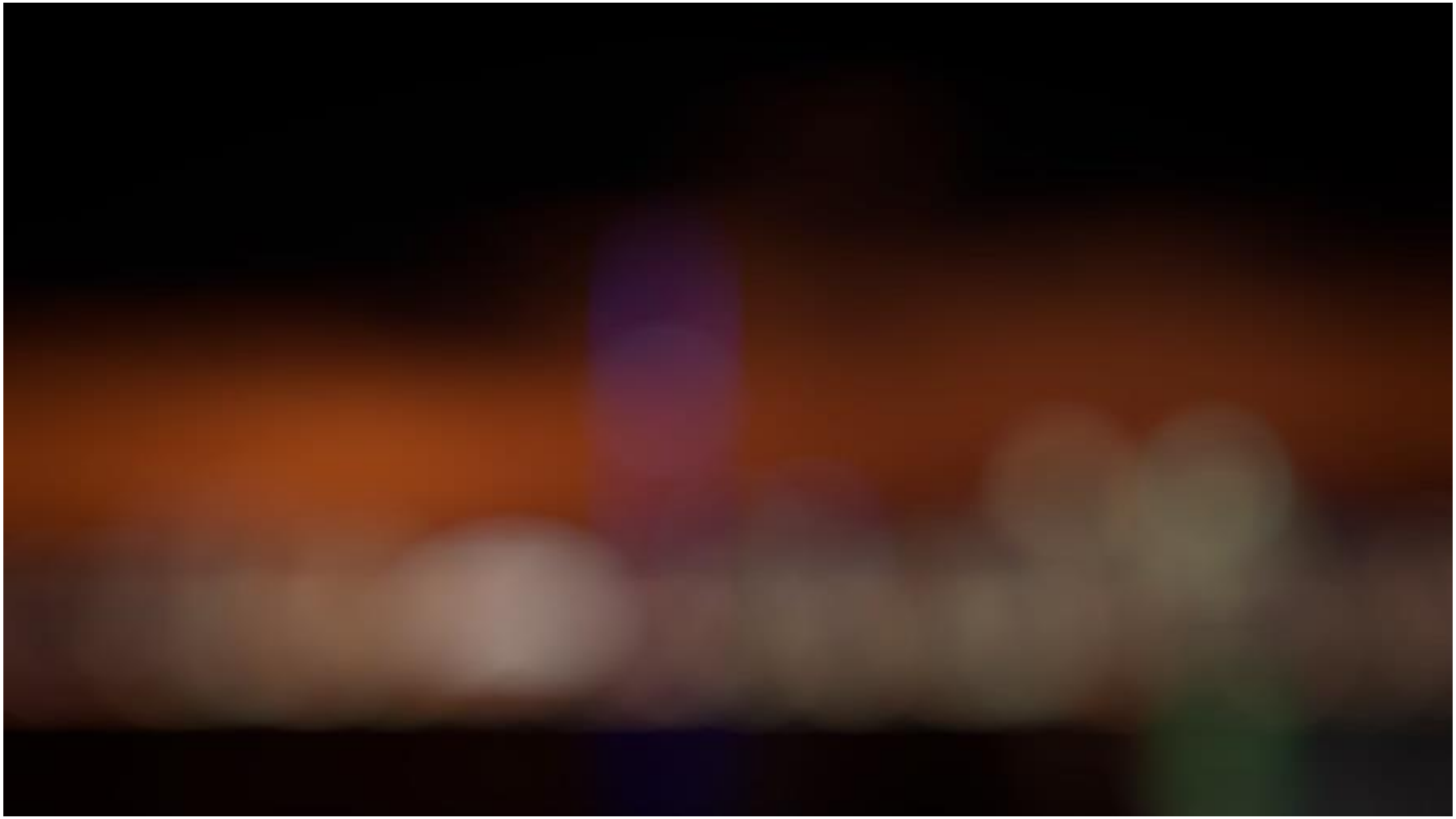
Synthetic **Full** System

Traffic Models Capturing Cache Coherence Behaviour

Mario Badr | Natalie Enright Jerger

UNIVERSITY OF TORONTO
FACULTY OF APPLIED SCIENCE & ENGINEERING

Amateur Photography



http://sevennine.net/archives/2009/10/01/torontos-skyline-at-night/

Photography 101

http://sevennine.net/archives/2009/10/01/torontos-skyline-at-night/

Hundreds of Pictures

SynFull

http://sevennine.net/archives/2009/10/01/torontos-skyline-at-night/

Pictures for Facebook

What Does SynFull Do?

- **Model** real application traffic to the NoC
- **Generate** realistic traffic synthetically for the NoC
- **Iterate** over several NoC designs quickly

Tool Available for Download

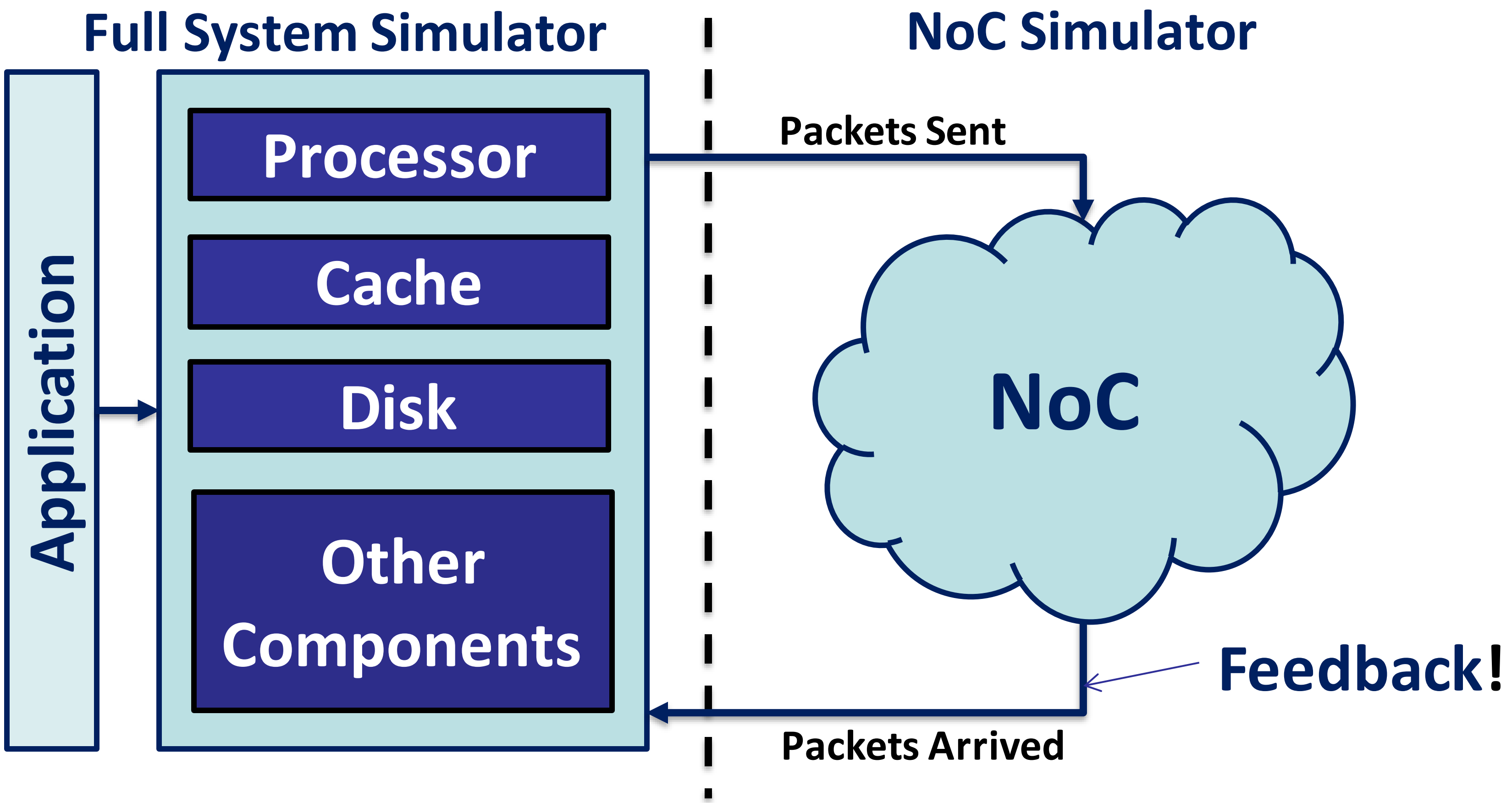
SynFull's Goals

- **Generic** – Current and future applications
  - 16 different benchmarks
- **Accurate** – Comparable performance metrics
  - 10.5% error
- **Fast** – Faster than full system and traces
  - 52x speed up

NoC Simulation Methodologies

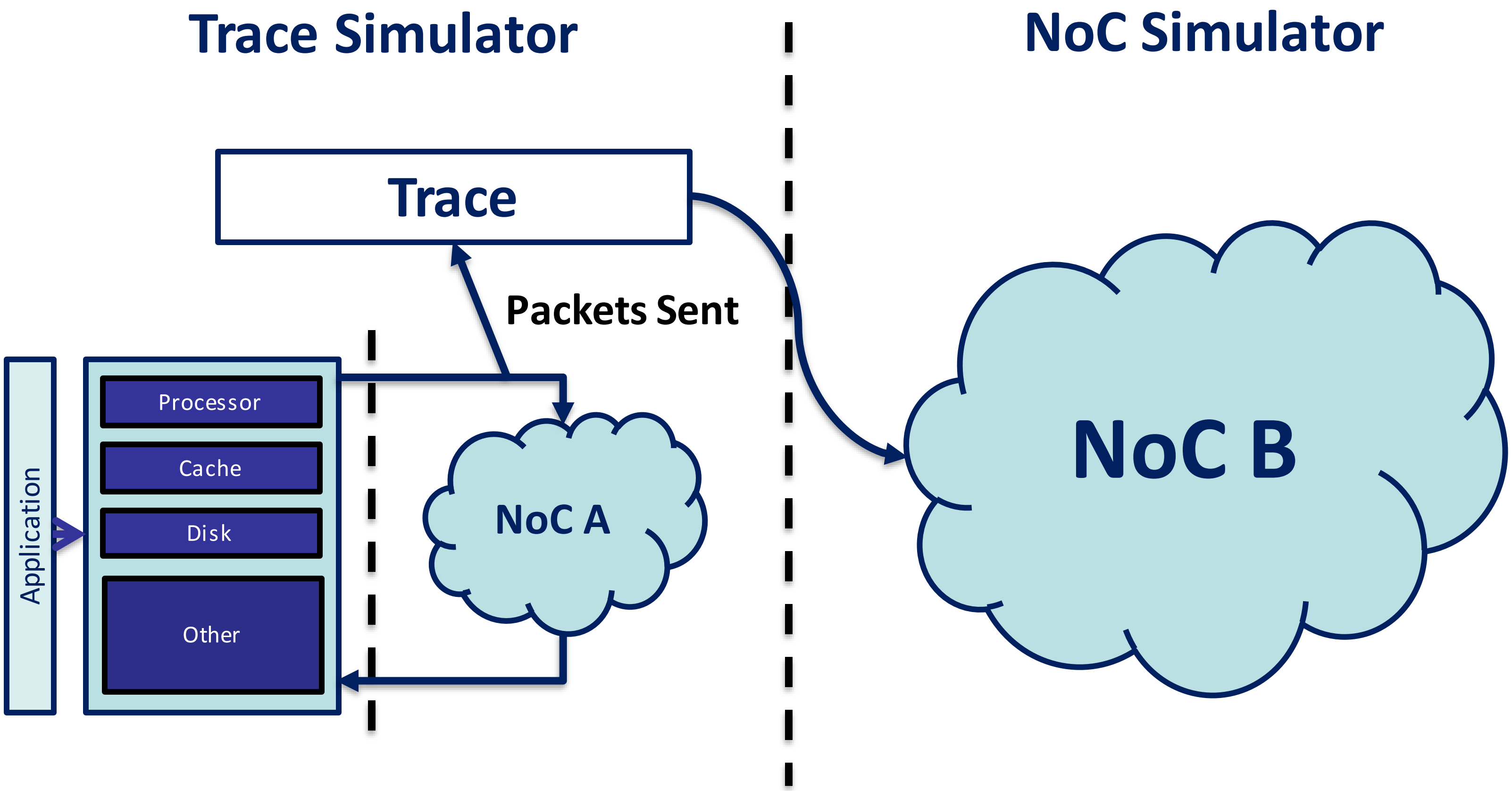
- **Full System**
- **Traces**
- **Traffic Patterns**

