Boosting Random Write Performance for Enterprise Flash Storage Systems

Tao Xie and Janak Koshia

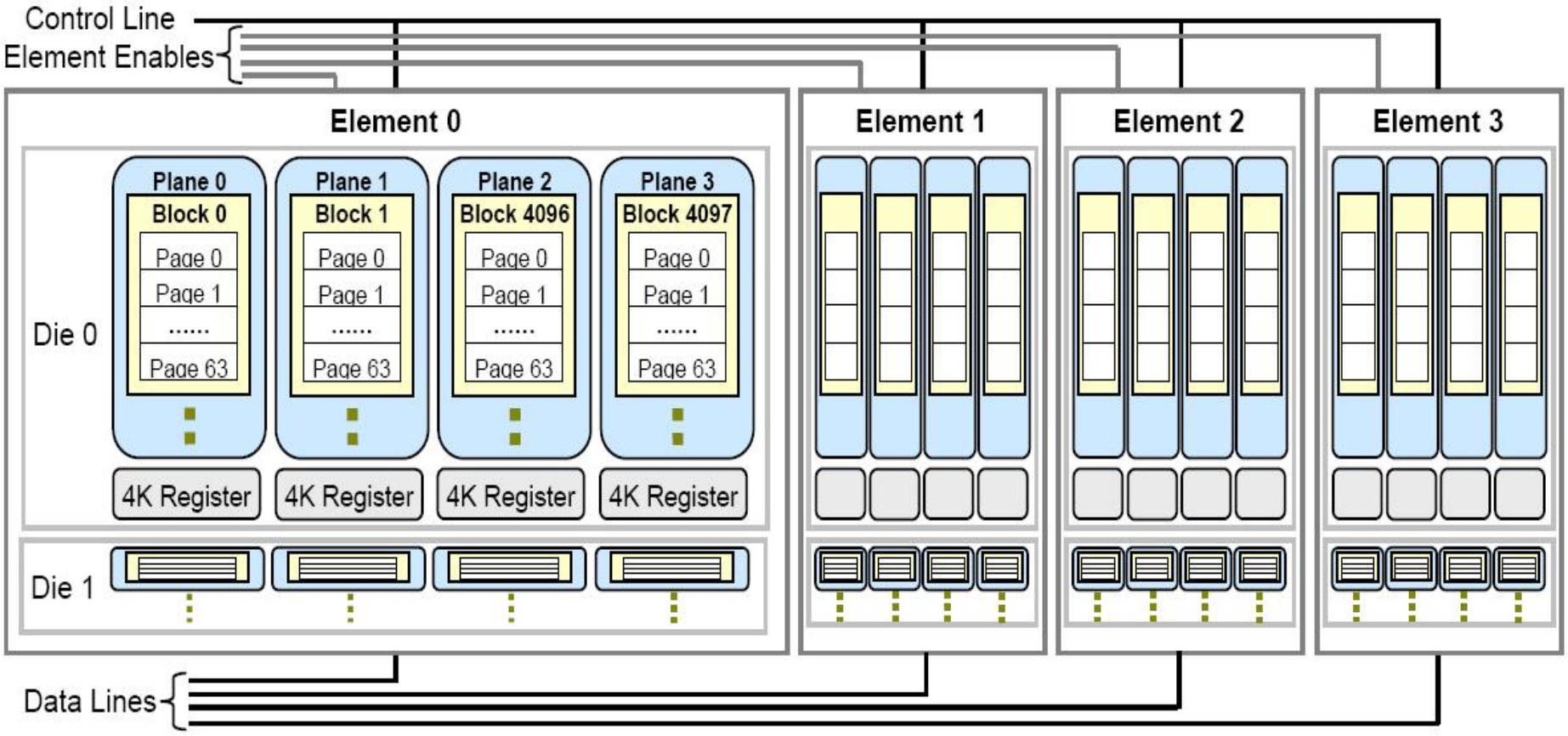
Computer Science Department
San Diego State University

SAN DIEGO STATE UNIVERSITY

Agenda

- Introduction
- Related Work and Motivation
- EPO (*E*lement-level *P*arallel *O*ptimization)
- Performance Evaluation
- Conclusion

Samsung's K9XXG08UXM series NAND-Flash



Flash Translation Layer

- Mapping between Logical address space to Physical address space
- Wear-leveling
- Garbage collecting

Why Random Write Is an Issue?

- Out-of-place update
- Time consuming erase and garbage collection
- Unit difference between RW and erase operation
- Complicated FTL (Flash Translation Layer) logic

Random Write Performance of SSD

- Depends mainly on two factors
  - Architecture of an SSD
  - Type of workload

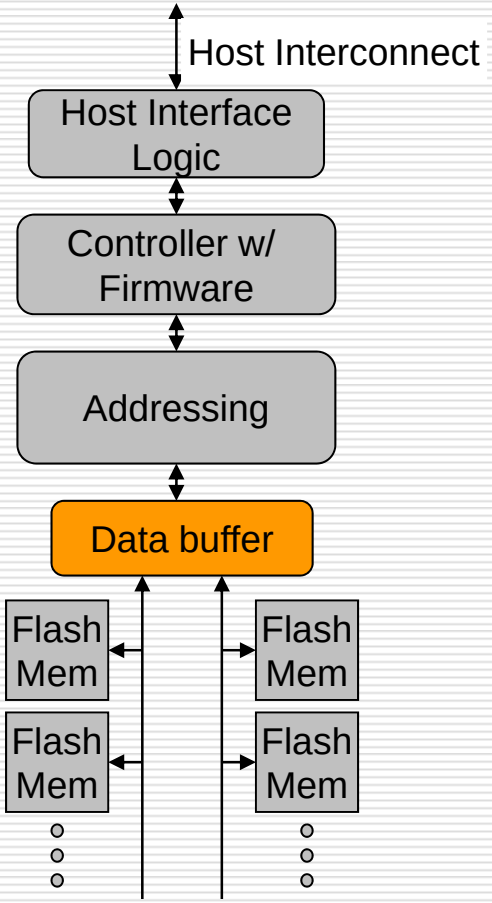
Related Work

- Add non-volatile RAM (NVRAM)
- DFTL (Demand based Flash Translation Layer) – FLT has to be changed
- BPLRU (Block Padding Least Recently Used) – Amplifies read and write requests

Motivation

- Exploiting the architecture of an SSD
  - Element-level Concurrency
  - Die-level parallelism
  - Plane-level interleaving

