13h	57

*input, output and cells

MEPS: Million node evaluations per second = $\frac{\#Nodes \cdot \#PatternPairs}{T \cdot 10^6}$

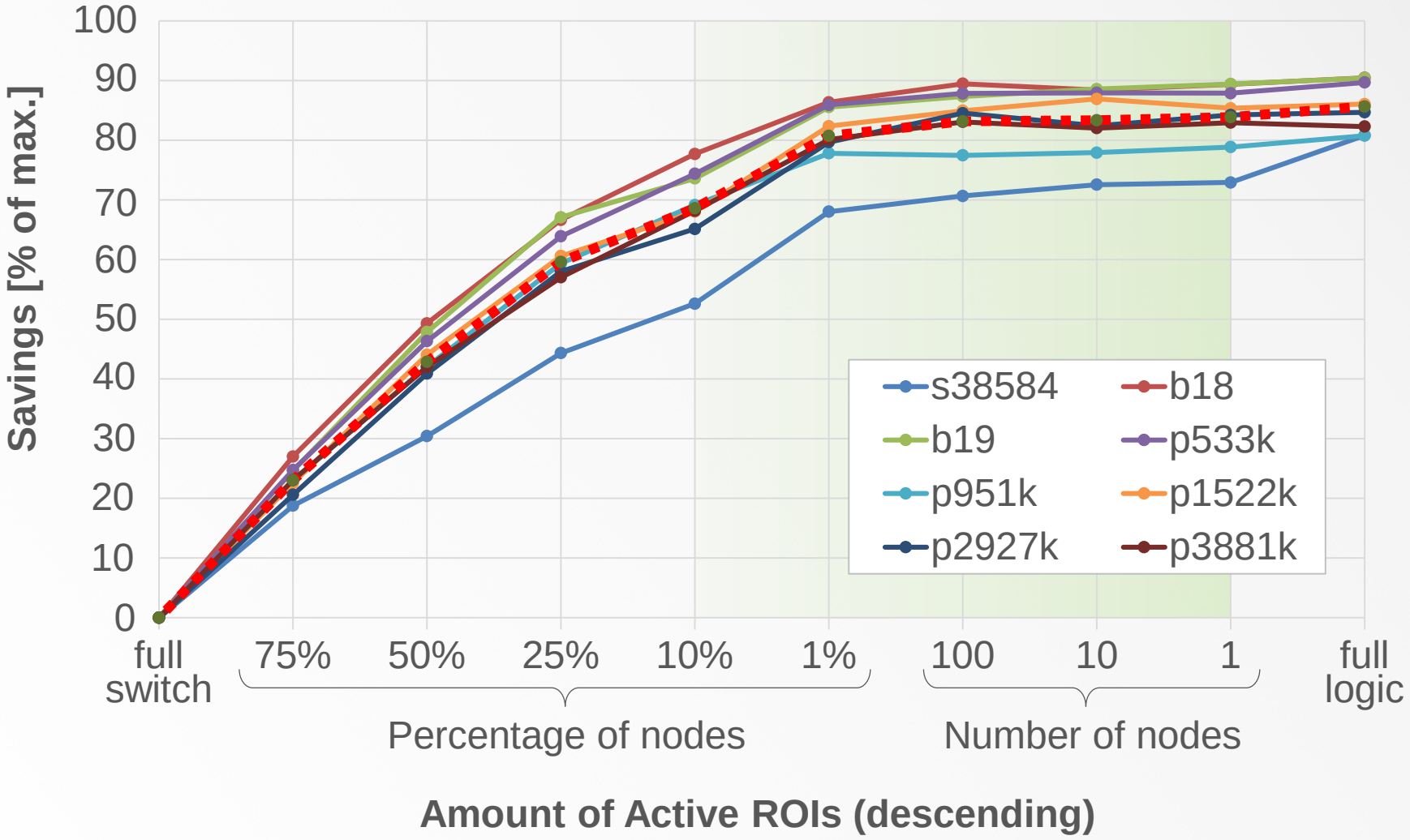
Simulation Speed-up

$$SpeedUp = \frac{T_{ref}}{T_{GPU}}$$



Runtime Savings

Savings compared to **full switch-level** simulation



Conclusion

- **First Multi-Level** Timing Simulation on GPUs
  - **High-Throughput** parallelization
  - Increased accuracy in arbitrary **regions of interest**
  - **Transparent** waveform transformation

- Scalable for **millions of cells**
- Speedups up to **three orders** of magnitude
  - Up to **444M** evaluations per second
- **Runtime savings** of up to 89% compared to full switch-level simulation