Full System Simulation



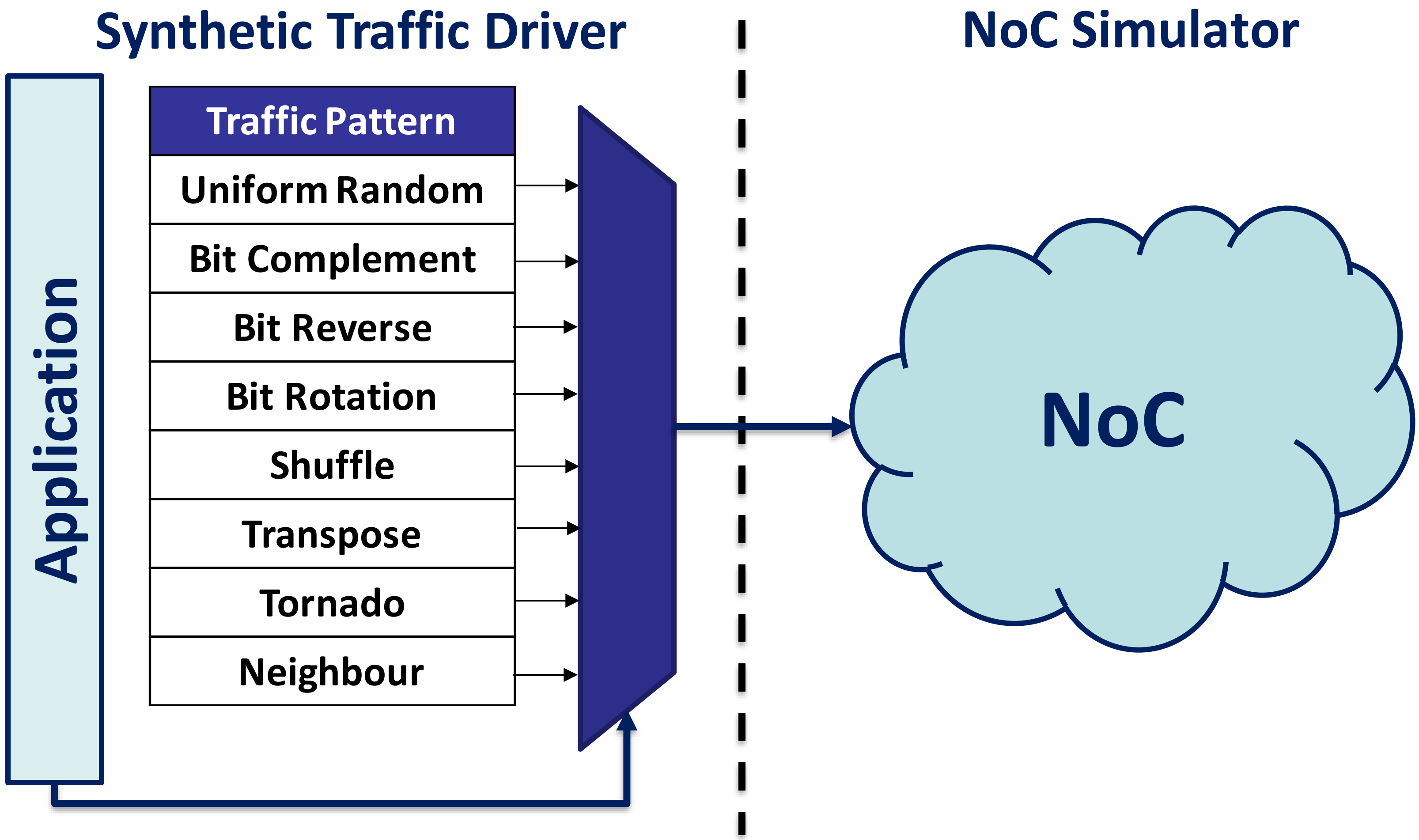
Accurate But ***Slow***

Trace Simulation



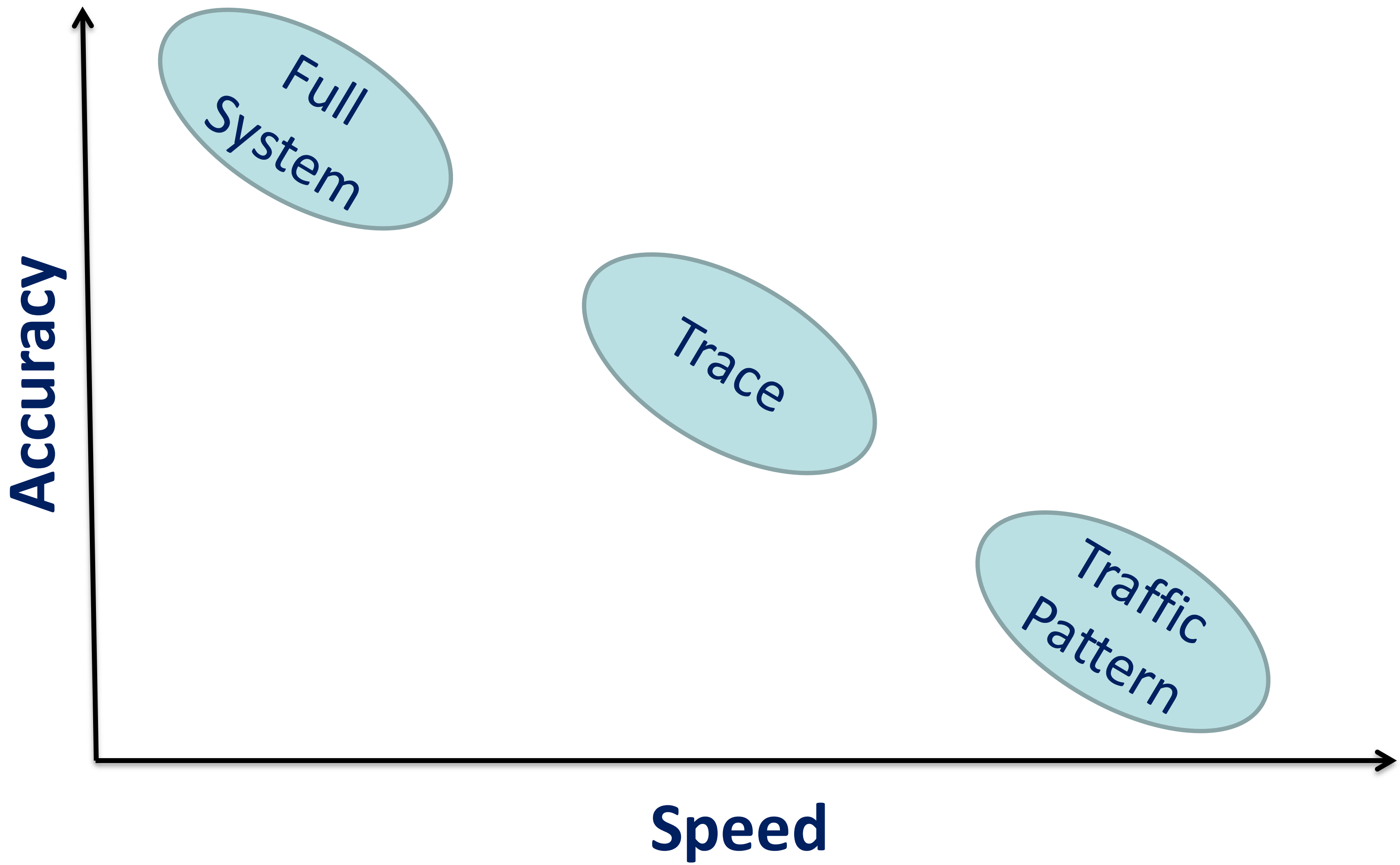
Faster But ***Less Accurate***

Traffic Patterns

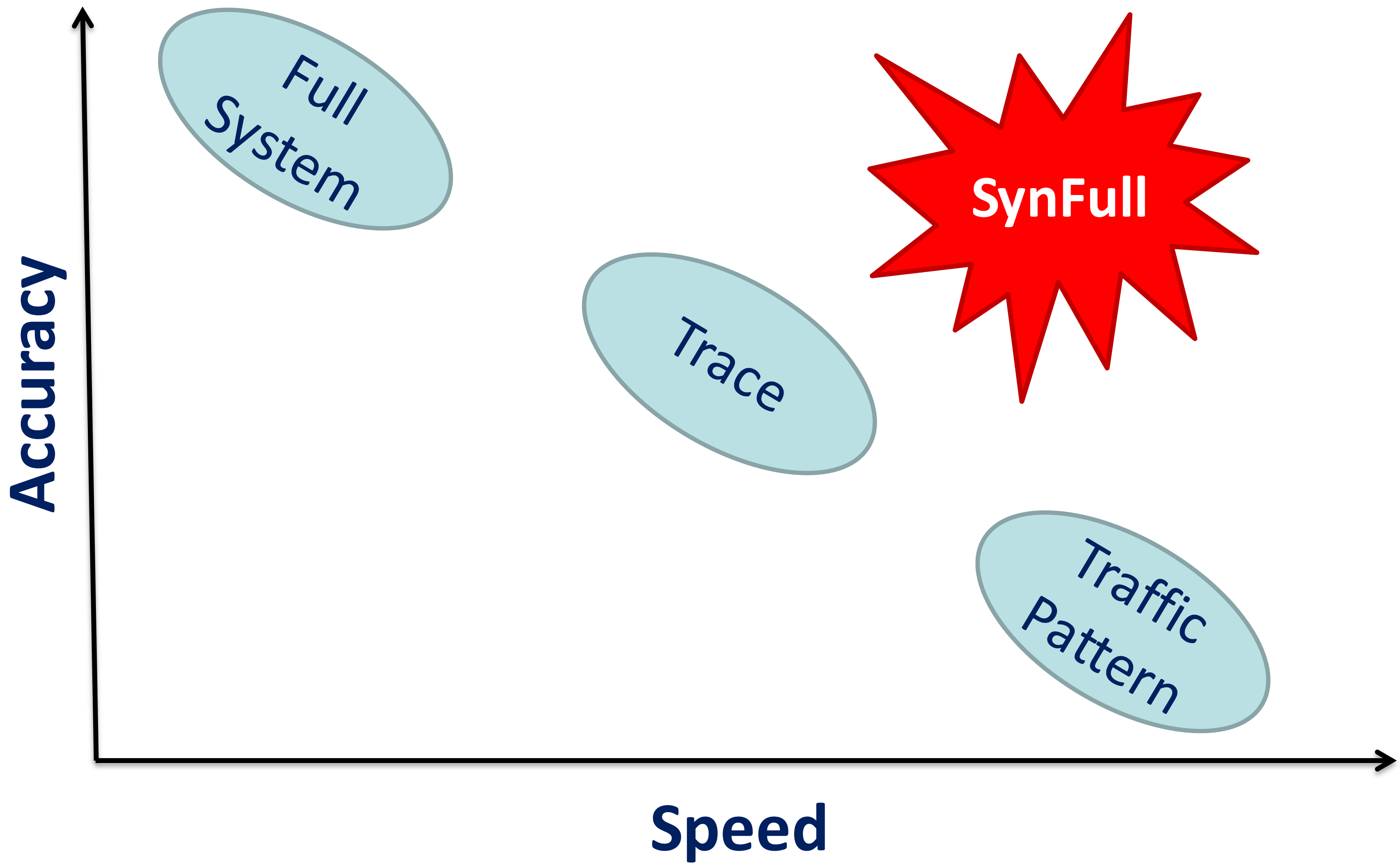


Very Fast But ***Inaccurate***

The Opportunity



The Opportunity



Achieving the Goals

- Synthetic Cache Coherence
  - Dependent Messages
  - Enable Research
- Time-Varying Behaviour
  - Short and Long Bursts of Traffic

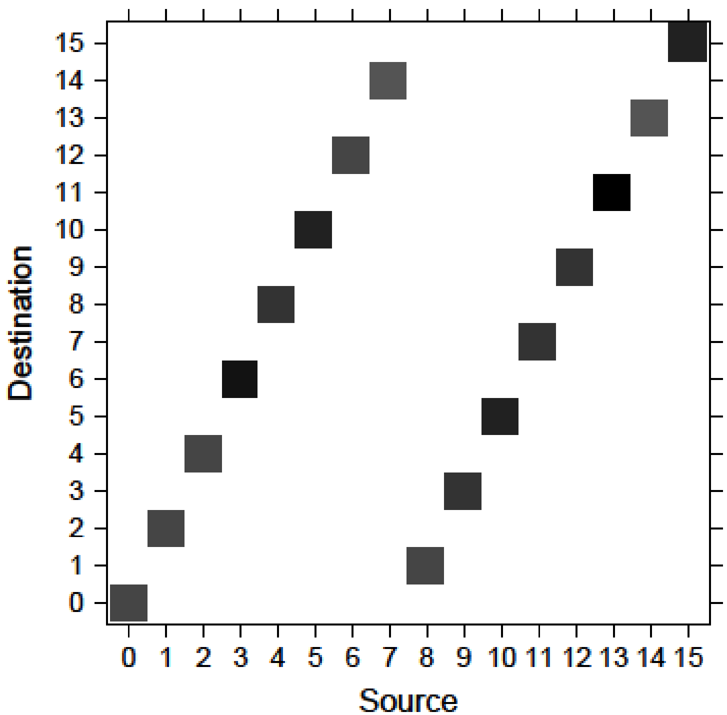
Accuracy

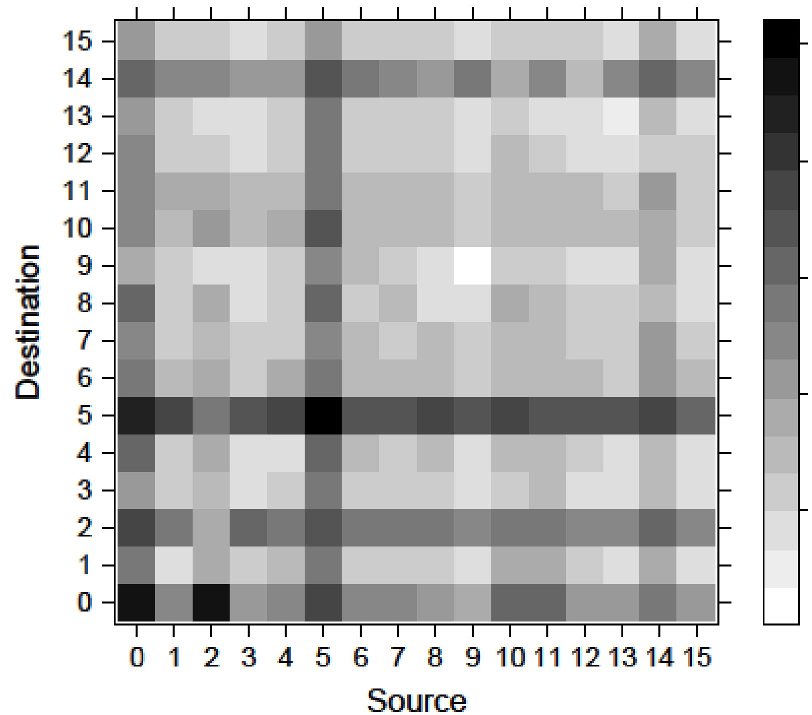
- Convergence
  - Simulation length?

Speed

Motivating Cache Coherence

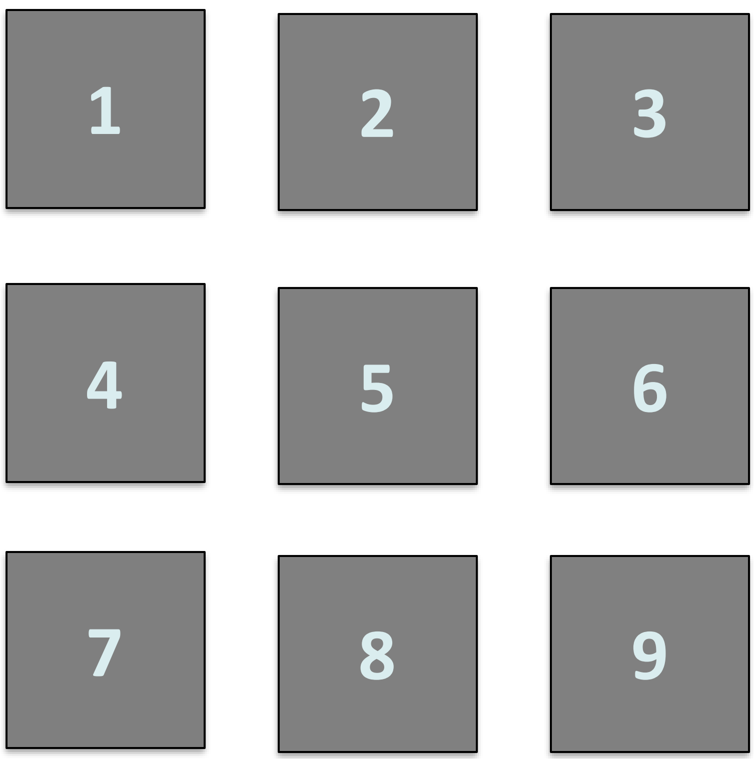
Shuffle → **Fast Fourier Transform**





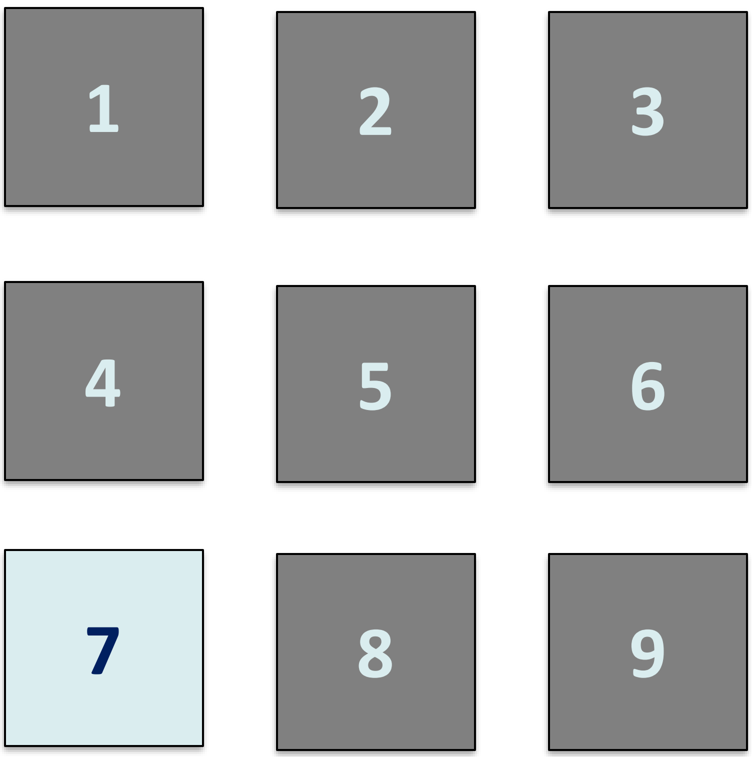
Cache Coherence Affects Traffic Behaviour

