Performance and energy analysis using transactional workloads

Anastasia Ailamaki

EPFL and RAW Labs SA

students: Danica Porobic, Utku Sirin, and Pinar Tozun

Online Transaction Processing

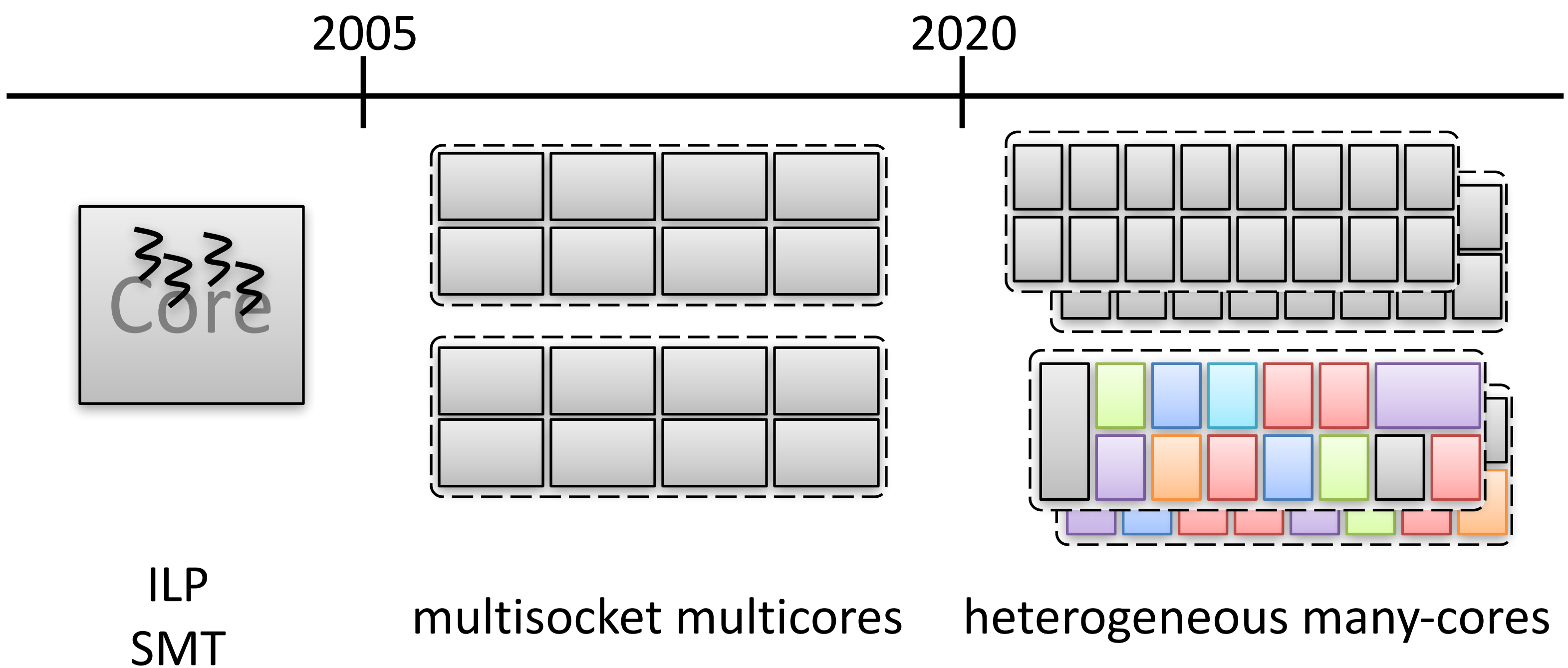
$20B+ industry

Characteristics:

- Has many concurrent requests
- Touch small part of whole data
- Need high & predictable performance

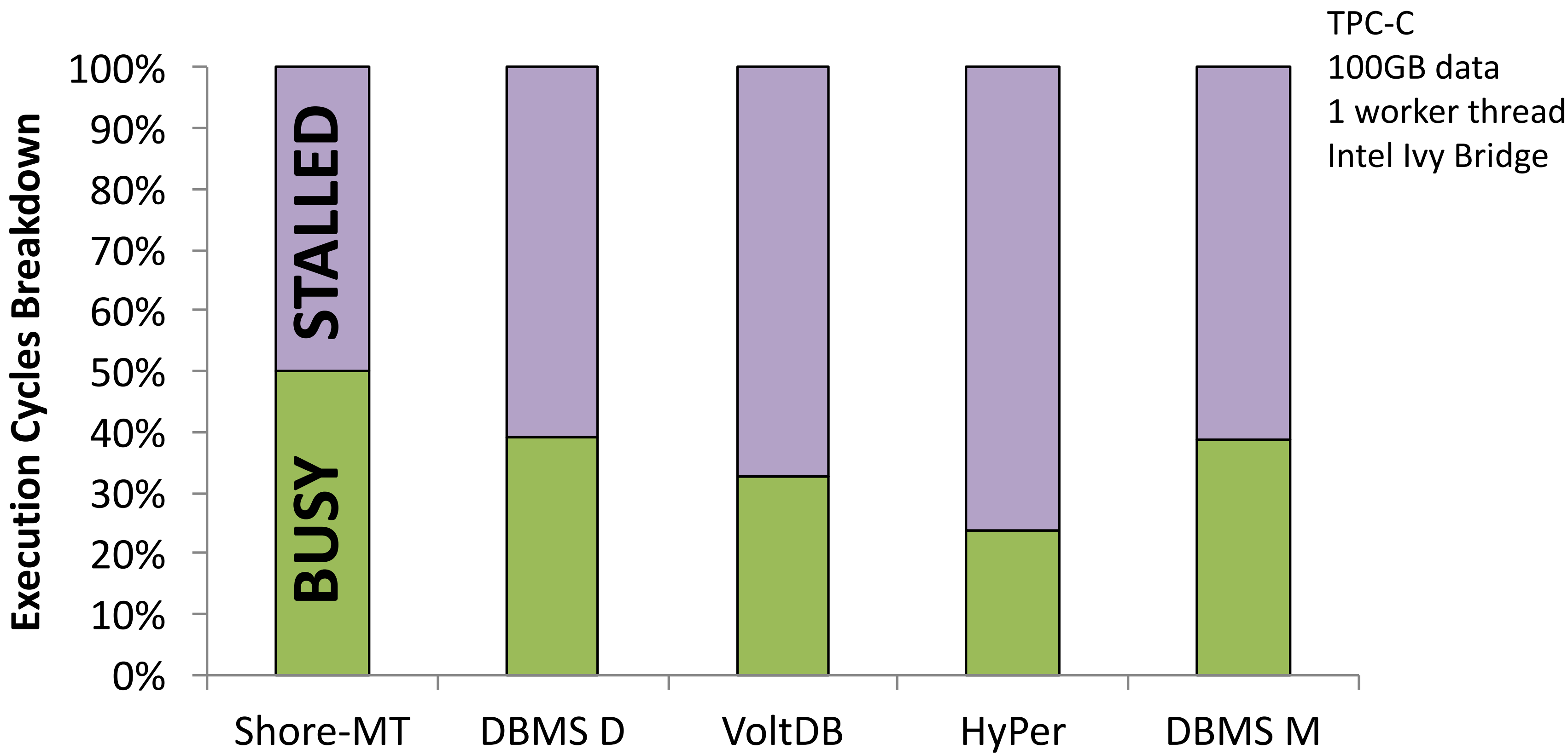
Primary application for databases

Hardware OLTP runs on



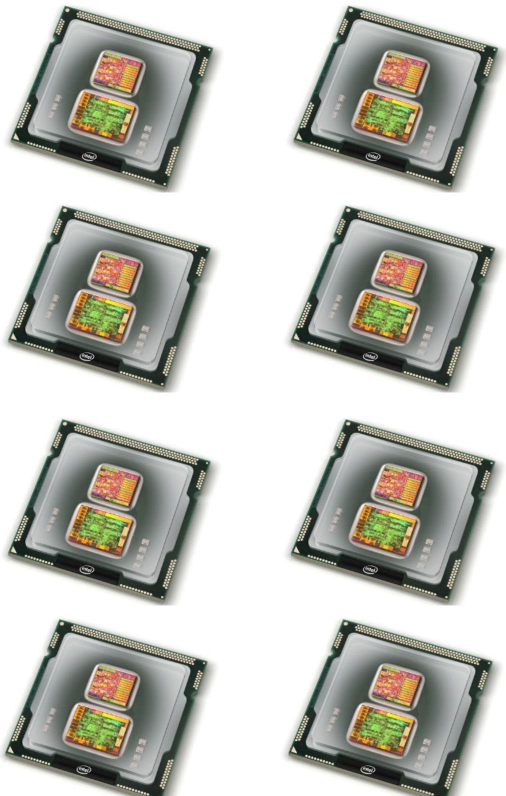
Hardware keeps providing new forms of parallelism
How’s the utilization?

