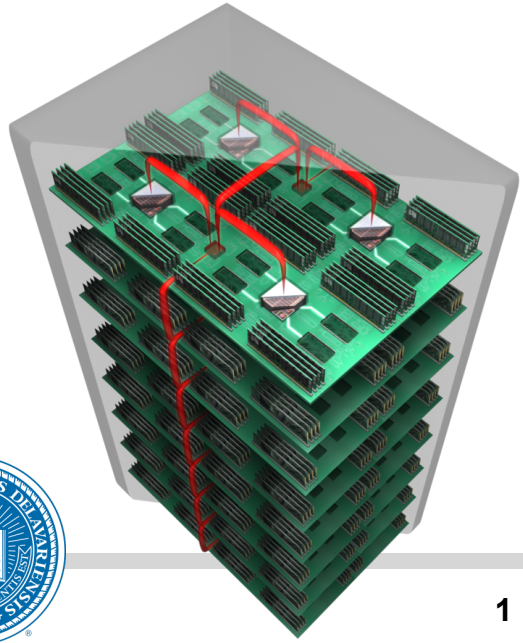
Reuse Distance Based Circuit Replacement in Silicon Photonic Interconnection Networks for HPC

HOTI 2014

Ke Wen, David Calhoun, Sébastien Rumley, Xiaoliang Zhu, Keren Bergman
Columbia University

Lian-Wee Luo, Michal Lipson
Cornell University

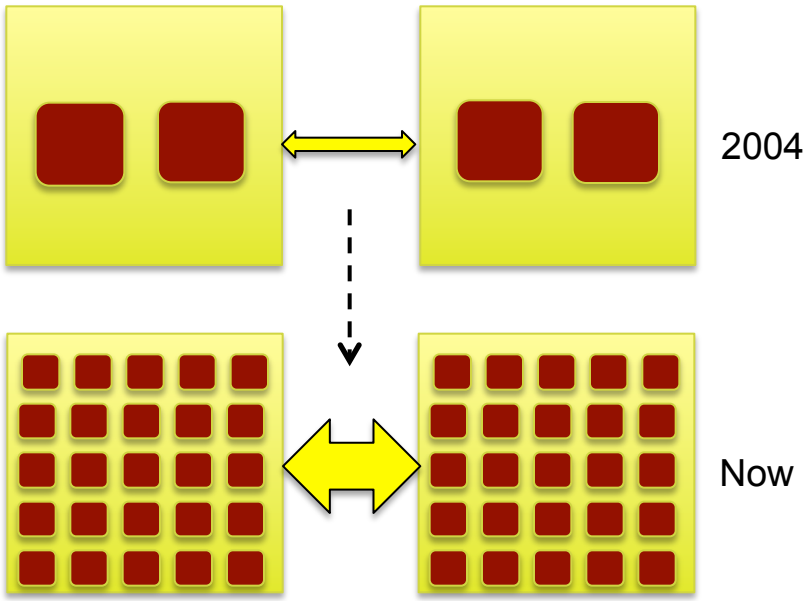
Yang Liu, Ran Ding, Tom Baehr-Jones, Michael Hochberg
University of Delaware



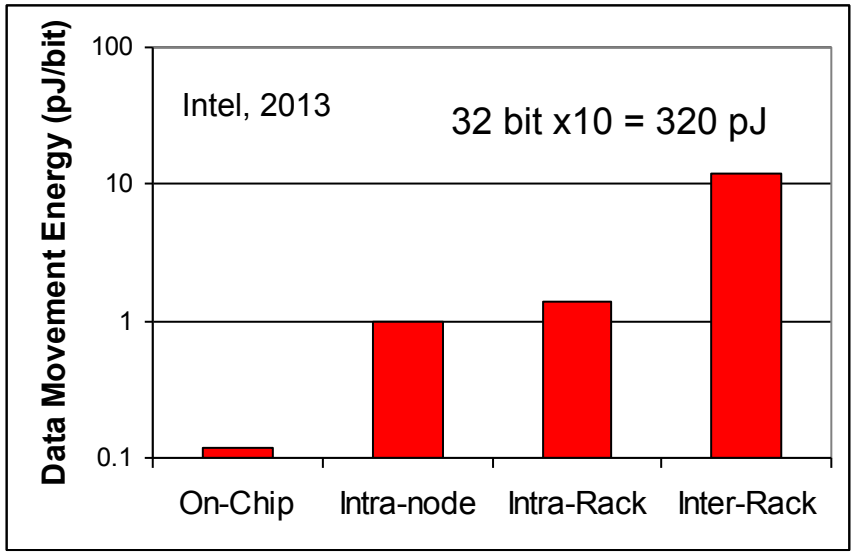
Data Movement Challenge in HPC

- Performance of HPC systems is no longer only determined by sheer FLOPS numbers, but also data movement capabilities.

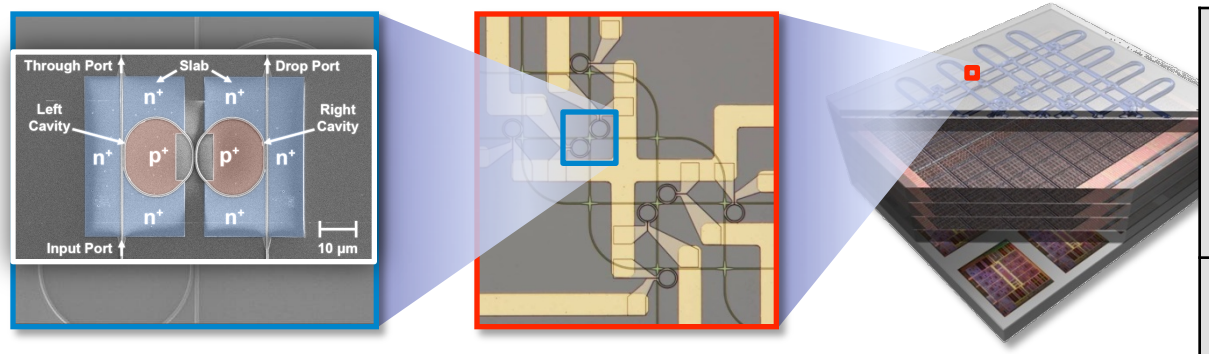
Gb/s: inter-chip/node communication significantly increases



Joule/bit:
20 MW / 1 ExaFLOPs
= 20 pJ/FLOP

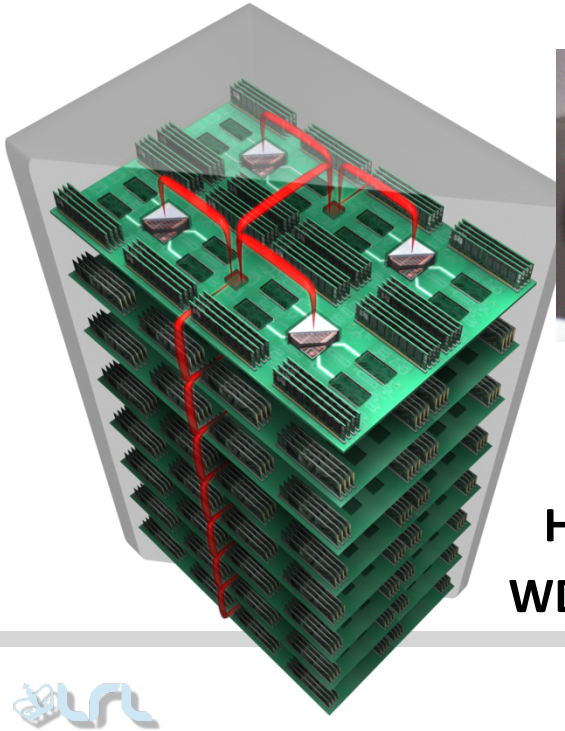


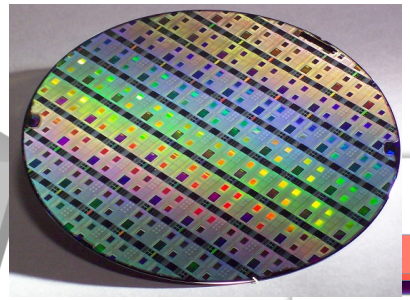
Systems Impact with Si-photonic interconnect