Capturing Coherence Traffic



- **Example**
  - **MOESI Protocol**
  - **Can be adapted**

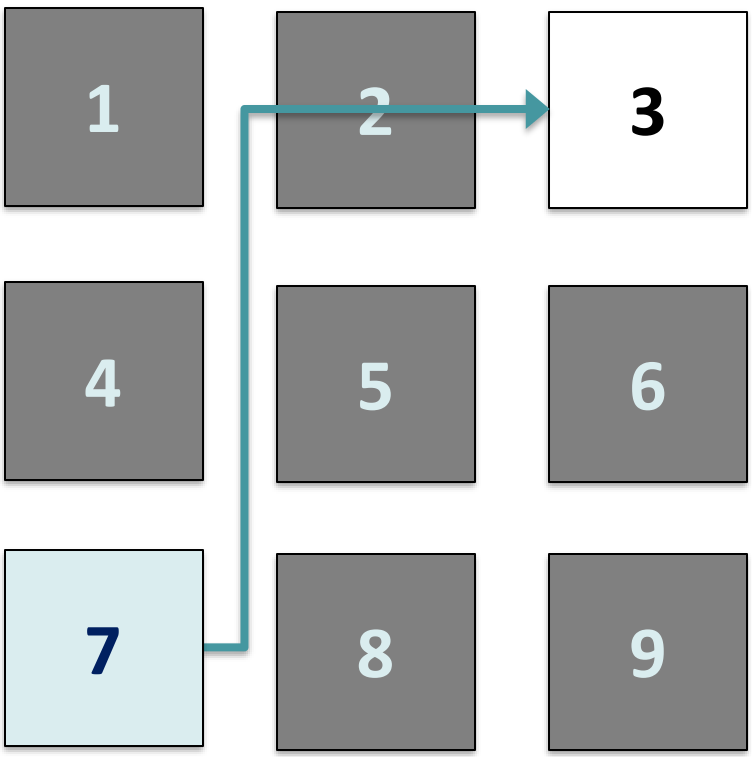
Capturing Coherence Traffic



Store Miss 7

- **Initiate Transaction**
- **Store Miss**
  - **Source**

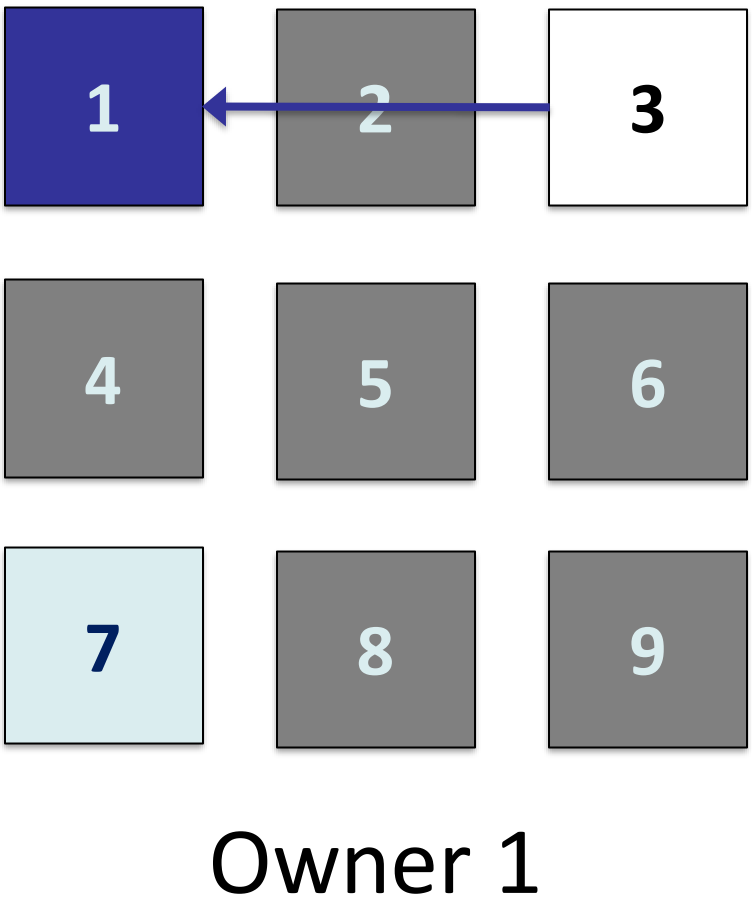
Capturing Coherence Traffic



Directory 3

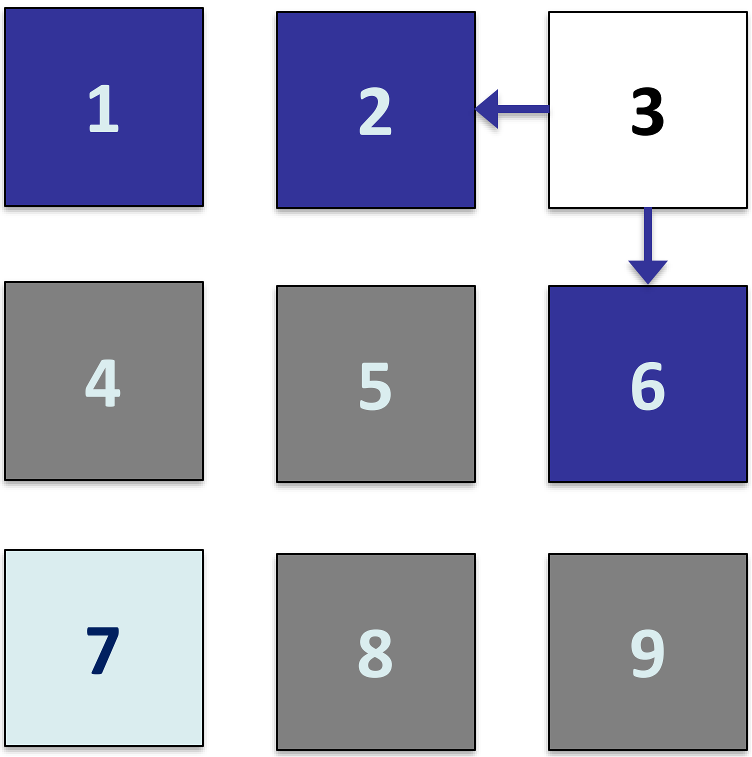
- **Store Miss**
  - **Source**
  - **Destination**

Capturing Coherence Traffic



- **Store Miss**
- **Forwarded Request**
  - **Destination**

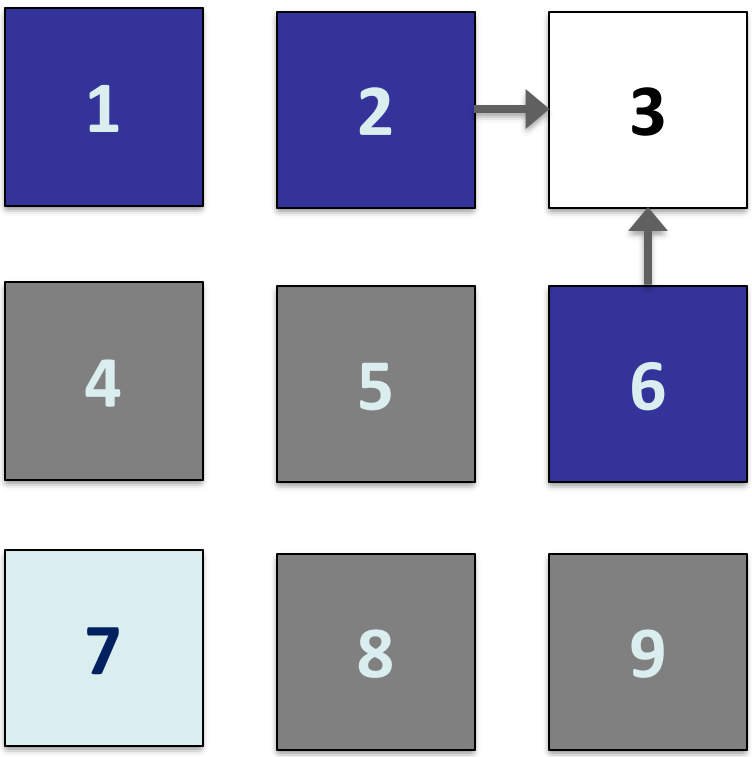
Capturing Coherence Traffic



Invalidate 2, 6

- Store Miss
- Forwarded Request
- Invalidations
  - Quantity
  - Destinations

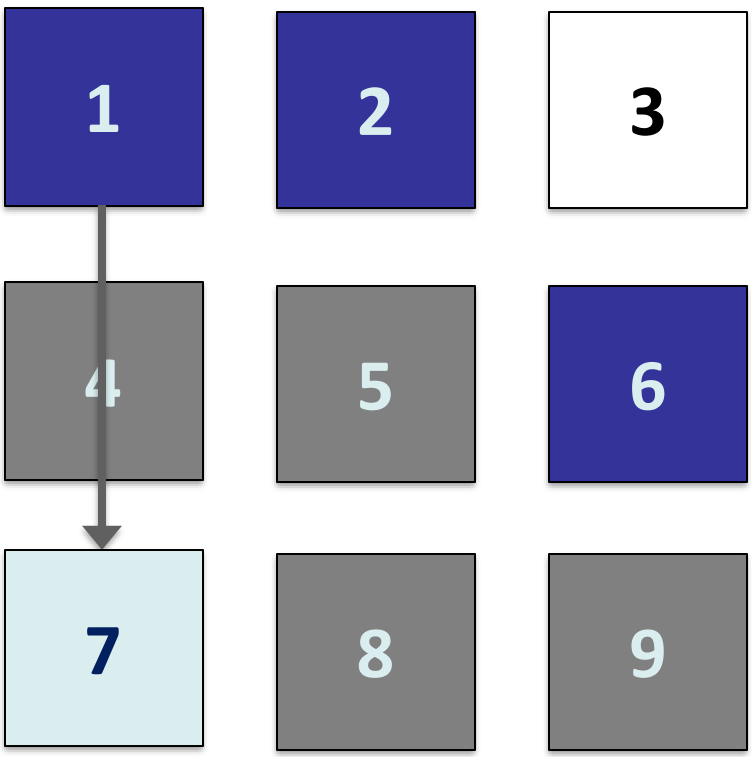
Capturing Coherence Traffic



ACKs 2, 6

- Store Miss
- Forwarded Request
- Invalidations
- Acknowledgements

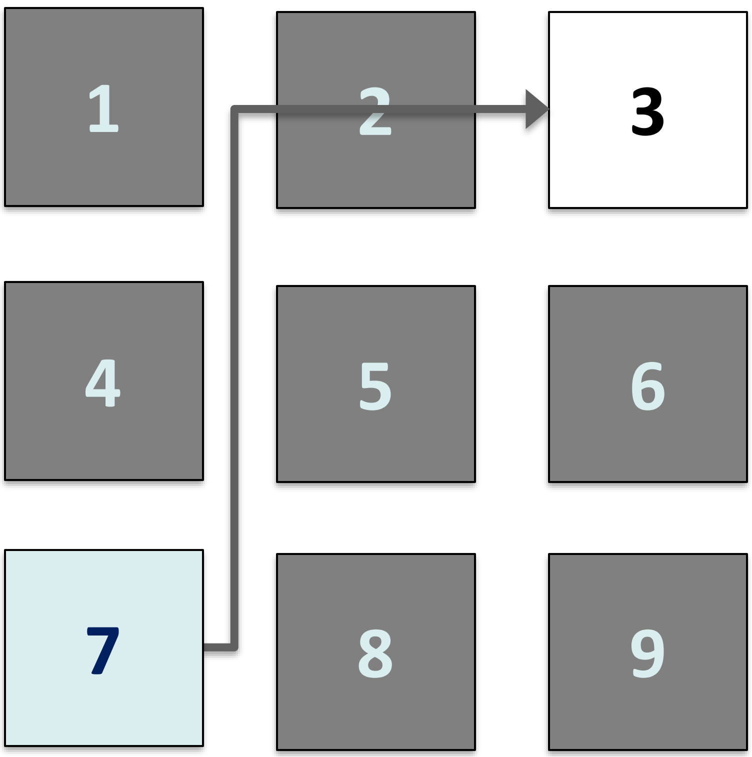
Capturing Coherence Traffic



Data to 7

- **Store Miss**
- **Forwarded Request**
- **Invalidations**
- **Acknowledgements**
- **Data Response**

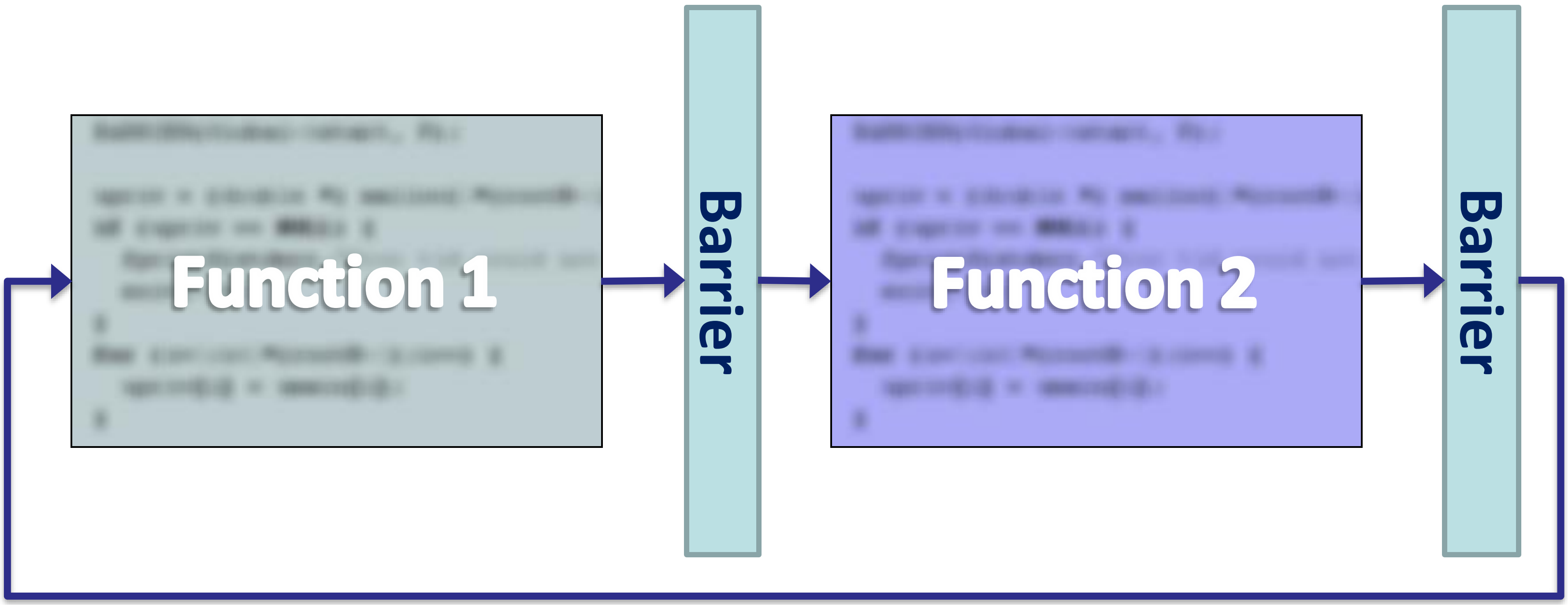
Capturing Coherence Traffic



Transaction Complete

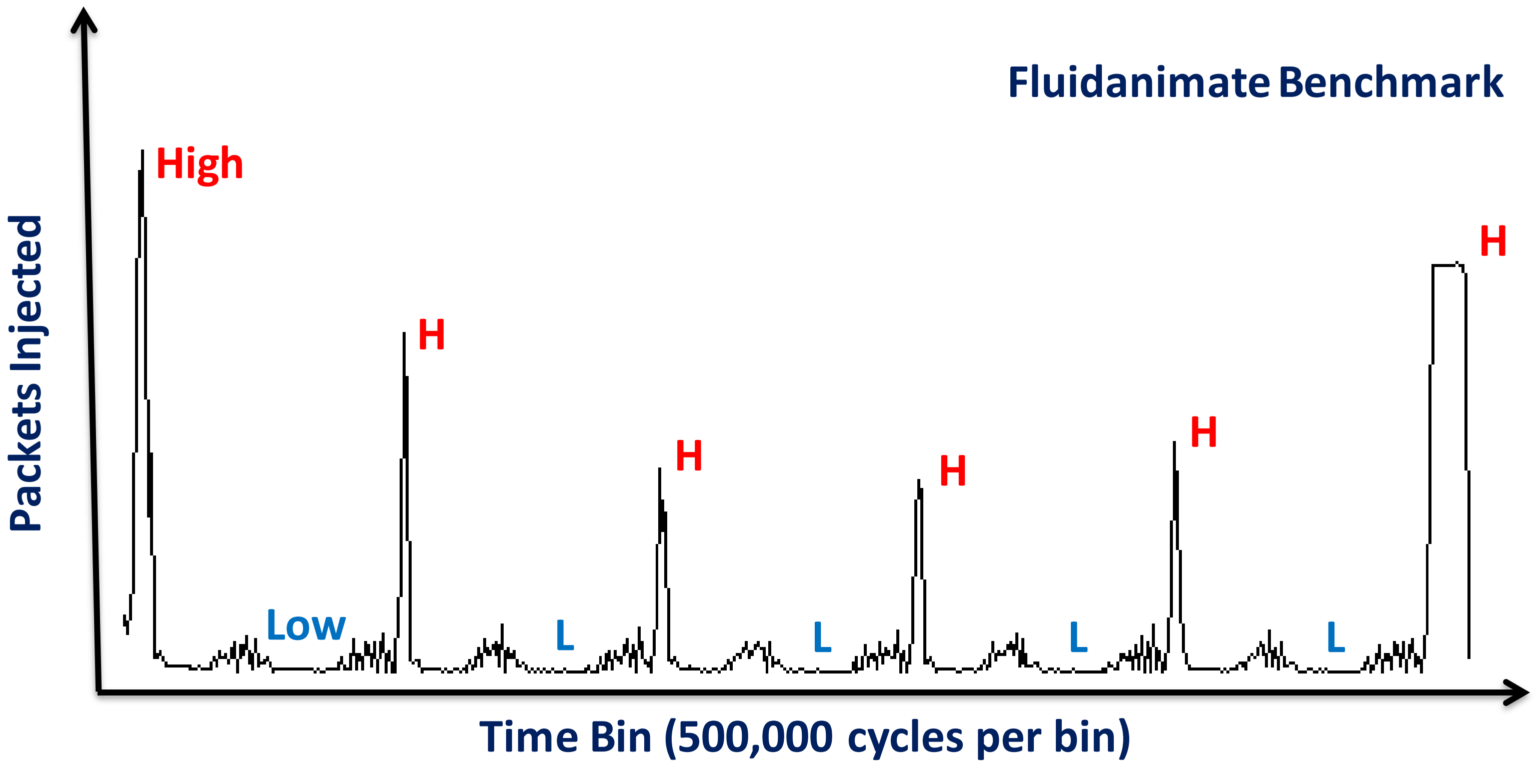
- **Store Miss**
- **Forwarded Request**
- **Invalidations**
- **Acknowledgements**
- **Data Response**
- **Unblock**

