Low Latency meets Large Scale when you CRANK UP THE AMPS

Website: www.crankuptheamps.com

60East TECHNOLOGIES

Achieving Killer Performance with Storage, Networking and Compute in a NUMA World

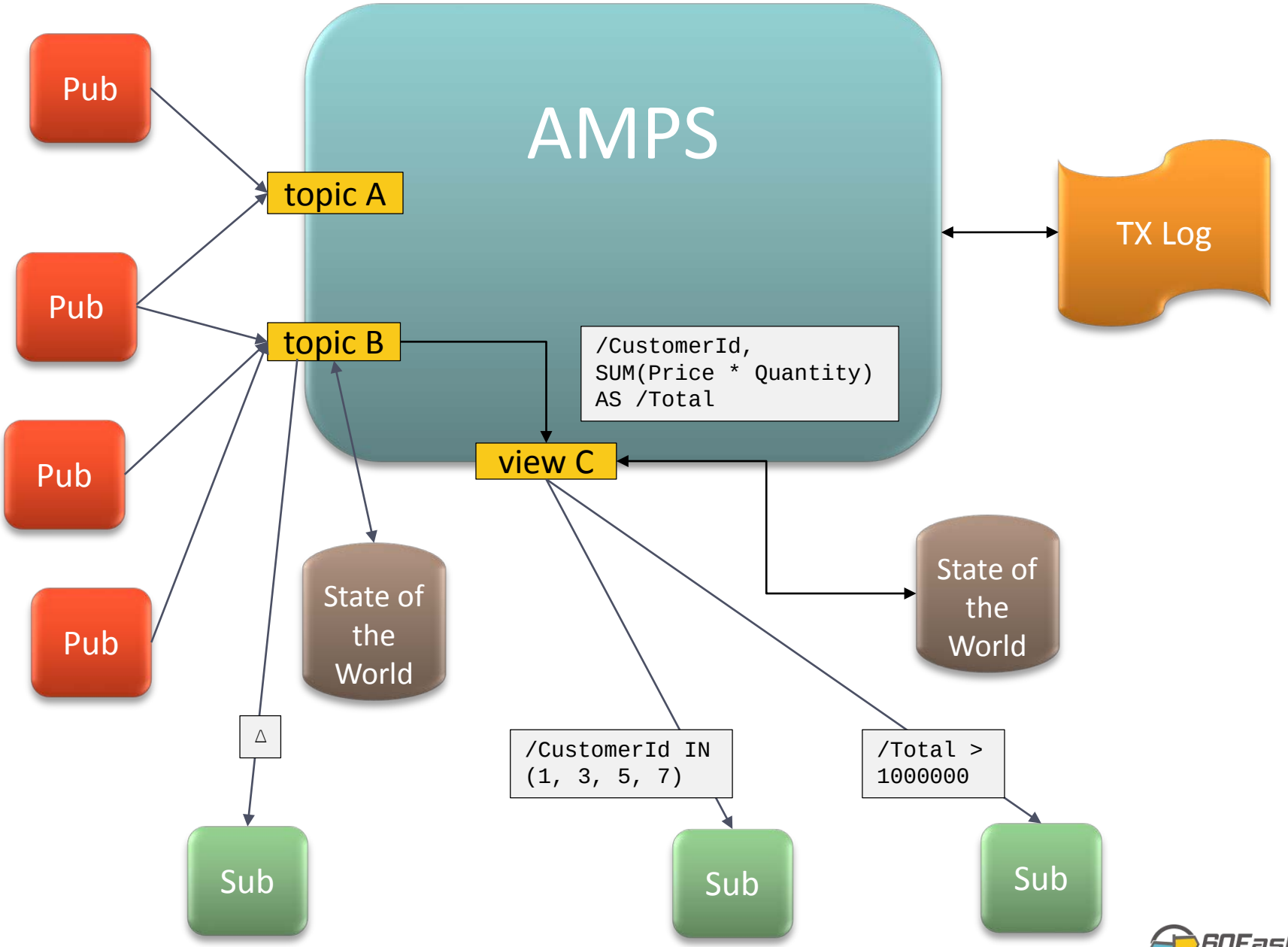
PUSHING AMPS FURTHER

Jeffrey M. Birnbaum jmb@crankuptheamps.com

Website http://crankuptheamps.com/

60East TECHNOLOGIES

Overview of AMPS



60East TECHNOLOGIES

Fast Publish/Subscribe Solution

High Performance Content Filtering

- Filters resemble SQL-92 + Xpath
- Sub-microsecond processing latencies
- Capacity to do >1M messages/sec/core

Example subscription filters:

XML:

```
/FIXML/Order@Sym = “IBM” and
    /FIXML/Order/OrdQty@Qty >= 5000
```

FIX:

```
/55 = “IBM” and /35 in (‘D’, ‘C’)
```

State of the World (Database)

- Content filtered queries
- Atomic query + subscribe
- Message deltas (both in and out)
- Focus Tracking

Analytics Engine (Real-time Aggregation)

- Casts one topic into another
- Parallel and lock-free design

Analytics Engine (Real-time Aggregation)

- Projects one topic into another
  - Think: Real-time SQL-92 “Materialized View”

Example:

- Project:
  - /11 as /customer
  - /55 as /symbol
  - sum(/14 * /99)/sum(/14) AS /vwap
- GroupBy: /11, 55
- New Topic Name: VWAP

This:

- 11=c01;55=INTC;14=1000;99=34.50;
- 11=c01;55=INTC;14=5000;99=34.75;
- 11=c01;55=INFA;14=100;99=18.75;

Becomes:

- customer=c01;symbol=INTC;vwap=34.70833;
- customer=c01;symbol=INFA;vwap=18.75;

Factors in Performance

Network

Memory

CPU

Storage

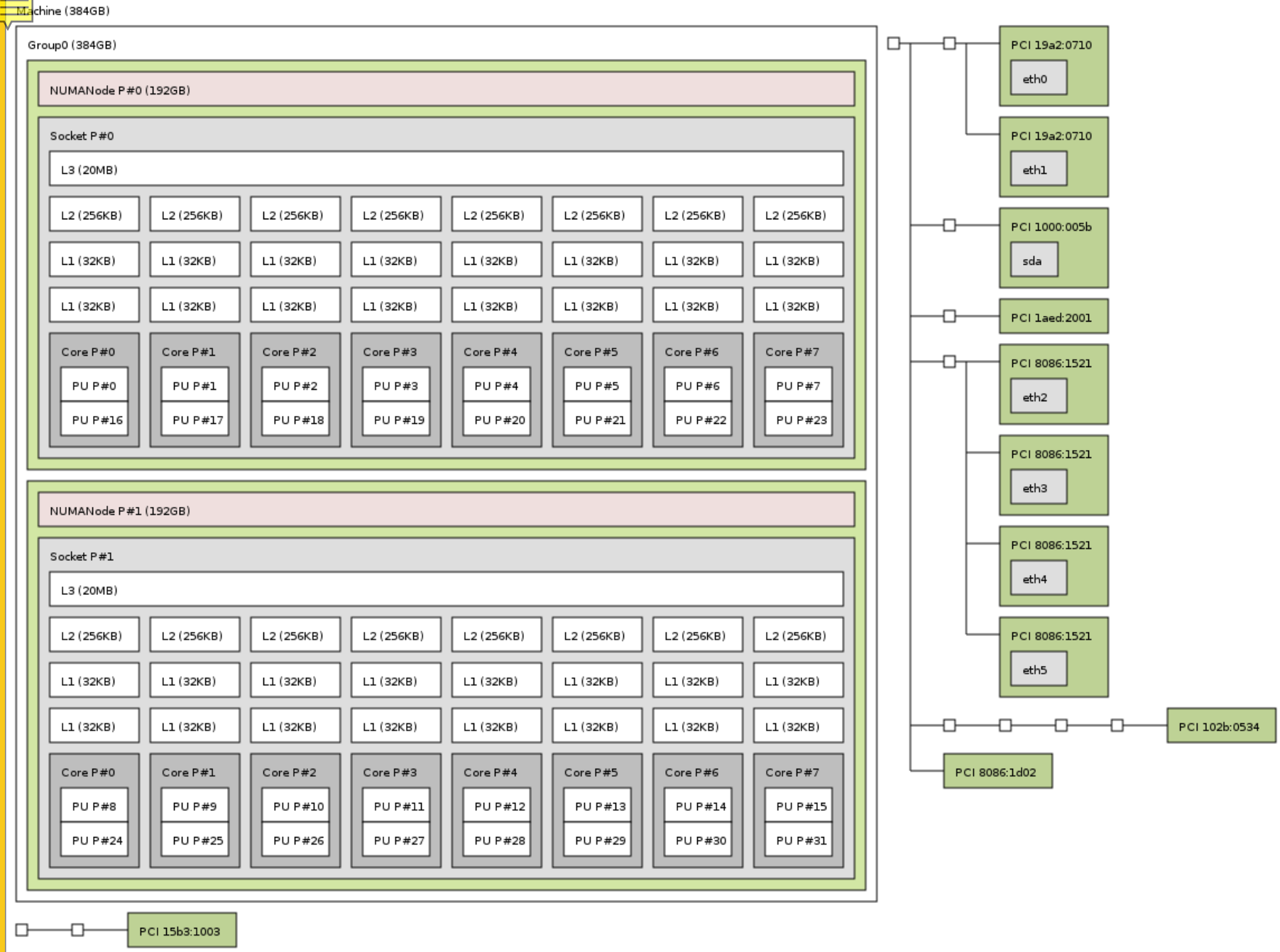
GbE
RDMA

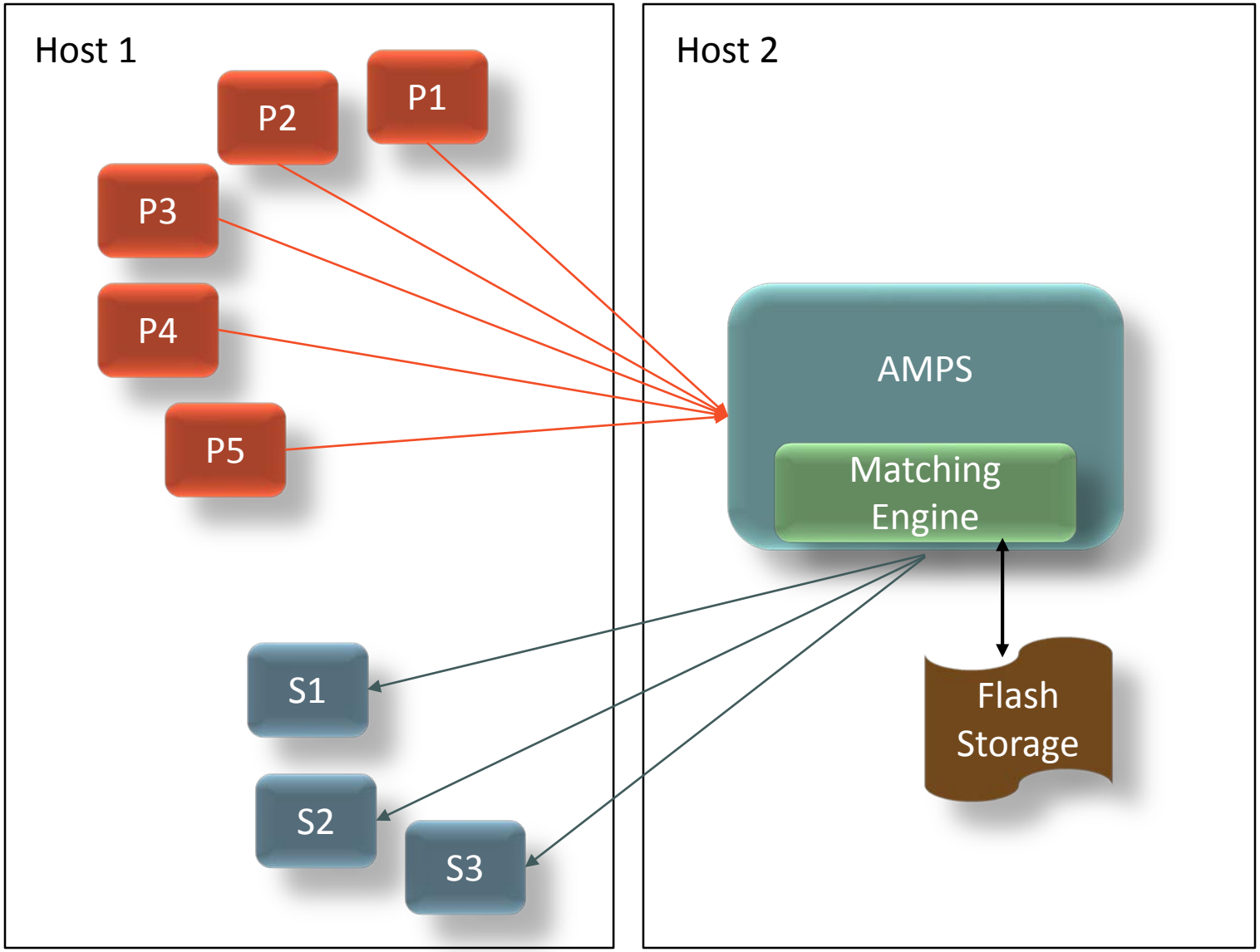
NUMA

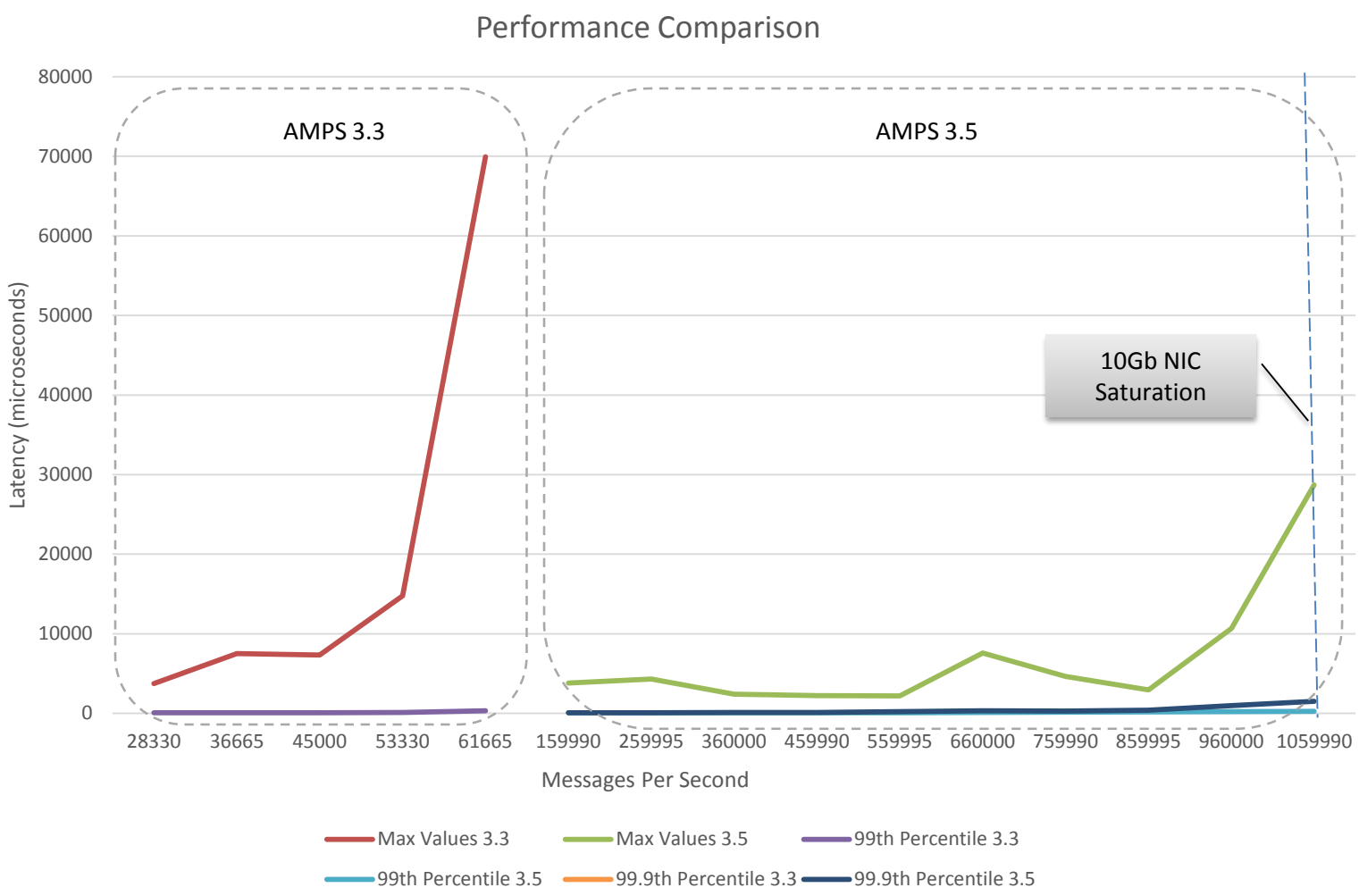
FLASH

60East TECHNOLOGIES

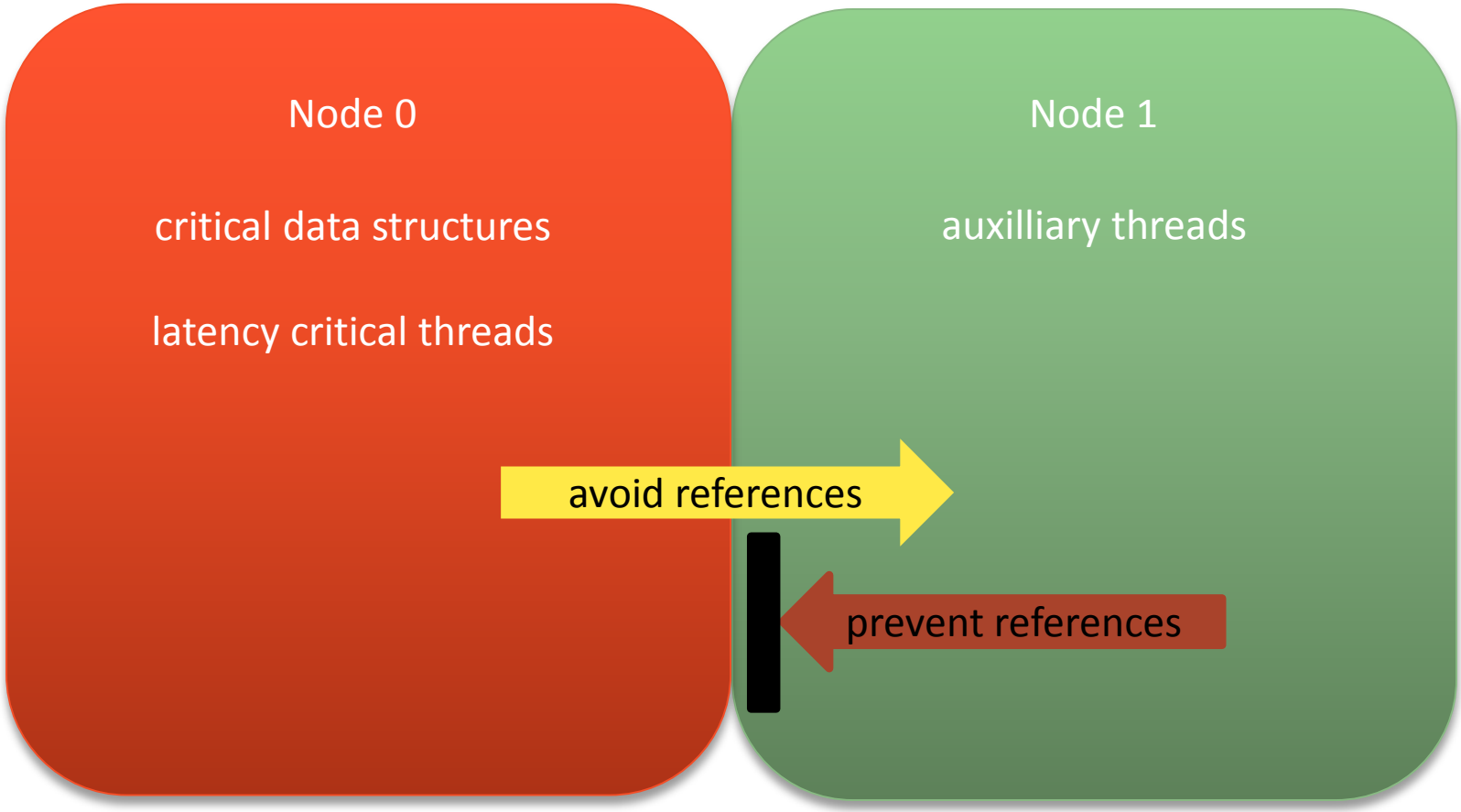
NUMA Architecture (Sandy Bridge)




Host 1
P1
P2
P3
P4
P5
S1
S2
S3
Host 2
AMPS
Matching Engine
Flash Storage



60East TECHNOLOGIES



60EAST TOP SECRET ORIGINAL RECIPE SECRET SAUCE

60East TECHNOLOGIES

Manual instrumentation using pieces in libnuma ("man numa")

set affinity

set allocation to prefer:0

look up node for datastructures with move_pages()

We do this in AMPS for assertion level debugging and guarding against regressions

Verification of all memory references using pintool

(http://software.intel.com/en-us/articles/pin-a-dynamic-binary-instrumentation-tool)

We have a pintool to watch cross node memory reads/writes from threads

We're trying to find the best way to share our pintools at the moment