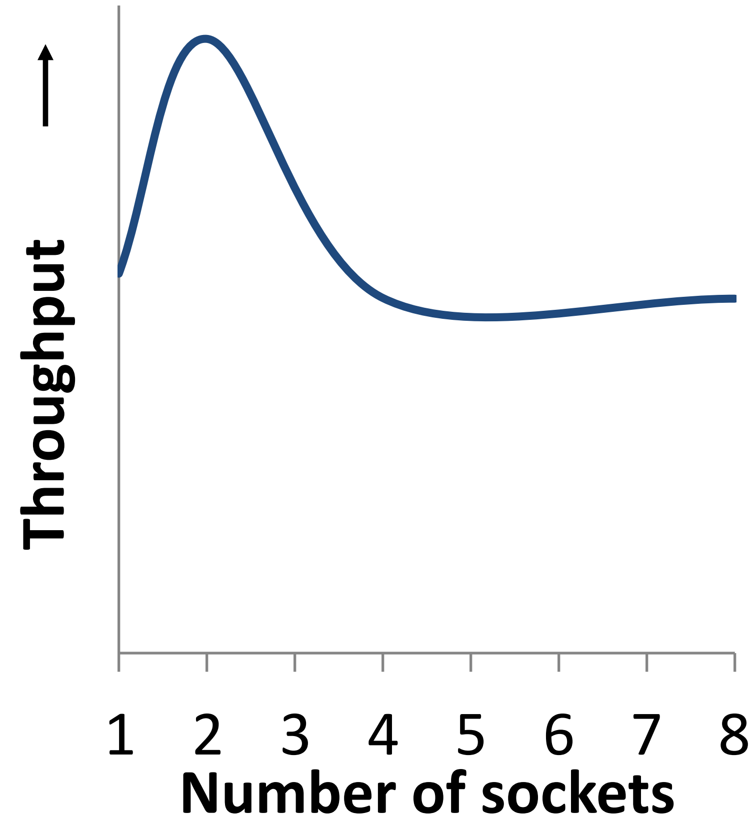
Utilizing modern processors



Processor stalled most of the time

Scaling up OLTP on multisockets

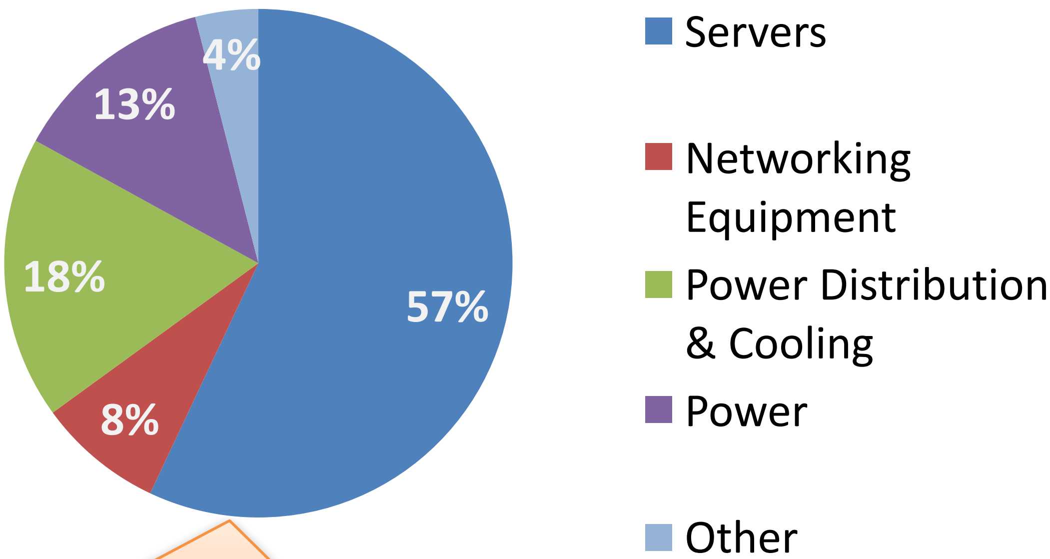




Multisocket servers severely under-utilized

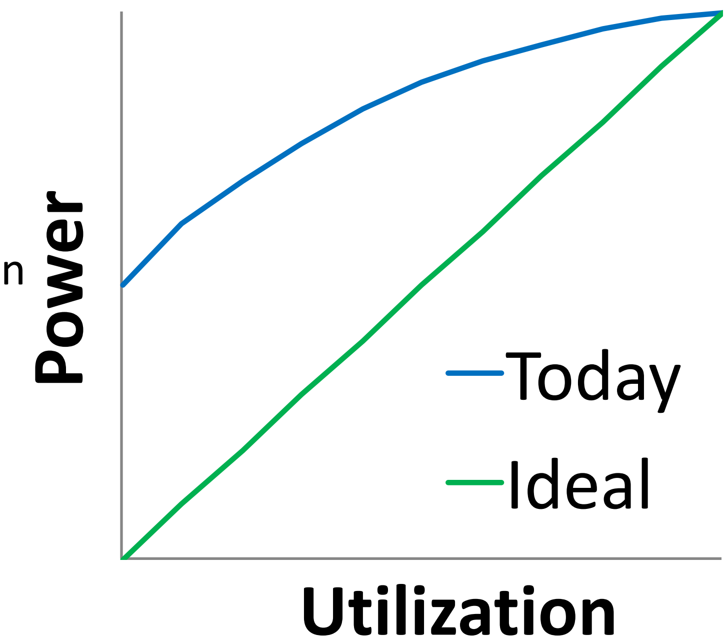
Why care about power?

Monthly datacenter costs [J. R. Hamilton]



30% power-related
Dynamic fraction increasing

Energy proportionality



Energy efficiency as important as performance

- Why is my system under-utilizing hardware?
- Why isn’t my system faster on new hardware?
- Are new processors more energy-efficient?


बस अड्डा
I.S.B.T.
लाल किला
RED FORT
दिल्ली गेट
DELHI GATE

Analyzing performance and energy

- Macrobenchmarks or Microbenchmarks?
- Execution time breakdowns
- Measuring energy efficiency

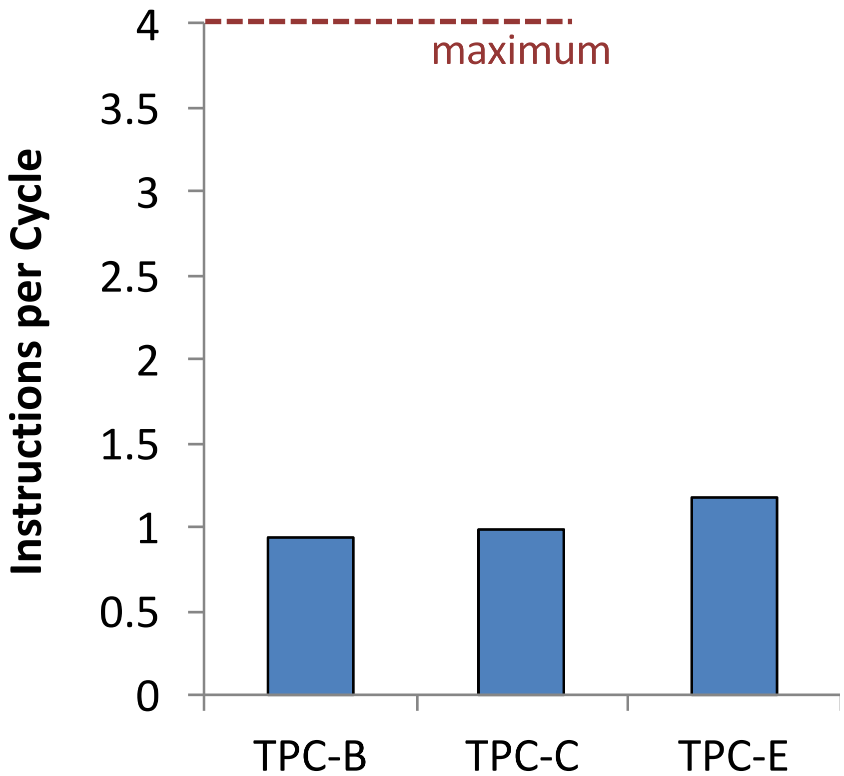
Analyzing performance and energy

- **Macrobenchmarks or Microbenchmarks?**
- Execution time breakdowns
- Measuring energy efficiency

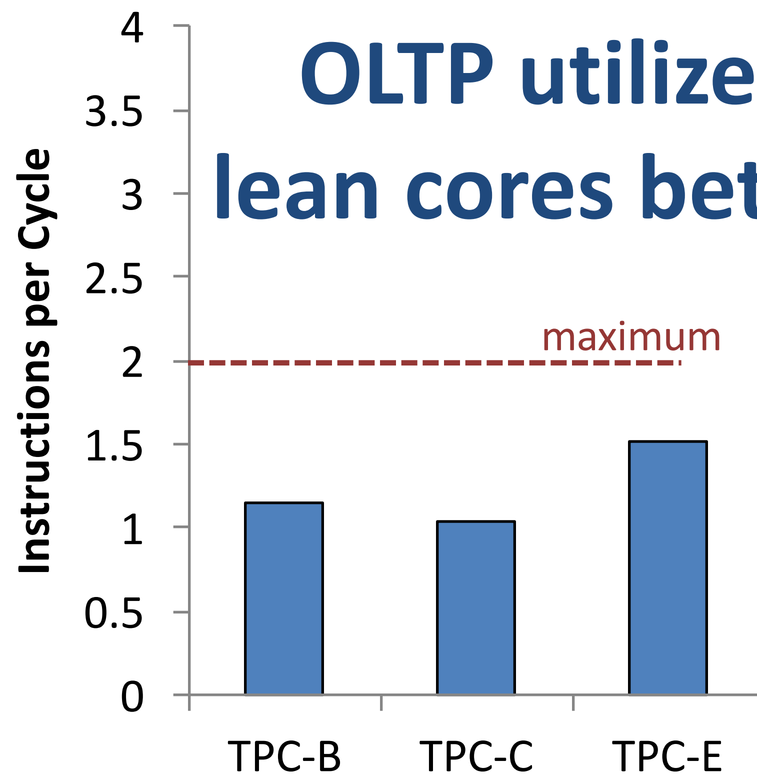
Utilization (microarchitecture level)

[EDBT13]

Intel Xeon X5660



Sun Niagara T2

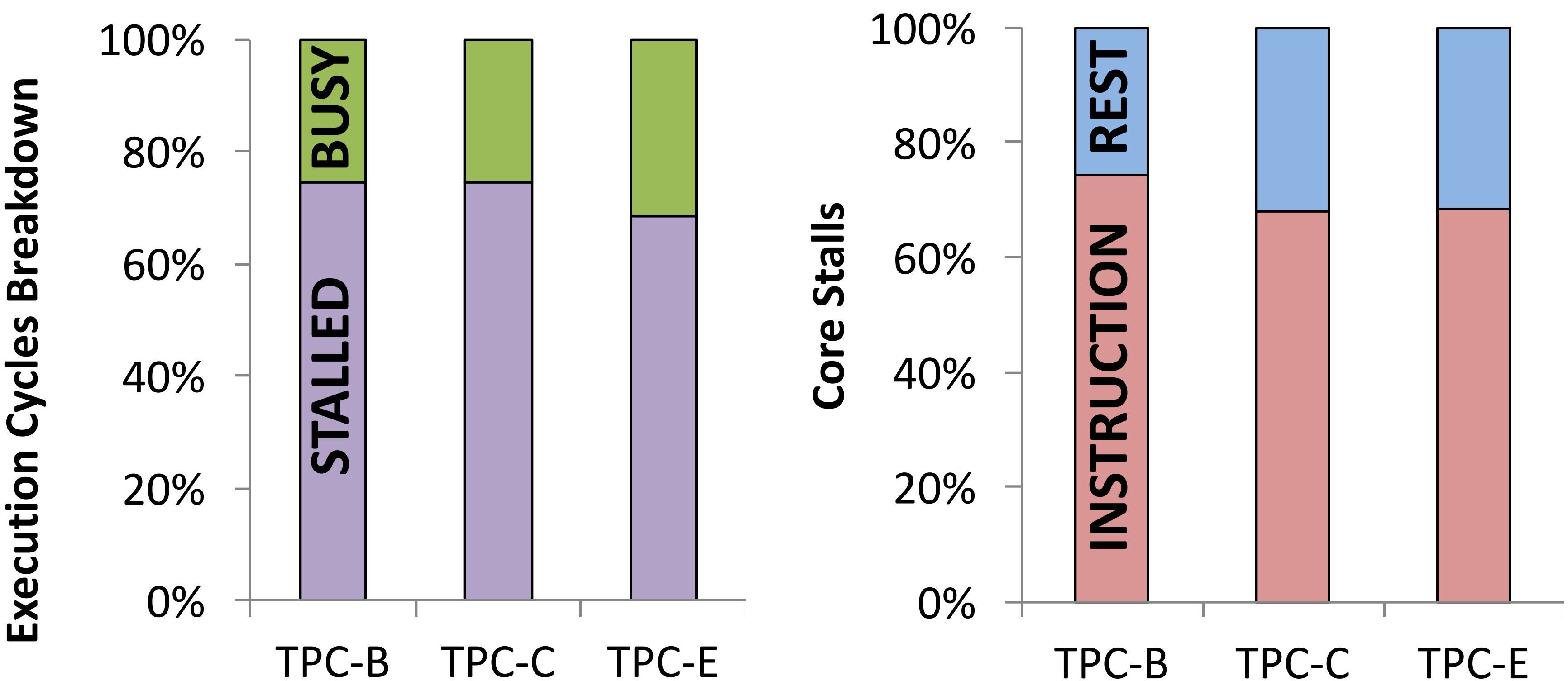


OLTP utilizes lean cores better

TPC-E has higher IPC

Macrobenchmark: Execution Cycles & Stalls

Intel Xeon X5660 [EDBT13]