Time Varying Behaviour



Initiating Transactions and Sharing Patterns ***Can Change***

Time-Varying Behaviour

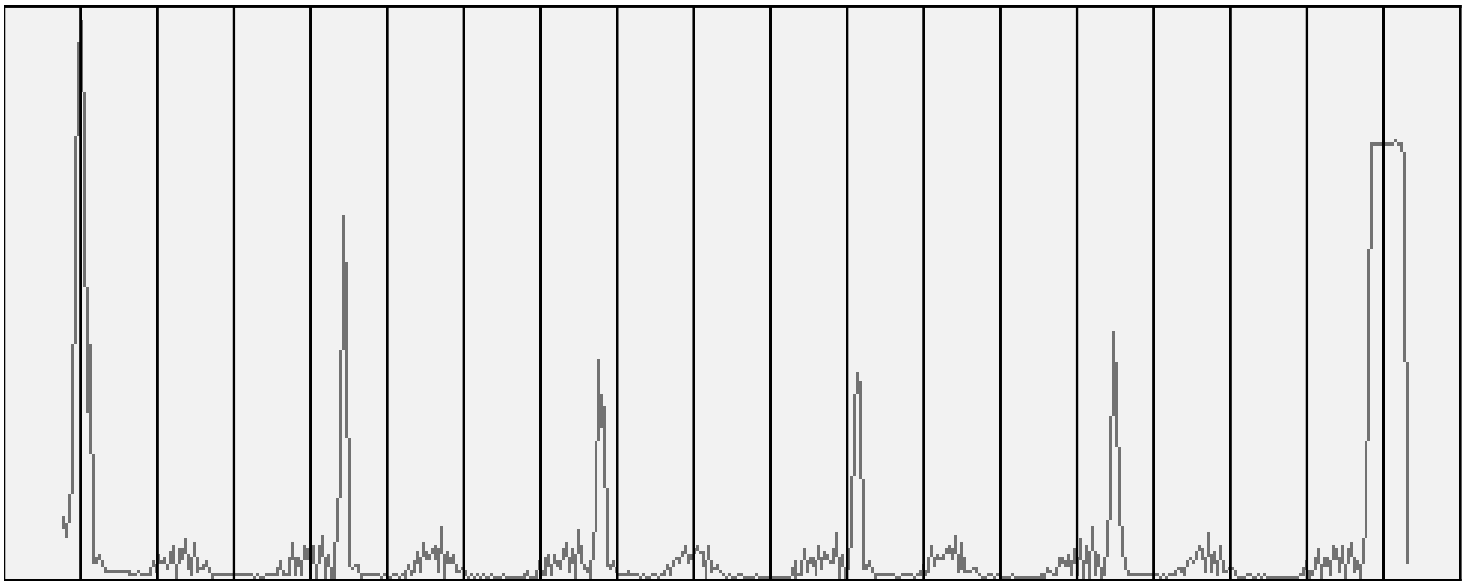


Applications go through ***phases***

Modelling Time-Varying Behaviour

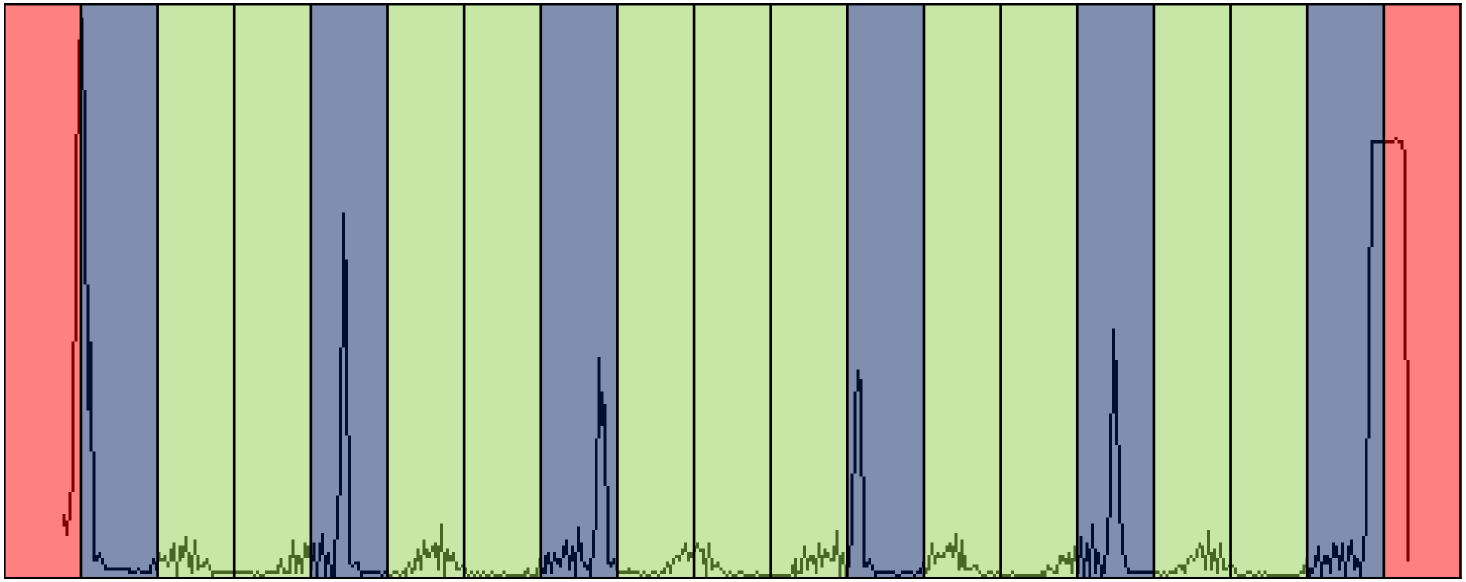
- Create and group phases
  - *Clustering*
- Transition from one phase to another
  - *Markov Chains*

Dividing Into Intervals



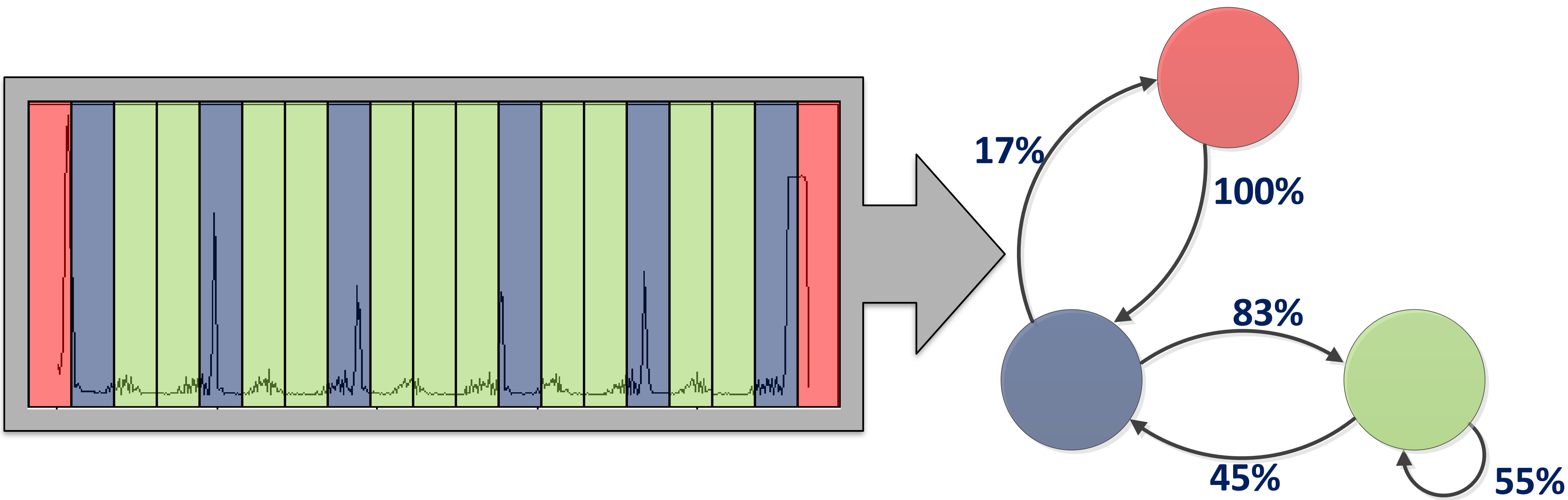
Intervals are a fixed size

Dividing Into Intervals



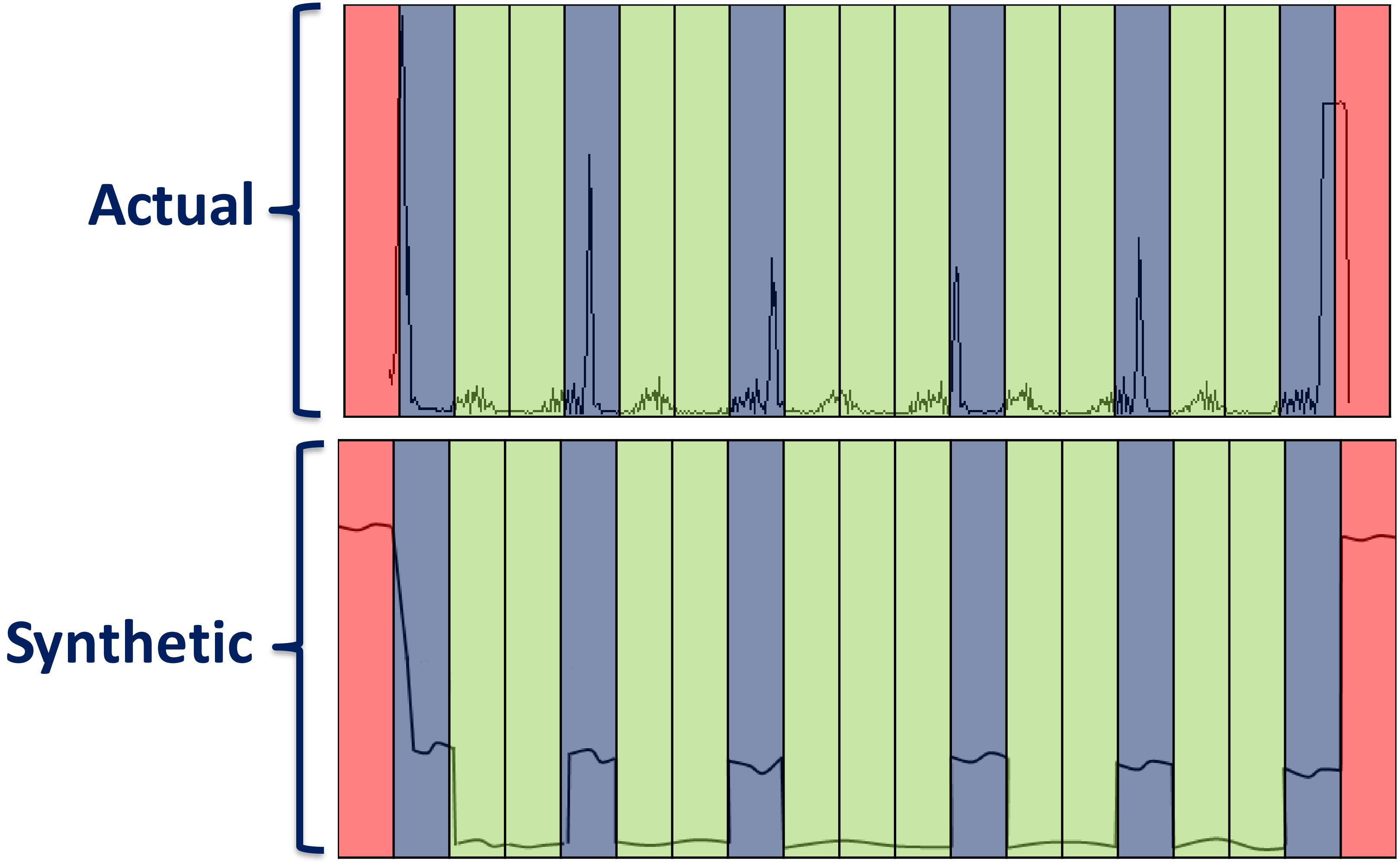
Visually we see: **High**, **Low + High**, and **Low** Intervals

Phase Transitions: Markov Chains



P[Next State | Current State]

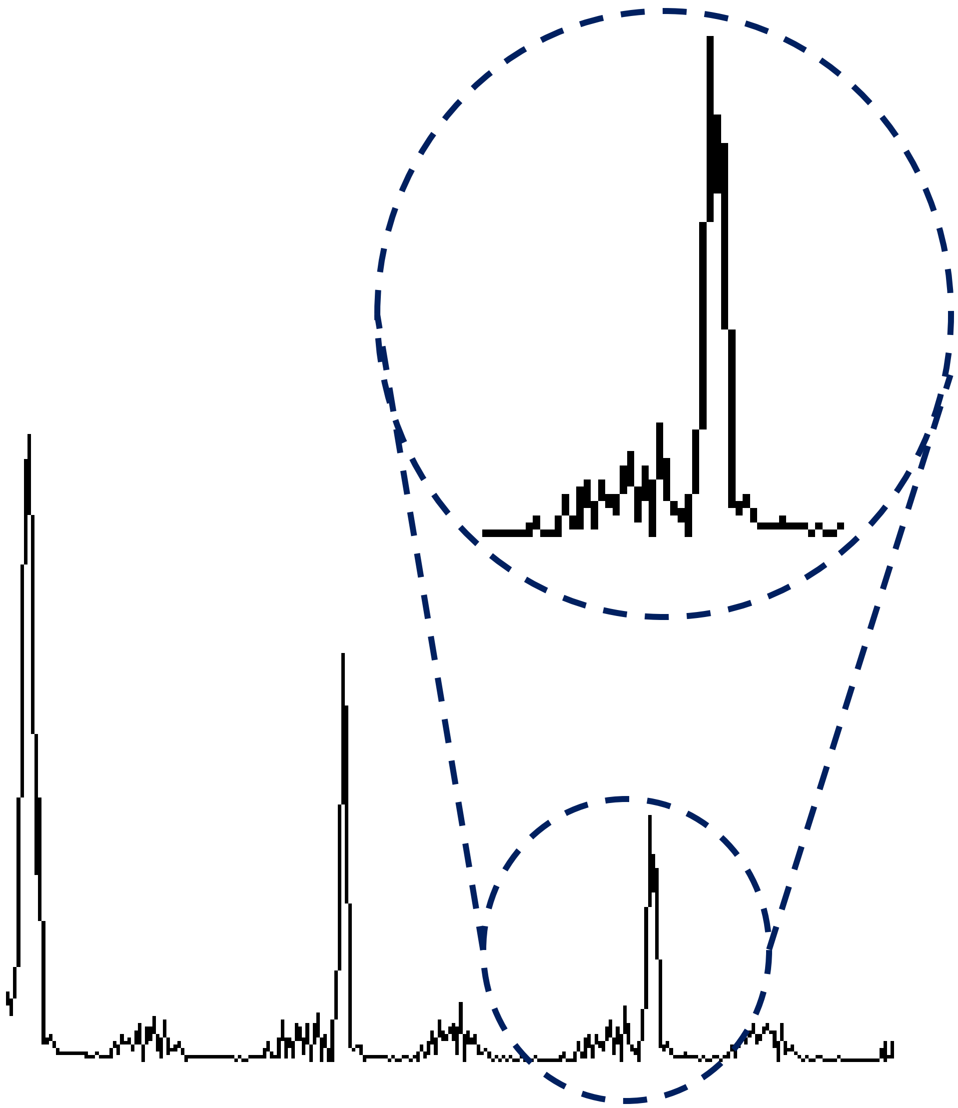
Traffic Comparison



Coarse Granularity → *Average* Behaviour

Capturing Short Bursts

- Macro Level
  - 100,000s of Cycles
  - Long phases
  - Outer-Loops
- Micro Level
  - 100s of Cycles
  - Short Bursts
  - Inner-Loops
- **Hierarchical Model**