PMU tools (http://github.com/andikleen/pmu-tools)

have you ever run "sudo ./ocperf.py top"? Mind blowing.

the csv lists shipped with pmutools have full list of available counters

numatop (http://01.org/numatop)

Early glimpse of tools of the future

great tool, requires a patch, but may make it into Linux 3.9 kernel.

numatop sample

```
                    NumaTOP v1.0, (C) 2012 Intel Corporation

Monitoring 304 processes and 428 threads (interval: 5.0s)

   PID          PROC      RMA(K)      LMA(K)     RMA/LMA        CPI      *CPU%
  7111          mgen     33097.5         3.7      8835.4      89.61        3.1
  7113       numatop         0.2         0.5         0.3       1.53        0.0
  4510    irqbalance         1.5         1.2         1.3       1.17        0.0
  1289   kworker/9:1         0.0         0.0         0.8       1.22        0.0
     1          init         0.0         0.0         0.0       0.00        0.0
     2      kthreadd         0.0         0.0         0.0       0.00        0.0
     3   ksoftirqd/0         0.0         0.0         0.0       0.00        0.0
     4   kworker/0:0         0.0         0.0         0.0       0.00        0.0
     5   kworker/0:0         0.0         0.0         0.0       0.00        0.0
     7   kworker/u:0         0.0         0.0         0.0       0.00        0.0
     8   migration/0         0.0         0.0         0.0       0.00        0.0
     9        rcu_bh         0.0         0.0         0.0       0.00        0.0
    10     rcu_sched         0.0         0.0         0.0       0.00        0.0
    11   ksoftirqd/1         0.0         0.0         0.0       0.00        0.0

<- Hotkey for sorting: 1(RMA), 2(LMA), 3(RMA/LMA), 4(CPI), 5(CPU%) ->
CPU% = system CPU utilization

Q: Quit; H: Home; R: Refresh; I: IR Normalize; N: Node
```

