Microsoft Research Faculty Summit 2014

15TH ANNUAL

Microsoft Research
Faculty Summit
2014 15TH ANNUAL

Thin Servers with Smart Pipes: SoC Accelerators for Memcached

Thomas Wenisch
Assoc. Prof. of CSE
University of Michigan

Acknowledgements:
Kevin Lim (HP Labs)
Ali Saidi (ARM)
David Meisner (Facebook)
Partha Ranganathan (Google)

Microsoft

Key-Value caching layers are pervasive

- Memcached: Distributed, in-memory key-value cache
  - Critical in many scale-out, interactive workloads
  - Low-latency, high throughput
- Deployments have grown organically
  - Cache must keep up with user growth
  - Incremental SW refinement on commodity HW
- Individual server performance is important
  - All-to-all communication can make any server a bottleneck
  - Power constraints demand efficiency

Are commodity deployments efficient?

Commodity microarchitectures: poor match for workload characteristics

- Execution dominated by ill-behaved OS code
  - Mostly pushing small packets through network stack
- Xeons: power-hungry, ILP features do not help
  - Front end bottlenecks: Poor i-cache, branch predictability
  - Wasted cache: Little data locality beyond L1
- Atoms: too much performance sacrificed
  - Up to 4x worse performance per peak Watt than Xeon
  - Latency worsens rapidly under increased load

Commodity architectures ill-suited for frequent, small network requests

TSSP: Thin server with smart pipes

- Observation: Most memcached workloads are GET dominated
  - Facebook reports 30-to-1 GET:SET ratio [Atikoglu 2012]

- Thin Servers with Smart Pipes
  - Accelerate GETs fully in hardware
    - Networking stack, packet decipher, response gen. pipeline
    - HW/SW co-managed hash table

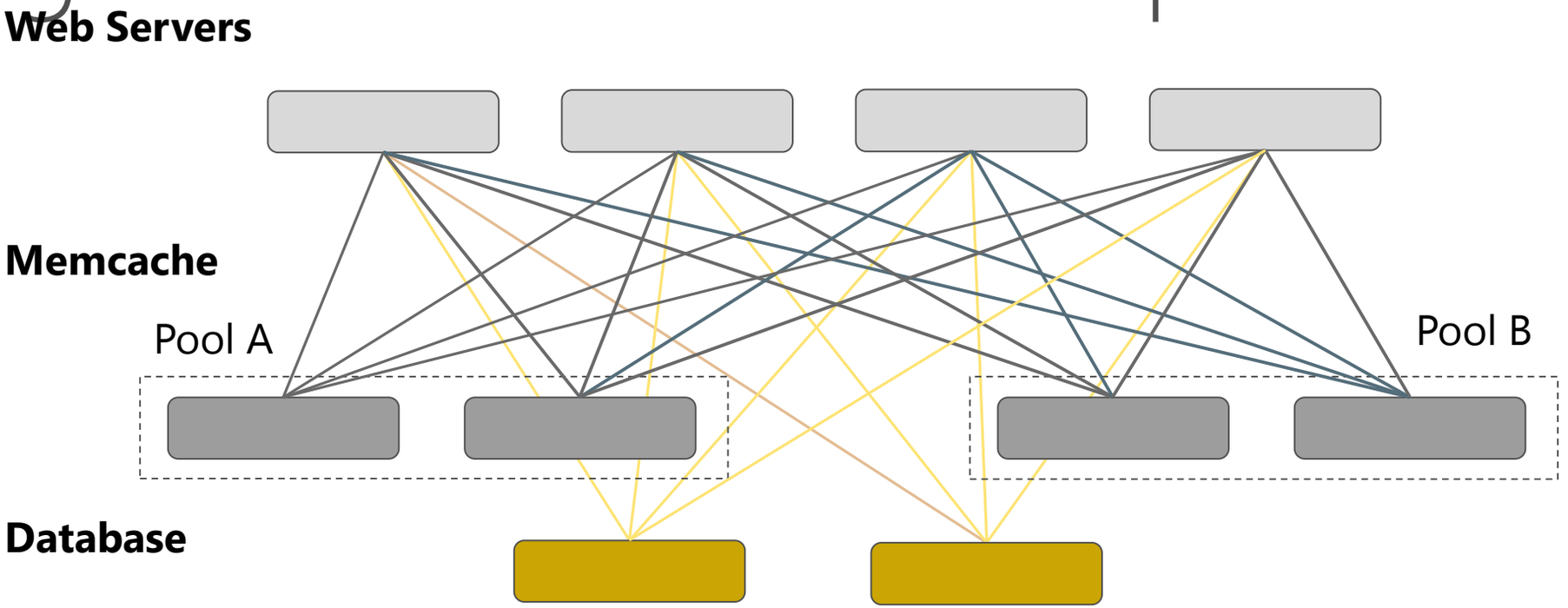
  - But, handle complex functionality in software
    - Memory alloc., key replacement, logging tough in HW
    - Atom-class core can keep up with remaining requests