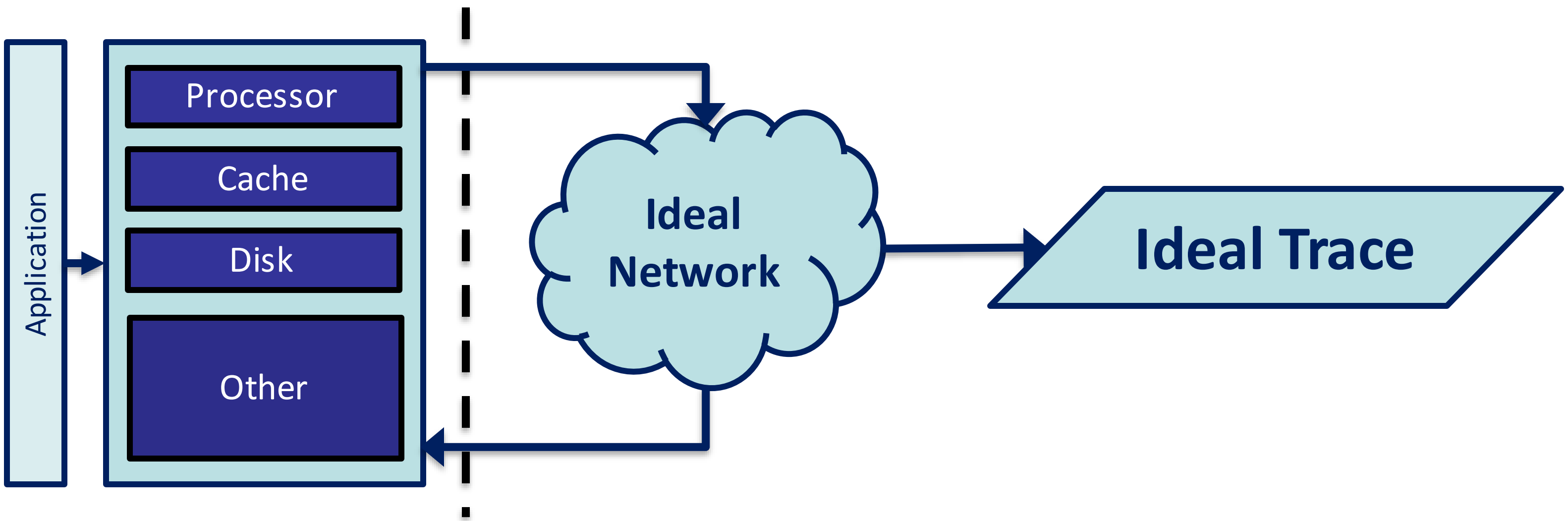


Modelling Parameters

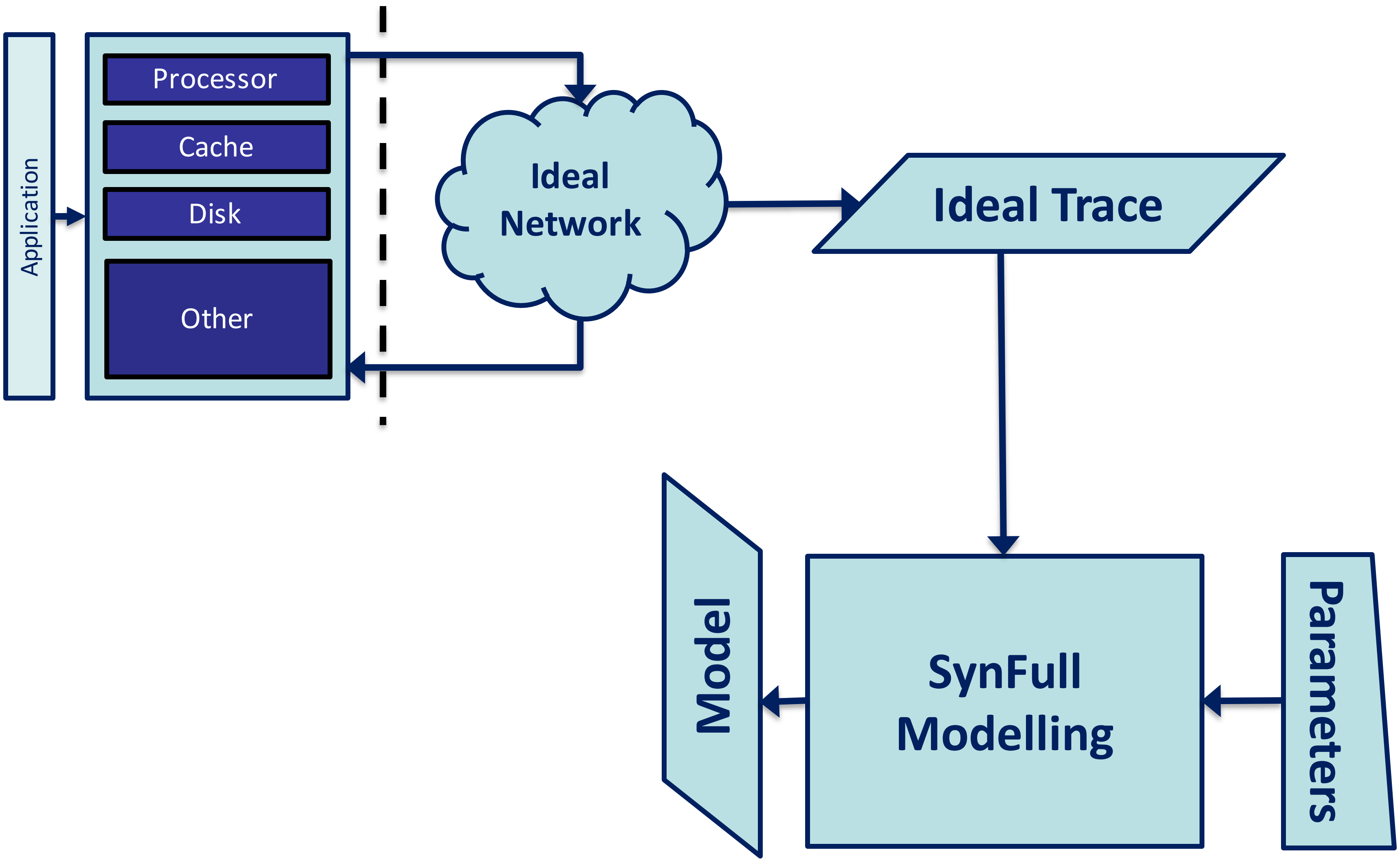
- Model accuracy affected by:
  - Interval Size
  - Interval Similarity
  - Number of Clusters

See Paper for Parameter Sweep & Recommendations

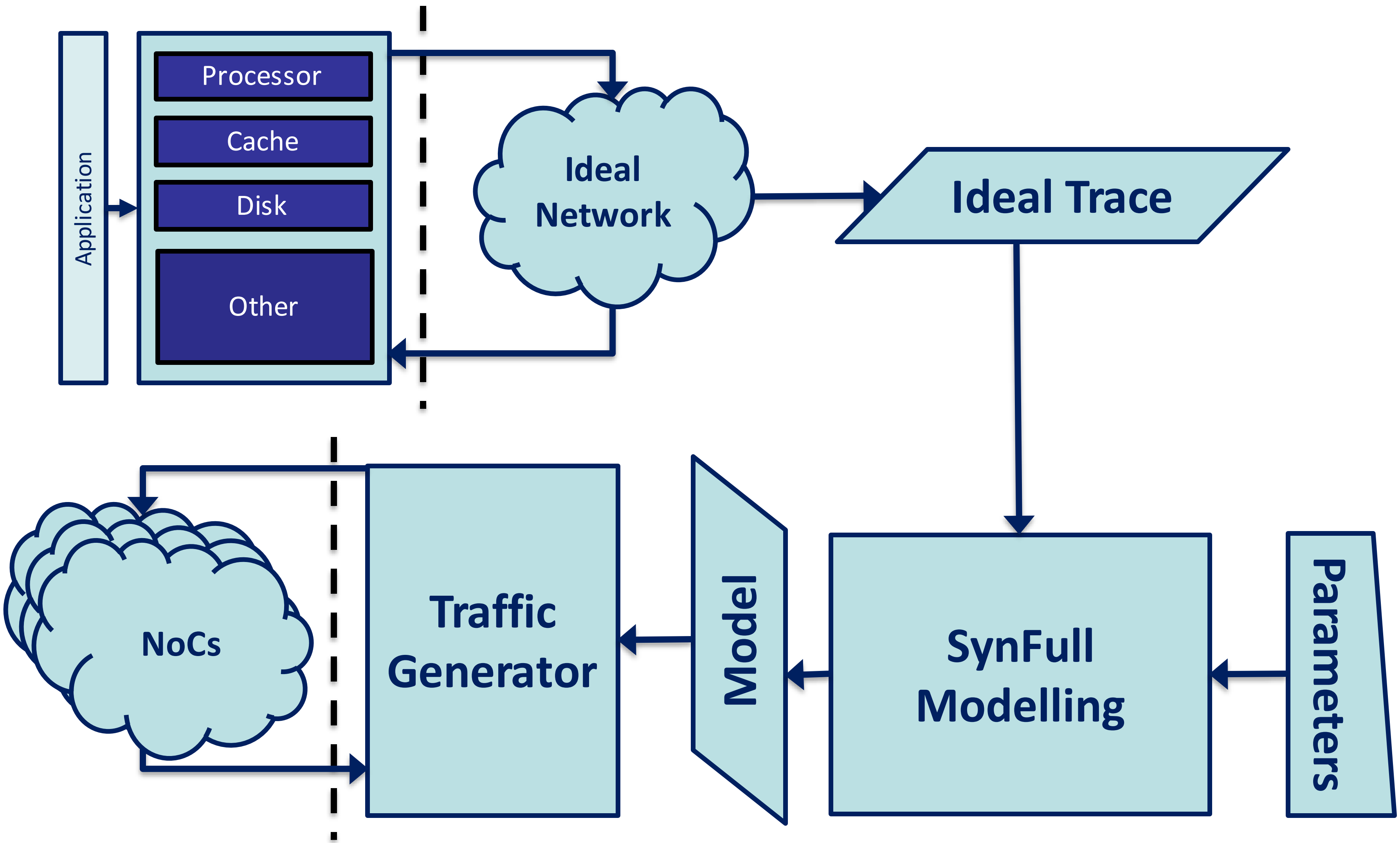
Creating The Models



Creating The Models



Creating The Models

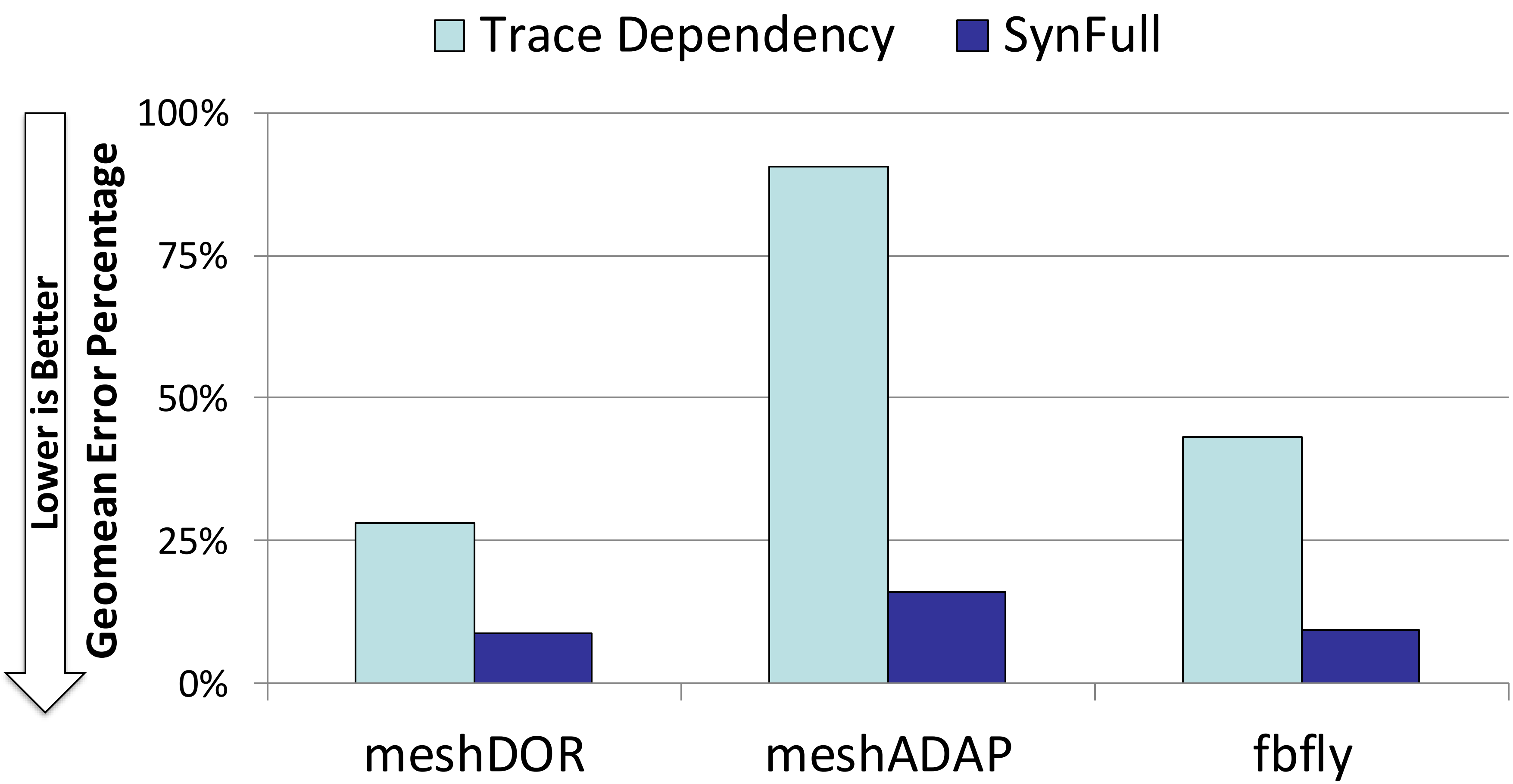


Evaluation Methodology

Network	meshDOR	meshADAP	fbfly
Topology	Mesh	Mesh	Flattened Butterfly
Channel Width	8 bytes	4 bytes	4 bytes
Virtual Channels	2 per port	2 per port	4 per port
Routing	XY	Adaptive YX-XY	UGAL

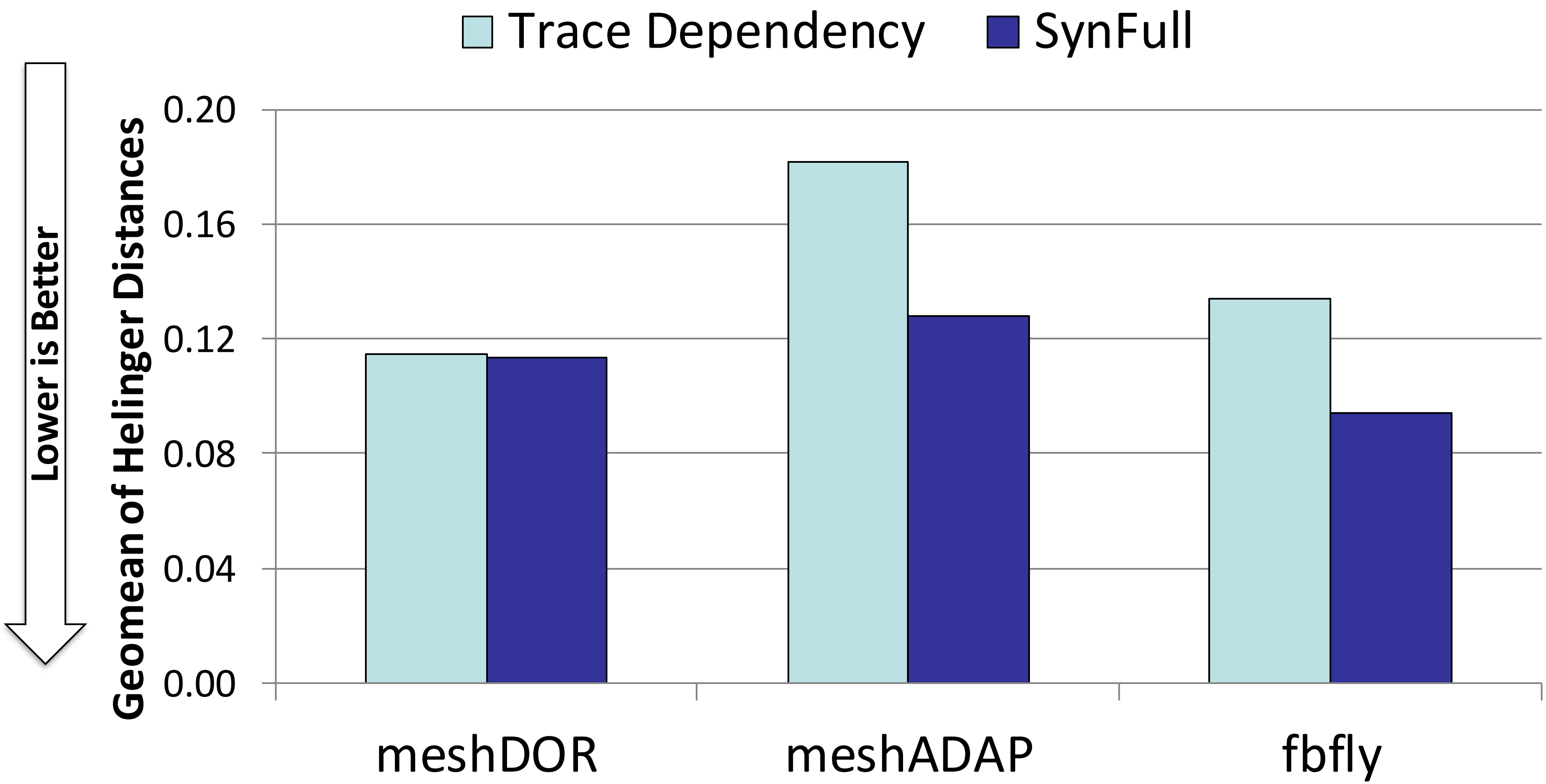
- 16 Out-of-Order Cores
- MOESI Protocol
- 16 Benchmarks (Splash-2, PARSEC)
- Traces with Dependencies Comparison