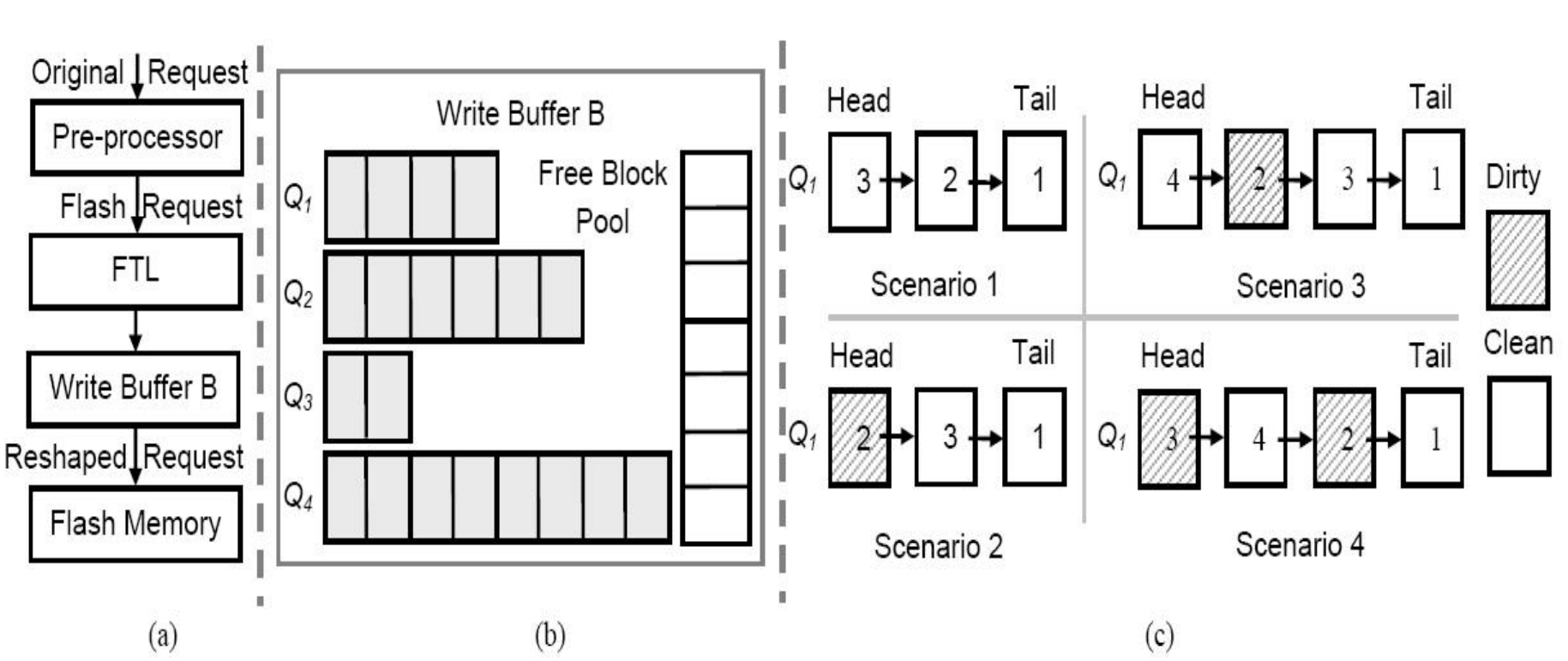
Architectural of an SSD



Important assumptions

- Workload contains only write requests
- Each request is one page size
- LBA of each request is page aligned

The EPO Strategy



E.g. 1, 2, 3, 2, 4, 3

Experimental Setup

- Simulator DiskSim 4.0 and SSD add-on by Microsoft
  - Command line tool
    - disksim <parfile> <outfile> <tracetype> <tracefile> <synthgen> [par override [...
- Traces
  - Real-world traces (OLTP)
    - TPC-C
    - Financial1
    - Financial2
- Comparing results with other three schemes
  - No cache
  - LRU (Least Recently Used)
  - BPLRU (Block Padding Least Recently Used)

Parameters Tested

- Configuration
  - Varying cache size
  - Varying page size
  - Varying number of elements

Real World Trace Statistics

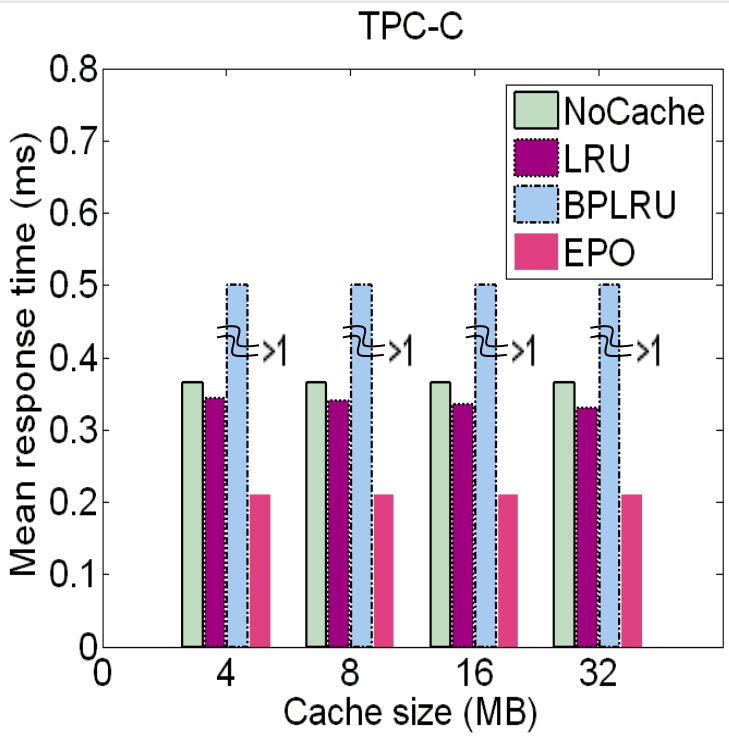
Workloads	Financial1	Financial2	TPC-C
Number of writes	2,000,000	650,000	2,000,000
Mean write size (KB)	3.9	2.9	10.2
Write per second	62.76	10.52	4337.83
Write size range (KB)	0.5-3148.5	0.5-256.5	0.5-1024