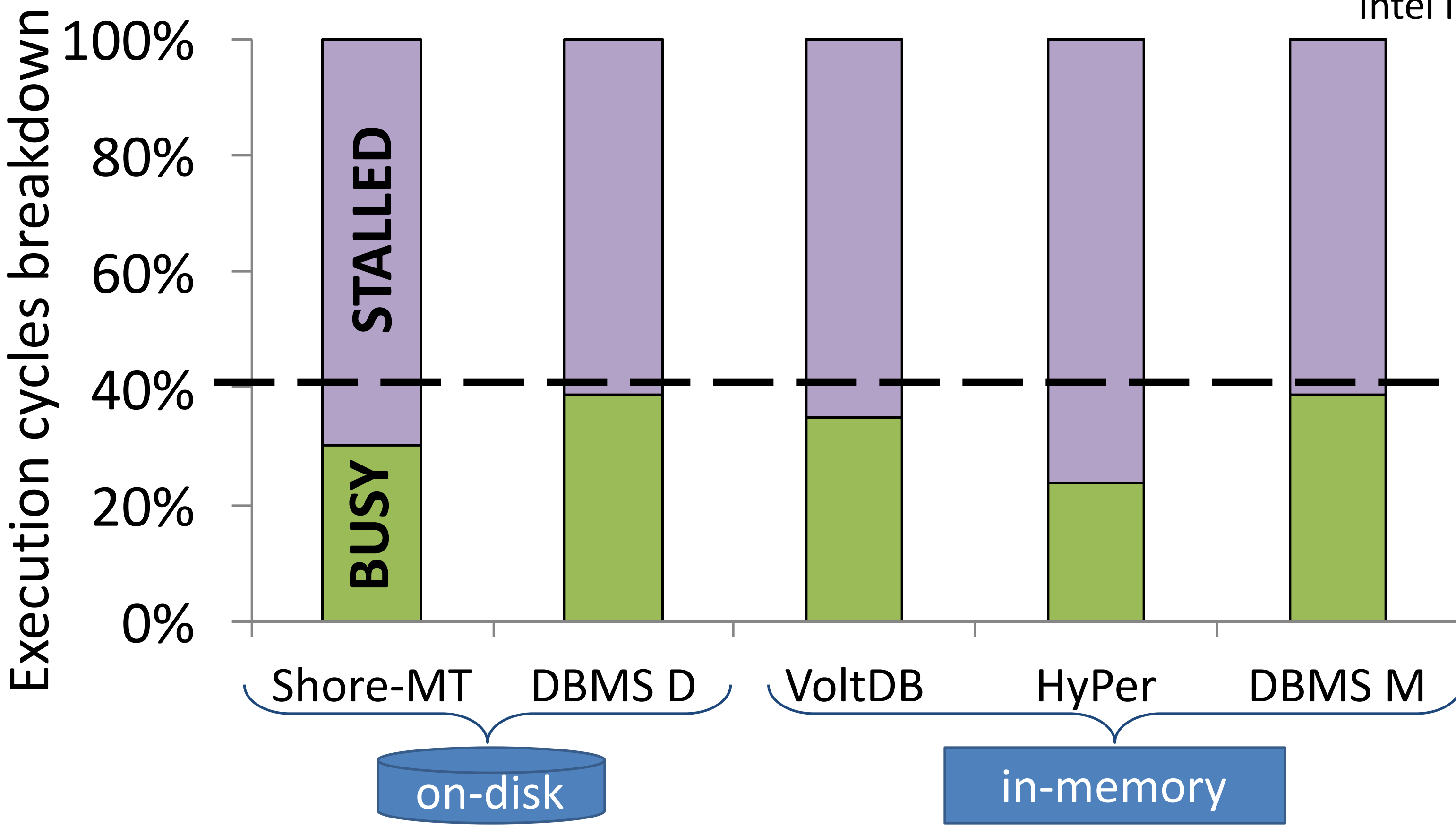


Over 70% of time goes to stalls

Instruction stalls are the main problem

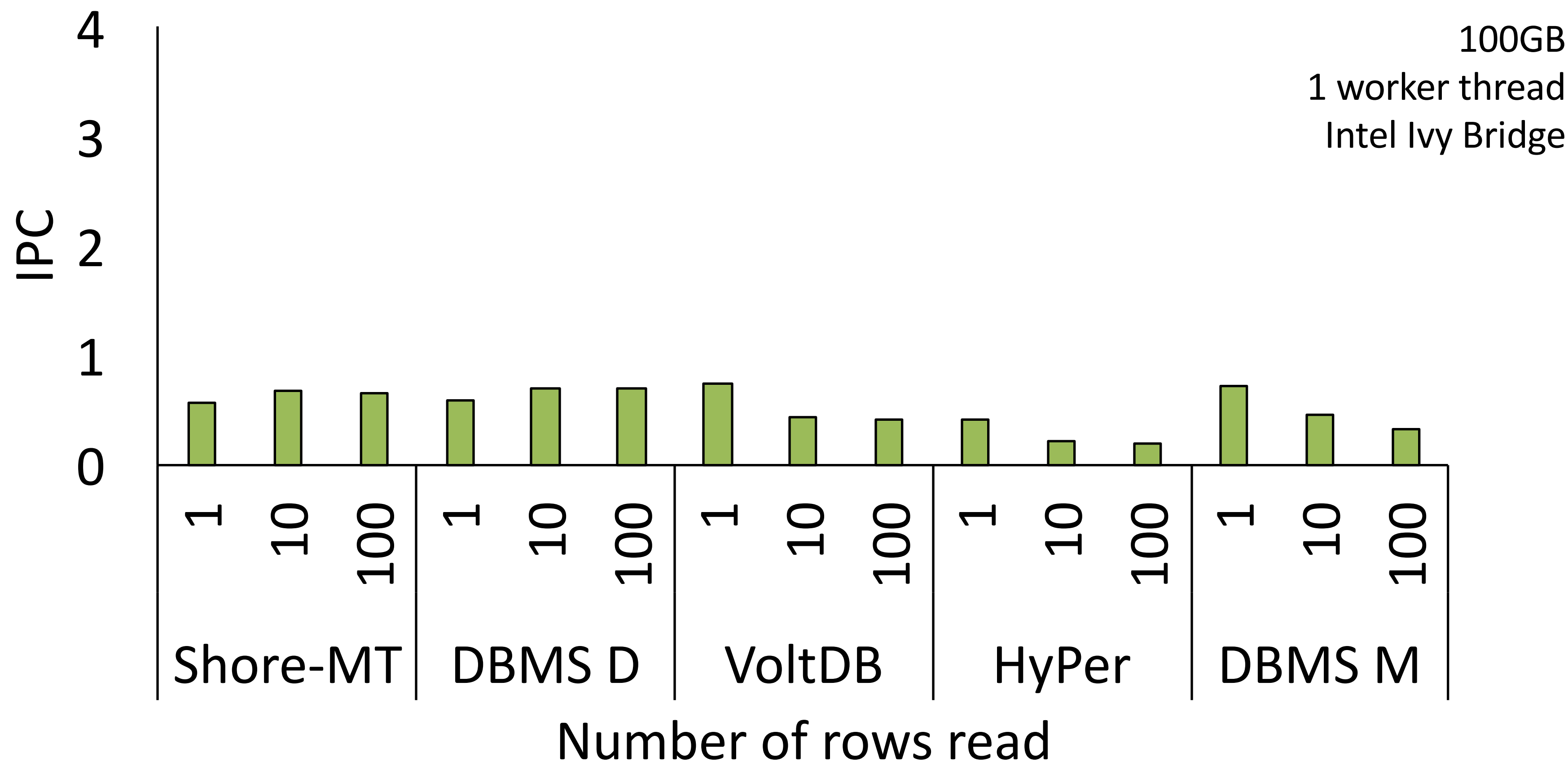
Utilization across systems

[SIGMOD16]
TPC-C, 100GB
1 worker thread
Intel Ivy Bridge



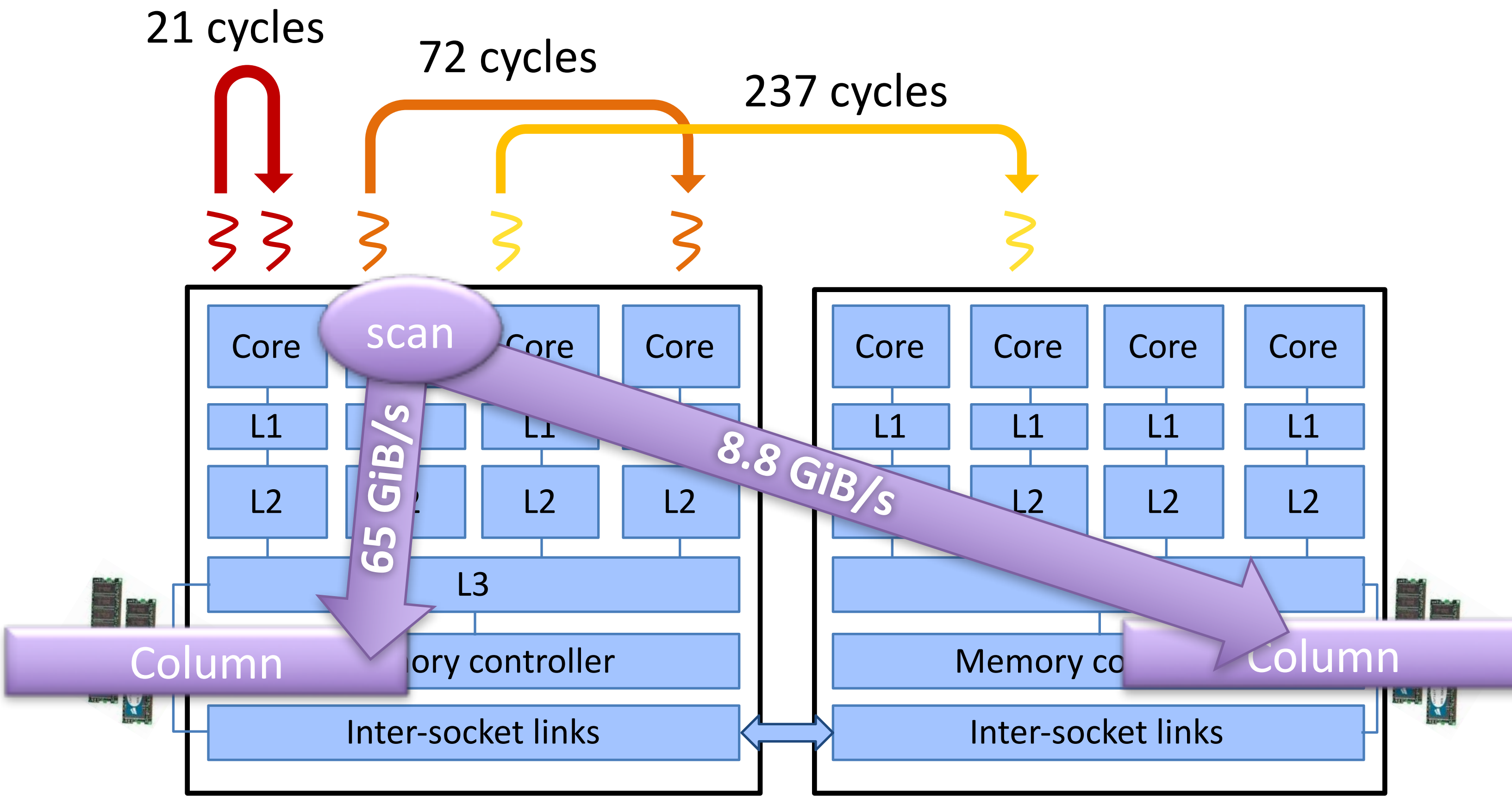
Even in-memory systems stall > 60% time

Microbenchmarks for what-if analysis



Lower data locality → low IPC for some systems

OLTP on hardware islands

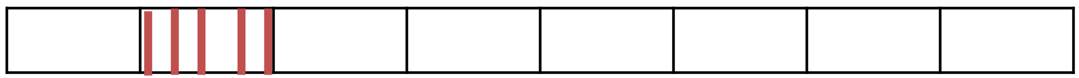


How measure impact?

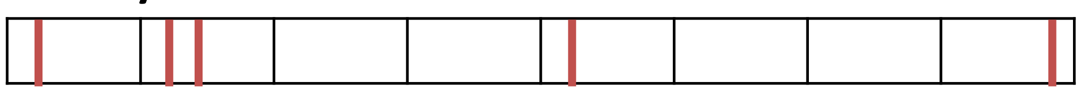
[PVLDB12]

Partition sensitive microbenchmark

- Single site version
  - probe/update N rows from the local site

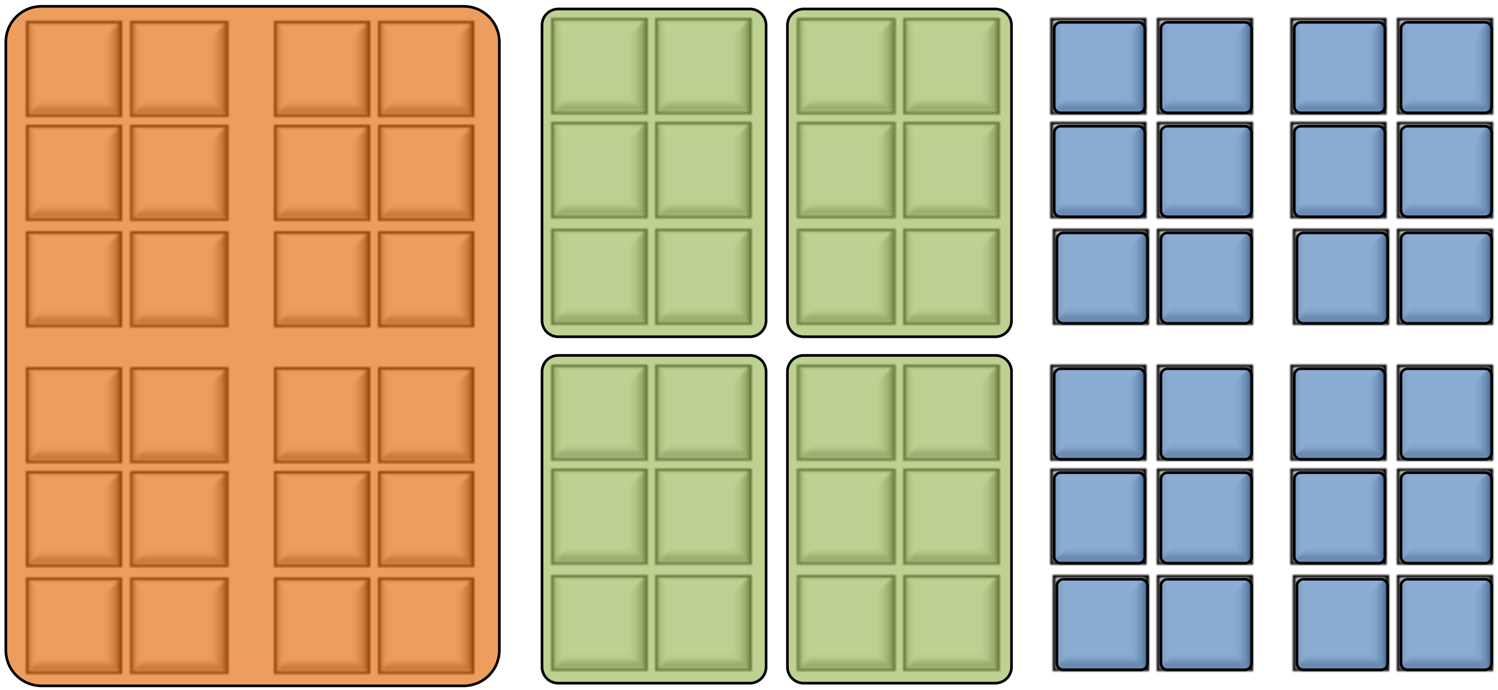


- Multisite version
  - probe/update 1 row from the local site
  - probe/update N-1 rows uniformly from any site
  - sites may reside on the same instance



[PVLDB12]

OLTP deployment configurations

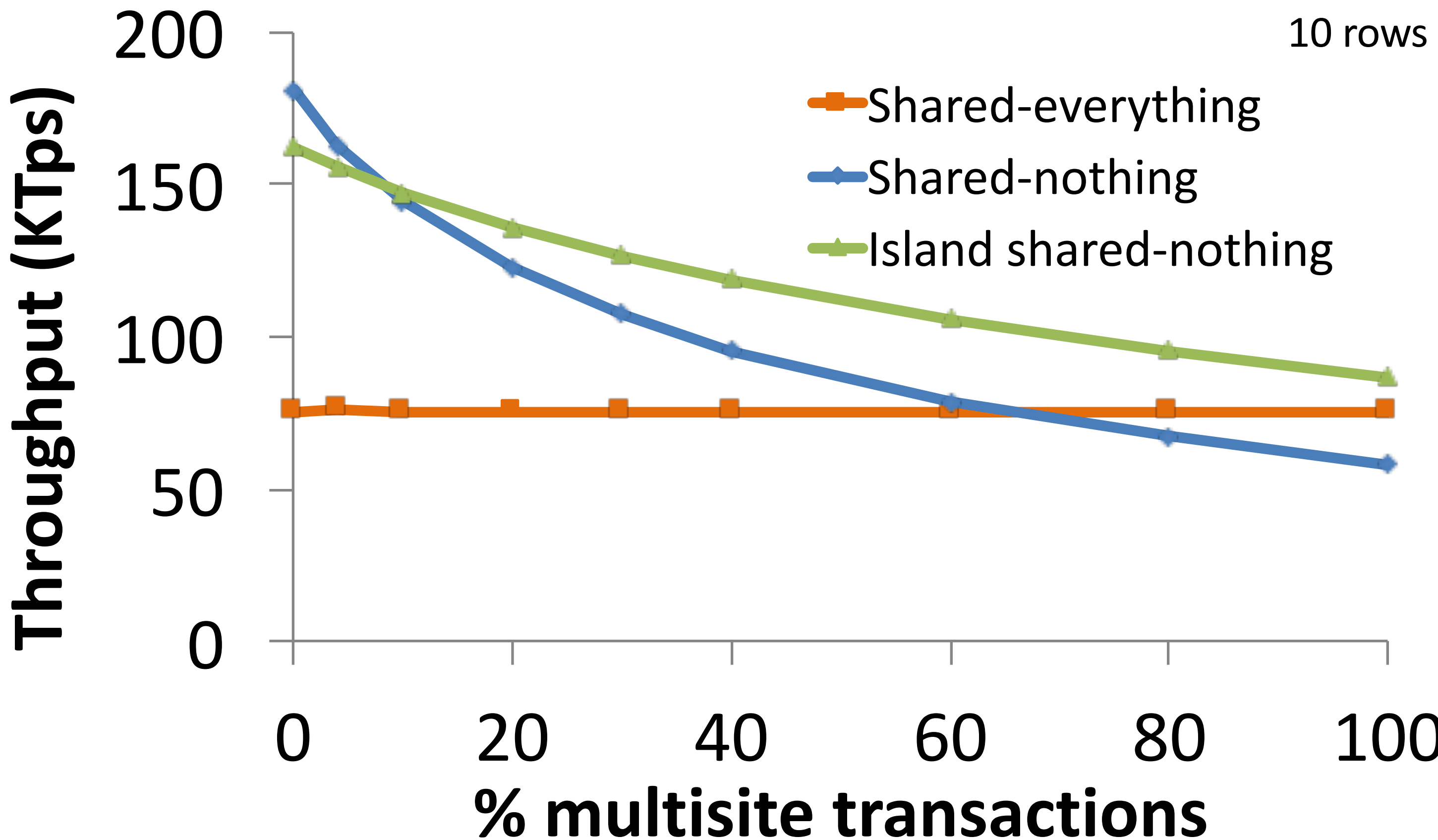


Shared-everything Island shared-nothing Shared-nothing

[PVLDB12]