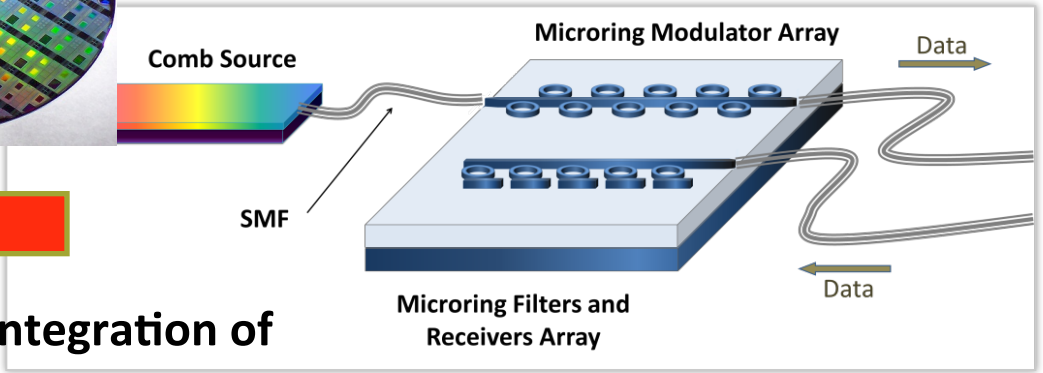


Bandwidth	10 Gb/s * 100 wavelengths = 1Tb/s
Efficiency	~ 1pJ/bit (end-to-end)

Reconfigurable photonic switches



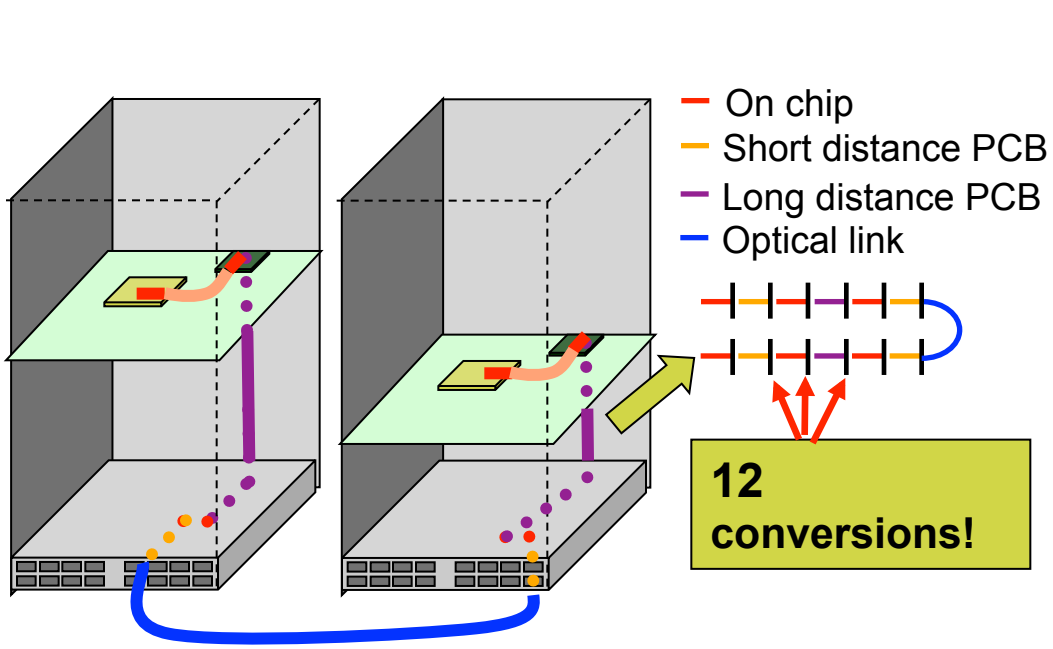




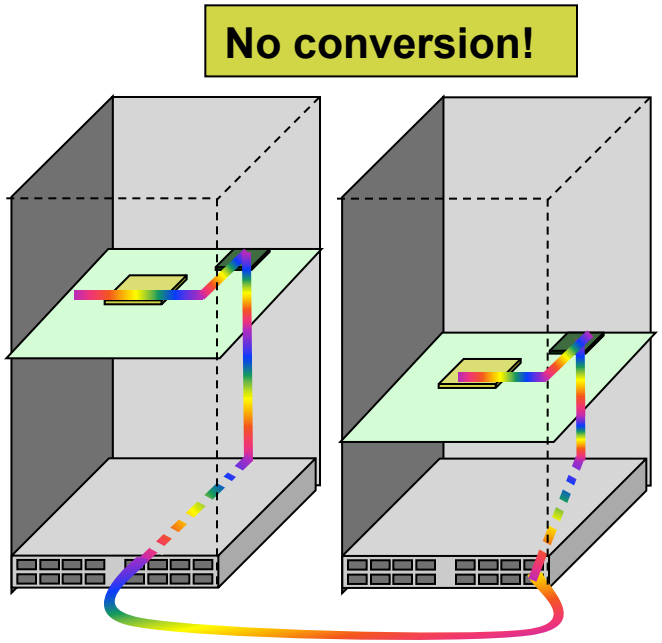
High density integration of WDM Transceivers with CMOS

Optical Data Movement Beyond Wire Replacement

- Optics-enabled system architecture transformations:
  - distance-independent, cut-through, bufferless



Conventional hop-by-hop data movement

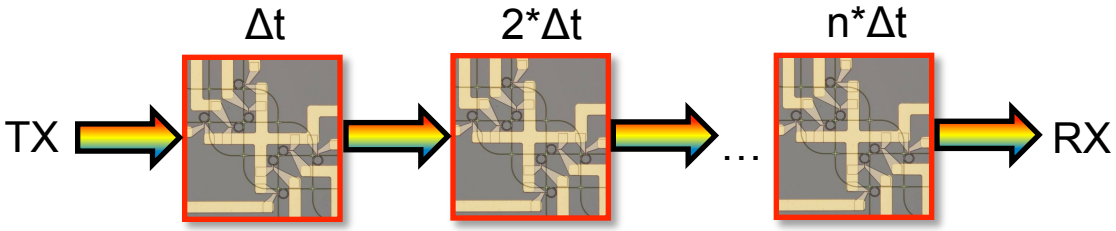


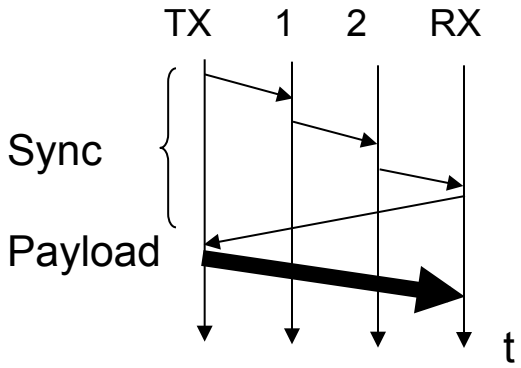
Fully flattened end-to-end data movement