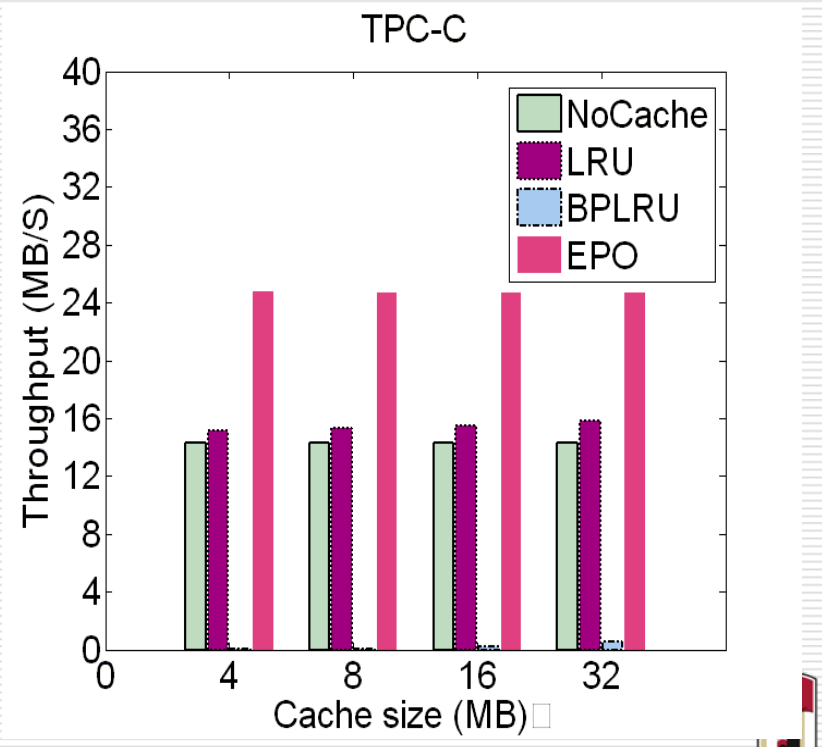
Simulation Parameter

Parameter	Value (Default) – (Varied)
Write buffer capacity (MB)	(8) – (4, 8, 16, 32)
Number of elements	(48) – (16, 32, 48, 64)
Number of planes in an element	(8)
Page size (KB)	(4) – (1, 2, 4)
Flash block size (page)	(64)
Element capacity (GB)	(4)
Flash SSD capacity (GB)	(192) – (64, 128, 192, 256)
Block erase latency (μs)	(1500)
Page read latency (μs)	(25)
Page write latency (μs)	(200)
Chip transfer latency per byte (μs)	(0.025)