Performance / Watt: 6x over Xeon, 16x over Atom

Outline

- Introduction
- Memcached on Commodity Hardware
- Thin Servers with Smart Pipes
- Conclusion

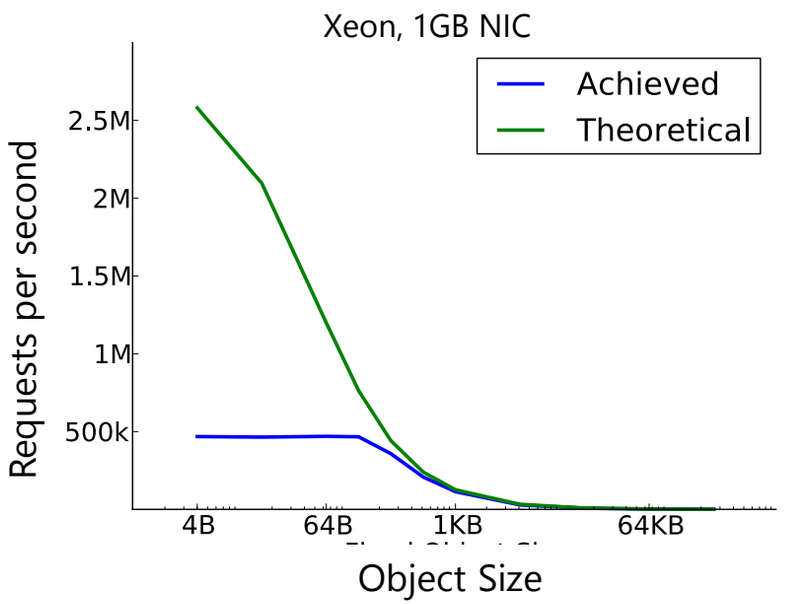
Background: memcached operation

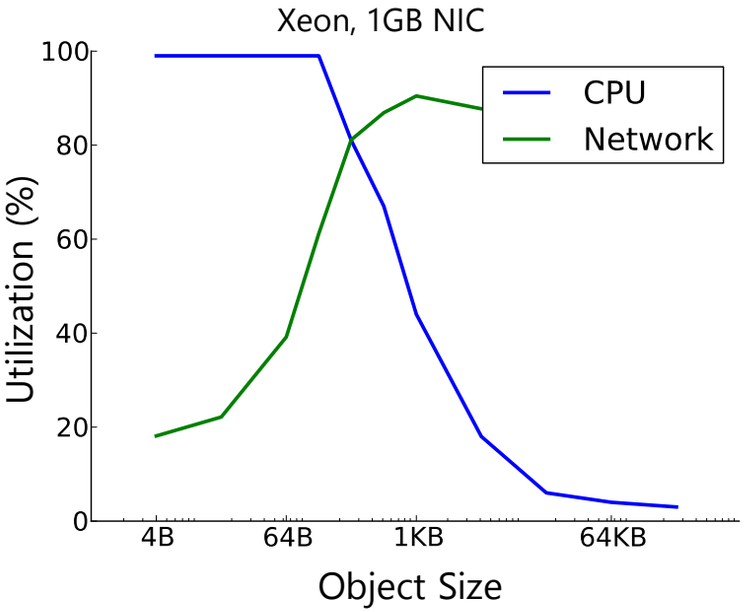


- Memcache shields database from majority of throughput
- Pools optimized for specific characteristics (pop. and size dists.)

How can we design systems for these use cases?

Poor performance for small objects





- Large values → easy to saturate network bandwidth
- Small values → CPU can't process packets fast enough

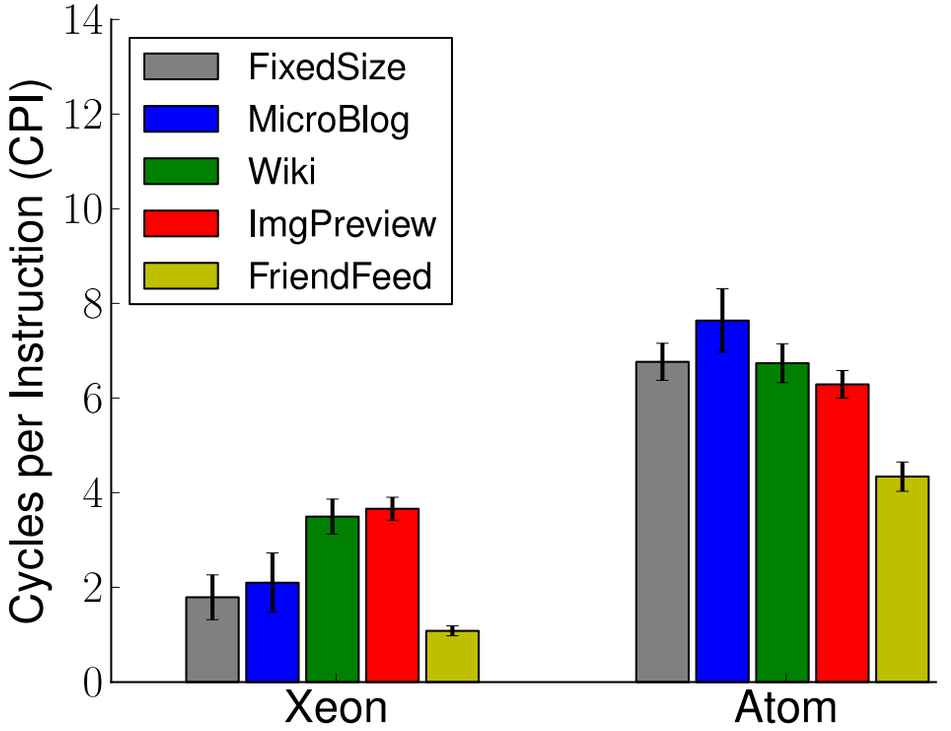
CPU becomes the bottleneck. Why?

