Everything You Always Wanted to Know About Synchronization but Were Afraid to Ask

Tudor David, Rachid Guerraoui, Vasileios Trigonakis

École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

{tudor.david, rachid.guerraoui, vasileios.trigonakis}@epfl.ch

The problem

- The designer of a concurrent system still has little indication of:
  - a priori, of whether a given synchronization scheme will scale on a given modern many-core architecture
  - a posteriori, about exactly why a given scheme did, or did not, scale.

The Analysis

depth	breadth
concurrent software hash table, Memcached, STM	**single-socket** uniform (Sun Niagara 2) non-uniform (Tilera TILE-Gx36)
primitives locks, message passing	
atomic operations CAS, FAI, TAS, SWAP	**multi-socket** directory-based (AMD Opteron) broadcast-based (Intel Xeon)
cache coherence loads, stores	

“The paper presents the most exhaustive study of synchronization on many-cores”

Background

- Hardware-based synchronization
- Software-based synchronization

Background

- Hardware-based synchronization
  - Cache-coherence is the most used protocol to maintain consistency data on the cache memory
  - It implements the read and write operations on cache
  - Most processors use the MESI cache coherence protocol
    - **M**odified, **E**xclusive, **S**hared, **I**nvalid

Background

- Hardware-based synchronization
  - Cache-coherence is the most used protocol to maintain consistency data on the cache memory
  - It implements the read and write operations on cache
  - Most processors use the MESI cache coherence protocol
    - **M**odified, **E**xclusive, **S**hared, **I**nvalid
      - The data are stale in the memory and no other cache has a copy of this line

Background

- Hardware-based synchronization
  - Cache-coherence is the most used protocol to maintain consistency data on the cache memory
  - It implements the read and write operations on cache
  - Most processors use the MESI cache coherence protocol
    - **M**odified, **E**xclusive, **S**hared, **I**nvalid
      - The data are up-to-date in the memory and no other cache has a copy of this line

Background

- Hardware-based synchronization
  - Cache-coherence is the most used protocol to maintain consistency data on the cache memory
  - It implements the read and write operations on cache
  - Most processors use the MESI cache coherence protocol
    - **M**odified, **E**xclusive, **S**hared, **I**nvalid
      - The data are up-to-date in the memory and other caches might have copies of the line

Background

- Hardware-based synchronization
  - Cache-coherence is the most used protocol to maintain consistency data on the cache memory
  - It implements the read and write operations on cache
  - Most processors use the MESI cache coherence protocol
    - **M**odified, **E**xclusive, **S**hared, **I**nvalid
      - The data are invalid

Background

- Software-based synchronization
  - The most popular technique are locks
    - Locks can be implemented as spinlocks, queue locks, hierarchical locks, suspended locks, among others.

Background

- Software-based synchronization
  - The most popular technique are locks
    - Locks can be implemented as spinlocks, queue locks, hierarchical locks, suspended locks, among others.
  - Message passing emerge as an alternative to locks

Platforms

	Multi-socket		Single-socket	
Name	**Opteron**	**Xeon**	**Niagara**	**Tilera**
System	AMD Magny Cours	Intel Westmere-EX	SUN SPARC-T5120	Tilera TILE-Gx36
Processors	4× AMD Opteron 6172	8× Intel Xeon E7-8867L	SUN UltraSPARC-T2	TILE-Gx CPU
# Cores	48	80 (no hyper-threading)	8 (64 hardware threads)	36
Core clock	2.1 GHz	2.13 GHz	1.2 GHz	1.2 GHz
L1 Cache	64/64 KiB I/D	32/32 KiB I/D	16/8 KiB I/D	32/32 KiB I/D
L2 Cache	512 KiB	256 KiB		256 KiB
Last-level Cache	2×6 MiB (shared per die)	30 MiB (shared)	4 MiB (shared)	9 MiB Distributed
Interconnect	6.4 GT/s HyperTransport (HT) 3.0	6.4 GT/s QuickPath Interconnect (QPI)	Niagara2 Crossbar	Tilera iMesh
Memory #Channels / #Nodes	128 GiB DDR3-1333 4 per socket / 8	192 GiB Sync DDR3-1067 4 per socket / 8	32 GiB FB-DIMM-400 8 / 1	16 GiB DDR3-800 4 / 2
OS	Ubuntu 12.04.2 / 3.4.2	Red Hat EL 6.3 / 2.6.32	Solaris 10 u7	Tilera EL 6.3 / 2.6.40

Table 1: The hardware and the OS characteristics of the target platforms.

The SSYNC suite

- Libraries
- Microbenchmarks
- Concurrent Software

The SSYNC suite

- Libraries
  - **libslock:** the implementation of 9 widely used locks and interfaces for atomic operations
  - **libsmp**: the implementation of message passing technique

The SSYNC suite

- Microbenchmarks
  - **ccbench:** tool for measuring the cost of operations on a cache line.
  - **stress tests**: tests for the primitives in libslock and libsmp

The SSYNC suite

- Concurrent Software
  - **Hash Table (ssht):** an efficient implementation of a hash table with put, get and remove operations.
  - **Transactional Memory (TM2C)**: is a transactional memory system for many-cores

Results

- Hardware-Level Analysis
- Software-Level Analysis

Hardware-Level Analysis

System	Opteron				Xeon			Niagara		Tilera	
State \ Hops	same die	same MCM	one hop	two hops	same die	one hop	two hops	same core	other core	one hop	max hops
loads											
Modified	81	161	172	252	109	289	400	3	24	45	65
Owned	83	163	175	254	-	-	-	-	-	-	-
Exclusive	83	163	175	253	92	273	383	3	24	45	65
Shared	83	164	176	254	44	223	334	3	24	45	65
Invalid	136	237	247	327	355	492	601	176	176	118	162
stores											
Modified	83	172	191	273	115	320	431	24	24	57	77
Owned	244	255	286	291	-	-	-	-	-	-	-
Exclusive	83	171	191	271	115	315	425	24	24	57	77
Shared	246	255	286	296	116	318	428	24	24	86	106
atomic operations: CAS (C), FAI (F), TAS (T), SWAP (S)											
Operation	all	all	all	all	all	all	all	C/F/T/S	C/F/T/S	C/F/T/S	C/F/T/S
Modified	110	197	216	296	120	324	430	71/108/64/95	66/99/55/90	77/51/70/63	98/71/89/84
Shared	272	283	312	332	113	312	423	76/99/67/93	66/99/55/90	124/82/121/95	142/102/141/115

Table 2: Latencies (cycles) of the cache coherence to load/store/CAS/FAI/TAS/SWAP a cache line depending on the MESI state and the distance. The values are the average of 10000 repetitions with $< 3\%$ standard deviation.

Hardware-Level Analysis

Access to an invalid line are access to the main memory

System	Opteron				Xeon			Niagara		Tilera	
State \ Hops	same die	same MCM	one hop	two hops	same die	one hop	two hops	same core	other core	one hop	max hops
loads											
Modified	81	161	172	252	109	289	400	5	24	45	65
Owned	83	163	175	254	-	-	-	-	-	-	-
Exclusive	83	163	175	253	92	273	383	3	24	45	65
Shared	83	164	176	254	44	223	334	3	24	45	65
Invalid	136	237	247	327	355	492	601	176	176	118	162
stores											
Modified	83	172	191	273	115	320	431	24	24	57	77
Owned	244	255	286	291	-	-	-	-	-	-	-
Exclusive	83	171	191	271	115	315	425	24	24	57	77
Shared	246	255	286	296	116	318	428	24	24	86	106
atomic operations: CAS (C), FAI (F), TAS (T), SWAP (S)											
Operation	all	all	all	all	all	all	all	C/F/T/S	C/F/T/S	C/F/T/S	C/F/T/S
Modified	110	197	216	296	120	324	430	71/108/64/95	66/99/55/90	77/51/70/63	98/71/89/84
Shared	272	283	312	332	113	312	423	76/99/67/93	66/99/55/90	124/82/121/95	142/102/141/115

Table 2: Latencies (cycles) of the cache coherence to load/store/CAS/FAI/TAS/SWAP a cache line depending on the MESI state and the distance. The values are the average of 10000 repetitions with $< 3\%$ standard deviation.

Hardware-Level Analysis

A load has basically the same latency regardless of the previous state of the line.