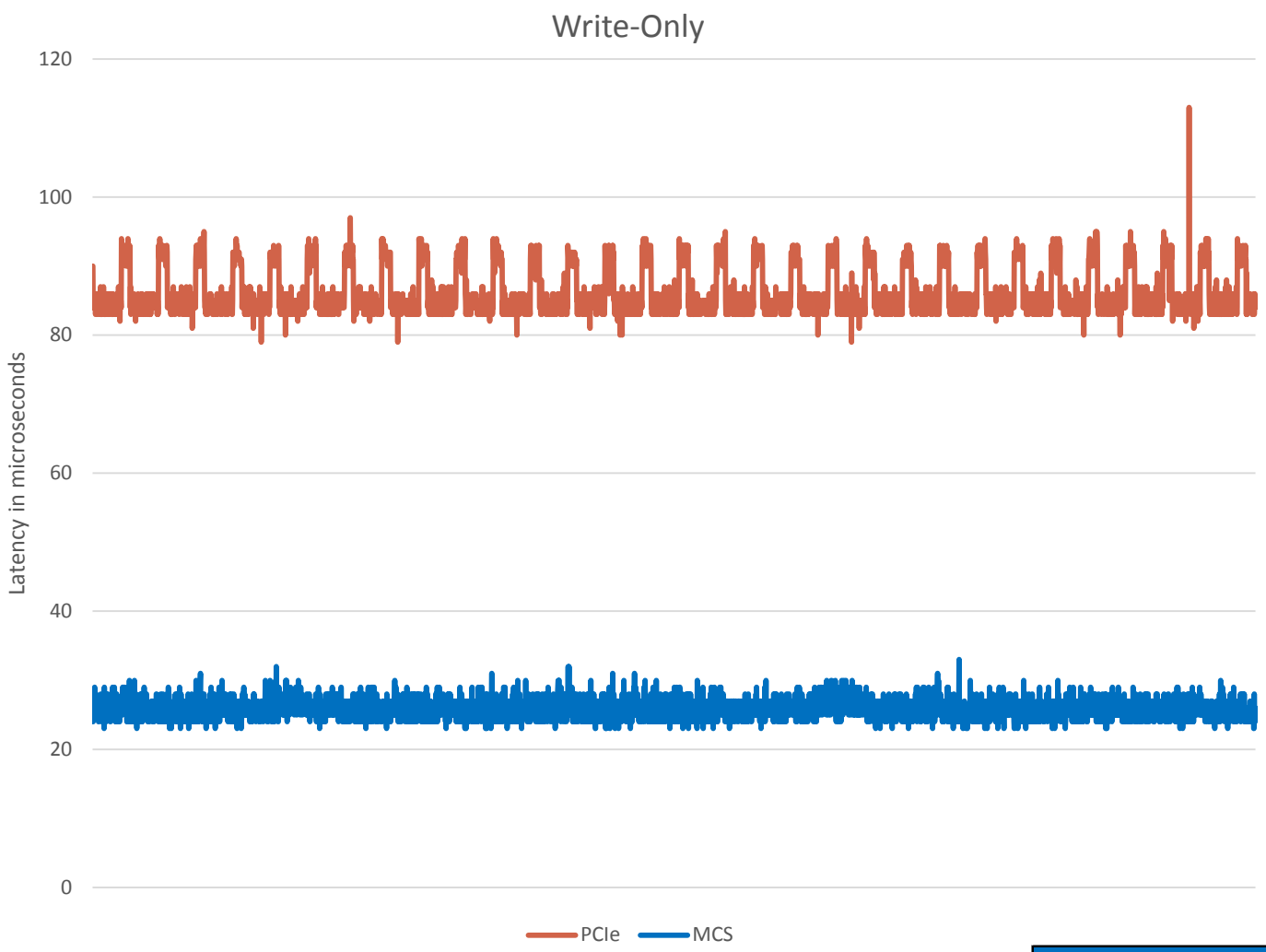
60East TECHNOLOGIES

- Experiment
- Read and Learn
  - Dave Dice Blog
  - https://blogs.oracle.com/dave/entry/numa_aware_reader_writer_locks
- Portable Hardware Locality (hwloc)
  - lstopo – display system topology
  - numactl – control NUMA policy
  - numstat – observe cross-node memory requests
  - libnuma – control affinity of threads and memory
- Design with non-uniform access in mind
  - Locality of threads and memory is critical so design processing paths accordingly
  - Try to reduce inter-package communication especially wrt memory access pattterns