Impacts of Cache Size

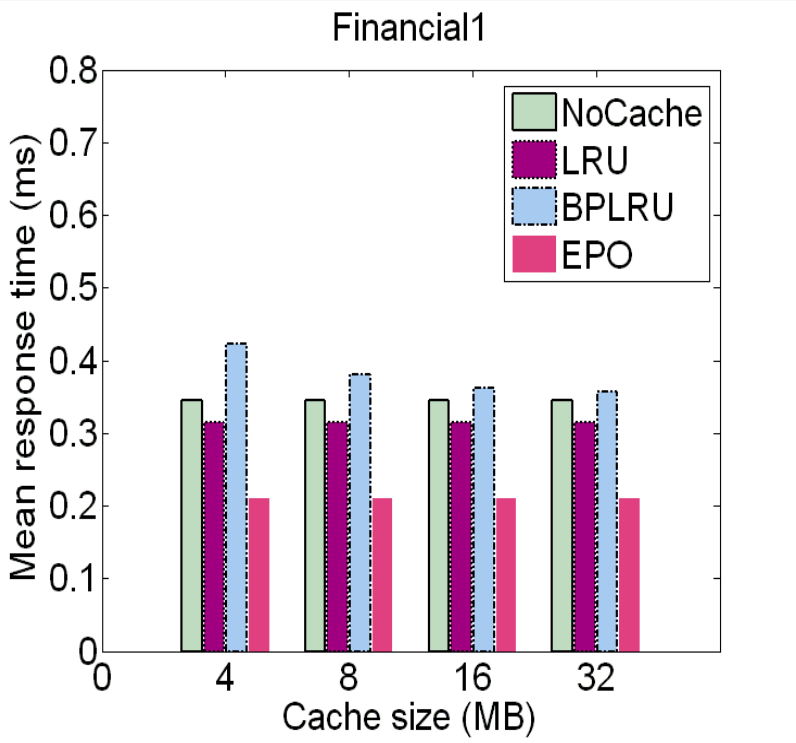
- TPC-C

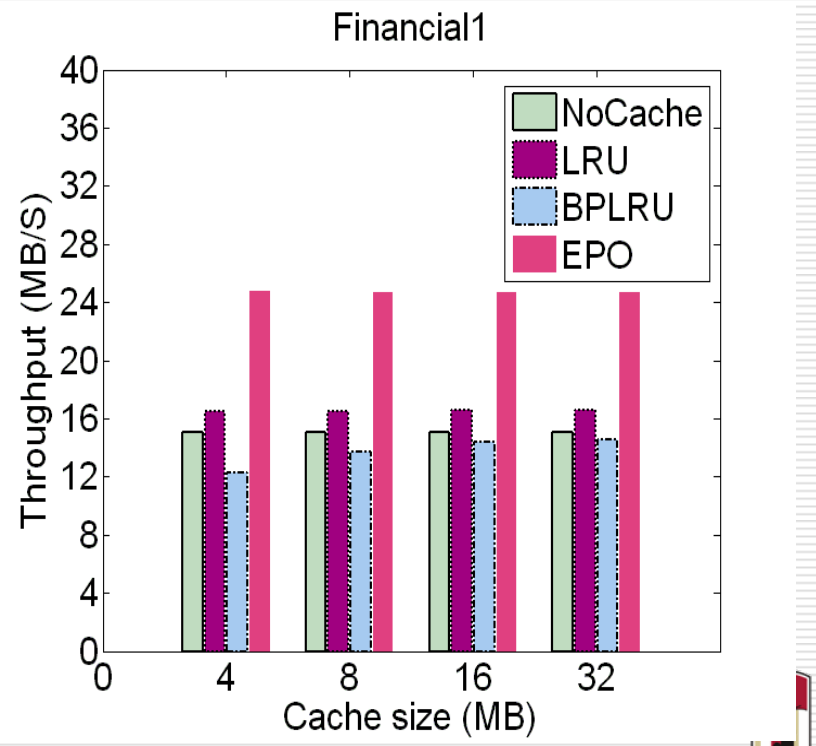




Continue…

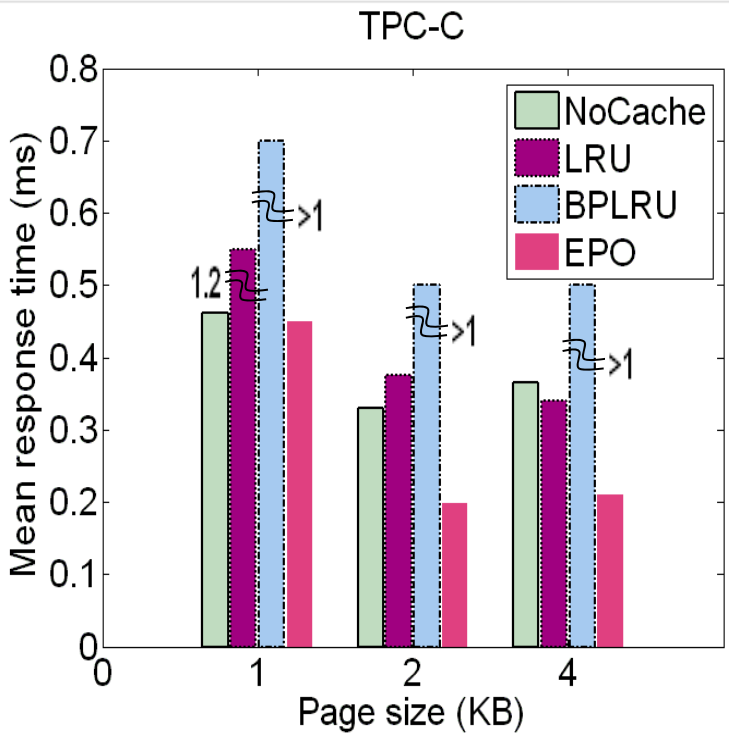
- Financial1

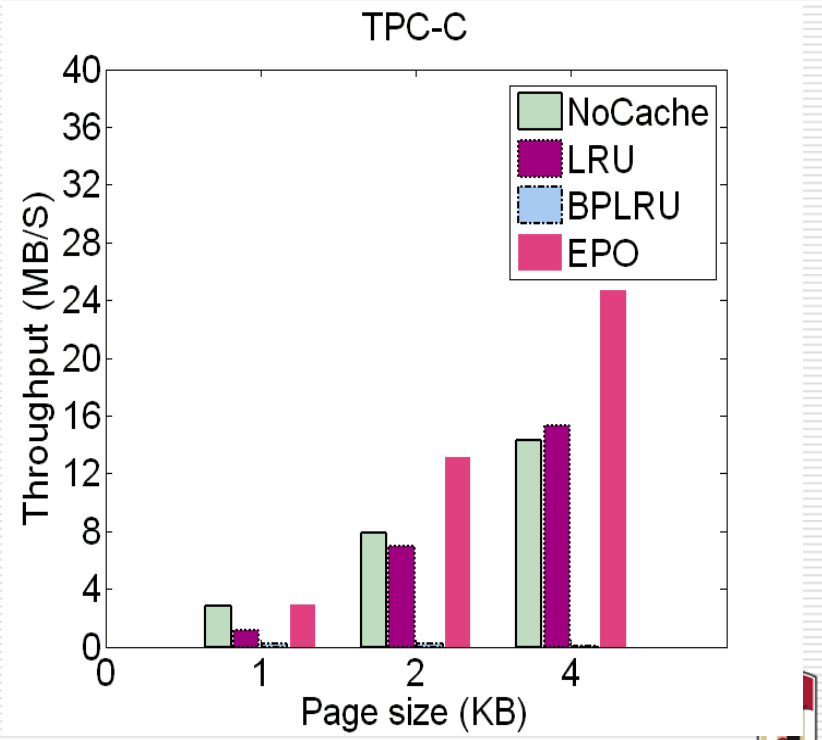




Impacts of Page Size

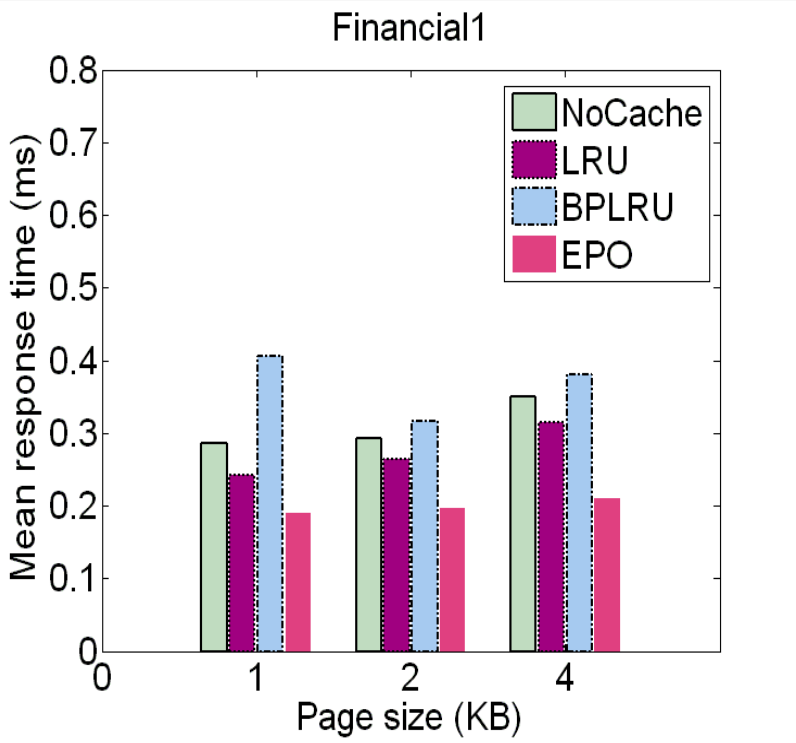
- TPC-C

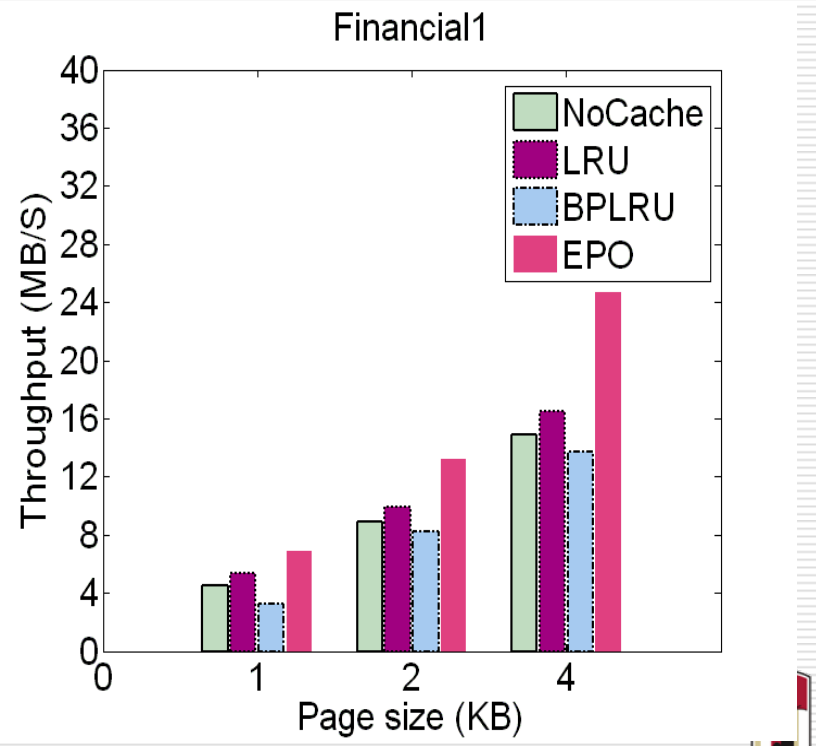




Continue…

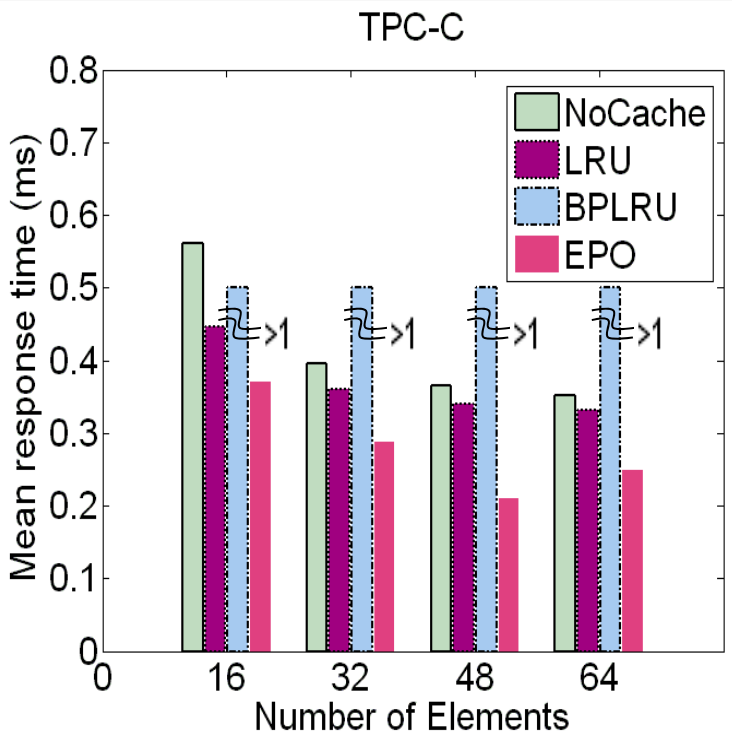
- Financial1