System	Opteron				Xeon			Niagara		Tilera	
Hops / State	same die	same MCM	one hop	two hops	same die	one hop	two hops	same core	other core	one hop	max hops
loads											
Modified	81	161	172	252	109	289	400	3	24	45	65
Owned	83	163	175	254	-	-	-	-	-	-	-
Exclusive	83	163	175	253	92	273	383	3	24	45	65
Shared	83	164	176	254	44	223	334	3	24	45	65
Invalid	136	237	247	327	355	492	601	176	176	118	162
stores											
Modified	83	172	191	273	115	320	431	24	24	57	77
Owned	244	255	286	291	-	-	-	-	-	-	-
Exclusive	83	171	191	271	115	315	425	24	24	57	77
Shared	246	255	286	296	116	318	428	24	24	86	106
atomic operations: CAS (C), FAI (F), TAS (T), SWAP (S)											
Operation	all	all	all	all	all	all	all	C/F/T/S	C/F/T/S	C/F/T/S	C/F/T/S
Modified	110	197	216	296	120	324	430	71/108/64/95	66/99/55/90	77/51/70/63	98/71/89/84
Shared	272	283	312	332	113	312	423	76/99/67/93	66/99/55/90	124/82/121/95	142/102/141/115

Table 2: Latencies (cycles) of the cache coherence to load/store/CAS/FAI/TAS/SWAP a cache line depending on the MESI state and the distance. The values are the average of 10000 repetitions with $< 3\%$ standard deviation.

Hardware-Level Analysis

The latencies between two dies in an MCM, and two dies that are directly connected differ by roughly 12 cycles.

System	Opteron				Xeon			Niagara		Tilera	
State \ Hops	same die	same MCM	one hop	two hops	same die	one hop	two hops	same core	other core	one hop	max hops
loads											
Modified	81	161	172	252	109	289	400	3	24	45	65
Owned	83	163	175	254	-	-	-	-	-	-	-
Exclusive	83	163	175	253	92	273	383	3	24	45	65
Shared	83	164	176	254	44	223	334	3	24	45	65
Invalid	136	237	247	327	355	492	601	176	176	118	162
stores											
Modified	83	172	191	273	115	320	431	24	24	57	77
Owned	244	255	286	291	-	-	-	-	-	-	-
Exclusive	83	171	191	271	115	315	425	24	24	57	77
Shared	246	255	286	296	116	318	428	24	24	86	106
atomic operations: CAS (C), FAI (F), TAS (T), SWAP (S)											
Operation	all	all	all	all	all	all	all	C/F/T/S	C/F/T/S	C/F/T/S	C/F/T/S
Modified	110	197	216	296	120	324	430	71/108/64/95	66/99/55/90	77/51/70/63	98/71/89/84
Shared	272	283	312	332	113	312	423	76/99/67/93	66/99/55/90	124/82/121/95	142/102/141/115

Table 2: Latencies (cycles) of the cache coherence to load/store/CAS/FAI/TAS/SWAP a cache line depending on the MESI state and the distance. The values are the average of 10000 repetitions with $< 3\%$ standard deviation.

Hardware-Level Analysis

One extra hop adds an additional overhead of 80 cycles.

System	Opteron				Xeon			Niagara		Tilera	
Hops / State	same die	same MCM	one hop	two hops	same die	one hop	two hops	same core	other core	one hop	max hops
loads											
Modified	81	161	172	252	109	289	400	3	24	45	65
Owned	83	163	175	254	-	-	-	-	-	-	-
Exclusive	83	163	175	253	92	273	383	3	24	45	65
Shared	83	164	176	254	44	223	334	3	24	45	65
Invalid	136	237	247	327	355	492	601	176	176	118	162
stores											
Modified	83	172	191	273	115	320	431	24	24	57	77
Owned	244	255	286	291	-	-	-	-	-	-	-
Exclusive	83	171	191	271	115	315	425	24	24	57	77
Shared	246	255	286	296	116	318	428	24	24	86	106
atomic operations: CAS (C), FAI (F), TAS (T), SWAP (S)											
Operation	all	all	all	all	all	all	all	C/F/T/S	C/F/T/S	C/F/T/S	C/F/T/S
Modified	110	197	216	296	120	324	430	71/108/64/95	66/99/55/90	77/51/70/63	98/71/89/84
Shared	272	283	312	332	113	312	423	76/99/67/93	66/99/55/90	124/82/121/95	142/102/141/115

Table 2: Latencies (cycles) of the cache coherence to load/store/CAS/FAI/TAS/SWAP a cache line depending on the MESI state and the distance. The values are the average of 10000 repetitions with $< 3\%$ standard deviation.

Hardware-Level Analysis

Loading from a shared state is 7.5 times more expensive over two hops than loadings within the socket

System	Opteron				Xeon			Niagara		Tilera	
State \ Hops	same die	same MCM	one hop	two hops	same die	one hop	two hops	same core	other core	one hop	max hops
loads											
Modified	81	161	172	252	109	289	400	3	24	45	65
Owned	83	163	175	254	-	-	-	-	-	-	-
Exclusive	83	163	175	253	92	273	383	3	24	45	65
Shared	83	164	176	254	44	223	334	3	24	45	65
Invalid	136	237	247	327	355	492	601	176	176	118	162
stores											
Modified	83	172	191	273	115	320	431	24	24	57	77
Owned	244	255	286	291	-	-	-	-	-	-	-
Exclusive	83	171	191	271	115	315	425	24	24	57	77
Shared	246	255	286	296	116	318	428	24	24	86	106
atomic operations: CAS (C), FAI (F), TAS (T), SWAP (S)											
Operation	all	all	all	all	all	all	all	C/F/T/S	C/F/T/S	C/F/T/S	C/F/T/S
Modified	110	197	216	296	120	324	430	71/108/64/95	66/99/55/90	77/51/70/63	98/71/89/84
Shared	272	283	312	332	113	312	423	76/99/67/93	66/99/55/90	124/82/121/95	142/102/141/115

Table 2: Latencies (cycles) of the cache coherence to load/store/CAS/FAI/TAS/SWAP a cache line depending on the MESI state and the distance. The values are the average of 10000 repetitions with $< 3\%$ standard deviation.

Hardware-Level Analysis

The load results have much lower variability on a single-socket.

System	Opteron				Xeon			Niagara		Tilera	
State \ Hops	same die	same MCM	one hop	two hops	same die	one hop	two hops	same core	other core	one hop	max hops
loads											
Modified	81	161	172	252	109	289	400	3	24	45	65
Owned	83	163	175	254	-	-	-	-	-	-	-
Exclusive	83	163	175	253	92	273	383	3	24	45	65
Shared	83	164	176	254	44	223	334	3	24	45	65
Invalid	136	237	247	327	355	492	601	176	176	118	162
stores											
Modified	83	172	191	273	115	320	431	24	24	57	77
Owned	244	255	286	291	-	-	-	-	-	-	-
Exclusive	83	171	191	271	115	315	425	24	24	57	77
Shared	246	255	286	296	116	318	428	24	24	86	106
atomic operations: CAS (C), FAI (F), TAS (T), SWAP (S)											
Operation	all	all	all	all	all	all	all	C/F/T/S	C/F/T/S	C/F/T/S	C/F/T/S
Modified	110	197	216	296	120	324	430	71/108/64/95	66/99/55/90	77/51/70/63	98/71/89/84
Shared	272	283	312	332	113	312	423	76/99/67/93	66/99/55/90	124/82/121/95	142/102/141/115

Table 2: Latencies (cycles) of the cache coherence to load/store/CAS/FAI/TAS/SWAP a cache line depending on the MESI state and the distance. The values are the average of 10000 repetitions with $< 3\%$ standard deviation.

Hardware-Level Analysis

On Opteron, both load and stores on a <u>modified</u> or an <u>exclusive</u> cache line have similar latencies.

System	Opteron				Xeon			Niagara		Tilera	
State \ Hops	same die	same MCM	one hop	two hops	same die	one hop	two hops	same core	other core	one hop	max hops
loads											
Modified	81	161	172	[illegible]	109	289	400	3	24	45	65
Owned	83	163	175	254	-	-	-	-	-	-	-
Exclusive	83	163	175	253	92	273	383	3	24	45	65
Shared	83	164	176	254	44	223	334	3	24	45	65
Invalid	136	237	247	327	355	492	601	176	176	118	162
stores											
Modified	83	172	191	273	115	320	431	24	24	57	77
Owned	244	255	286	291	-	-	-	-	-	-	-
Exclusive	83	171	191	271	115	315	425	24	24	57	77
Shared	246	255	286	296	116	318	428	24	24	86	106
atomic operations: CAS (C), FAI (F), TAS (T), SWAP (S)											
Operation	all	all	all	all	all	all	all	C/F/T/S	C/F/T/S	C/F/T/S	C/F/T/S
Modified	110	197	216	296	120	324	430	71/108/64/95	66/99/55/90	77/51/70/63	98/71/89/84
Shared	272	283	312	332	113	312	423	76/99/67/93	66/99/55/90	124/82/121/95	142/102/141/115

Table 2: Latencies (cycles) of the cache coherence to load/store/CAS/FAI/TAS/SWAP a cache line depending on the MESI state and the distance. The values are the average of 10000 repetitions with $< 3\%$ standard deviation.