diablo technologies

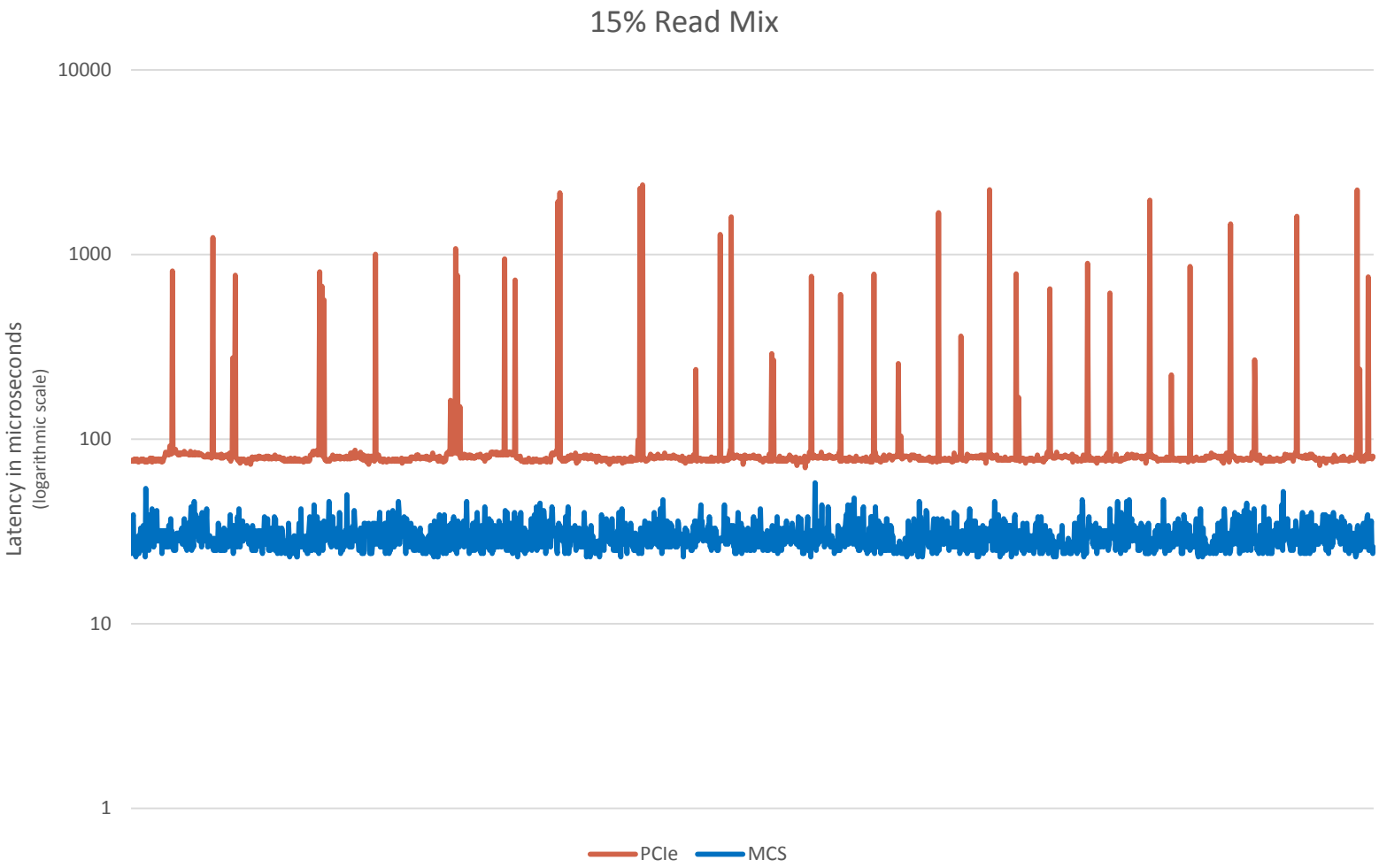
Memory Channel Storage™ Architecture

- Flash storage in DIMM package
- Puts storage on memory bus
- Low latency
- Consistent performance