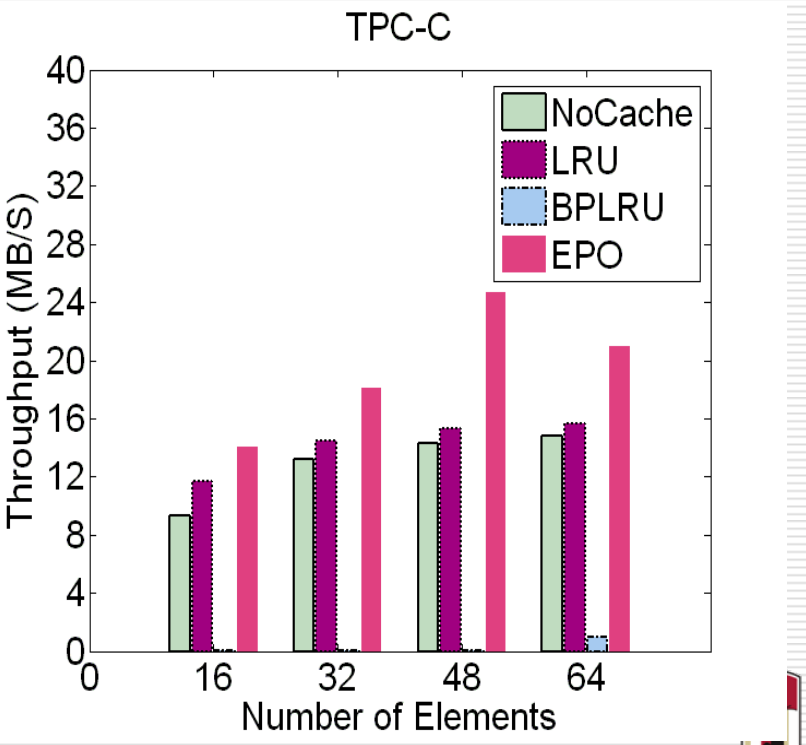




Scalability

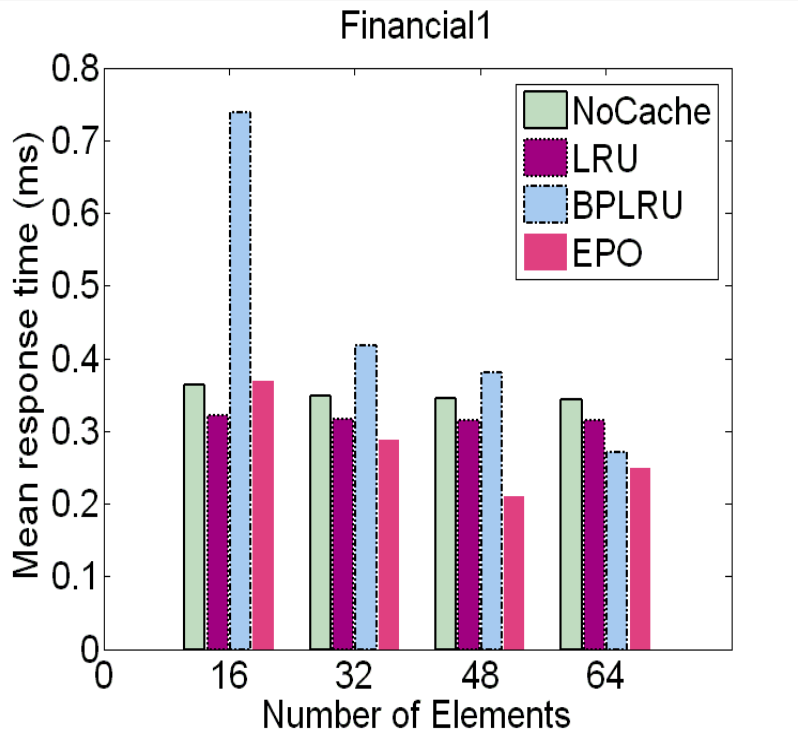
- TPC-C

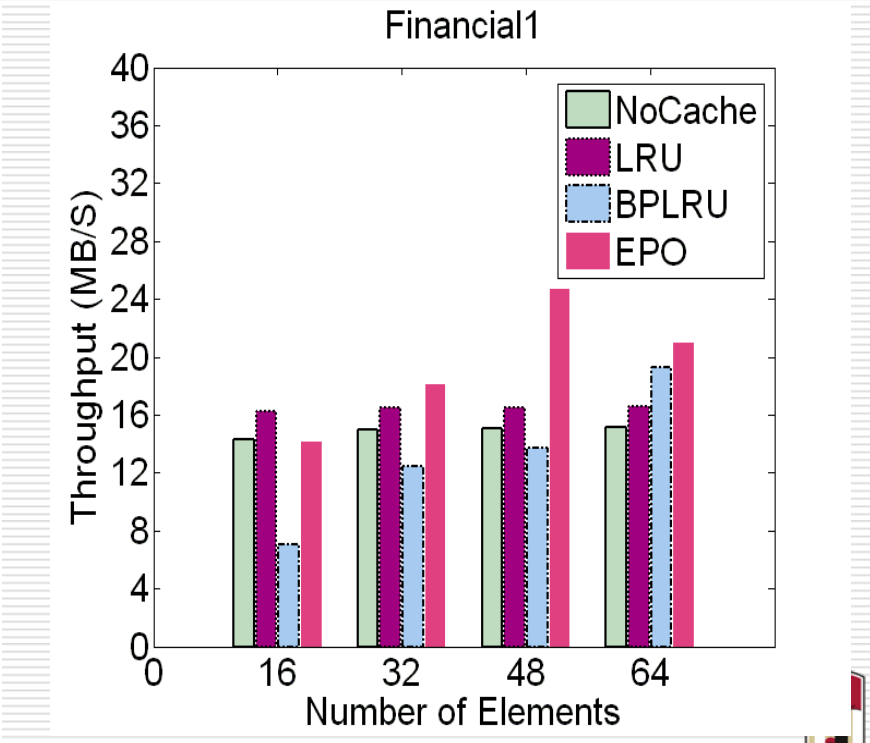




Continue…

- Financial1





Conclusion

- Write buffer is placed after FTL
- Easy to implement the algorithm
- Low cost hardware – small size DRAM
- Trivial change in FTL
- Gives good performance towards random write workload

Acknowledgements