Hardware-Level Analysis

However, every store on a <u>shared</u> or <u>owned</u> cache line incurs a broadcast message to all nodes.

This is because the cache coherency does not keep track of the sharers.

System	Opteron				Xeon						
State \ Hops	same die	same MCM	one hop	two hops	same die	one hop	two hops	same core			
loads											
Modified	81	161	172	252	109	[illegible]	400	3	24	45	65
Owned	83	163	175	254	[illegible]	-	-	-	-	-	-
Exclusive	83	163	175	253	[illegible]	273	383	3	24	45	65
Shared	83	164	176	254	44	223	334	3	24	45	65
Invalid	136	237	247	327	355	492	601	176	176	118	162
stores											
Modified	83	172	191	273	115	320	431	24	24	57	77
Owned	244	255	286	291	-	-	-	-	-	-	-
Exclusive	83	171	191	271	115	315	425	24	24	57	77
Shared	246	255	286	296	116	318	428	24	24	86	106
atomic operations: CAS (C), FAI (F), TAS (T), SWAP (S)											
Operation	all	all	all	all	all	all	all	C/F/T/S	C/F/T/S	C/F/T/S	C/F/T/S
Modified	110	197	216	296	120	324	430	71/108/64/95	66/99/55/90	77/51/70/63	98/71/89/84
Shared	272	283	312	332	113	312	423	76/99/67/93	66/99/55/90	124/82/121/95	142/102/141/115

Table 2: Latencies (cycles) of the cache coherence to load/store/CAS/FAI/TAS/SWAP a cache line depending on the MESI state and the distance. The values are the average of 10000 repetitions with $< 3\%$ standard deviation.

Hardware-Level Analysis

In general, stores behave similarly regardless of the previous state of the cache line

System	Opteron				Xeon			Niagara		Tilera	
State \ Hops	same die	same MCM	one hop	two hops	same die	one hop	two hops	same core	other core	one hop	max hops
loads											
Modified	81	161	172	252	109	289	400	3	[illegible]	45	65
Owned	83	163	175	254	-	-	-	-	-	-	-
Exclusive	83	163	175	253	92	273	383	3	24	45	65
Shared	83	164	176	254	44	223	334	3	24	45	65
Invalid	136	237	247	327	355	492	601	176	176	118	162
stores											
Modified	83	172	191	273	115	320	431	24	24	57	77
Owned	244	255	286	291	-	-	-	-	-	-	-
Exclusive	83	171	191	271	115	315	425	24	24	57	77
Shared	246	255	286	296	116	318	428	24	24	86	106
atomic operations: CAS (C), FAI (F), TAS (T), SWAP (S)											
Operation	all	all	all	all	all	all	all	C/F/T/S	C/F/T/S	C/F/T/S	C/F/T/S
Modified	110	197	216	296	120	324	430	71/108/64/95	66/99/55/90	77/51/70/63	98/71/89/84
Shared	272	283	312	332	113	312	423	76/99/67/93	66/99/55/90	124/82/121/95	142/102/141/115

Table 2: Latencies (cycles) of the cache coherence to load/store/CAS/FAI/TAS/SWAP a cache line depending on the MESI state and the distance. The values are the average of 10000 repetitions with $< 3\%$ standard deviation.

Hardware-Level Analysis

Similarly to load, the results for a store have much lower variability on a single-socket.

System	Opteron				Xeon			Niagara		Tilera	
State \ Hops	same die	same MCM	one hop	two hops	same die	one hop	two hops	same core	other core	one hop	max hops
loads											
Modified	81	161	172	252	109	289	400	3	24	45	65
Owned	83	163	175	254	-	-	-	-	-	-	-
Exclusive	83	163	175	253	92	273	383	3	[illegible]	45	65
Shared	83	164	176	254	44	223	334	3	[illegible]	45	65
Invalid	136	237	247	327	355	492	601	176	176	118	162
stores											
Modified	83	172	191	273	115	320	431	24	24	57	77
Owned	244	255	286	291	-	-	-	-	-	-	-
Exclusive	83	171	191	271	115	315	425	24	24	57	77
Shared	246	255	286	296	116	318	428	24	24	86	106
atomic operations: CAS (C), FAI (F), TAS (T), SWAP (S)											
Operation	all	all	all	all	all	all	all	C/F/T/S	C/F/T/S	C/F/T/S	C/F/T/S
Modified	110	197	216	296	120	324	430	71/108/64/95	66/99/55/90	77/51/70/63	98/71/89/84
Shared	272	283	312	332	113	312	423	76/99/67/93	66/99/55/90	124/82/121/95	142/102/141/115

Table 2: Latencies (cycles) of the cache coherence to load/store/CAS/FAI/TAS/SWAP a cache line depending on the MESI state and the distance. The values are the average of 10000 repetitions with $< 3\%$ standard deviation.

Hardware-Level Analysis

On the multi-socket, all atomic operations have essentially the same latencies.

System	Opteron				Xeon			Niagara		Tilera	
Hops / State	same die	same MCM	one hop	two hops	same die	one hop	two hops	same core	other core	one hop	max hops
loads											
Modified	81	161	172	252	109	289	400	3	24	45	65
Owned	83	163	175	254	-	-	-		-	-	-
Exclusive	83	163	175	253	92	273	383	3	24	45	65
Shared	83	164	176	254	44	223	334	3	24	45	65
Invalid	136	237	247	327	355	492	601	176	176	118	162
st s											
Modified	83	172	191	273	115	320	4	24	24	57	77
Owned	244	255	286	291	-	-		-	-	-	-
Exclusive	83	171	191	271	115	315	425	24	24	57	77
Shared	246	255	286	296	116	318	428	24	24	86	106
atomic operations: CAS (C), FAI (F), TAS (T), SWAP (S)											
Operation	all	all	all	all	all	all	all	C/F/T/S	C/F/T/S	C/F/T/S	C/F/T/S
Modified	110	197	216	296	120	324	430	71/108/64/95	66/99/55/90	77/51/70/63	98/71/89/84
Shared	272	283	312	332	113	312	423	76/99/67/93	66/99/55/90	124/82/121/95	142/102/141/115

Table 2: Latencies (cycles) of the cache coherence to load/store/CAS/FAI/TAS/SWAP a cache line depending on the MESI state and the distance. The values are the average of 10000 repetitions with $< 3\%$ standard deviation.

Hardware-Level Analysis

On a single-socket, some operations clearly have different hardware implementations, e. g. FAI (Fetch-and-inc.) on Tilera is faster than the others.

System	Opteron				Xeon			Niagara		Tilera	
Hops / State	same die	same MCM	one hop	two hops	same die	one hop	two hops	same core	other core	[illegible]e [illegible]op	max hops
loads											
Modified	81	161	172	252	109	289	400	3	24	45	65
Owned	83	163	175	254	-	-	-	-	-	-	-
Exclusive	83	163	175	253	92	273	383	3	24	45	65
Shared	83	164	176	254	44	223	334	3	24	45	65
Invalid	136	237	247	327	355	492	601	176	176	118	162
stores											
Modified	83	172	191	273	115	320	431	24	24	57	77
Owned	244	255	286	291	-	-	-	-	-	-	-
Exclusive	83	171	191	271	115	315	425	24	24	57	77
Shared	246	255	286	296	116	318	428	24	24	86	106
atomic operations: CAS (C), FAI (F), TAS (T), SWAP (S)											
Operation	all	all	all	all	all	all	all	C/F/T/S	C/F/T/S	C/F/T/S	C/F/T/S
Modified	110	197	216	296	120	324	430	71/108/64/95	66/99/55/90	77/51/70/63	98/71/89/84
Shared	272	283	312	332	113	312	423	76/99/67/93	66/99/55/90	124/82/121/95	142/102/141/115

Table 2: Latencies (cycles) of the cache coherence to load/store/CAS/FAI/TAS/SWAP a cache line depending on the MESI state and the distance. The values are the average of 10000 repetitions with $< 3\%$ standard deviation.

Hardware-Level Analysis: Summary

- Cross-socket communication is 2 to 7.5 times more expensive than intra-socket communication.

- A store on a <u>shared</u> or <u>owned</u> cache line induces an unnecessary broadcast messages. Then, a <u>modified</u> cache line should be favored.

- System designers should take advantage of the best performing atomic operation available on each platform.