Multisite transactions: read only

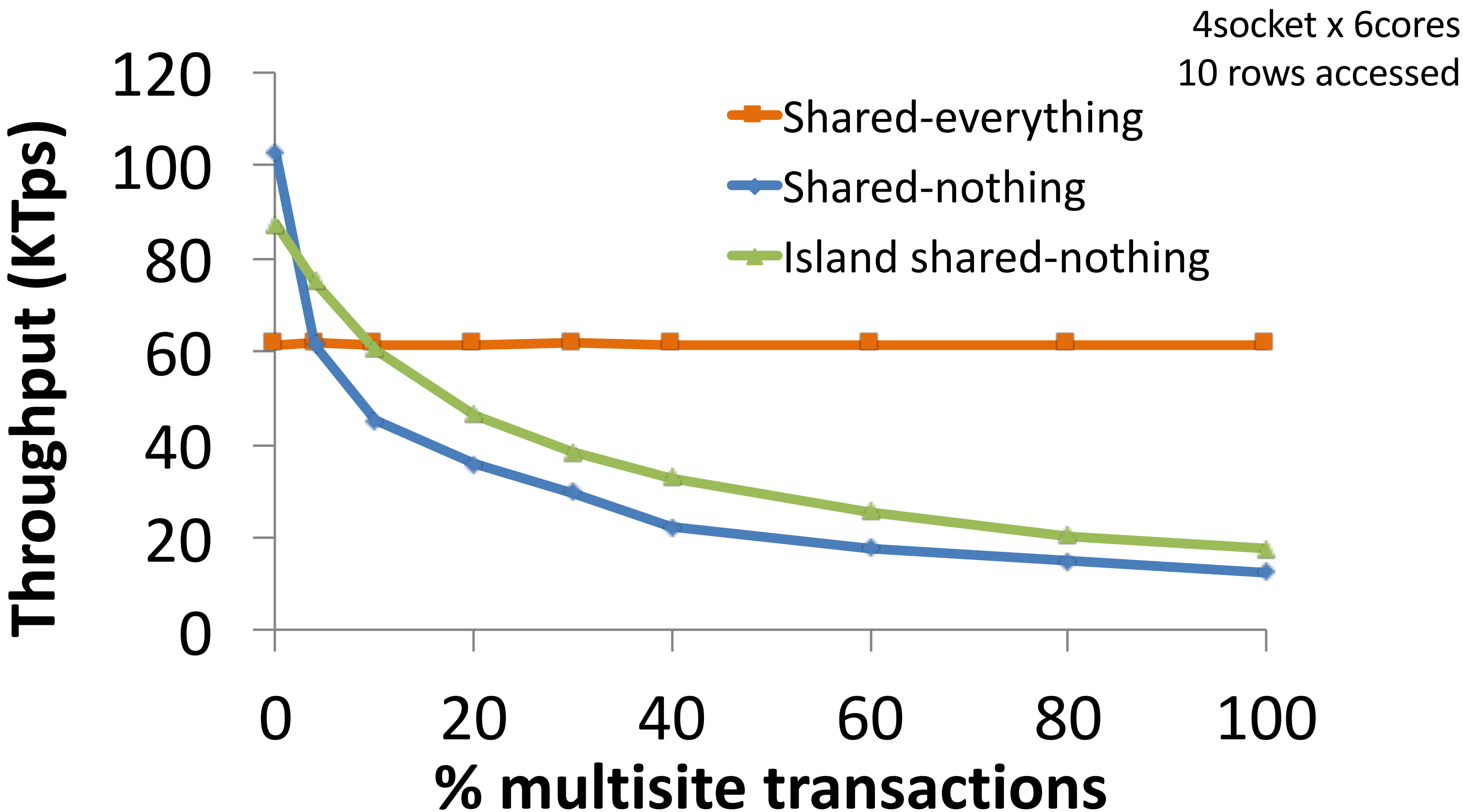
4socket x 6cores
10 rows accessed



More instances -> faster performance degradation

[PVLDB12]

Multisite transactions: updates



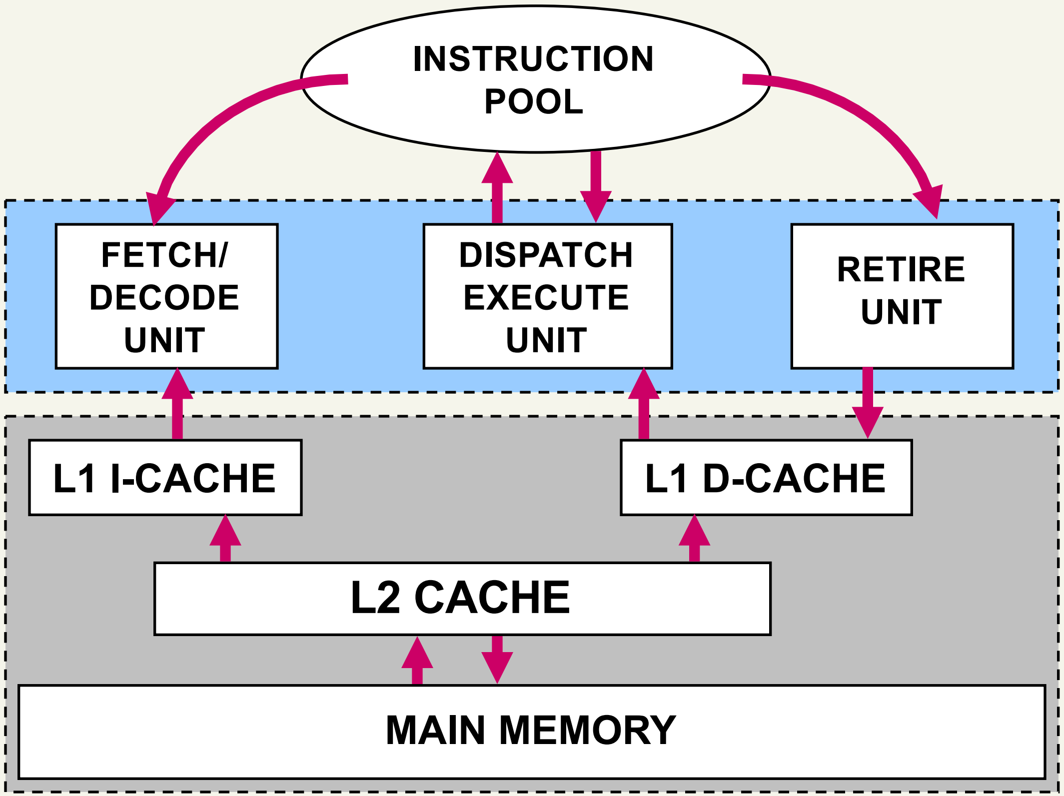
Update distributed transactions are more expensive

Analyzing performance and energy

- Macrobenchmarks or Microbenchmarks?
- **Execution time breakdowns**
- Measuring energy efficiency

My first toy: PII Xeon

[VLDB99]



+ Branch prediction, non-blocking caches, out-of-order

Where Does Time Go?

[VLDB99]

- Computation
- Stalls
  - Cache misses
  - Branch mispredictions
  - Other execution pipeline stalls

✪ Stall time and computation overlap

$$\textbf{Time} = T_{Computation} + T_{Memory} + T_{Branch} + T_{Resource} - T_{Overlap}$$

Setup and Methodology

[VLDB99]

Range Selection
(sequential, indexed)

```
select avg (a3)
from R
where a2 > Lo and a2 < Hi
```

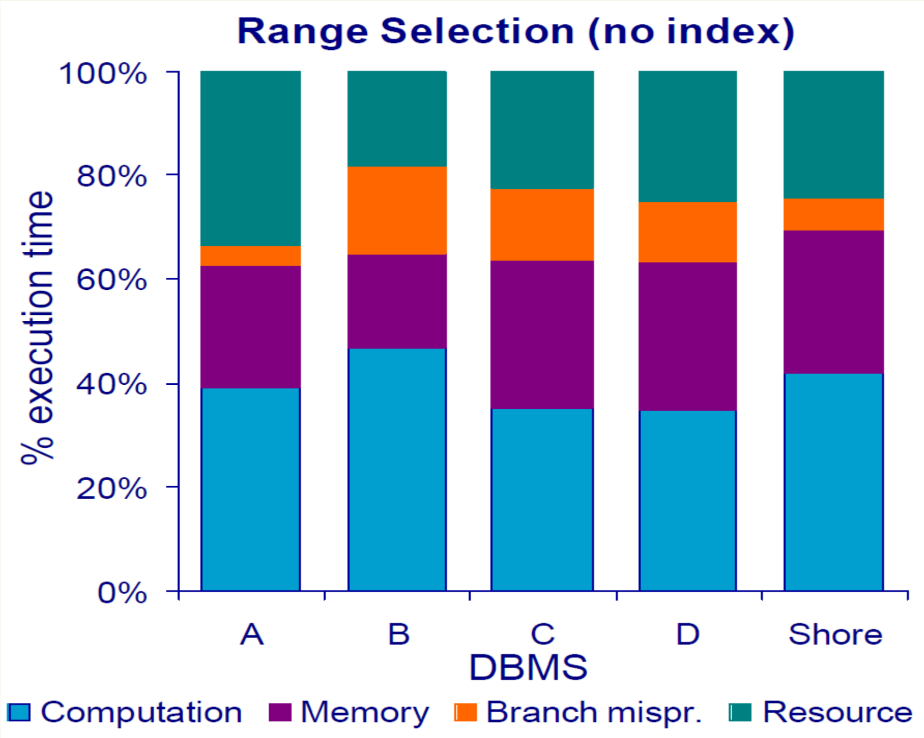
Equijoin
(sequential)

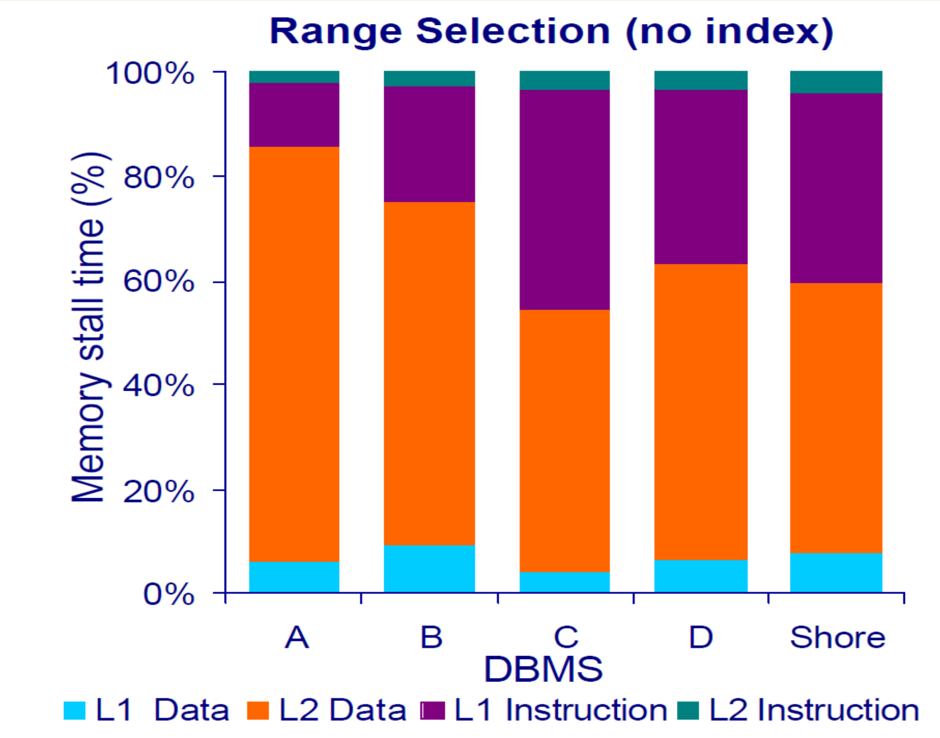
```
select avg (a3)
from R, S
where R.a2 = S.a1
```

- Four commercial DBMSs: A, B, C, D
- 6400 PII Xeon/MT running Windows NT 4
- Used PII counters
- Correctness: Measured & computed CPI