Silicon Photonic Network Challenges

1. Rely on circuit switching (bufferless)

→ Need to setup a lightpath before data transmission

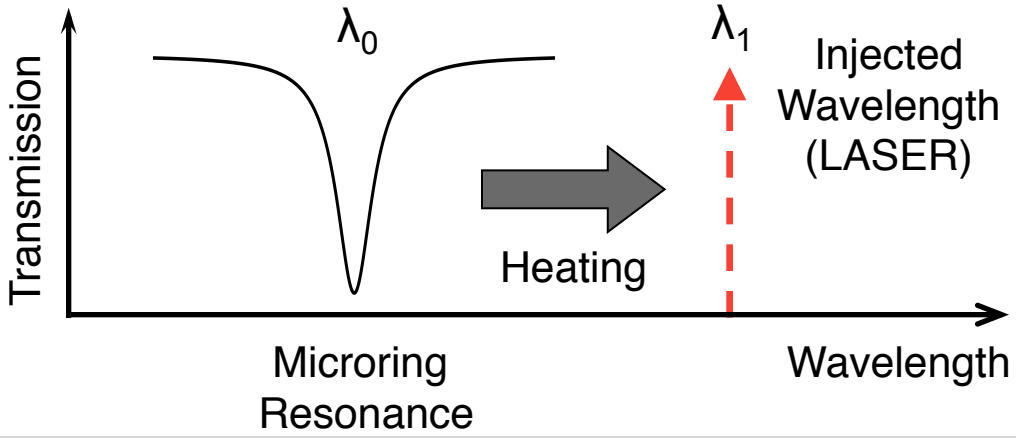


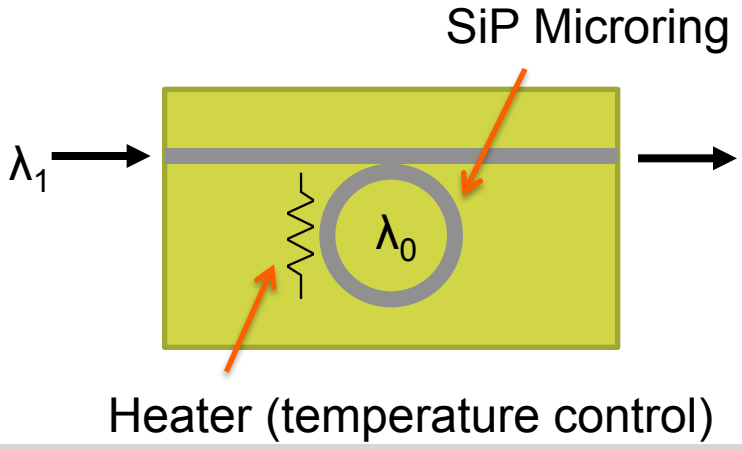


2. Microrings are sensitive to temperature

→ Need to thermally (re)-initialize microrings to work on correct wavelengths

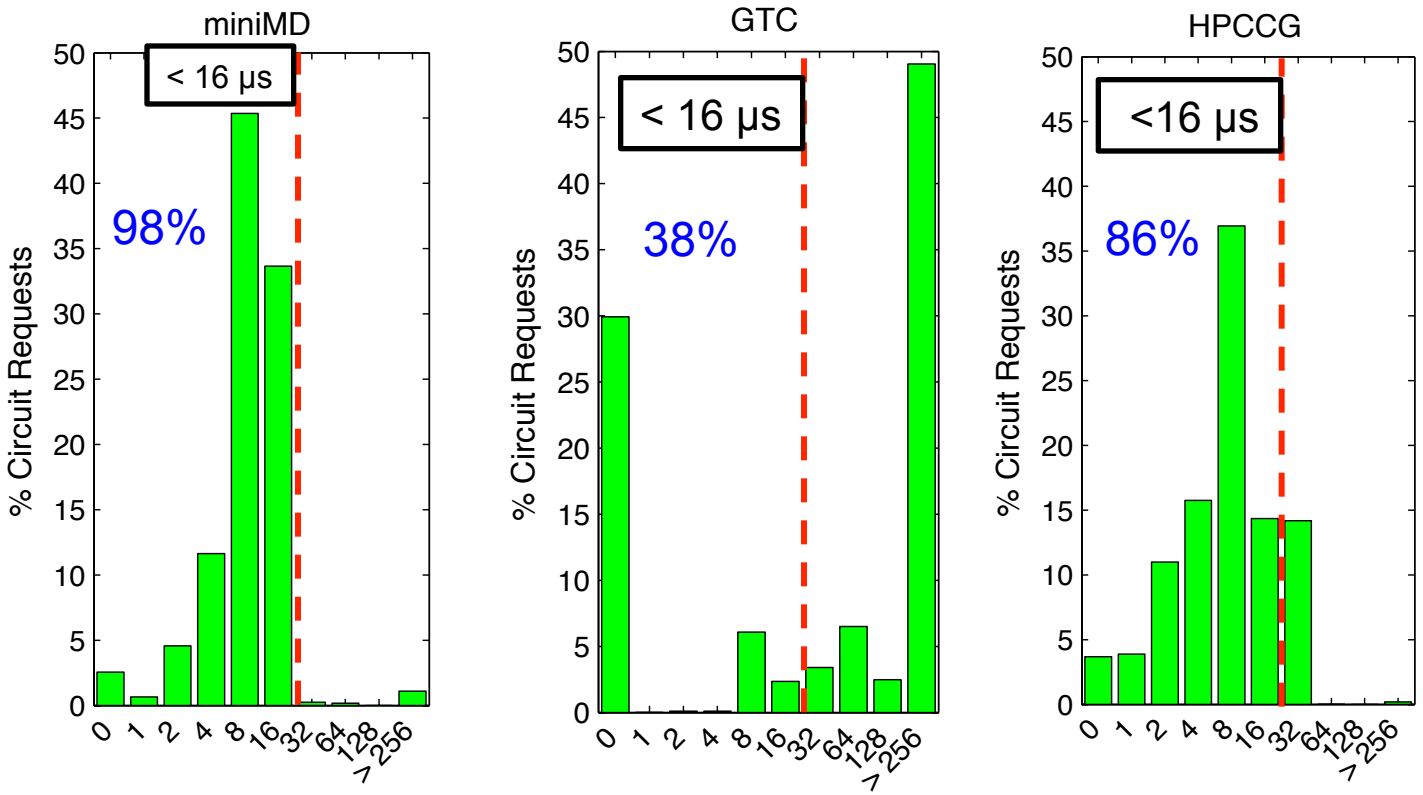
Thermal (re)-initialization delay: **~ 10-100 µs** ≈ **10^4-10^5 CPU cycles**





How Often is an Optical Circuit Needed?

* Profiled based on *Mantevo*[1] mini-apps (64 nodes, 1 process/node)