Software-Level Analysis

- Locks
- Message Passing
- Hash Table
- Key-value Store

Software-Level Analysis: Locks

Latency based on the location

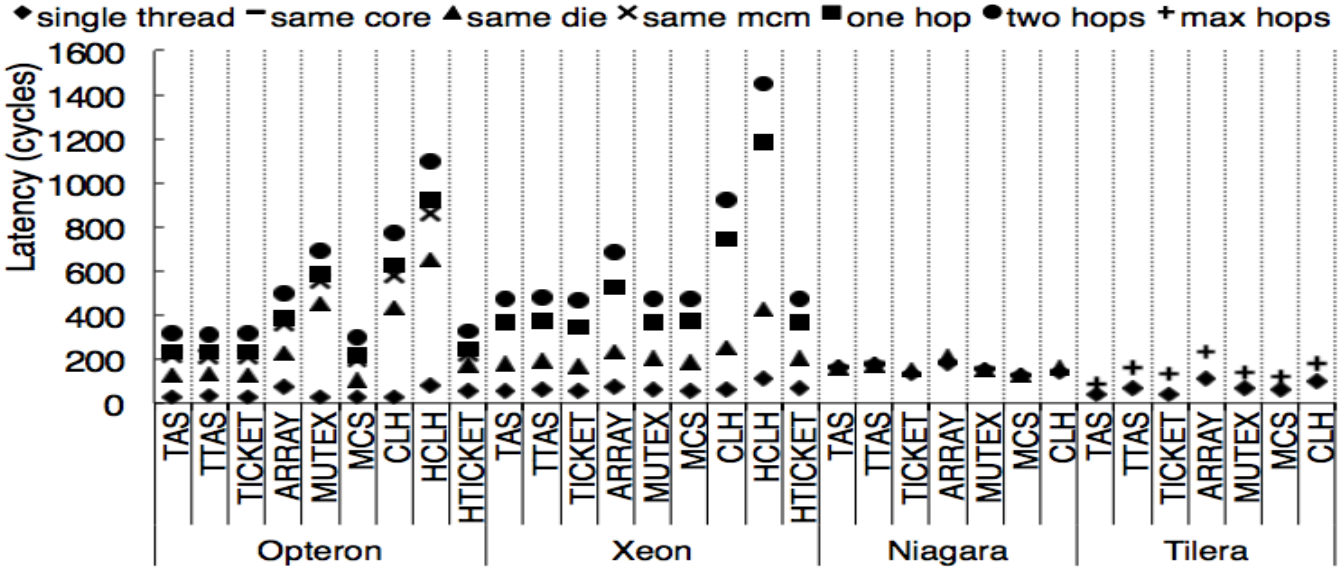
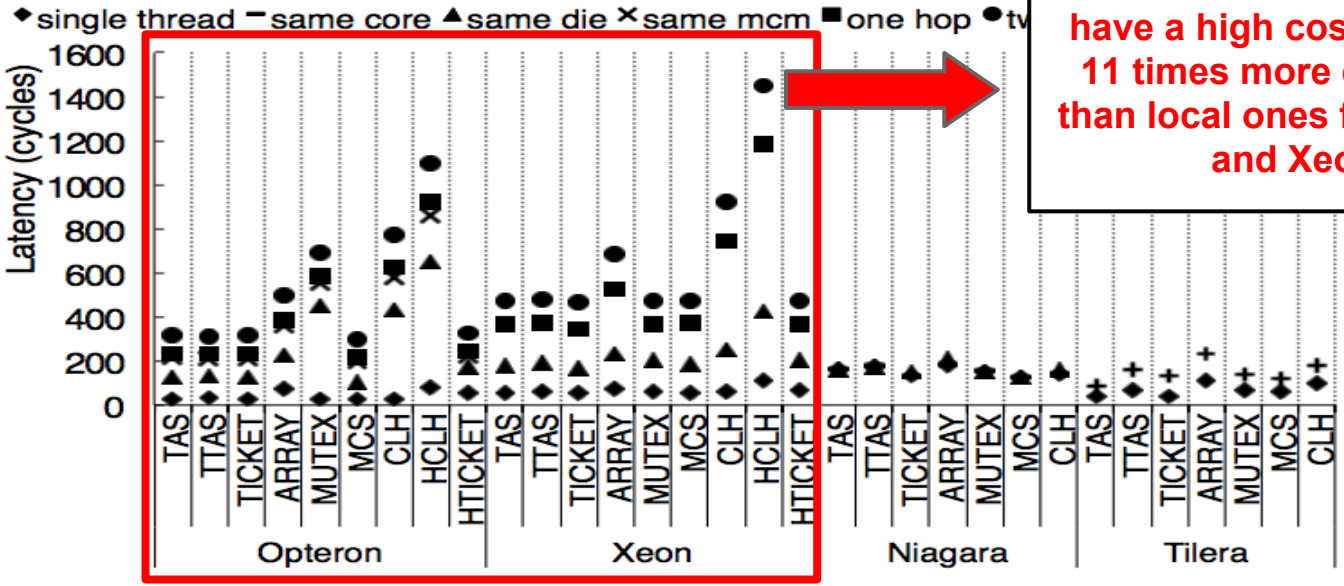


Figure 6: Uncontested lock acquisition latency based on the location of the previous owner of the lock.

Software-Level Analysis: Locks

Latency based on the location



Acquisitions that need to transfer data across sockets have a high cost (12.5 and 11 times more expensive than local ones for Opteron and Xeon)

Figure 6: Uncontested lock acquisition latency based on the location of the previous owner of the lock.

Software-Level Analysis: Locks

Latency based on the location

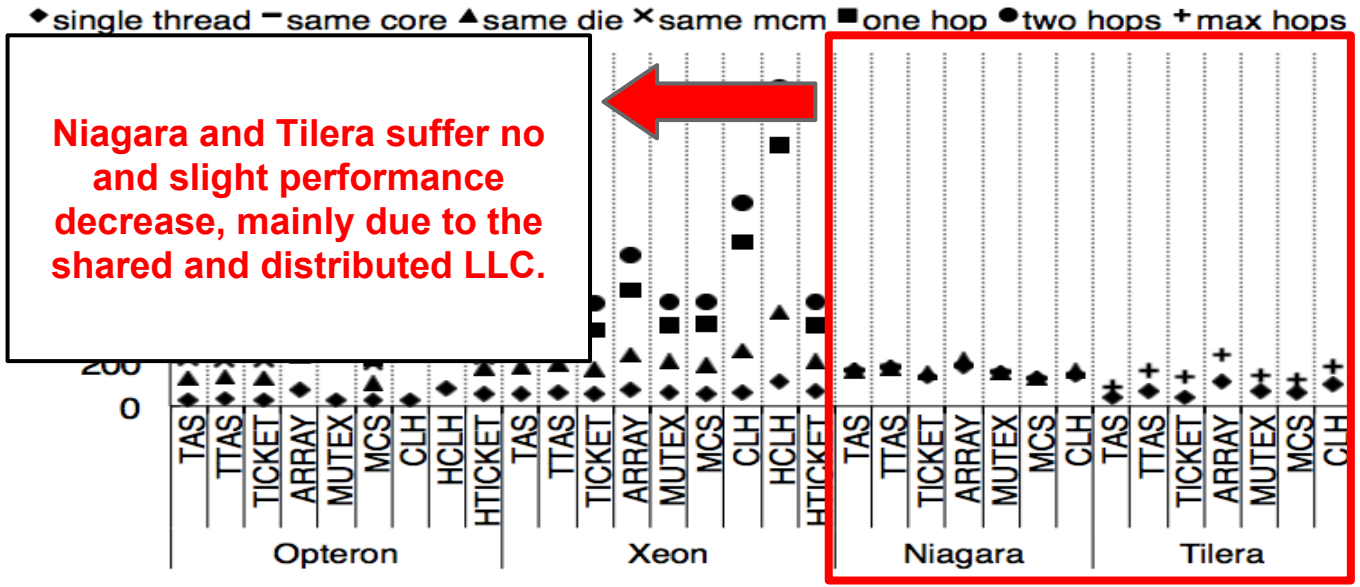


Figure 6: Uncontested lock acquisition latency based on the location of the previous owner of the lock.