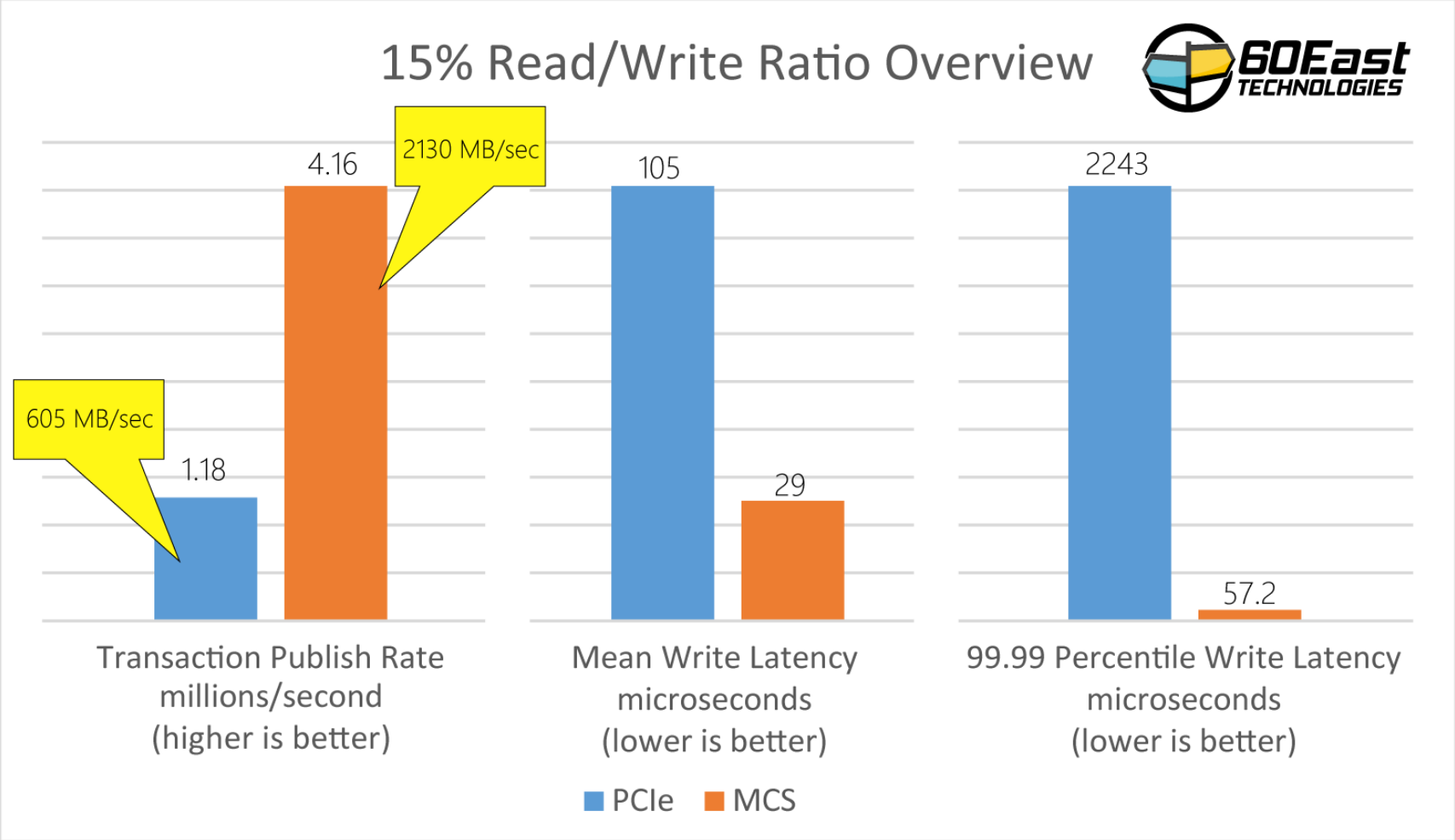
Average Latency		
	PCIe	86.09
	MCS	25.83
Maximum Latency		
	PCIe	113
	MCS	33

60East TECHNOLOGIES



Average Latency		
	PCIe	98.27
	MCS	29.74
Maximum Latency		
	PCIe	2382
	MCS	58

60East TECHNOLOGIES


15% Read/Write Ratio Overview
60East TECHNOLOGIES
605 MB/sec
2130 MB/sec
1.18
4.16
105
29
2243
57.2
Transaction Publish Rate
millions/second
(higher is better)
Mean Write Latency
microseconds
(lower is better)
99.99 Percentile Write Latency
microseconds
(lower is better)
PCIe
MCS

For More Information

- Slides for this talk
- Slides and video links for previous talks
- Evaluation version of AMPS
- 60East blog

www.crankuptheamps.com

60East TECHNOLOGIES