Two very useful breakdowns

[VLDB99]





processor stalled >50% of time

most stalls: L1I and L2D

Adapted formula

[SIGMOD16]

branch-misp *x* penalty

$$T_Q + T_{OVL} = T_C + T_M + T_B + T_R$$

$$T_M = T_{L1I} + T_{L1D} + T_{L2I} + T_{L2D} + T_{L3I} + T_{L3D}$$

+

$$T_{DTLB} + T_{ITLB}$$

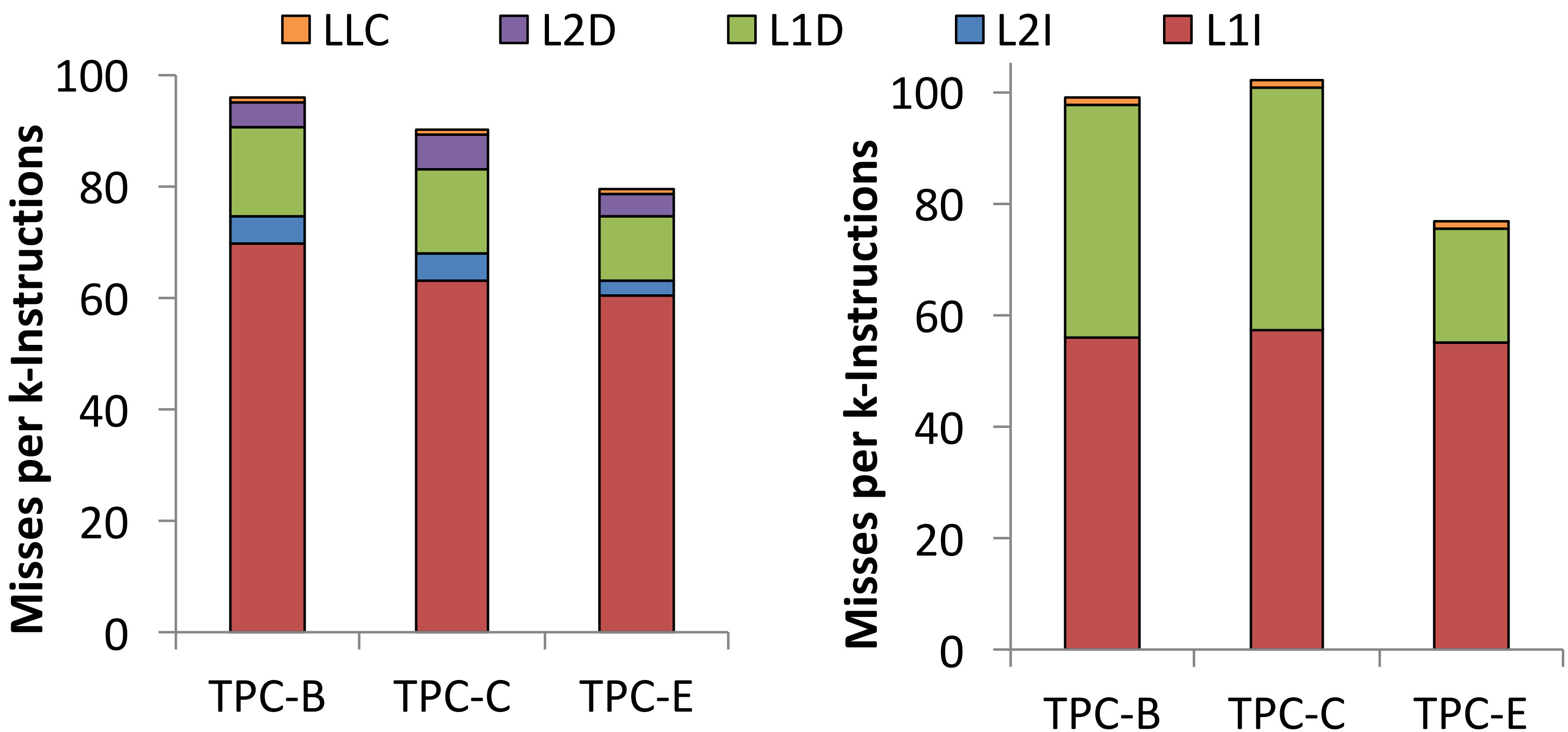
$$T_R = T_{FU} + T_{DEP} + T_{MISC}$$

[EDBT13]