Software-Level Analysis: Locks

Extreme contention

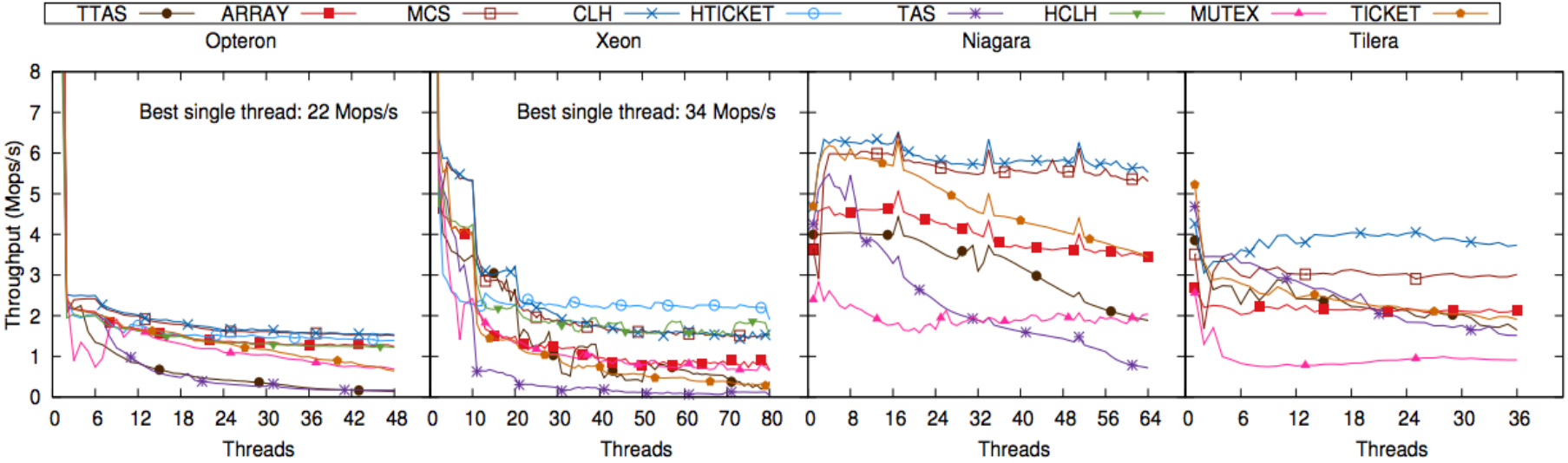


Figure 5: Throughput of different lock algorithms using a single lock.

Software-Level Analysis: Locks

Extreme contention

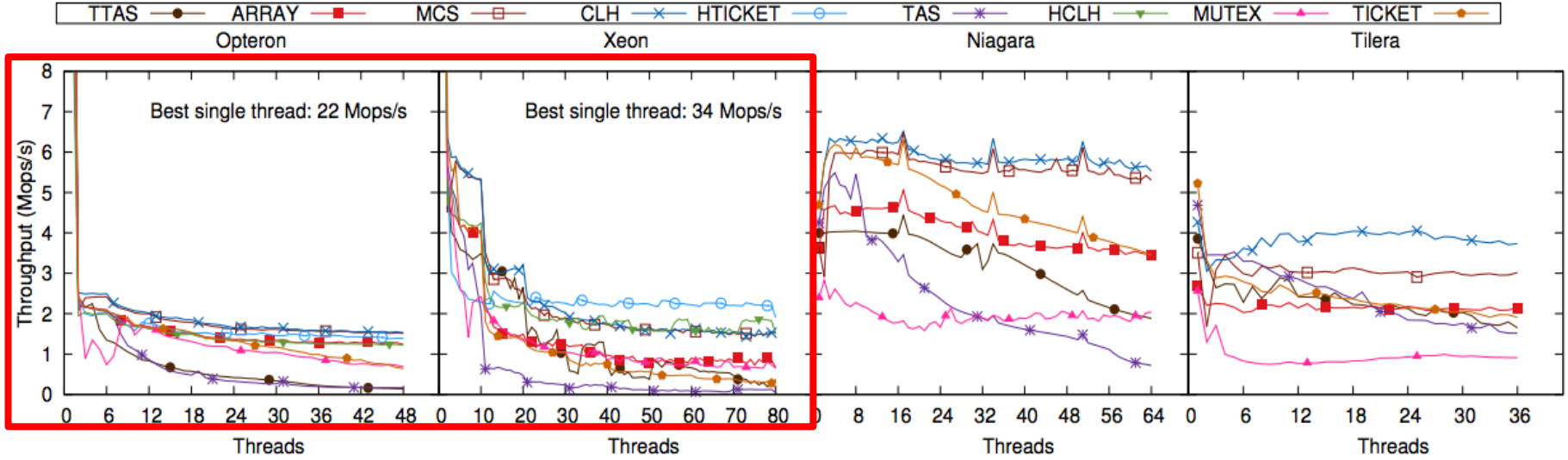


Figure 5: Throughput of different lock algorithms using a single lock.

Although there is a big drop from one to two cores on multi-sockets, within the same socket both Opteron and Xeon keep performance stable

Software-Level Analysis: Locks

Extreme contention

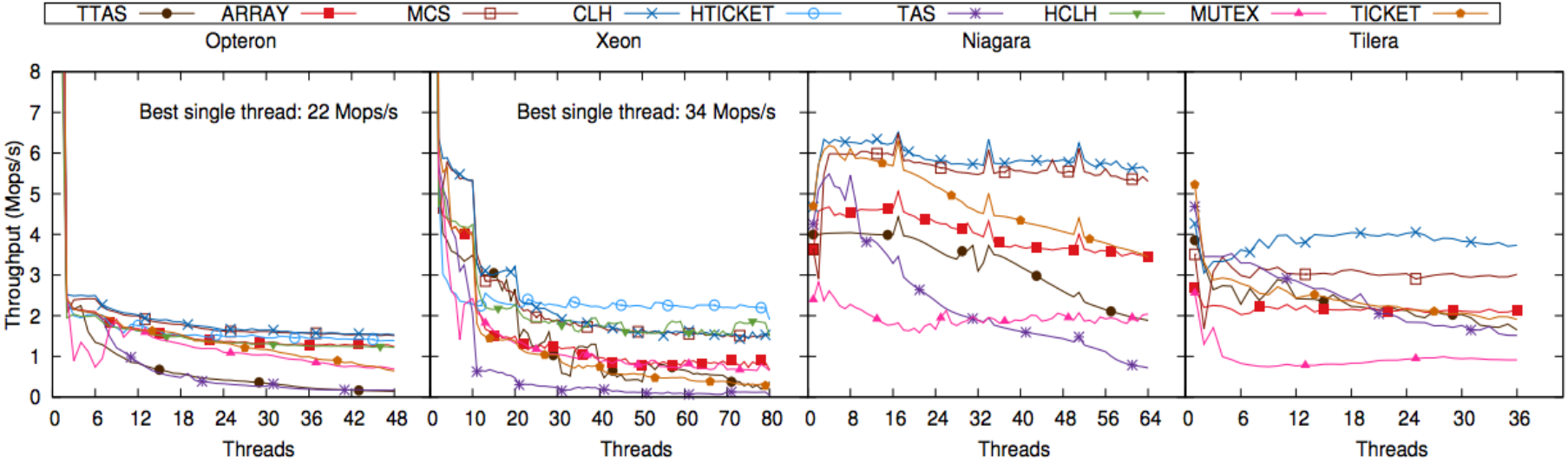


Figure 5: Throughput of different lock algorithms using a single lock.

Overall, the throughput on two or more cores on the multi-sockets is an order of magnitude lower than the single-core performance.

Software-Level Analysis: Locks

Extreme contention

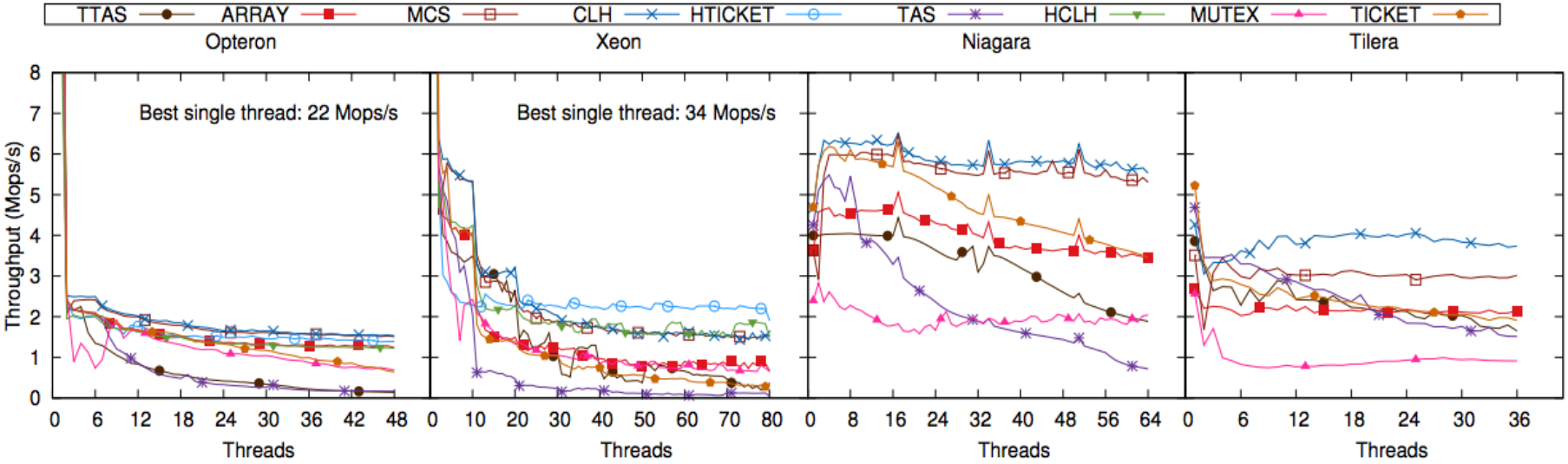


Figure 5: Throughput of different lock algorithms using a single lock.