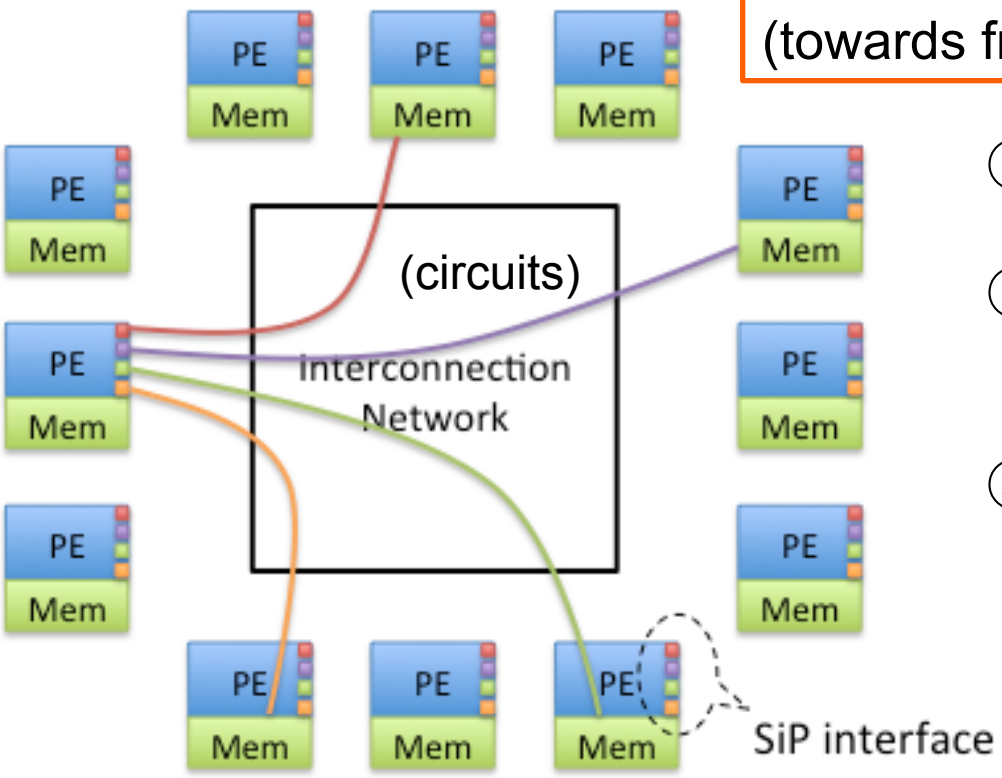
Time Interval (µs) Between Two Requests for the ***Same*** Circuit

Reuse interval ≤ circuit setup time ⇨ Need to keep circuits alive

[1] http://mantevo.org

Circuit-Maintained Architecture

Each node maintains X circuits.
(towards frequently accessed destinations).



① Circuit hit
→ time of flight
② Circuit miss
→ penalty from circuit setup & thermal initialization
③ Impossible to provision circuits towards all destinations
→ Need to replace circuits upon misses

- **Similar performance model to cache!**
- **Similar management requirement to cache!**

Architecture Design Space

- *Goal:* Maximize *circuit hit rate* in a cost effective way

  Def: % of requests that see an available circuit immediately

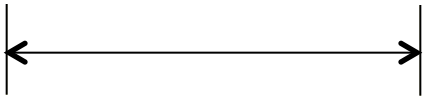
- *Design space*:
  - Number of circuits per node (cache size)
  - Replacement policy

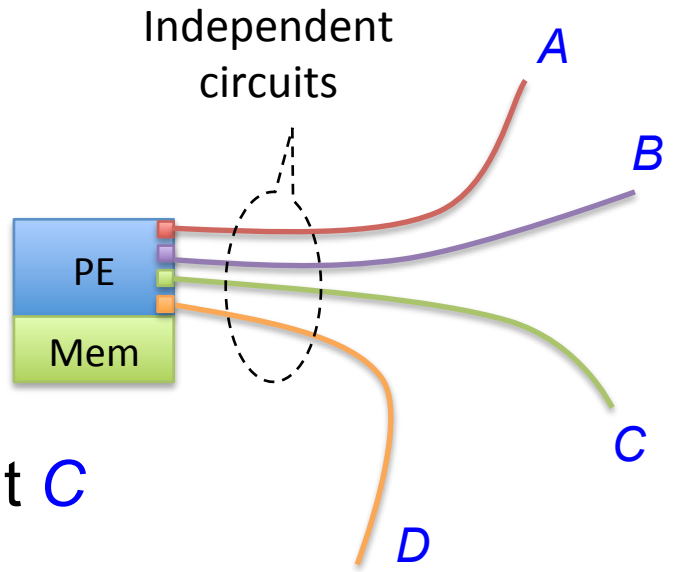
Inspired by Cache Modeling

- ***Reuse Distance*** captures how often a circuit is reused.
- Example:

  Circuits used by a node in order:

  C, A, B, D, B, C

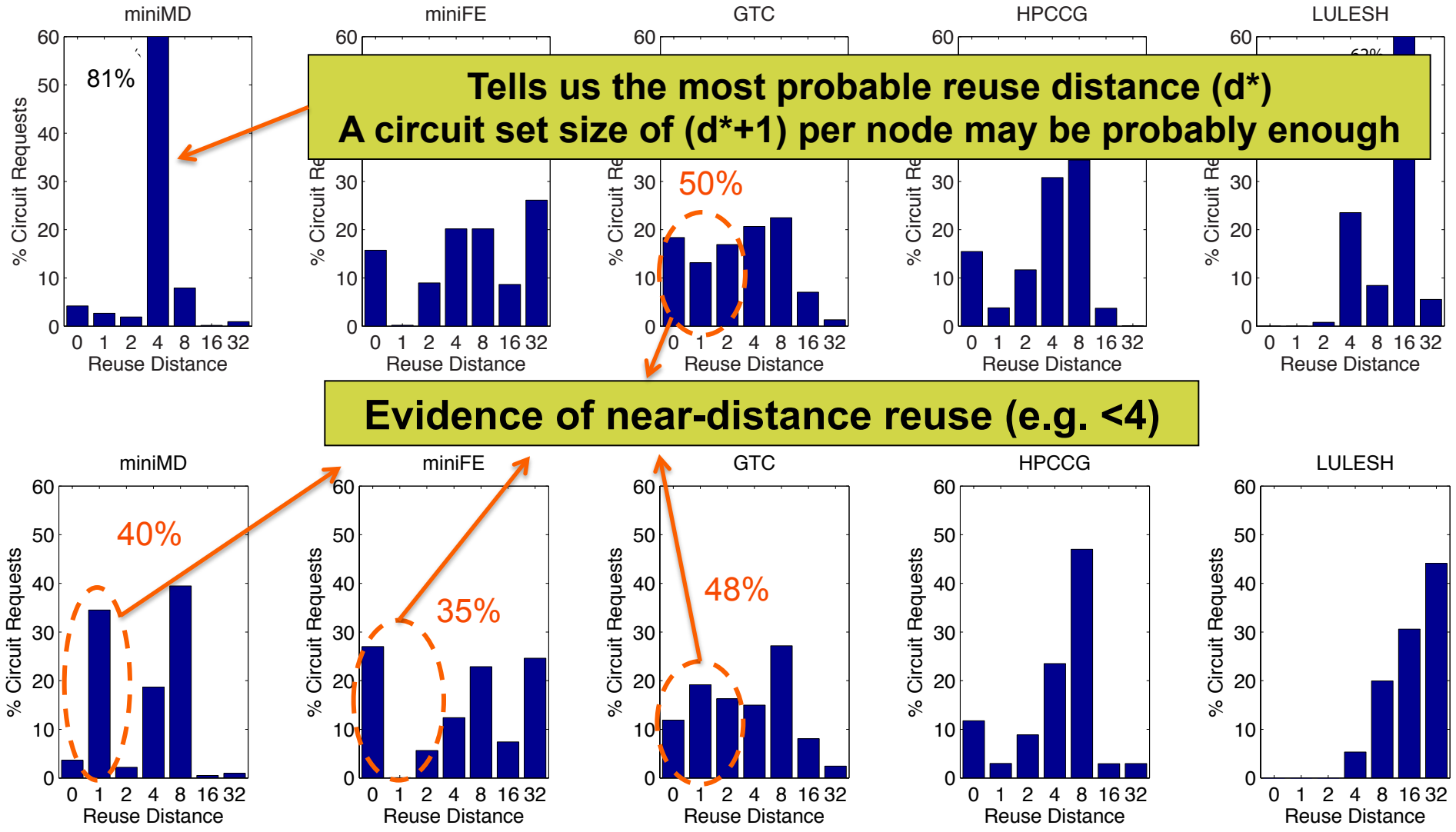




Reuse Distance (RD) = 4 for circuit *C*

- **Small reuse distance → The circuit is often reused.**
- **Large reuse distance → The circuit will not be used in the near future.**

