NTPT: On the End-to-End Traffic Prediction in the On-Chip Networks

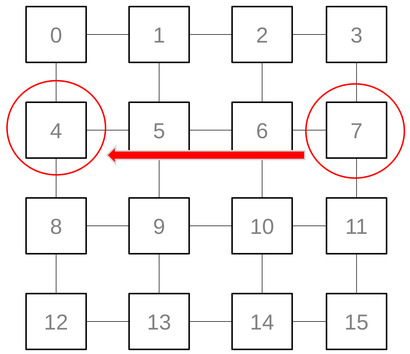
Yoshi Shih-Chieh Huang[1], Kaven Chun-Kai Chou[1], Chung-Ta King[1], and Shau-Yin Tseng[2]

[1]Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan

[2]SoC Technology Center, Industrial Technology Research Institute, Hsinchu, Taiwan

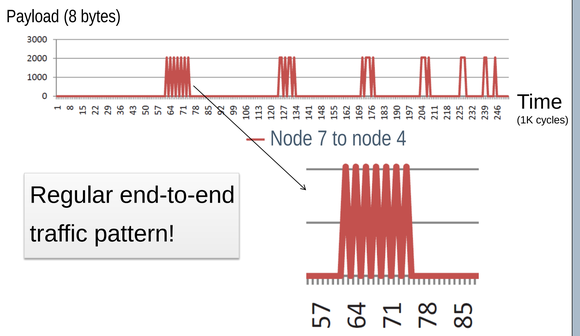
Motivation

- Consider LU Decomposition in SPLASH-2 running on 16-node Tilera TILE64
- Observe the communication behavior between any pair of nodes due to application execution
  - Amount of data transferred along time
    - → *end-to-end traffic* between a pair of nodes



Motivation (cont'd)

Consider the end-to-end traffic from node 7 to node 4



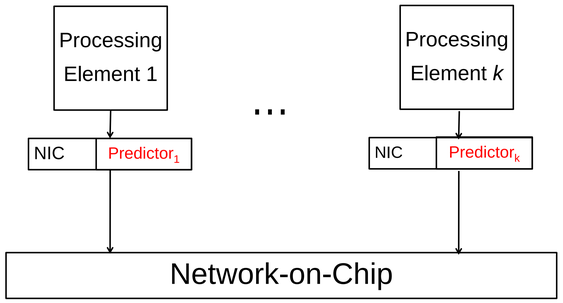
Regular end-to-end traffic pattern!

Key Questions

- Can the end-to-end traffic pattern be recognized?
- Can the end-to-end traffic pattern be predicted?
- What if we can?
  - Controlling the injection rate for E2E flow control
  - Performing dynamic routing based on workload
  - Remapping tasks
  - Controlling the power mode of switches and links

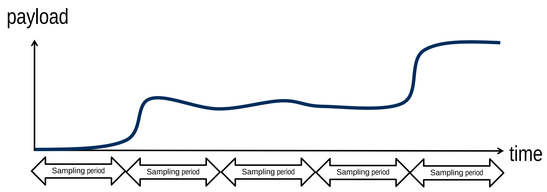
Domain: chip-multiprocessors with on-chip networks
Proposal: A hardware design for end-to-end traffic recognition and prediction

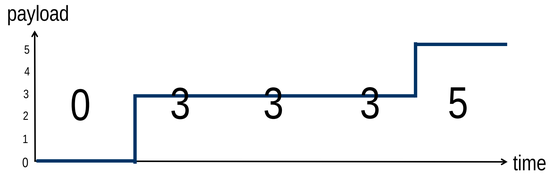
Overview of Design



Traffic Pattern Recognition and Quantization

- Record outgoing traffic to each destination node
- Accumulate for each sampling period

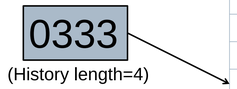




Traffic Prediction

- Intuition: more accurate prediction by tracking longer traffic history
- Pattern-oriented predictor
  - Traffic pattern register
  - Indexing into a table for prediction of traffic in the next sampling period