Cache Misses today

Intel Xeon X5660
32KB L1-I & 32 KB L1-D

Sun Niagara T2
16KB L1-I & 8KB L1-D

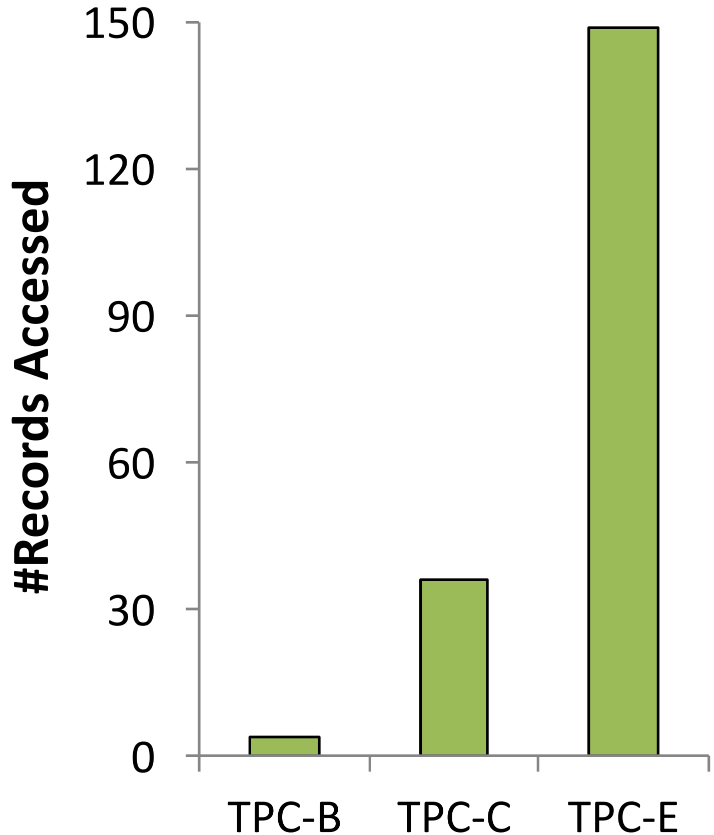


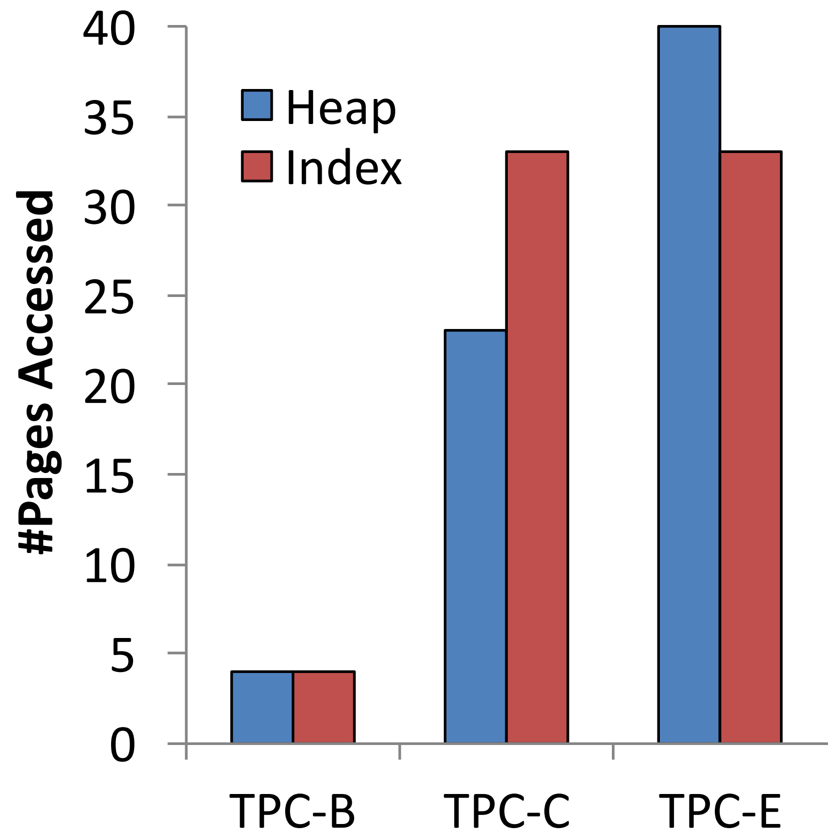
L1-I misses dominate

TPC-E has lower data miss ratio

TPC-E's lower miss ratio

Average per transaction

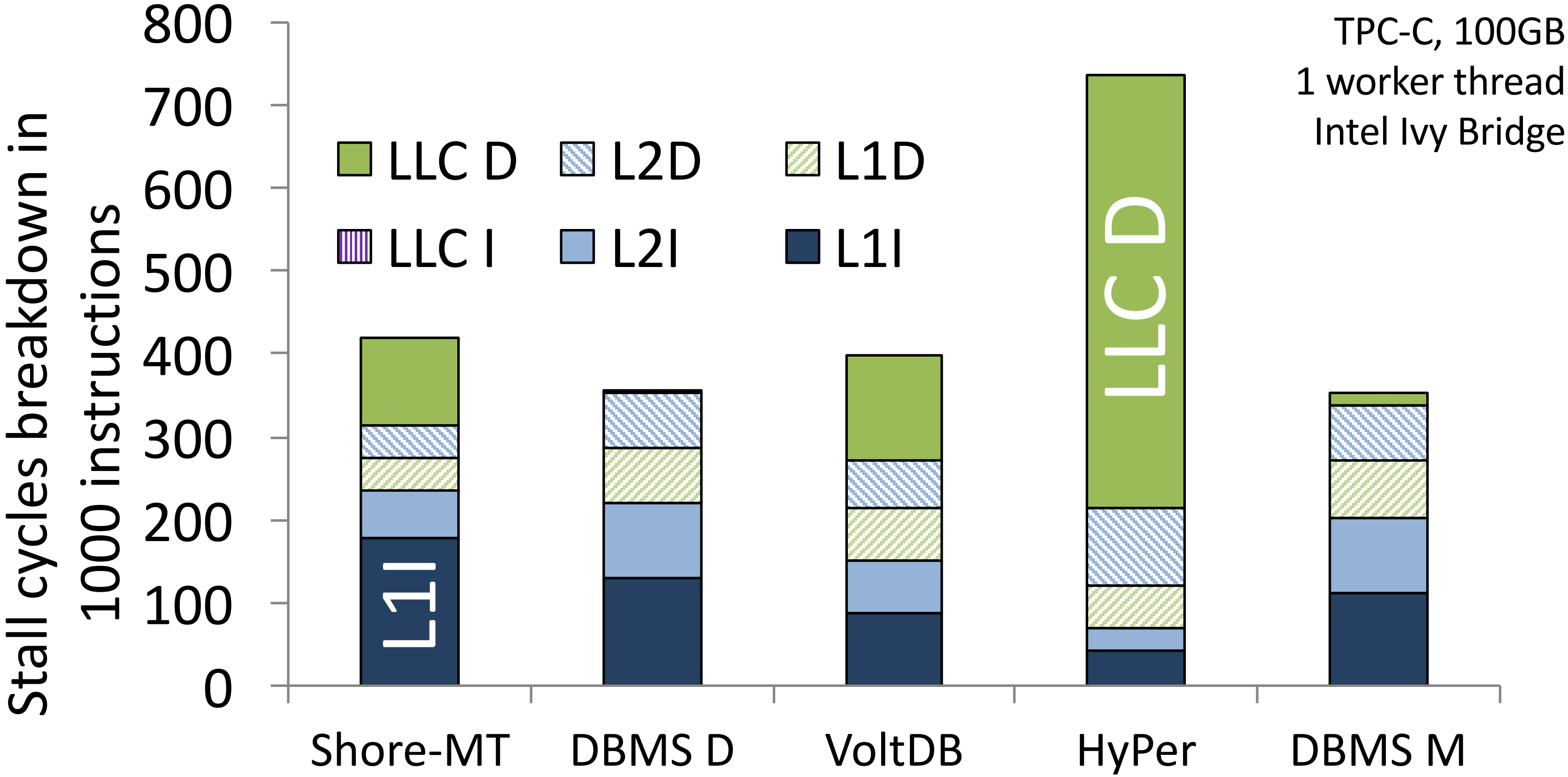




More scans → Increased page reuse

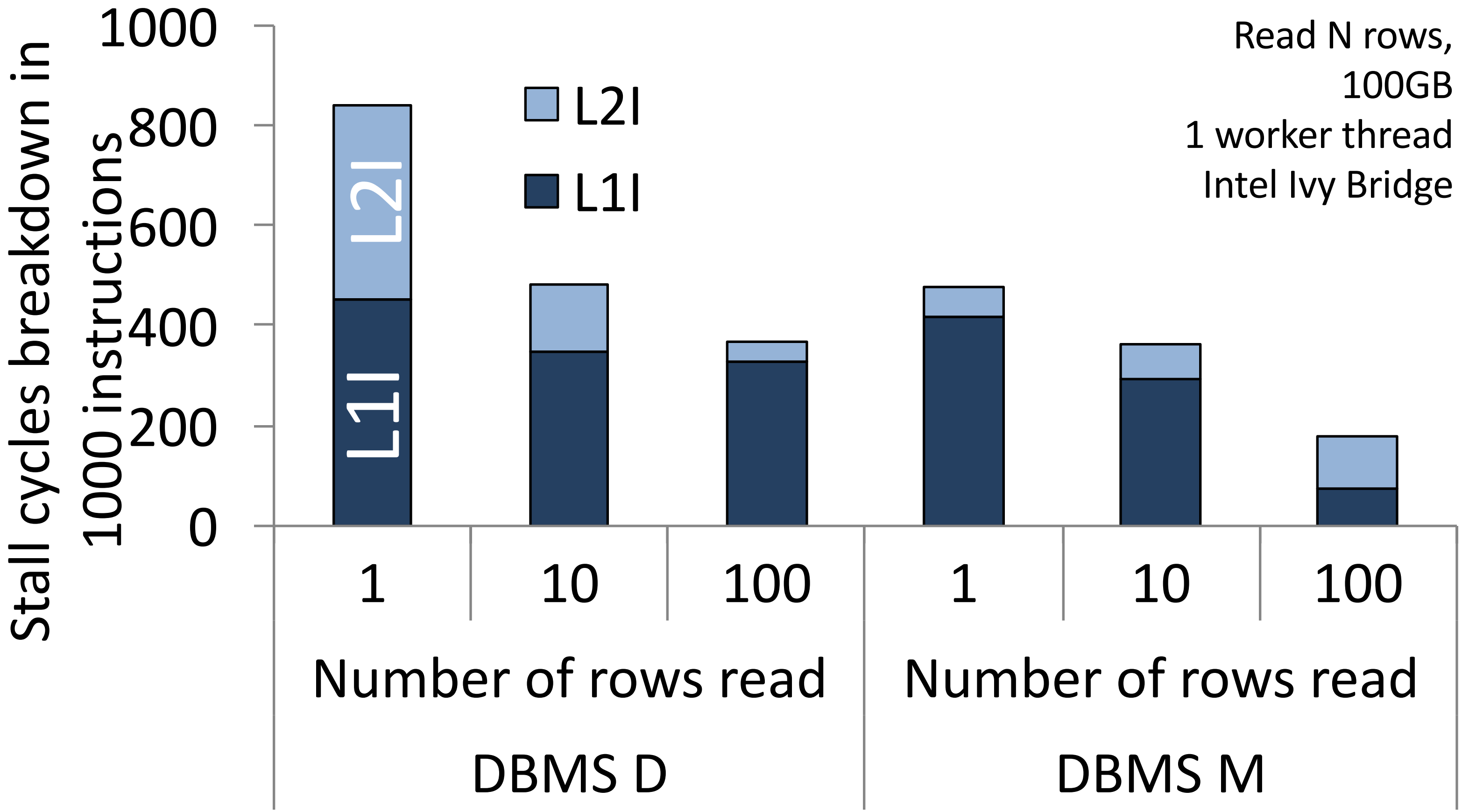
[SIGMOD16]