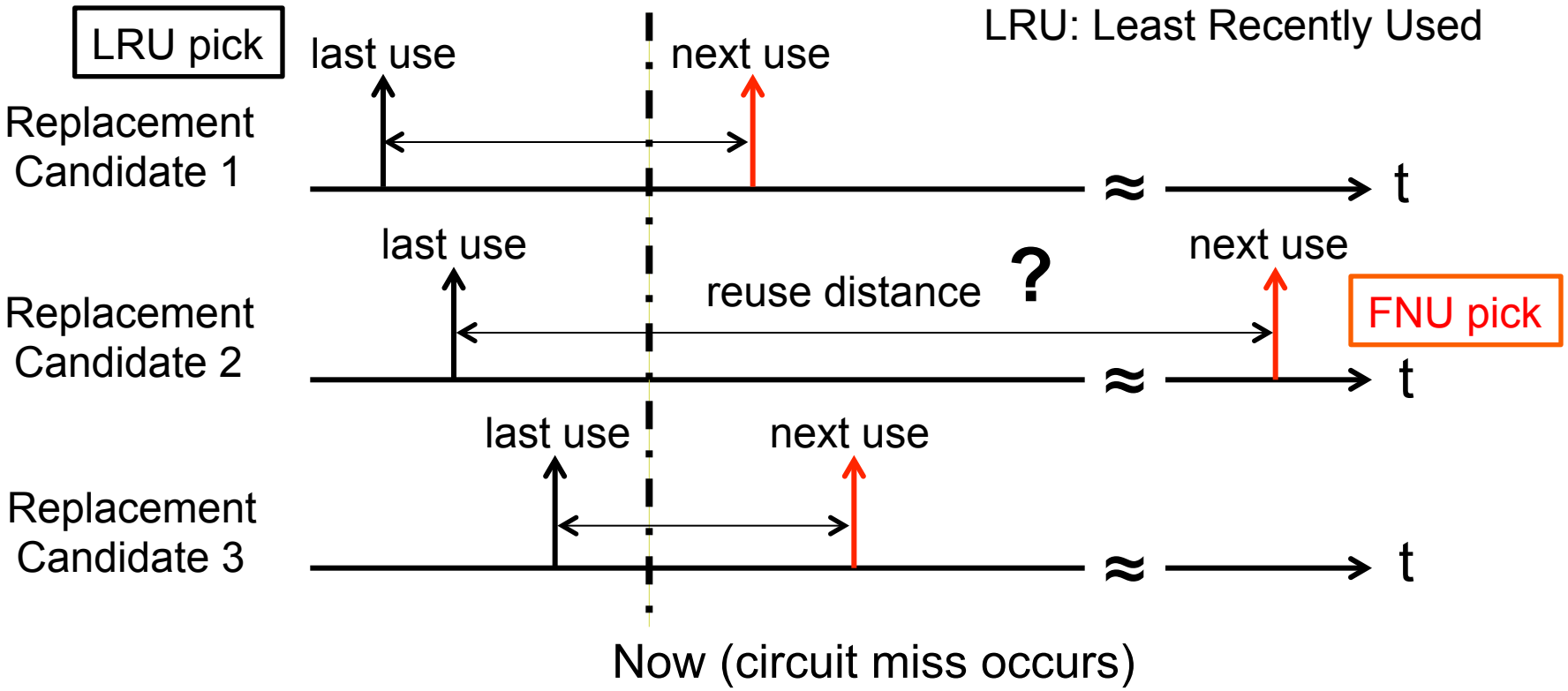
Distribution of Circuit Reuse Distance in Apps



Replacement policy: Farthest Next Use (FNU)

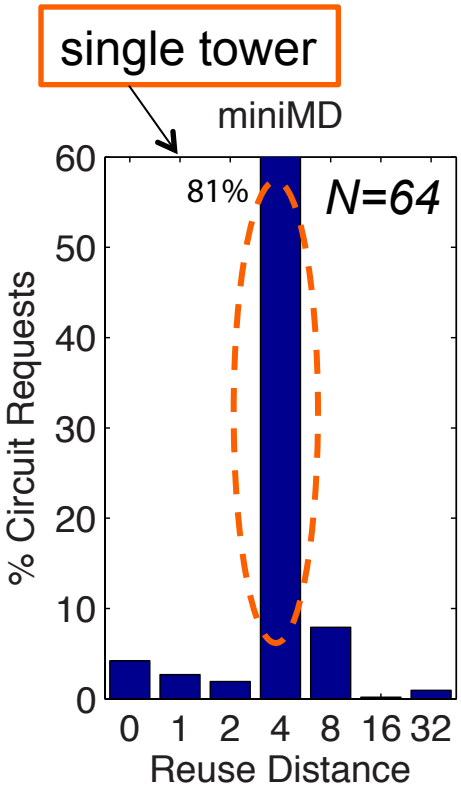
- Replace the circuit that is to be used in the farthest future