Packet Latency Error



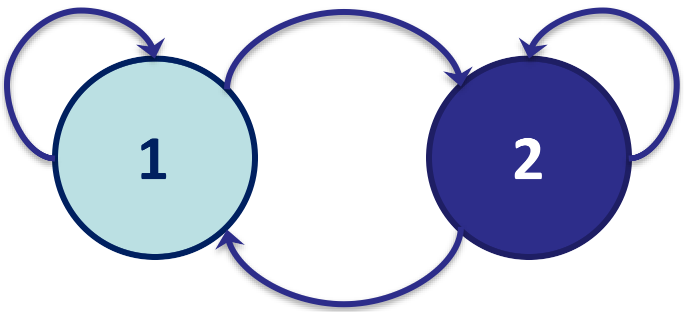
No Throttling For Initiating Transactions

Distribution Error



Captures Congestion

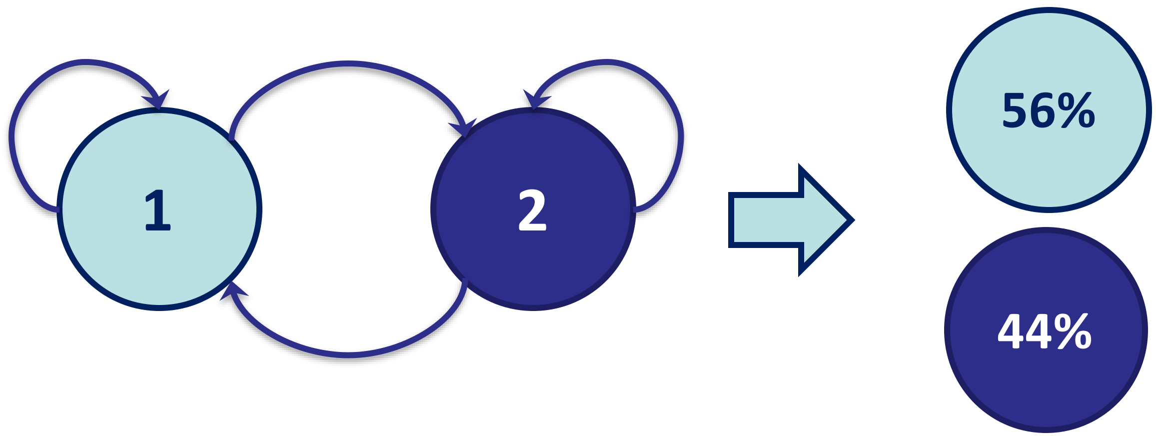
What About Speed?



$$\begin{bmatrix} 0.40 & 0.60 \\ 0.75 & 0.25 \end{bmatrix}$$

Markov Probability Matrix,

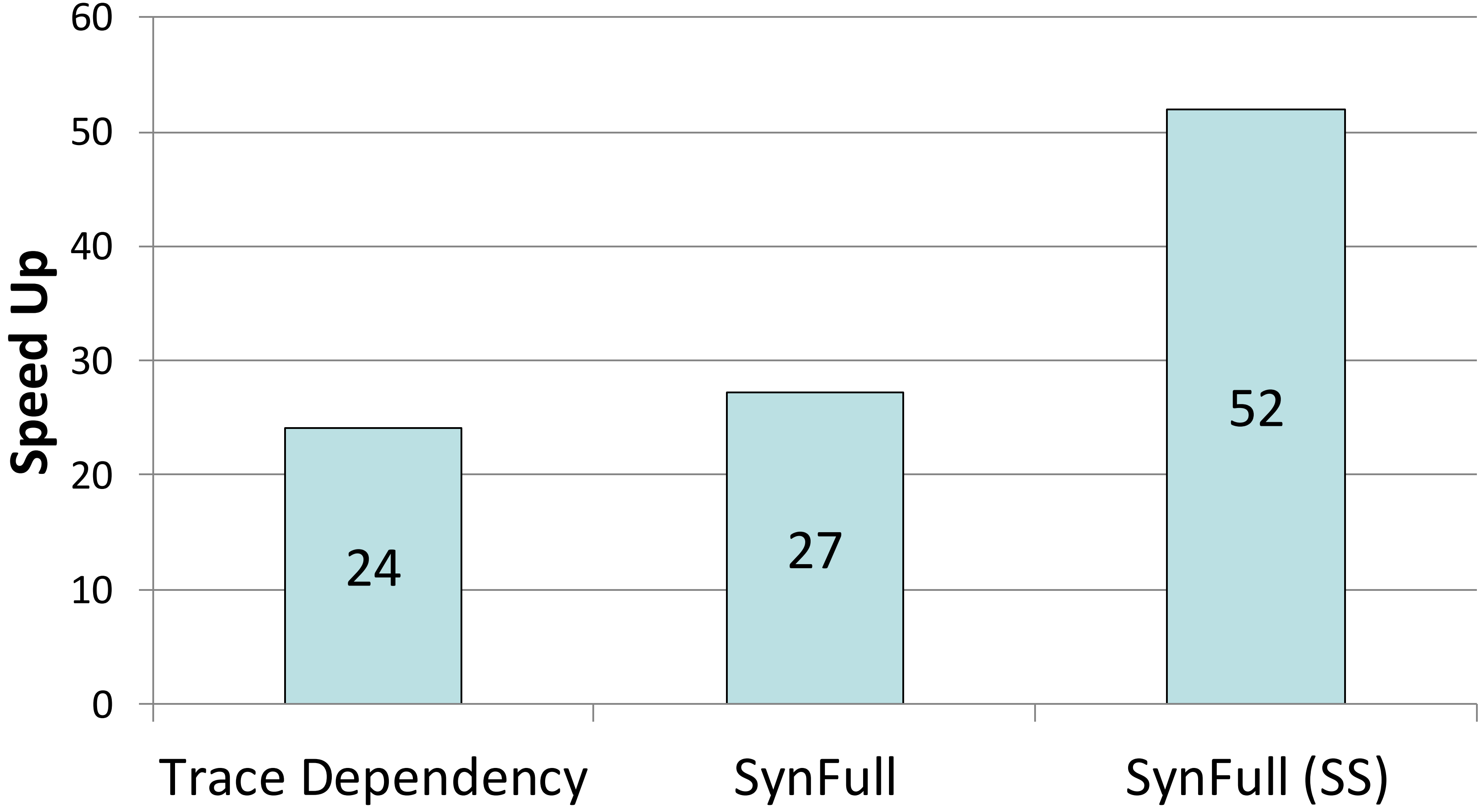
What About Speed?



$$\begin{bmatrix}0.40 & 0.60\\ 0.75 & 0.25\end{bmatrix}^{50} = \begin{bmatrix}0.56\\ 0.44\end{bmatrix}$$

Markov Probability Matrix, after a while... converges

Speed Up



52x Speed Up With 11.7% Error

Conclusion

- Implemented Synthetic Traffic Models that are
  - Accurate: **10.5%** error
  - Fast: Over **50x** average speed up
  - Generic: SynFull works for many applications

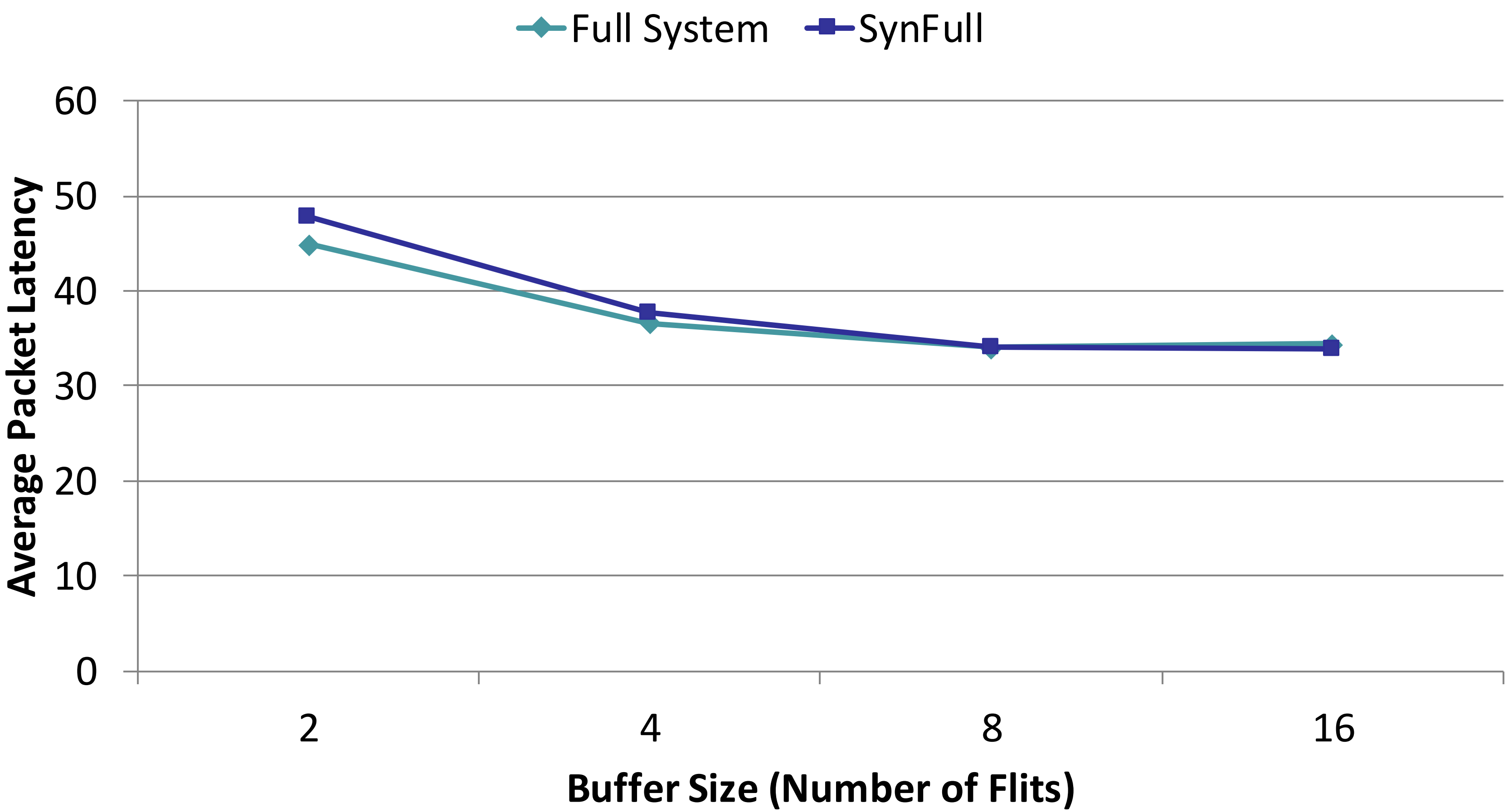
Try SynFull Out:

http://www.eecg.toronto.edu/~enright/items/synfull_download.html

Thank you for listening!