Breaking down clock cycles



L1I or LLC D stalls dominate

Where do L1I stalls come from?



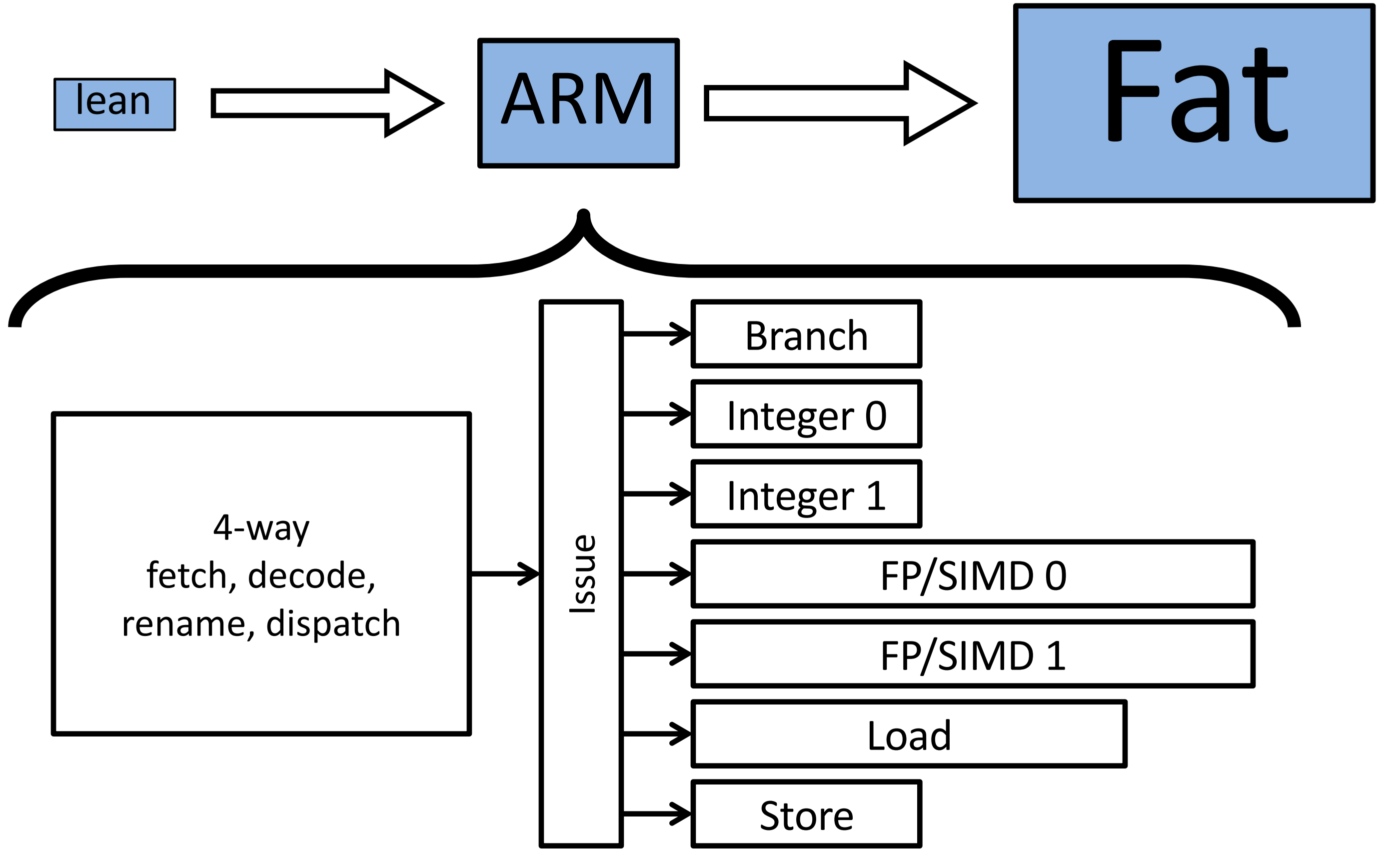
Code outside storage mgr → high L1I misses

Analyzing performance and energy

- Macrobenchmarks or Microbenchmarks?
- Execution time breakdowns
- **Measuring energy efficiency**

ARM server-grade processor

[DAMON16]



OLTP on ARM: performance & power?

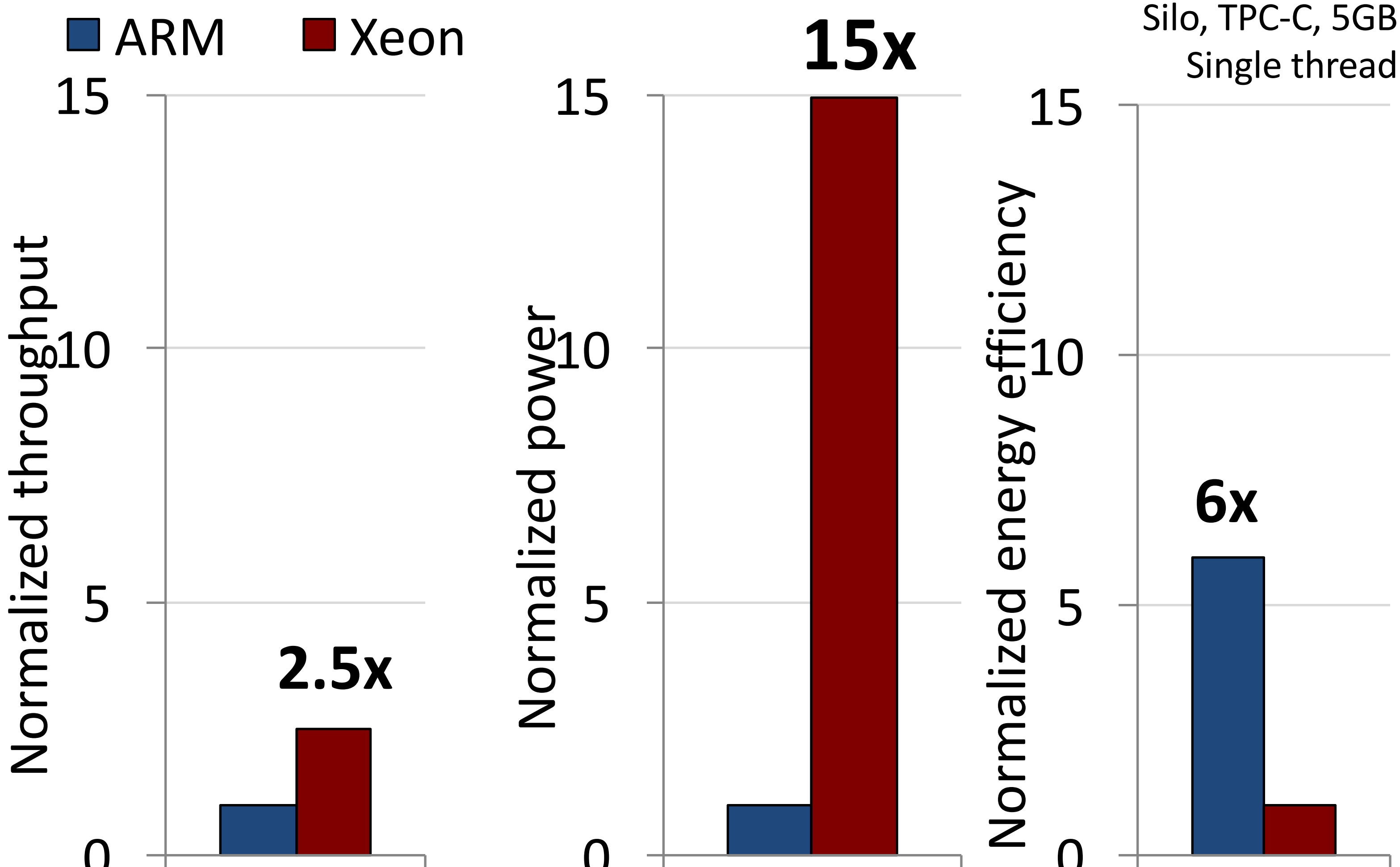
[DAMON16]