Overall, the throughput on two or more cores on the multi-sockets is an order of magnitude lower than the single-core performance.

In contrast, the single-sockets maintain a comparable performance on multiple cores

Software-Level Analysis: Locks

Very low contention

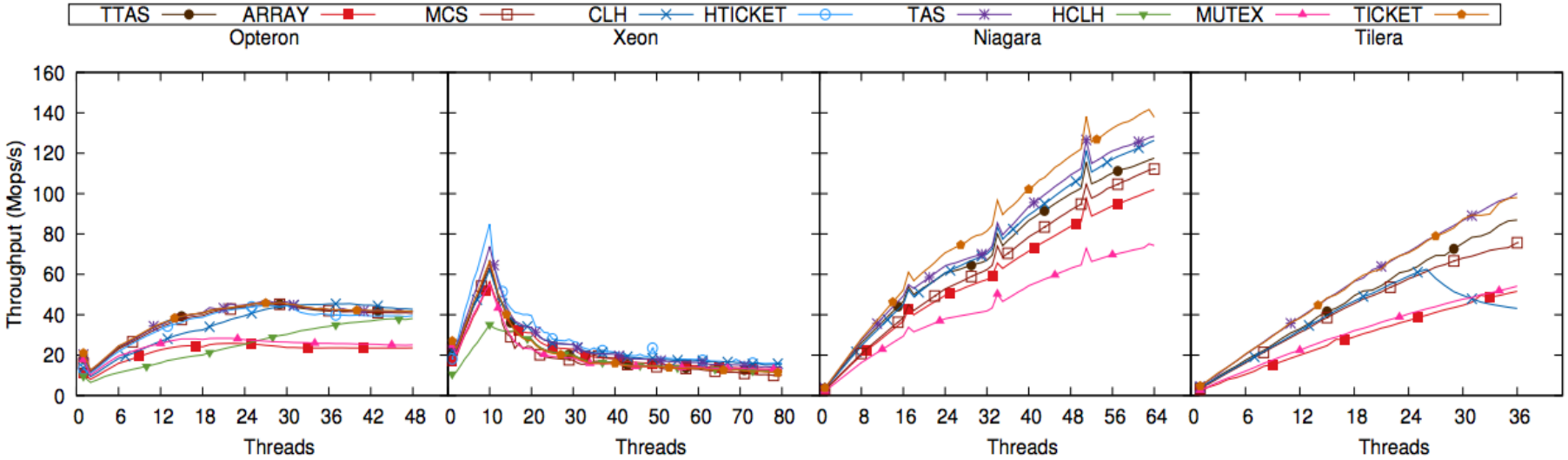


Figure 7: Throughput of different lock algorithms using 512 locks.

- **In general, simple locks match or even outperform more complex locks.**

Software-Level Analysis: Locks

Very low contention

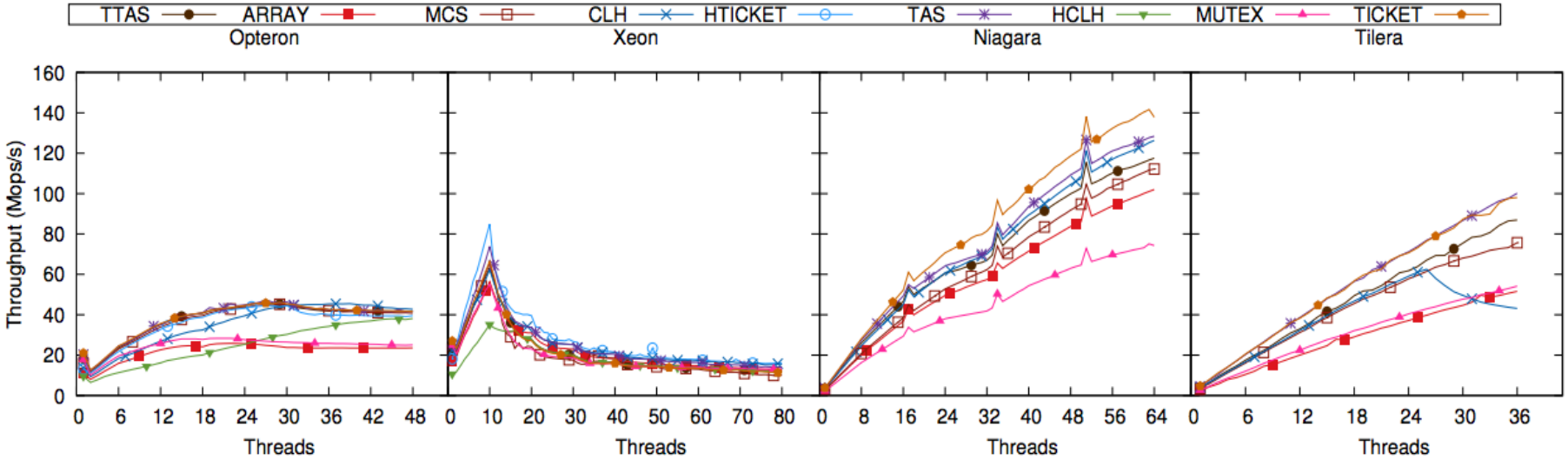


Figure 7: Throughput of different lock algorithms using 512 locks.

- **In general, simple locks match or even outperform more complex locks.**
- **The ticket lock performs the best of Opteron, Niagara and Tilera**

Software-Level Analysis: Locks

Summary

- **No lock is consistently the best on all platforms**
- **No lock is consistently the best within a platform**
- **Complex locks are generally the best under extreme contention**
- **Simple locks perform better under low contention**

Software-Level Analysis: Message Passing

One-to-One communication

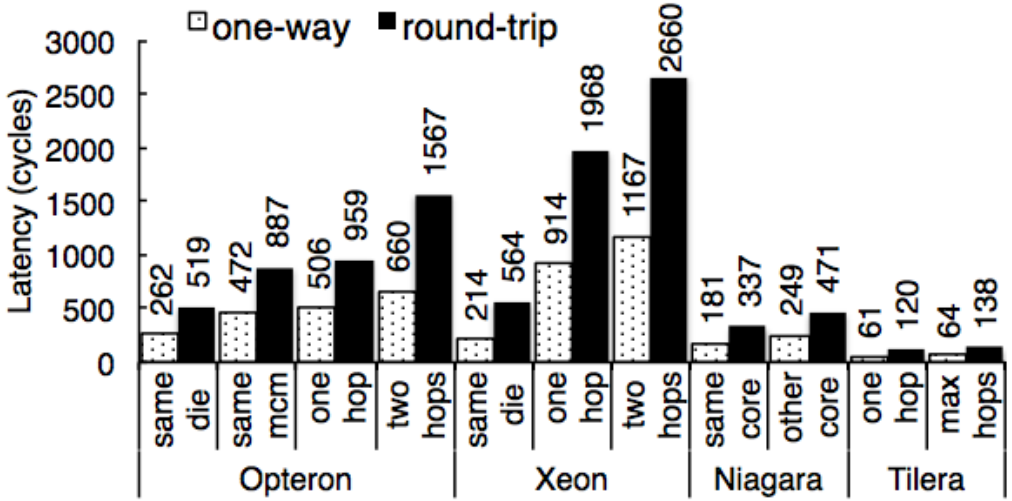


Figure 9: One-to-one communication latencies of message passing depending on the distance between the two cores.