Observation is Not Enough, Prediction is What Matters

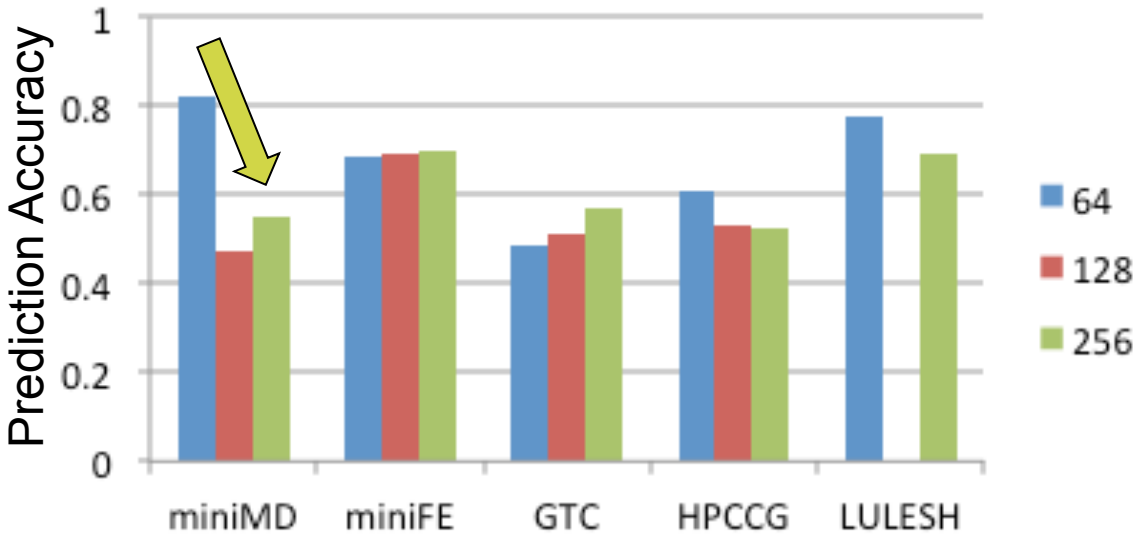


Prediction Method 1: Maximum Likelihood Prediction

Select reuse distance that has the highest frequency



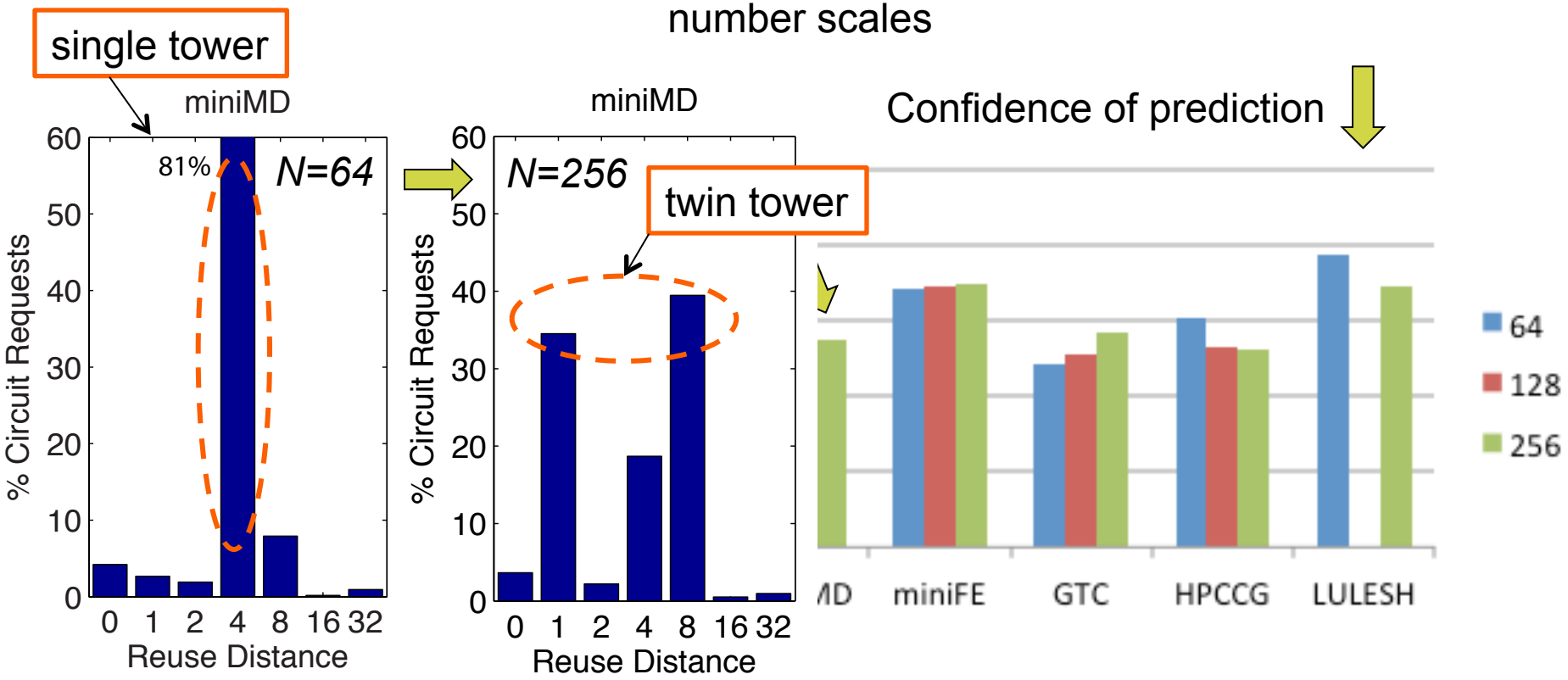
Prediction accuracy drops when node number scales



Prediction Method 1: Maximum Likelihood Prediction

Select reuse distance that has the highest frequency

Prediction accuracy drops when node number scales



Prediction Method 2: Temporal Transition Based Prediction

- Repeated communication patterns due to loops/iterations
- Example:

  RD samples observed over time for a circuit:

9 6 1 6 1 6 1 2 2 2 2 2 9 6 1 6 . . .

Repeated Transitions of RDs:

Prediction Method 2: Temporal Transition Based Prediction

- Repeated communication patterns due to loops/iterations
- Example:

  RD samples observed over time for a circuit:

9 6 1 6 1 6 1 2 2 2 2 2 9 6 1 6 . . .

Repeated Transitions of RDs:

Prediction Method 2: Temporal Transition Based Prediction

- Repeated communication patterns due to loops/iterations
- Example:

  RD samples observed over time for a circuit:

9 6 1 6 1 6 1 2 2 2 2 2 9 6 1 6 . . .

Repeated Transitions of RDs:

Prediction Method 2: Temporal Transition Based Prediction

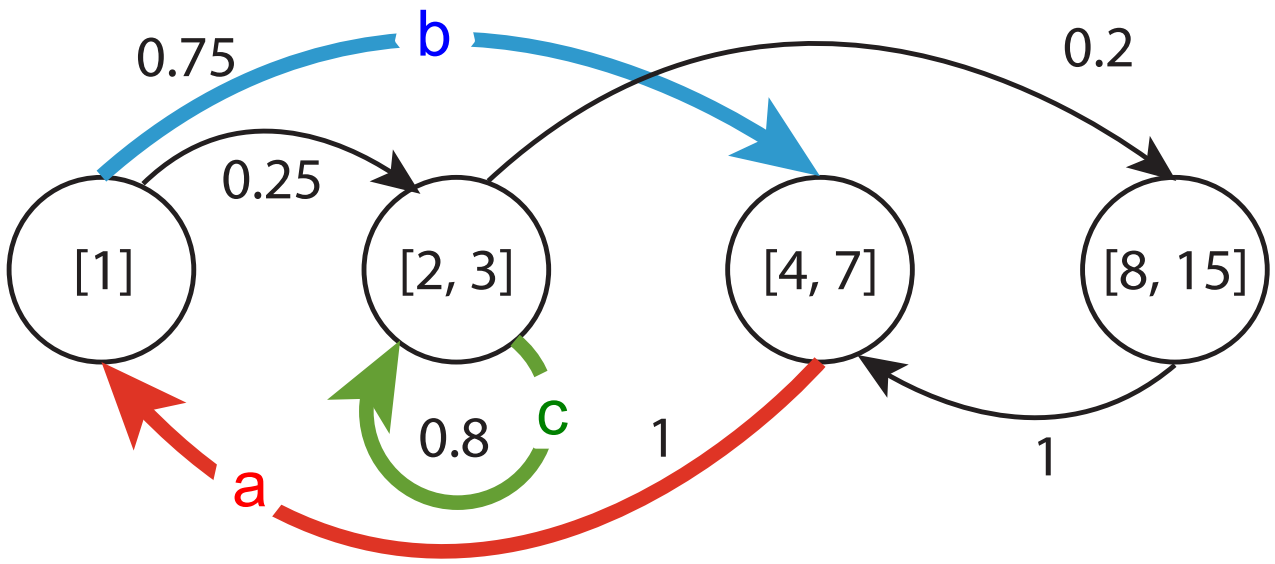
- Repeated communication patterns due to loops/iterations
- Example:

  RD samples observed over time for a circuit:

9 6 1 6 1 6 1 2 2 2 2 2 9 6 1 6 . . .

t

Repeated Transitions of RDs:



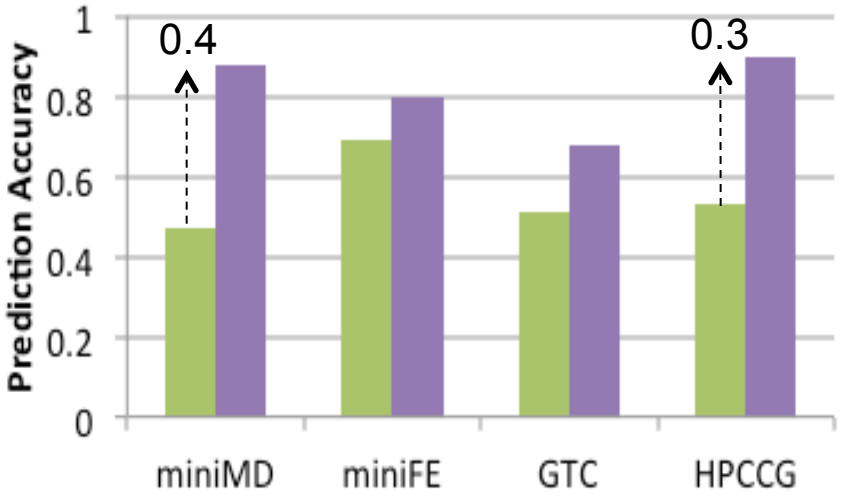
Upon prediction:

- Select the bin to which the current one has the highest transition probability

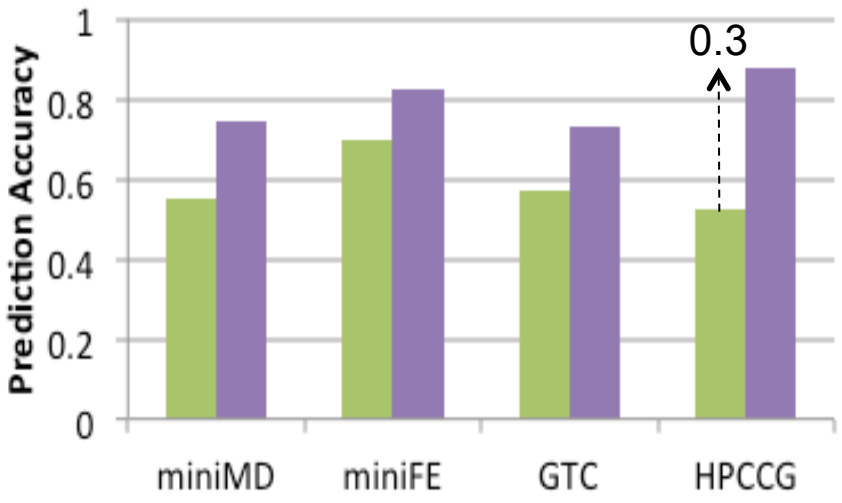
Prediction Accuracy Comparison

Max Likelihood Predictor

Temporal Transition Predictor



128 nodes



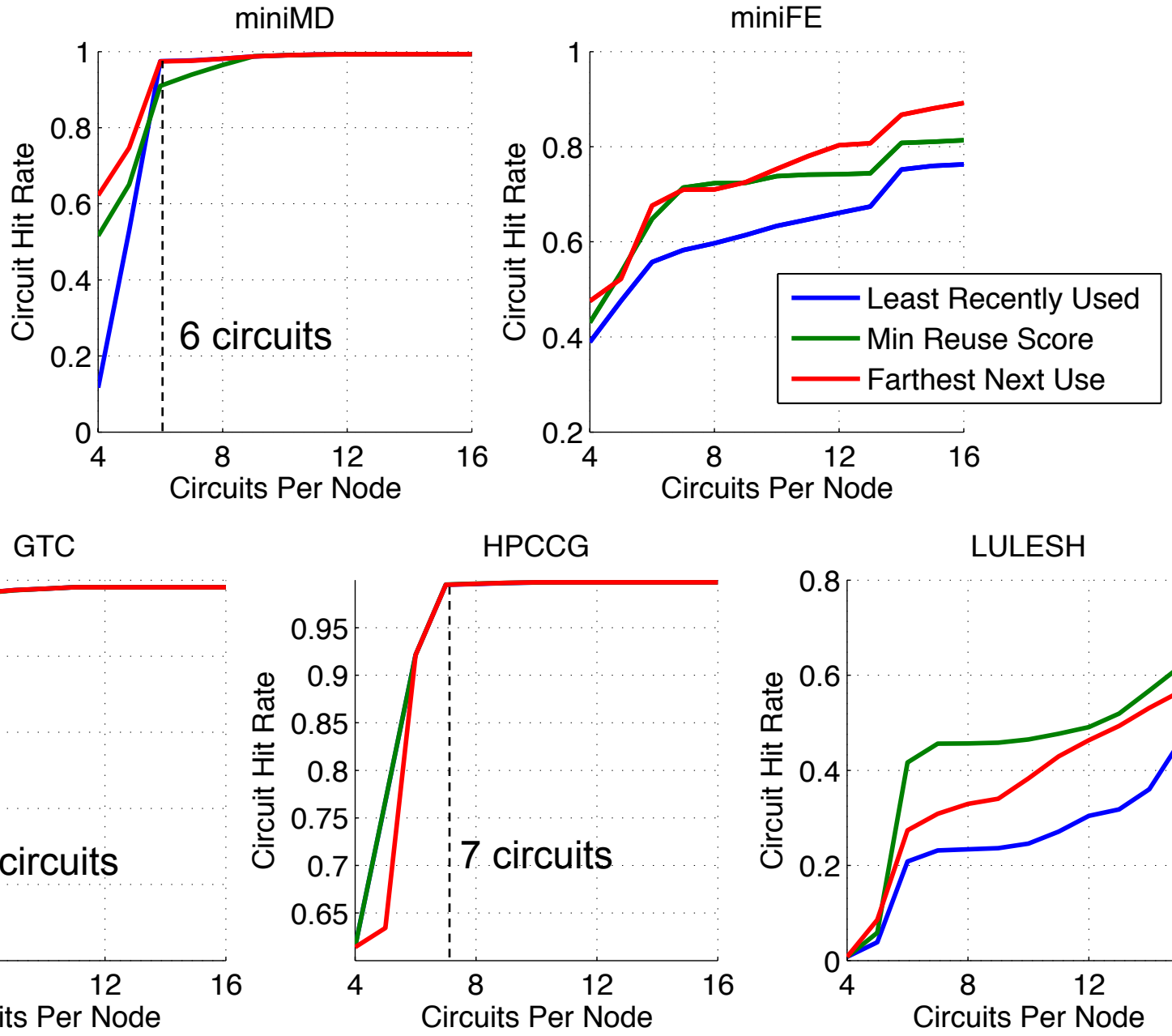
256 nodes

Replacement Policy Design and Comparison

- Base line: Least Recently Used (LRU) ***Based on Recent Past***
- Minimum Reuse Score ***Based on Full Past***
  - Each circuit accumulates scores based on number of uses
  - Close-distance reuse has a higher score than long-distance ones
  - Replace the circuit that has the minimum score
- Farthest Next Reuse ***Based on Prediction on Future***
  - Replace the circuit that is predicted to be used in the farthest future

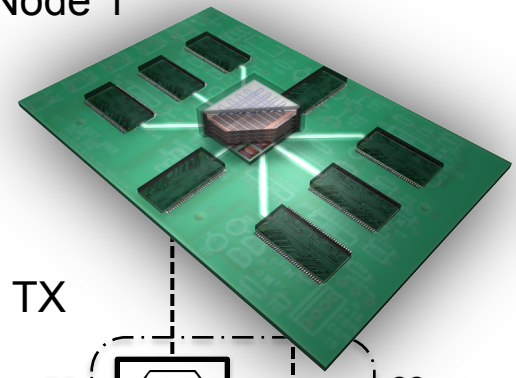
Evaluation:

- Co-simulate circuit management with mini-apps
- Leverage application skeletons to reduce simulation time
- Analyze impact of # of circuits per node


miniMD
Circuit Hit Rate
Circuits Per Node
6 circuits
miniFE
Circuit Hit Rate
Circuits Per Node
Least Recently Used
Min Reuse Score
Farthest Next Use
GTC
Circuit Hit Rate
Circuits Per Node
7 circuits
HPCCG
Circuit Hit Rate
Circuits Per Node
7 circuits
LULESH
Circuit Hit Rate
Circuits Per Node