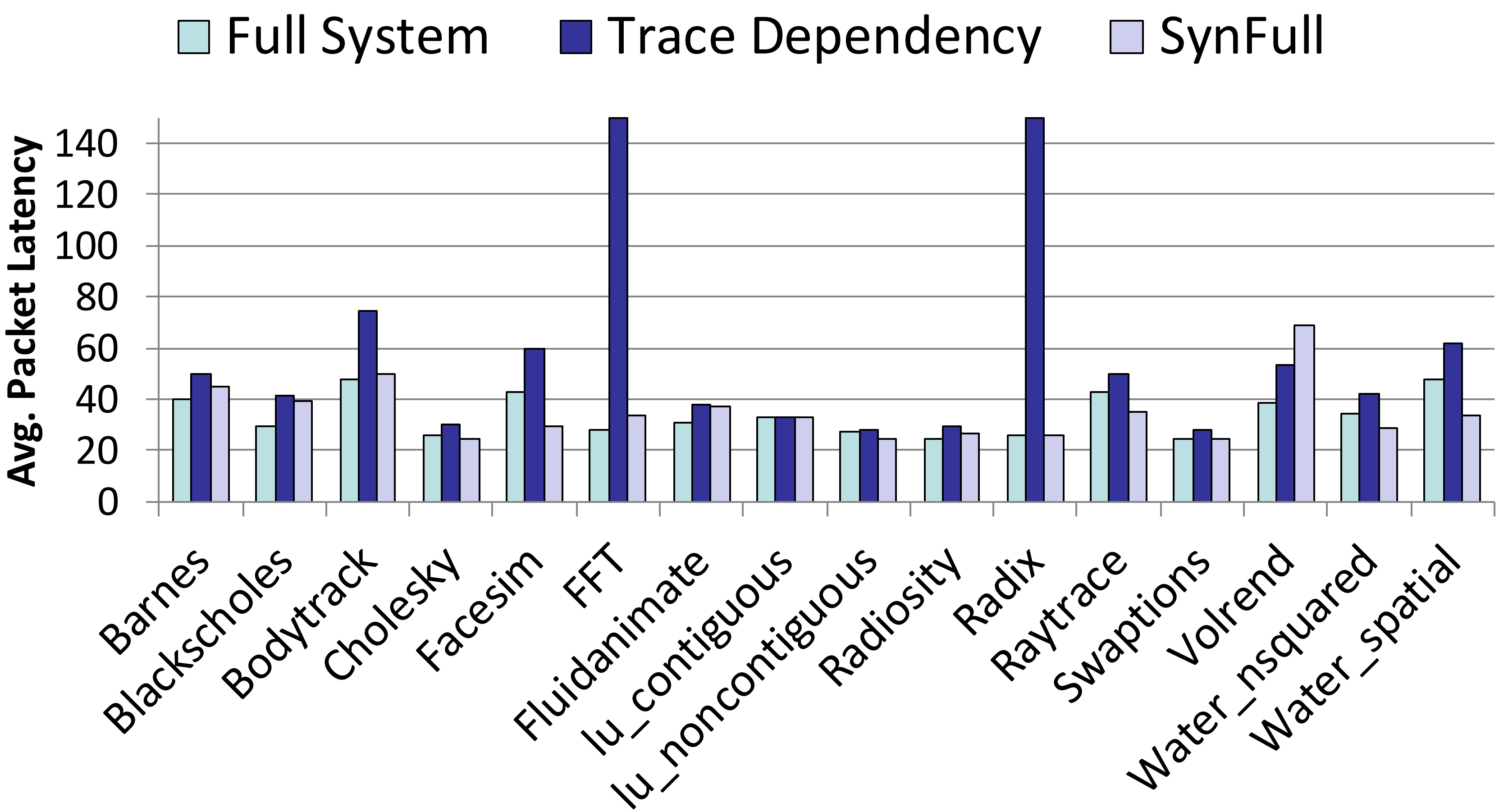
QUESTIONS & ANSWERS

Back Up: Design Space Exploration

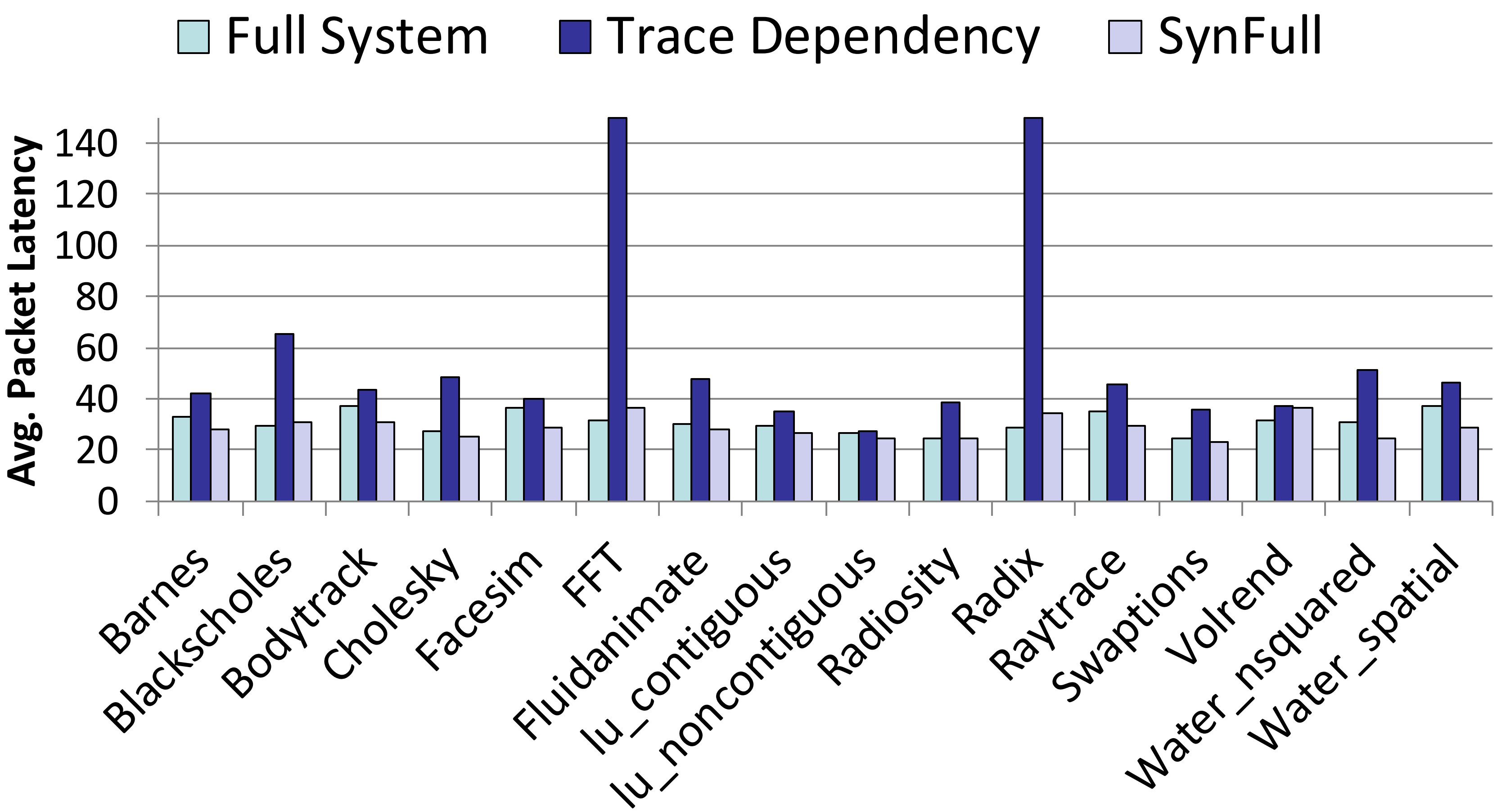


Same Conclusion, Less Time

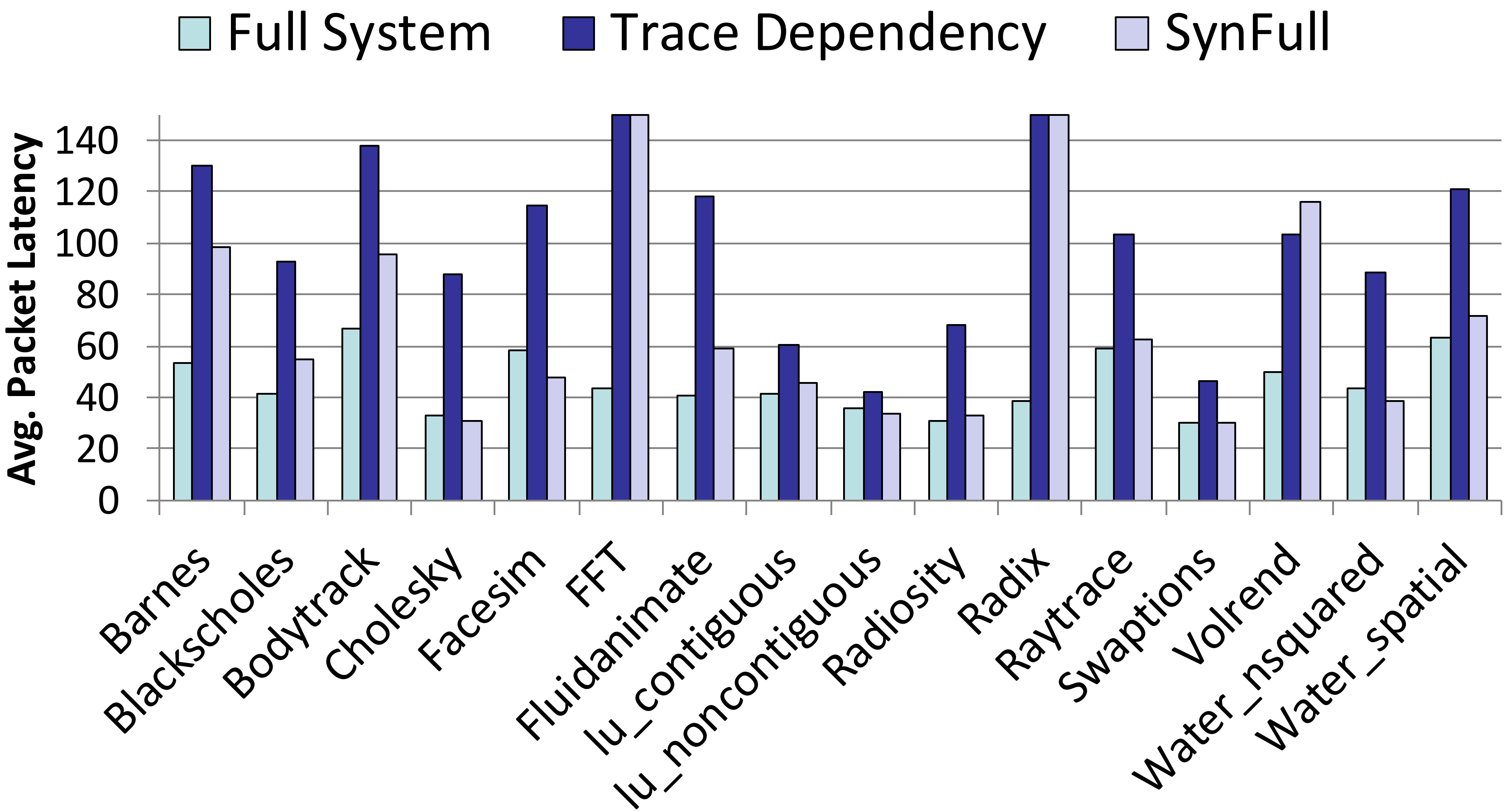
Back Up: meshDOR Packet Latency



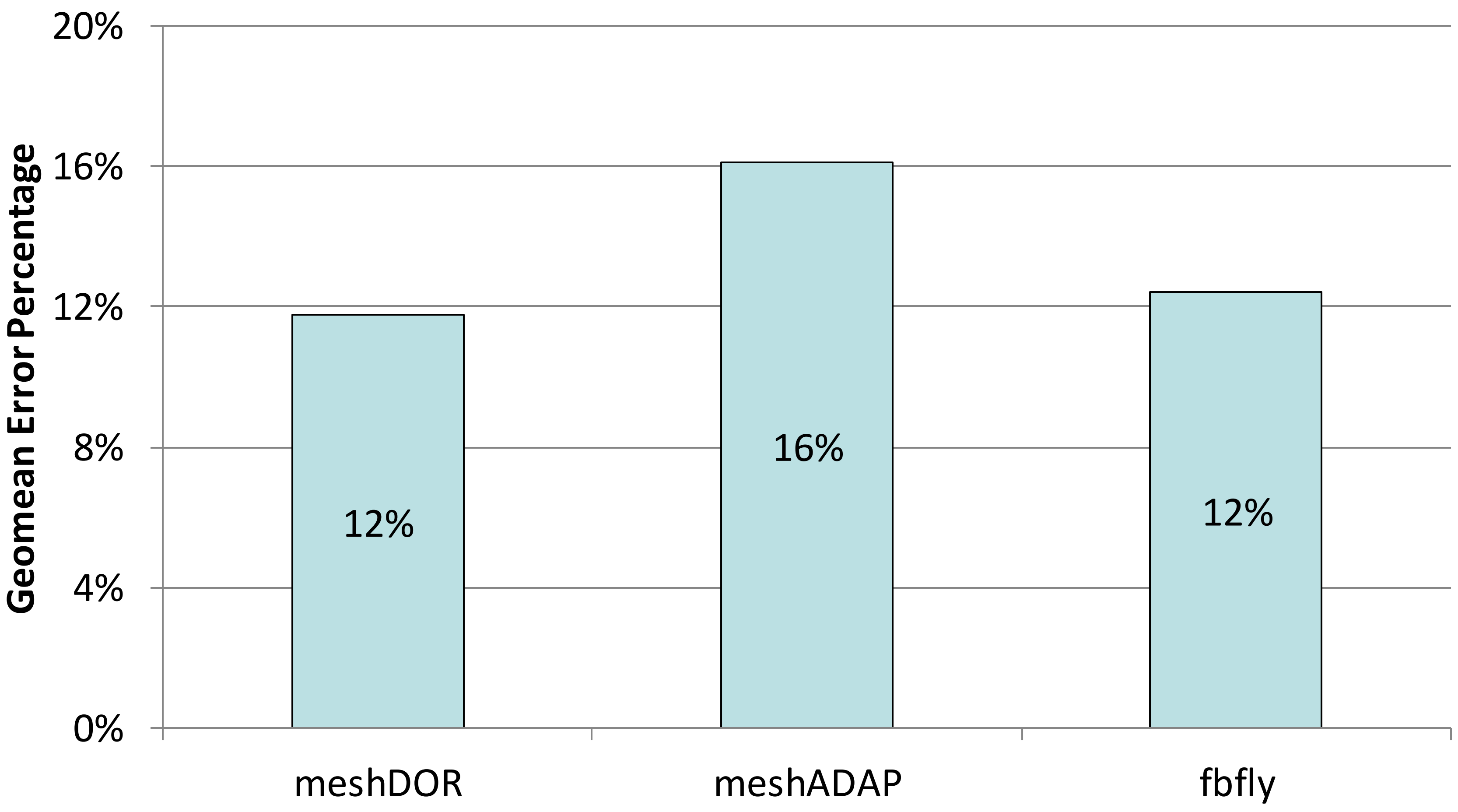
Back Up: fbfly Packet Latency



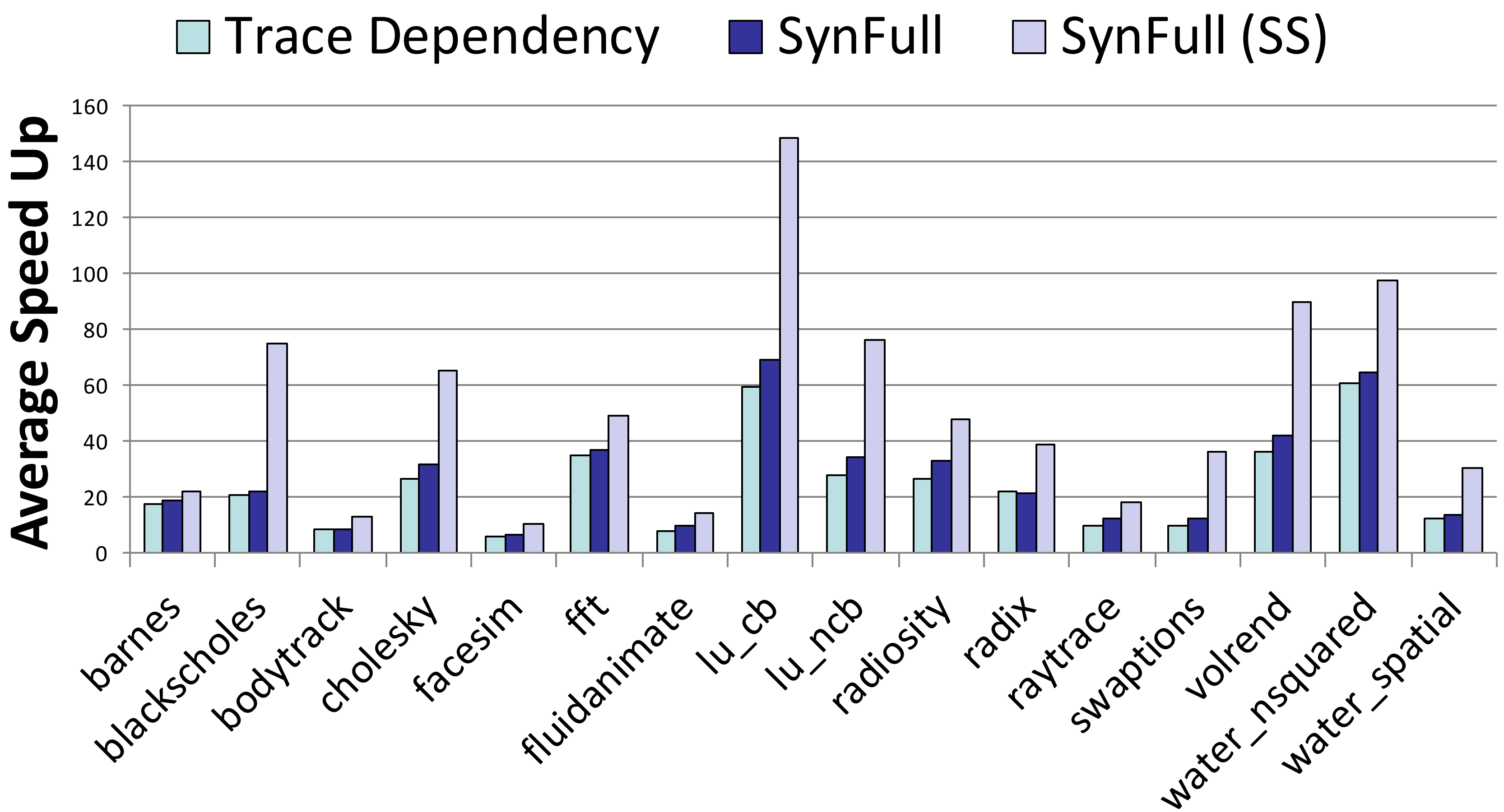
Back Up: meshADAP Packet Latency



Back Up: Average Throughput

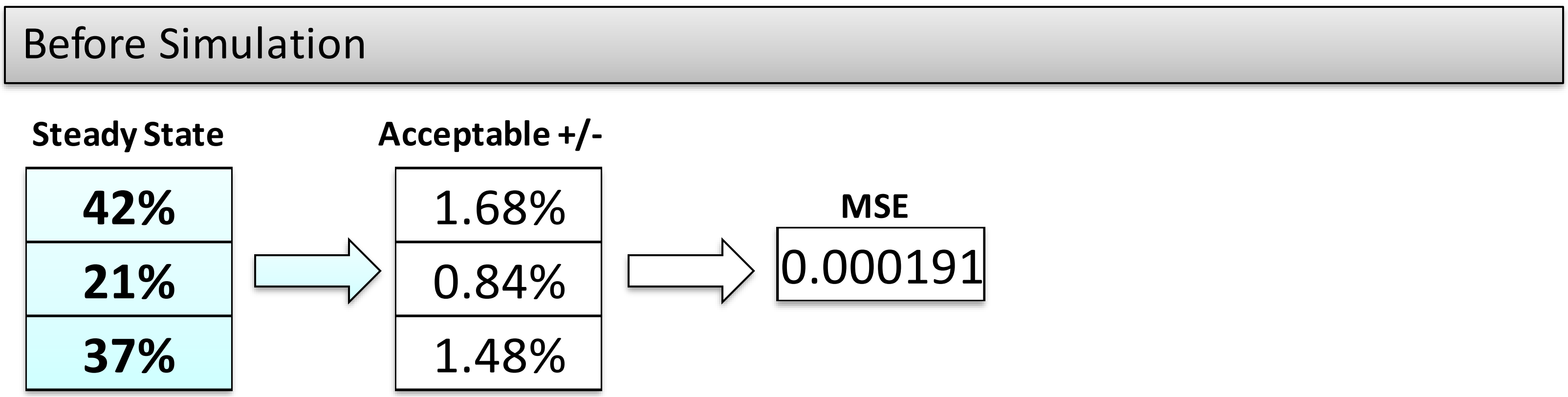


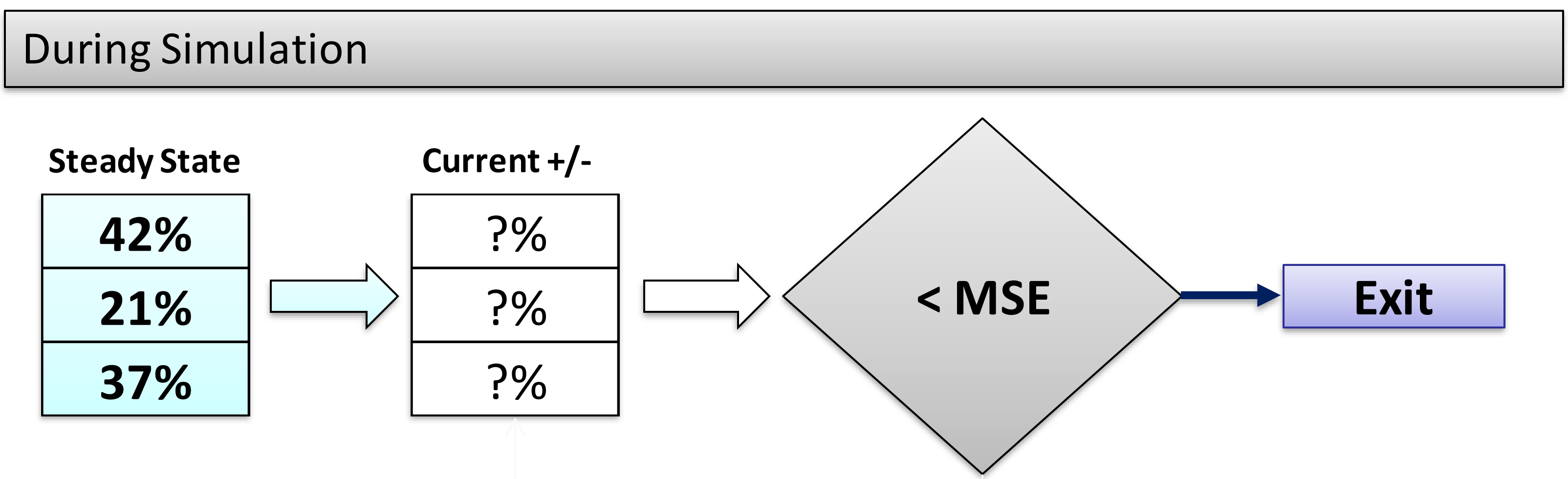