Software-Level Analysis: Message Passing

One-to-One communication

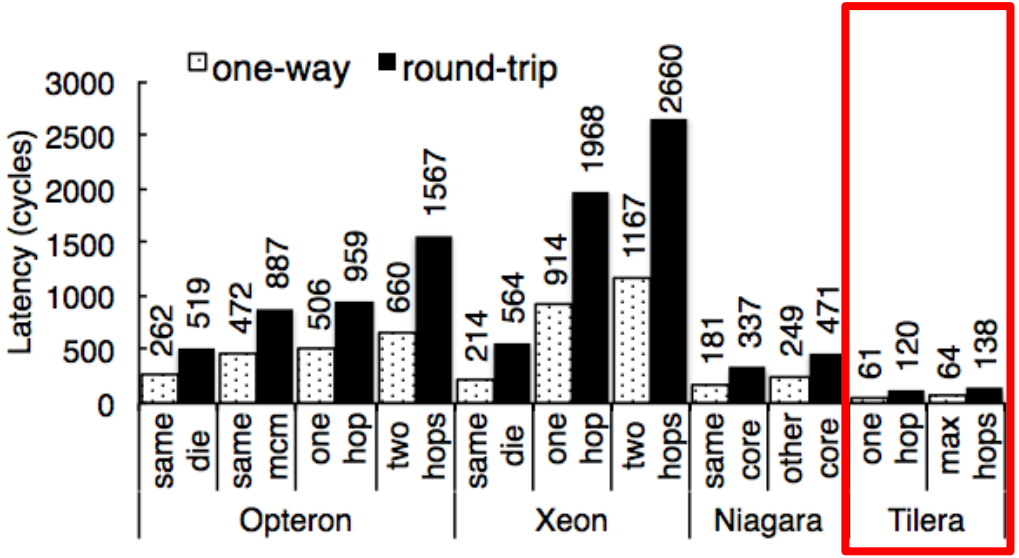


Figure 9: One-to-one communication latencies of message passing depending on the distance between the two cores.

Tilera performs the best because it uses its message passing implementation on hardware.

Software-Level Analysis: Message Passing

Client-server communication

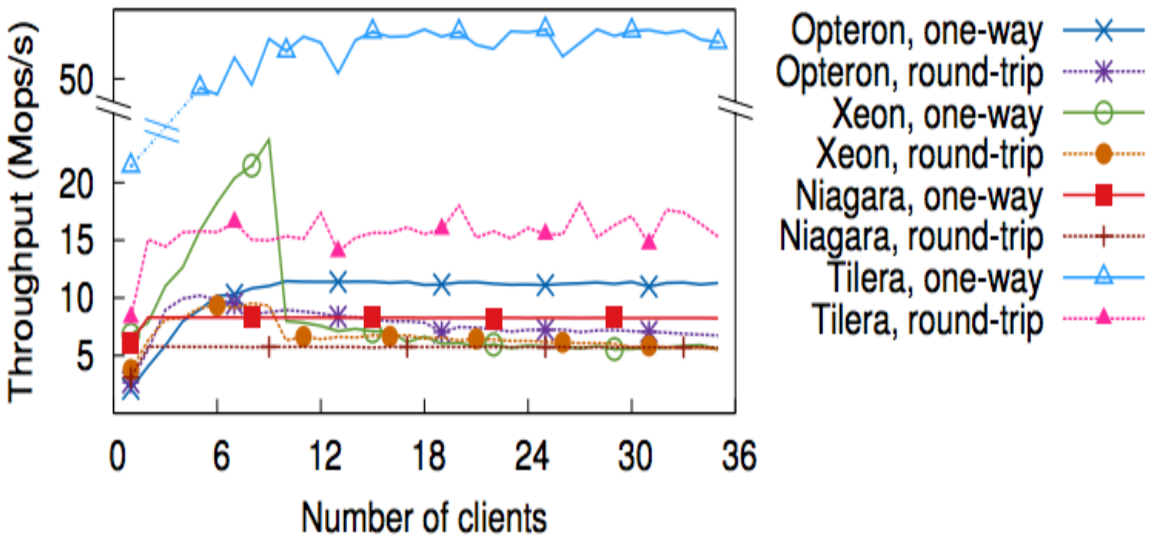
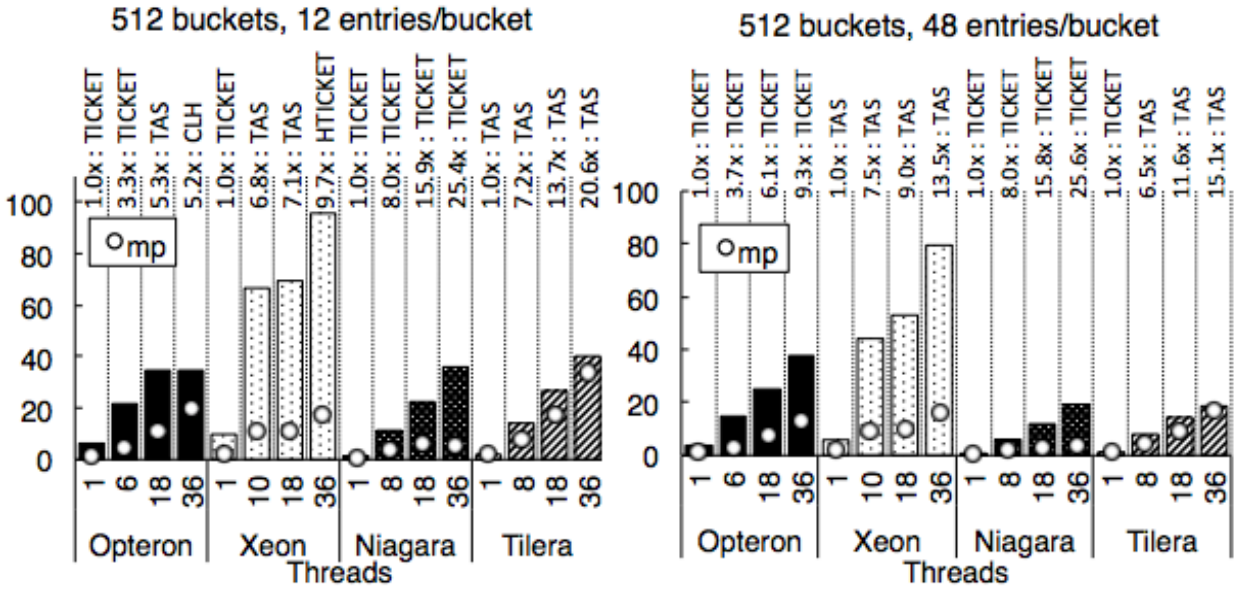


Figure 10: Total throughput of client-server communication.

Again, the hardware message passing on Tilera performs the best.

Software-Level Analysis: Hash Table

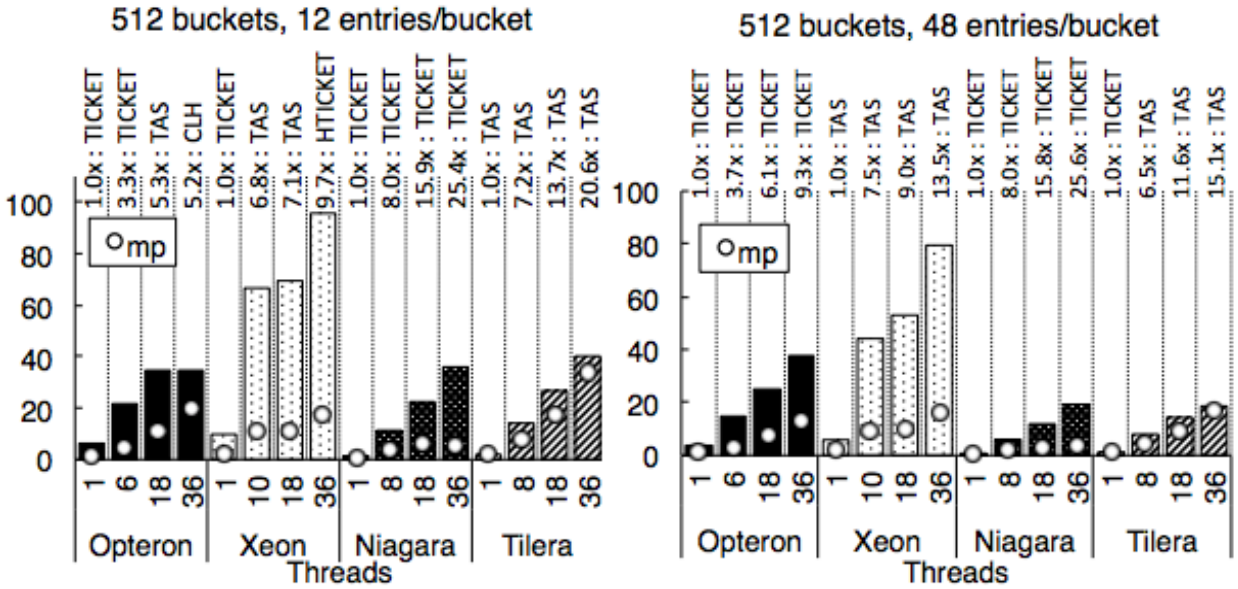
Low Contention



Throughput and scalability of the hash table on different configurations. The "X:Y" labels on top of each bar indicate the best-performing lock (Y) and the scalability over the single-thread execution (X).