- We thank DiskSim developers who provided an efficient, accurate, highly-configurable disk system simulator for storage research community.
- We would like to thank the researchers from Microsoft Research who developed the SSD model for DiskSim 4.0.
- This work was supported by the US National Science Foundation under grants CNS (CAREER)-0845105 and CNS-0834466.

SAN DIEGO STATE UNIVERSITY

Thank you

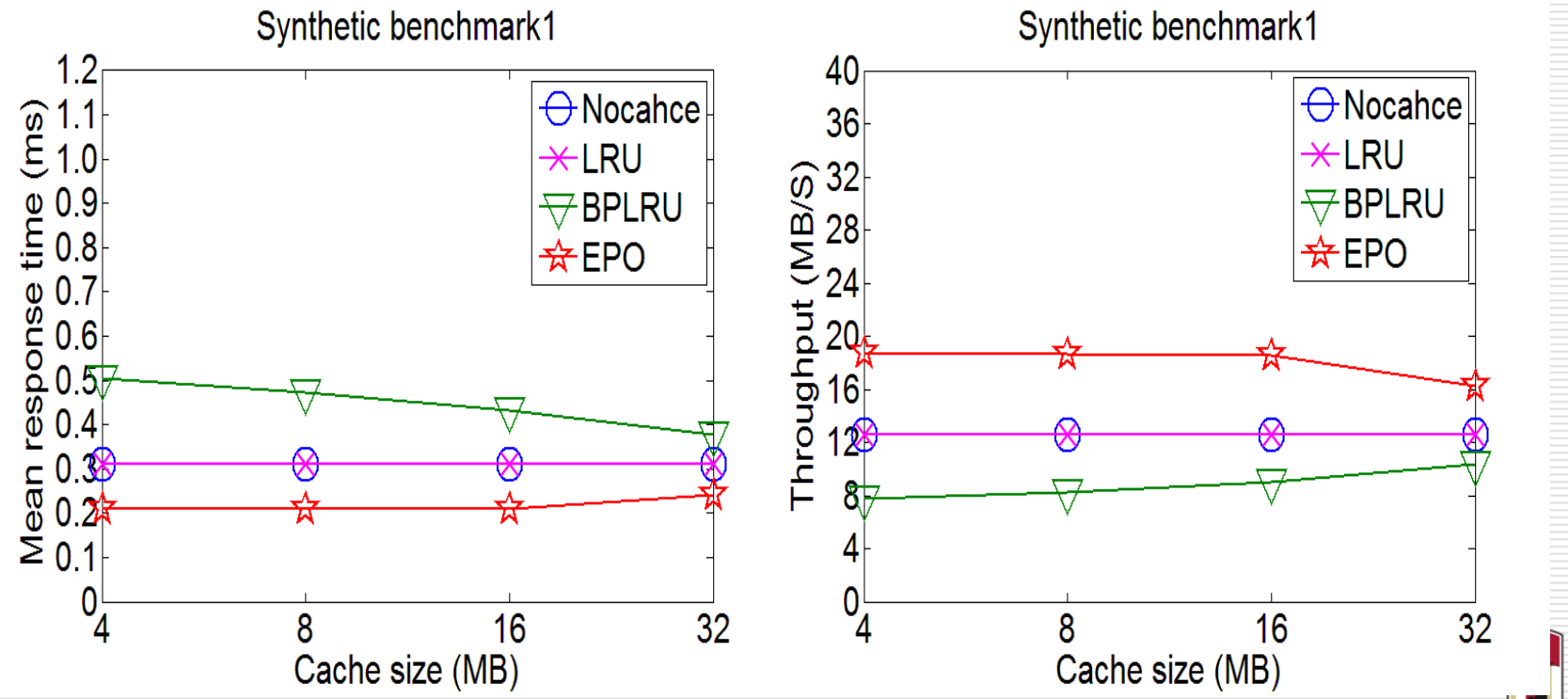
SAN DIEGO STATE UNIVERSITY

Questions?