Back Up: Speed Up Per Application



Averaged Over 3 Runs (Different NoCs)

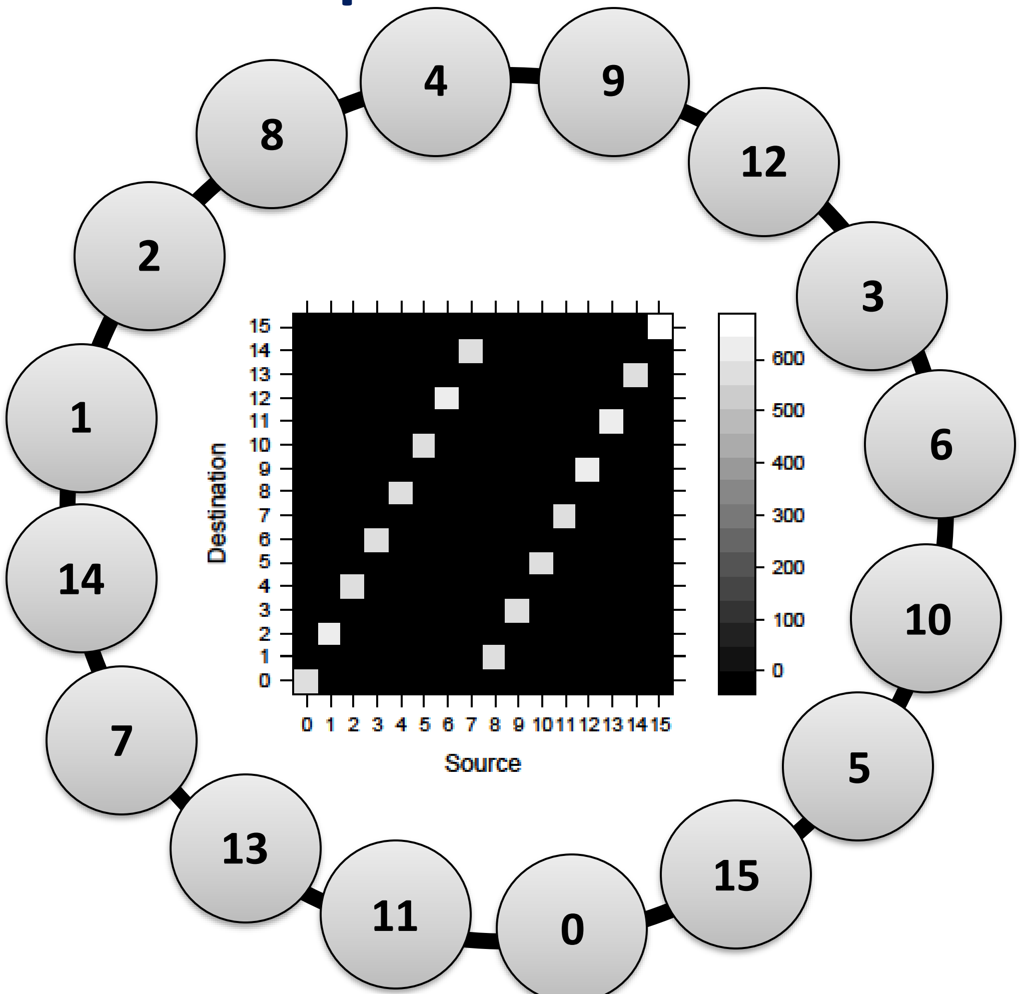
Back Up: Steady State



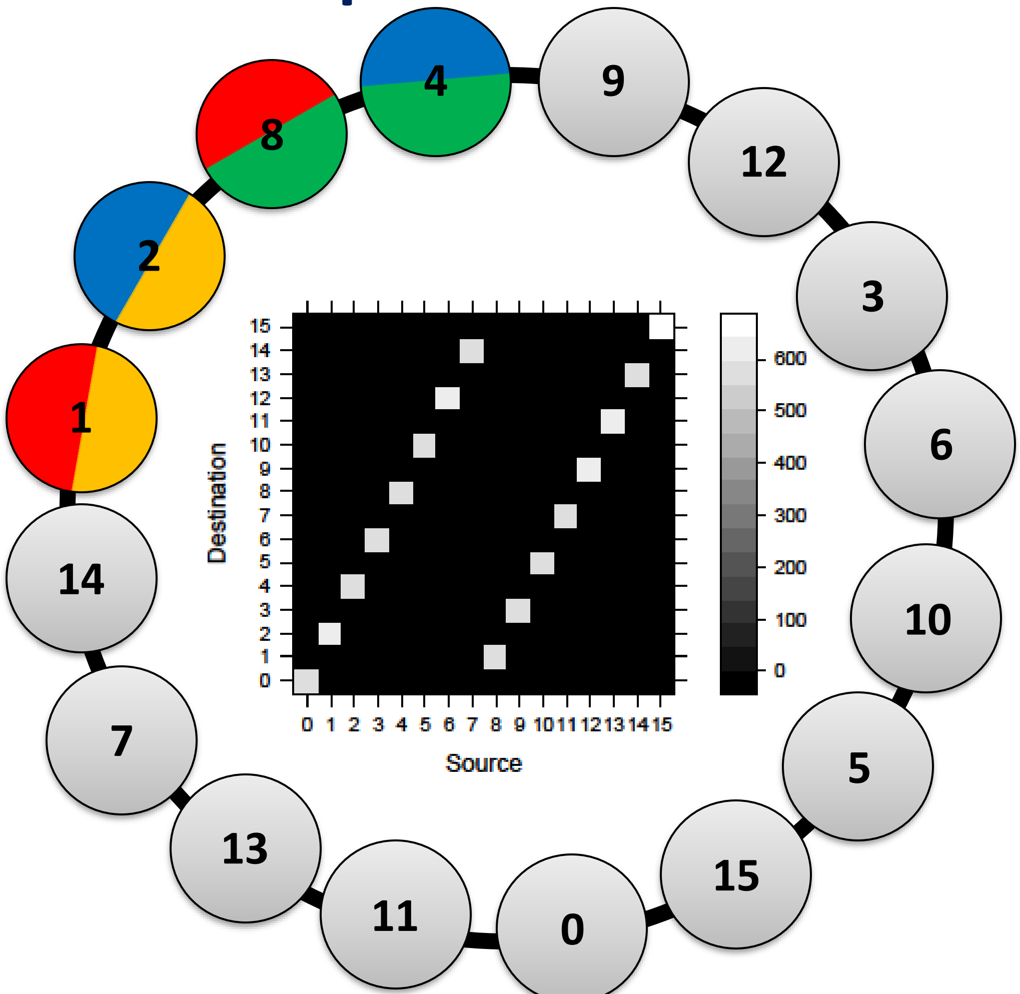


Current +/- Depends on State Transitions (RNG) and MSE

Back Up: Shuffle NoC



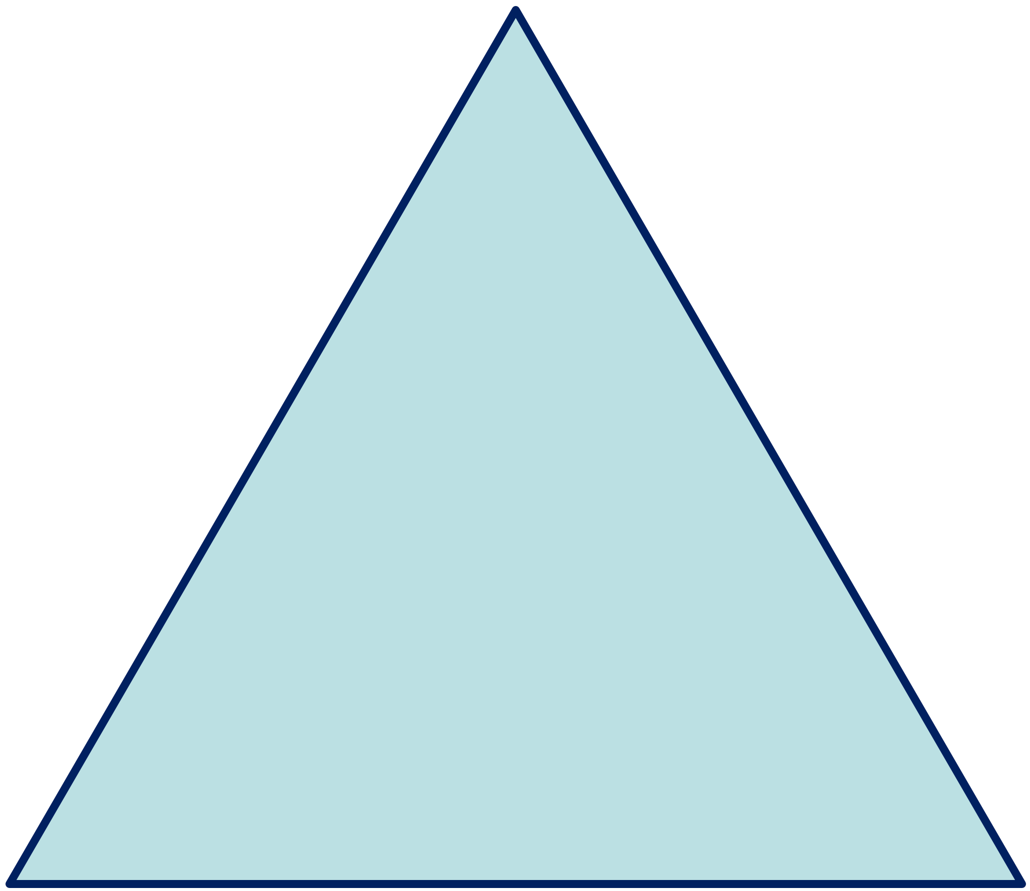
Back Up: Shuffle NoC



Ring Topology; Max. 2 Hops Needed

Triangle Score

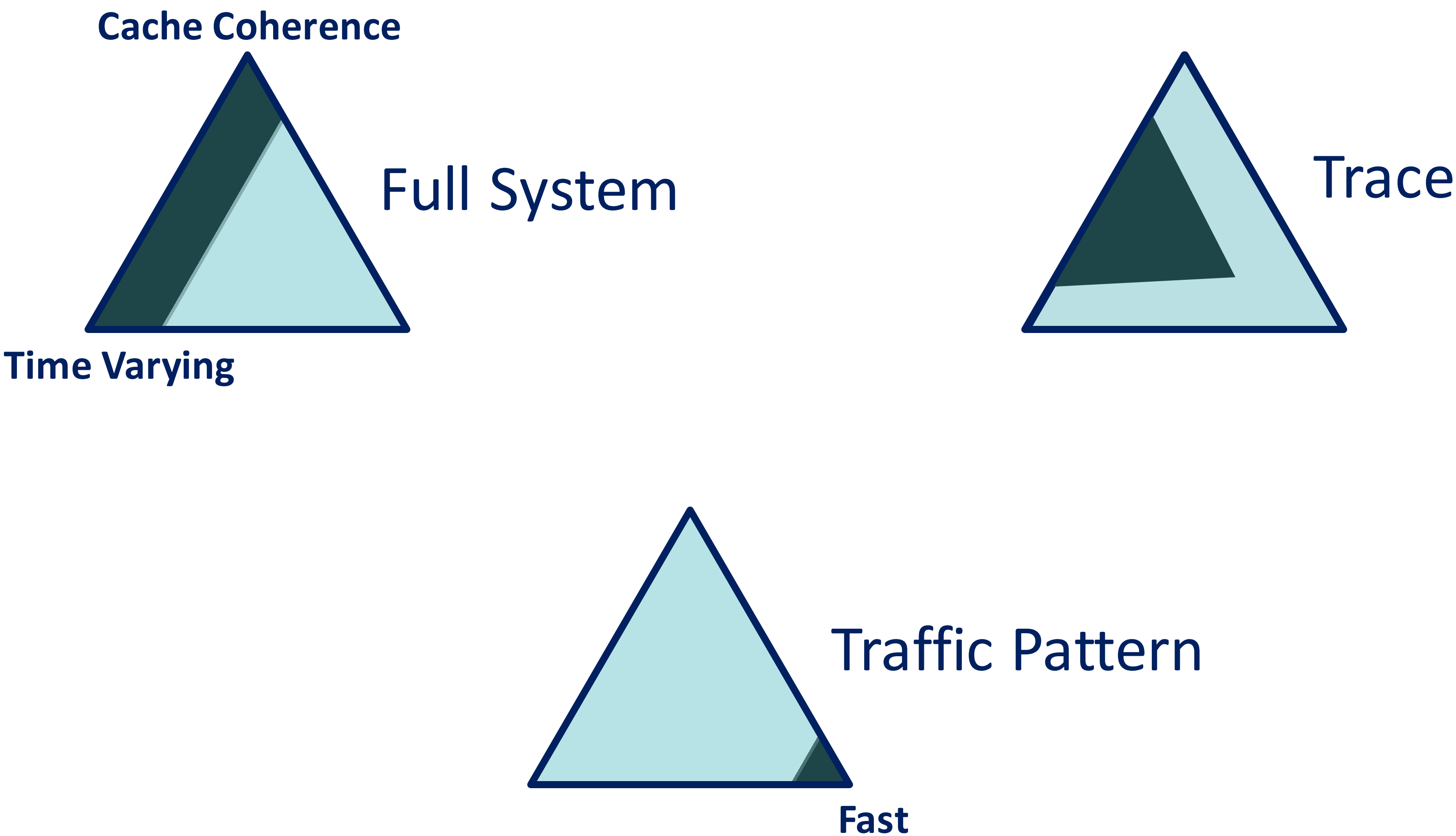
Cache Coherence



Time Varying

Fast

NoC Simulation Methodologies



SynFull