Software-Level Analysis: Hash Table

Low Contention

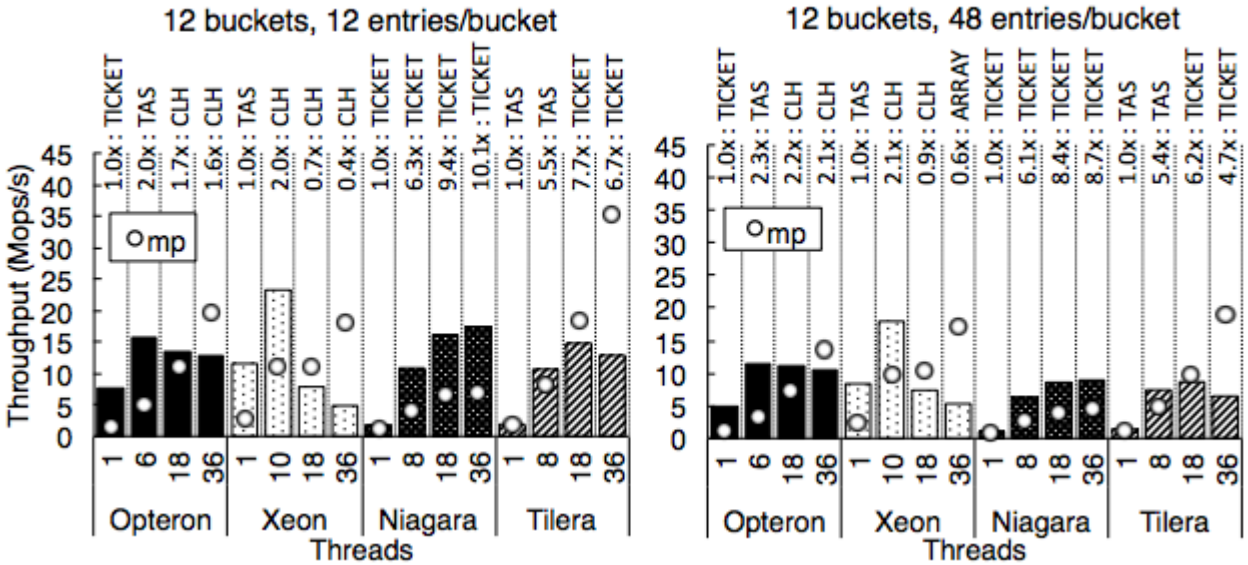


Throughput and scalability of the hash table on different configurations. The "X:Y" labels on top of each bar indicate the best-performing lock (Y) and the scalability over the single-thread execution (X).

Message passing is strictly slower than the lock-based ones, even on Tilera -- which has a hardware message passing implementation.

Software-Level Analysis: Hash Table

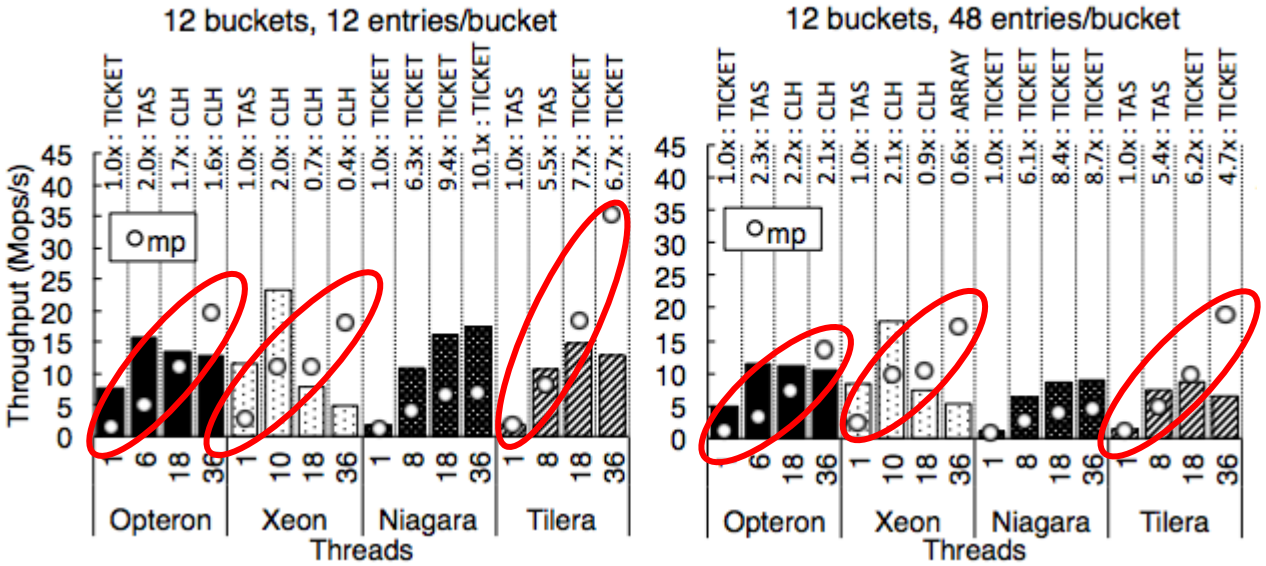
High Contention



Throughput and scalability of the hash table on different configurations. The "X:Y" labels on top of each bar indicate the best-performing lock (Y) and the scalability over the single-thread execution (X).

Software-Level Analysis: Hash Table

High Contention



Message passing outperforms in 3 out 4 platforms, and it also delivers by far the highest throughput

Throughput and scalability of the hash table on different configurations. The "X:Y" labels on top of each bar indicate the best-performing lock (Y) and the scalability over the single-thread execution (X).

Conclusions

- The paper dissects the cost of synchronization and studies its scalability along different directions
- Some of the results are:
  - Crossing socket is a killer
  - Load and stores can be expensive as atomic operations
  - Message passing shines when contention is very high
  - Every locking scheme has its fifteen minutes of fame
  - Simple locks are powerful


Q
&
A

Software-Level Analysis: Key-Value Store

Set-only test

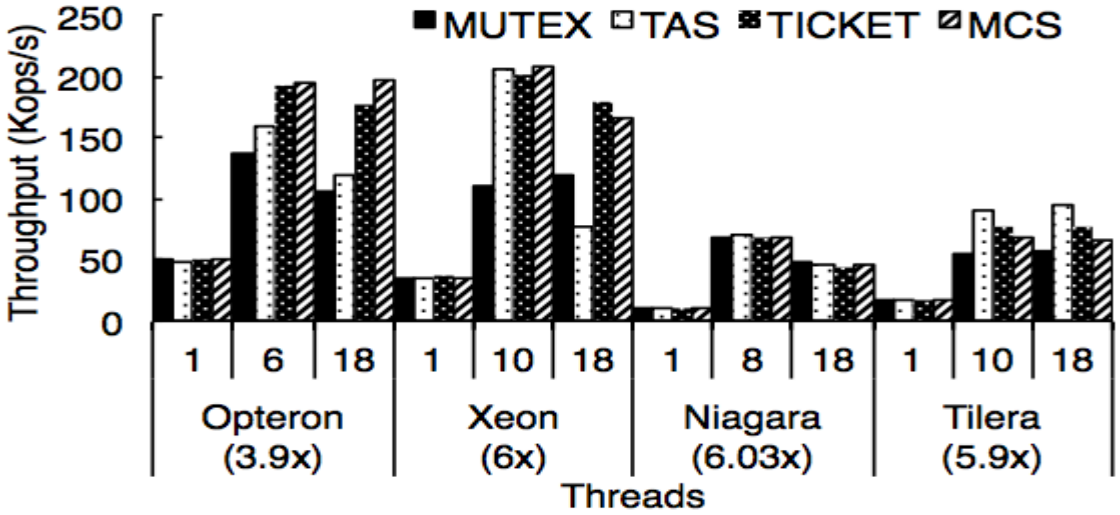


Figure 12: Throughput of Memcached using a set-only test. The maximum speed-up vs. single thread is indicated under the platform names.

Ticket, MCS or TAS locks achieves speedups between 29% and 50% on three of the four platforms.