Physical Layer Demonstration

FPGA-Emulated
Compute Node 1



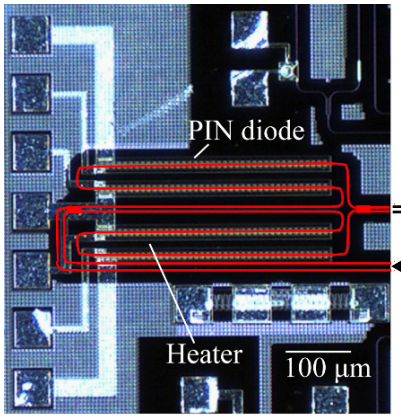
TX

1550 nm

1552 nm

50:50

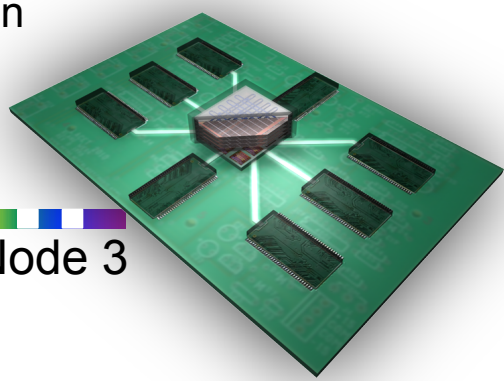
Commercial MZM



2x2 SiP
Mach Zehnder
Interferometer

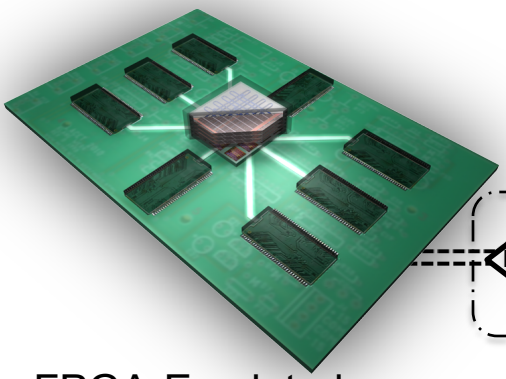
Out

In



To Node 3

FPGA
(Switch Control)



FPGA-Emulated
Compute Node 2

LA

PIN/TIA

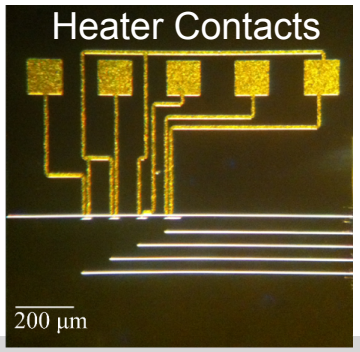
RX Chain

Voltage
Control

SiP WDM
Demux

In

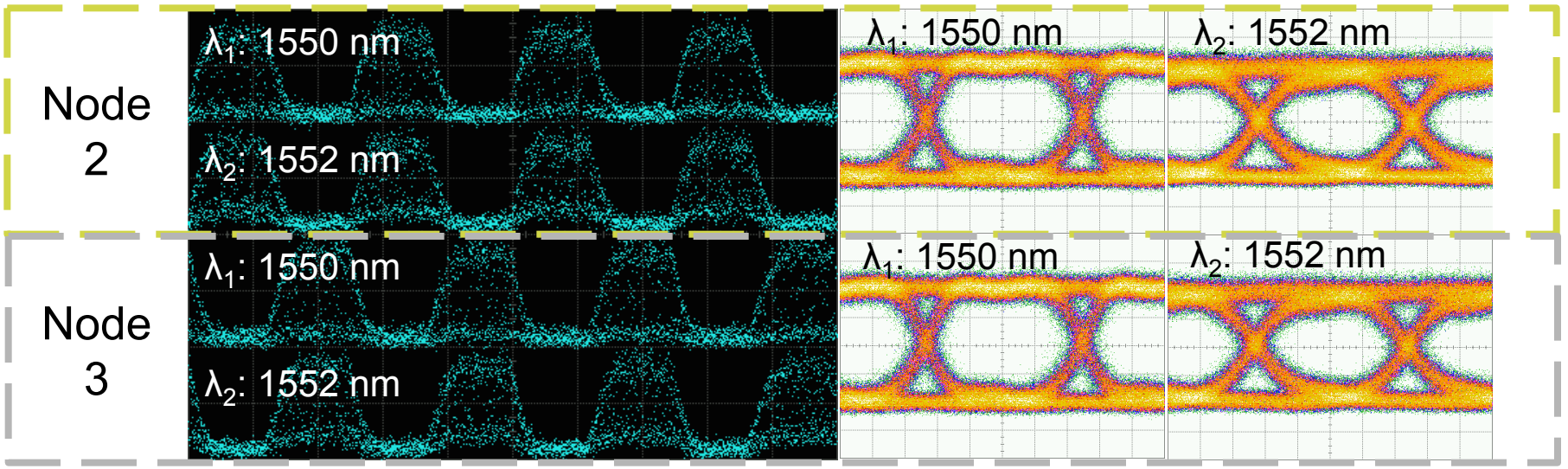
WDM
Demultiplexer



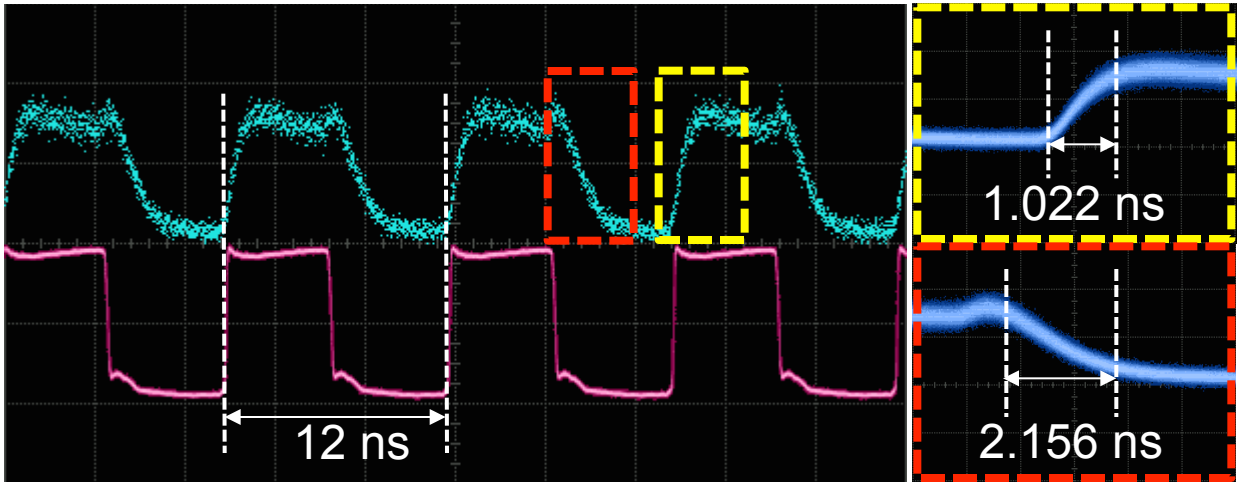
Through

Drop

Circuit Switching of Patterned Data



Optical vs. Electrical Response



Rise Settling Time

Fall Settling Time

Conclusion

- Silicon photonics can bring ultra-high bandwidth and energy efficiency to HPC in a cost-effective way.
- However, silicon photonics also has its challenges.
  - Circuit switching delay
  - Thermal sensitivity requires thermal initialization
  - Setup time couples with application reuse interval
- A circuit-maintained architecture mitigates such challenges.
  - Circuit reuses avoid setup penalty
  - Analogous to cache
  - Proposed prediction and replacement method show high circuit hit rate

Acknowledgement

- U.S. Department of Energy (DoE) National Nuclear Security Administration (NNSA) Advanced Simulation and Computing (ASC) program
- Sandia National Laboratories
- Portage Bay Photonics
- Gernot Pomrenke of the Air Force Office of Scientific Research