Xeon vs. ARM

Processor	Intel Xeon	ARM Cortex-A57
# Sockets	2 (one is active)	1
# Cores/socket	8	8
Issue width	4	4
Clock speed	2.00GHz	2.00Ghz
L1I / L1D	32KB / 32KB	32KB / 32KB
L2	256KB	256KB
L3 (shared)	20MB	8MB
RAM	256GB	16GB

[DAMON16]

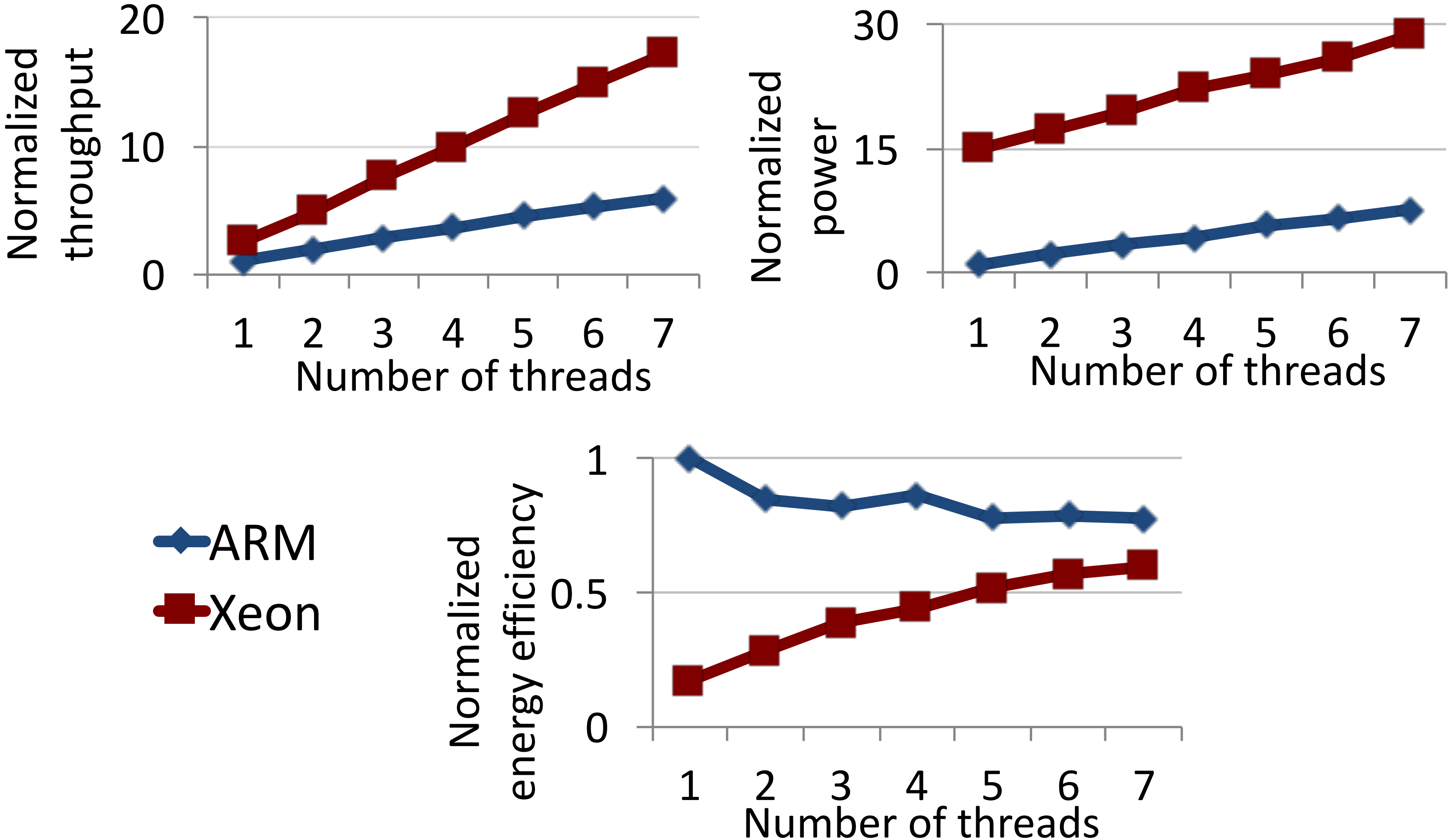
ARM is a promising alternative



[DAMON16]

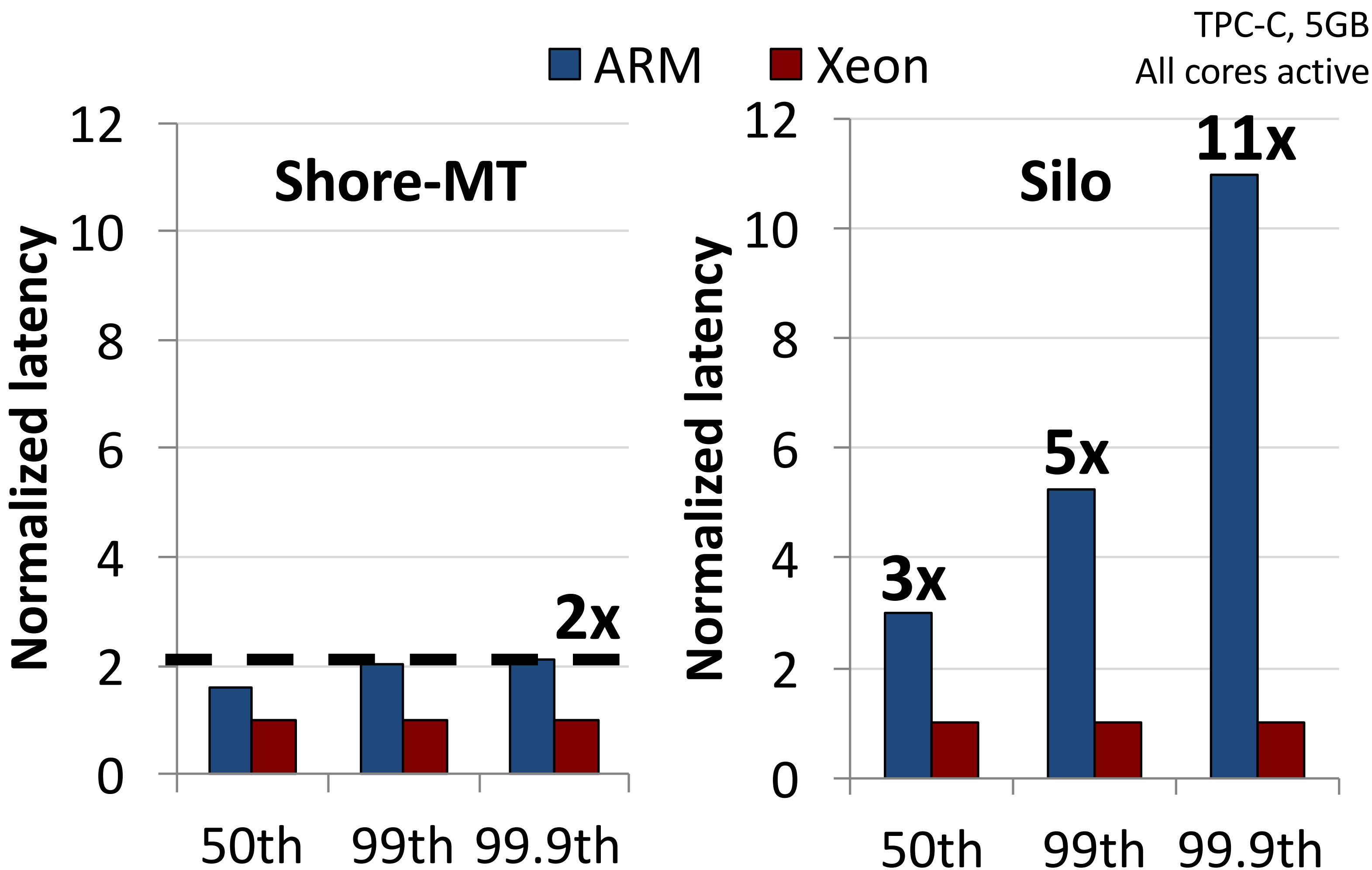
ARM achieves energy proportionality

Silo, TPC-C, 5GB



[DAMON16]

ARM is less suitable for low latency



Lessons learned

- Macrobenchmarks show big picture
- Microbenchmarks reveal details
- Breakdowns correlate numbers
- Sensitivity analysis highlights trends
- Right methodology is essential for understanding behavior

References

[DAMON16] U. Sirin, R. Appuswamy, and A. Ailamaki: OLTP on a server-grade ARM.

[EDBT13] P. Tözün, I. Pandis, C. Kaynak, D. Jevdjic, and A. Ailamaki: From A to E: Analyzing TPC's OLTP Benchmarks – The obsolete, the ubiquitous, the unexplored.

[PVLDB12] D. Porobic, I. Pandis, M. Branco, P. Tözün, and A. Ailamaki: OLTP on Hardware Islands.

[SIGMOD16] U. Sirin, P. Tözün, D. Porobic, and A. Ailamaki: Micro-architectural Analysis of In-memory OLTP

[VLDB99] A. Ailamaki, D. DeWitt, M. Hill, and D. Wood: DBMSs On A Modern Processor: Where Does Time Go?