Backup slides

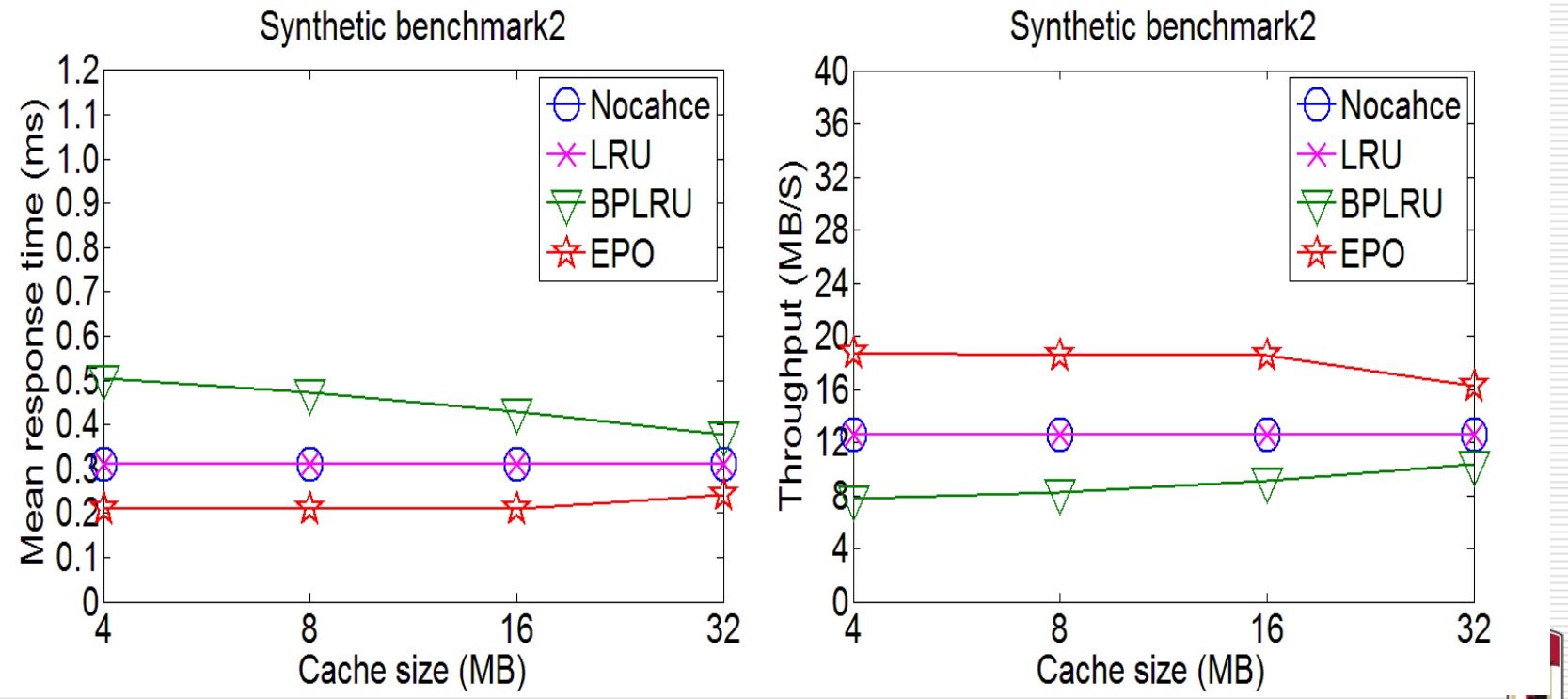
Varying cache size

- Synthetic benchmark1



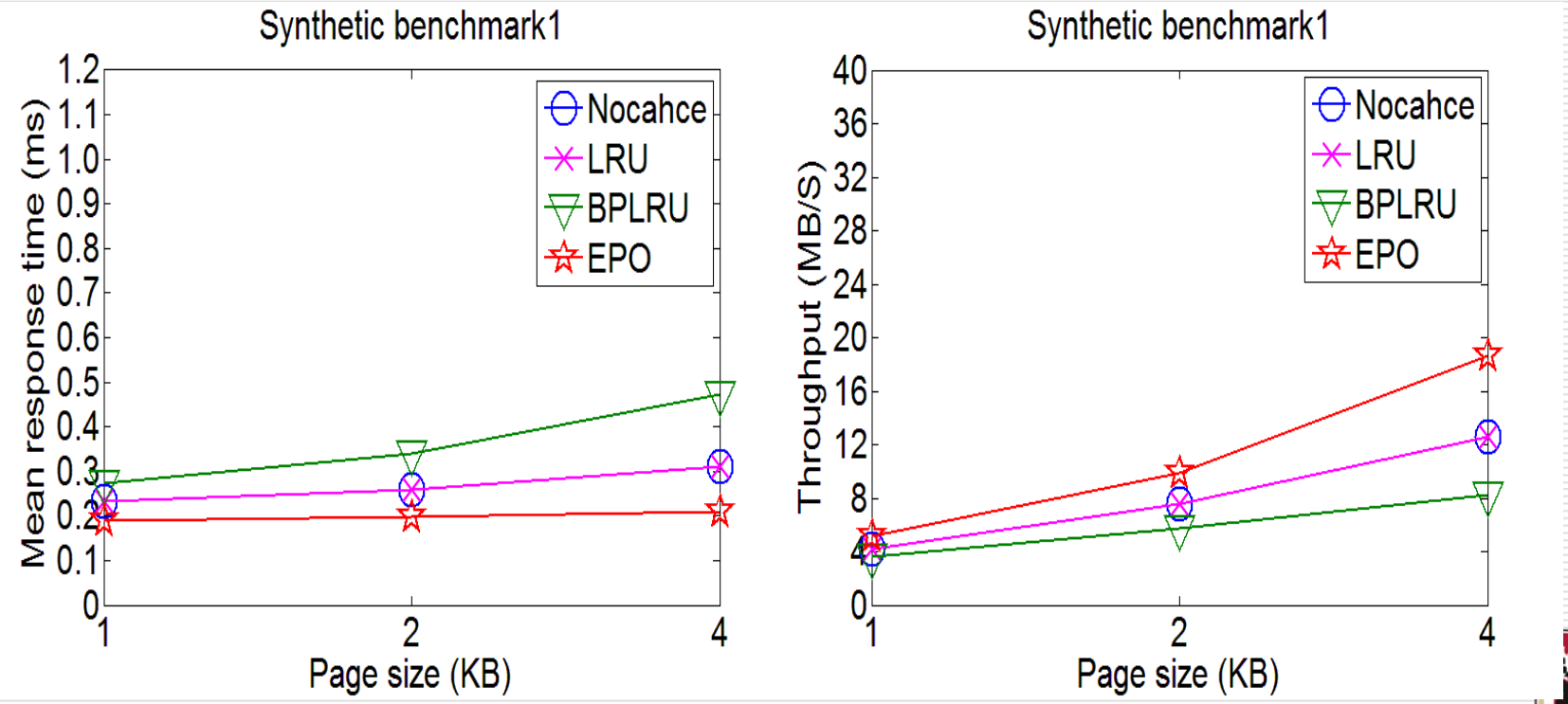
Continue…

- Synthetic benchmark2



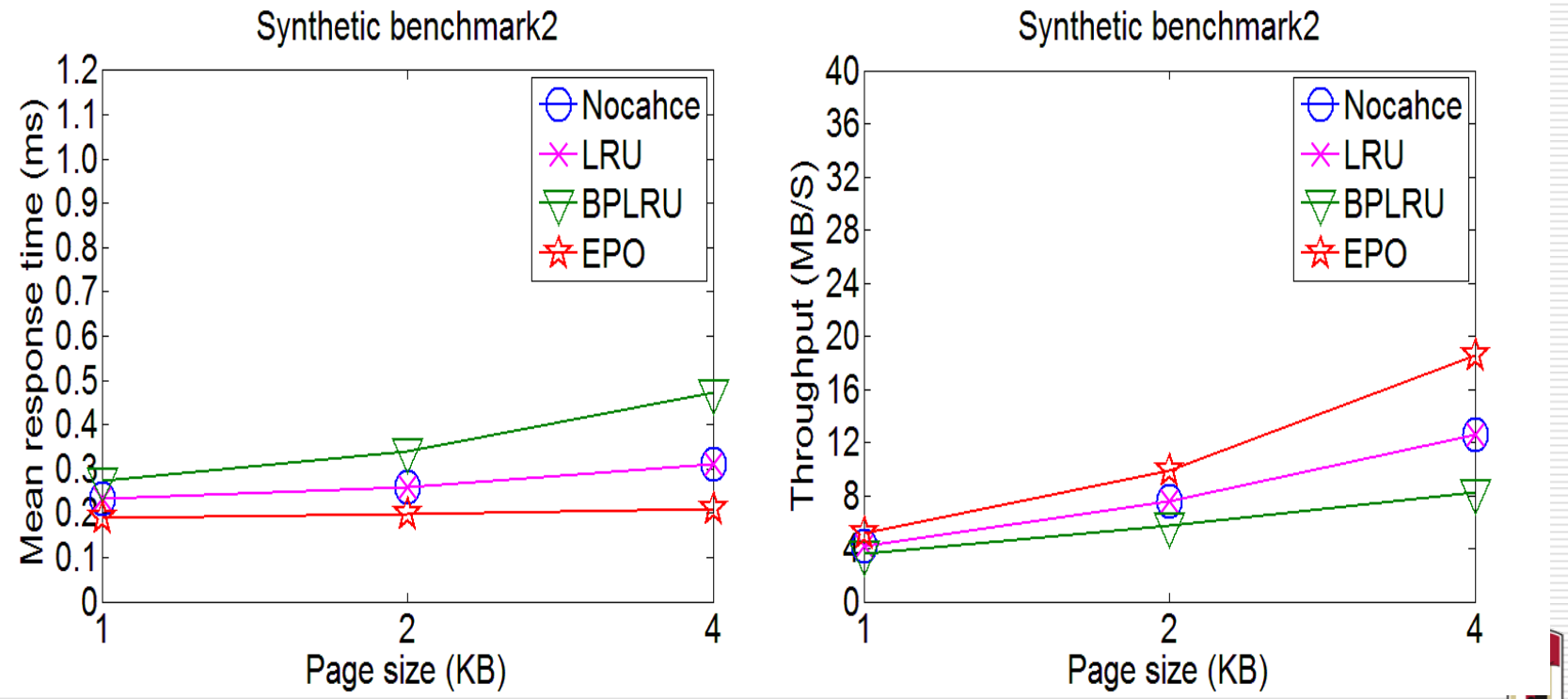