Pattern-Oriented Table Indexing

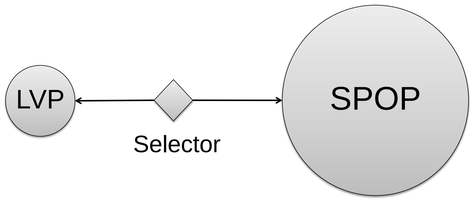


index	Prediction
0	nil
1	nil
2	3
3	nil
…	…
0332	2
0333	5
0334	nil
0335	3
…	…

Problem:
The prediction table may be huge and sparse!

NTPT Design

NTPT-based predictor =
last value predictor (LVP)
+ Simplified pattern-oriented predictor (SPOP)

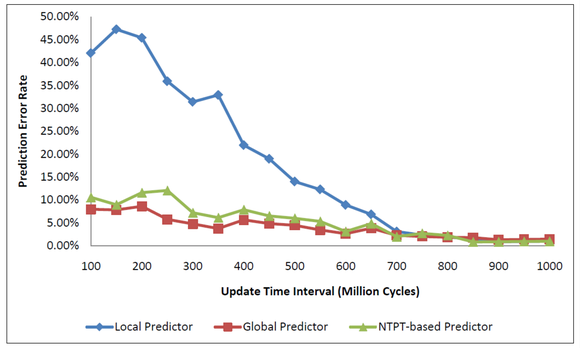


(Similar to branch prediction in computer architecture)

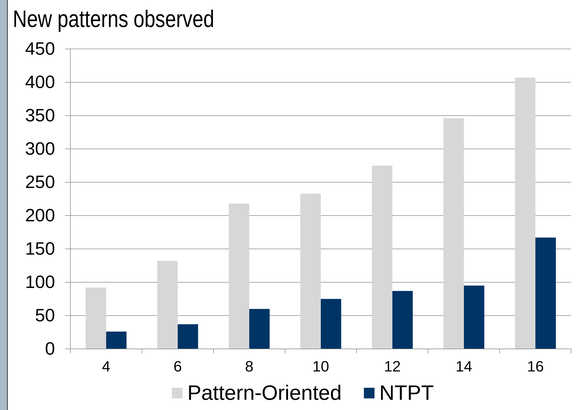
Evaluation Setup

- Emulate the NTPT-based predictor on the Tilera TILE64
- LU Decomposition in SPLASH-2 ported to TILE64
  - 16 nodes for executing application
  - 16 nodes for emulating NTPT-based predictors

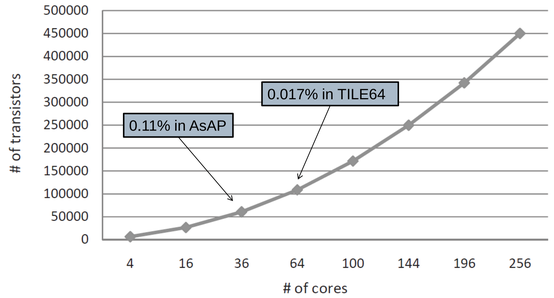
Prediction Error Rate



Prediction Table Size



Estimated Area Overhead



Conclusions and Future Works

- A hardware design for recognizing and predicting application-level end-to-end traffic on chip-multiprocessors using on-chip interconnection networks
  - Traffic patterns based on periodic accumulation
  - Traffic prediction based on prediction table
  - Good prediction accuracy with low hardware overhead
- Future works
  - Injection rate control for congestion avoidance in NoC
  - Dynamic voltage/frequency scaling for routers and links

Contact information

Yoshi Shih-Chieh Huang
yoshijava@gmail.com

Kaven Chun-Kai Chou
kavenc@gmail.com

Chung-Ta King
king@cs.nthu.edu.tw