Varying page size

- Synthetic benchmark1



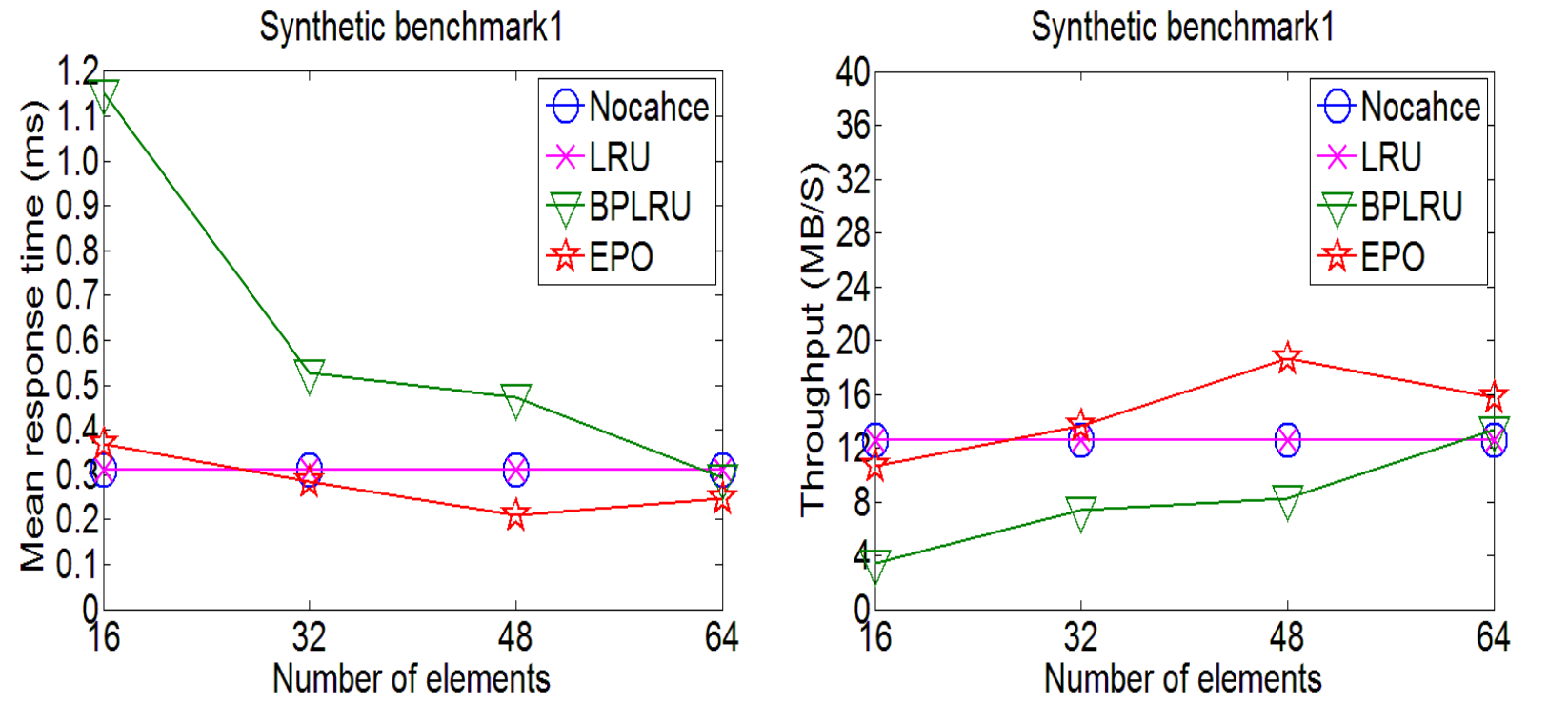
Continue…

- Synthetic benchmark2



Varying number of elements

- Synthetic benchmark1



Continue…

- Synthetic benchmark2