Memcached on commodity hardware

- Understand HW by loadtesting servers with representative traffic
  - Identify microarchitecture bottlenecks
  - Understand impact of NIC hardware

	Xeon	Atom
Processor	2.25Ghz 6-Core Xeon Westmere L5640 12 MB L3 Cache	1.6 Ghz Atom D510 Dual-core 2x SMT 1 MB L2 Cache
DRAM	3x 4GB DDR3-1066	2x 2GB DDR2-800 SDRAM
Commodity	Realtek RTL8111D Gigabit	-
Enterprise	Broadcom NetXtreme II Gigabit	Intel 82574L Gigabit
10GbE	Intel X520-T2 10GbE NIC	Intel X520-T2 10GbE NIC

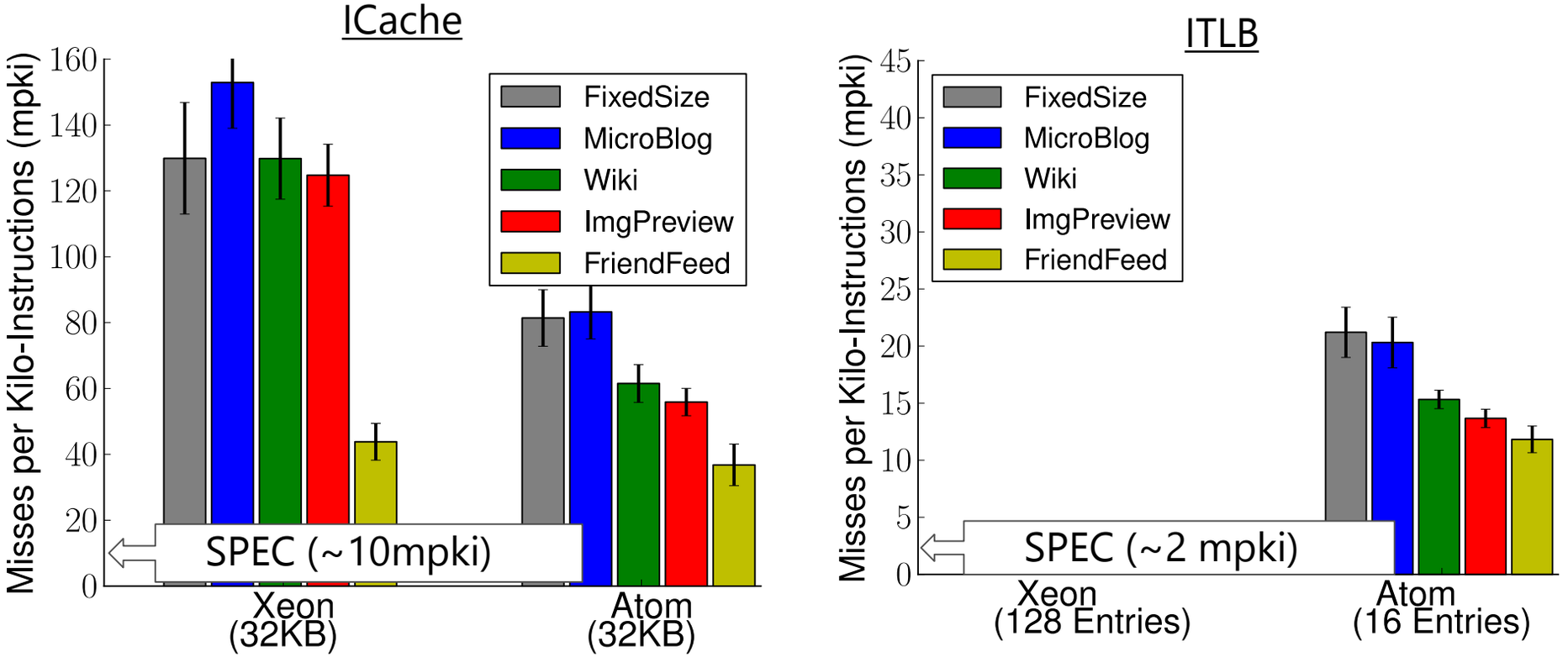
Massively inefficient use of microarchitecture



- Both achieve small fraction of peak performance
  - Percent of max IPC: 8.3% (Xeon) and 7% (Atom)

Why do instructions take 2-8 cycles to retire?

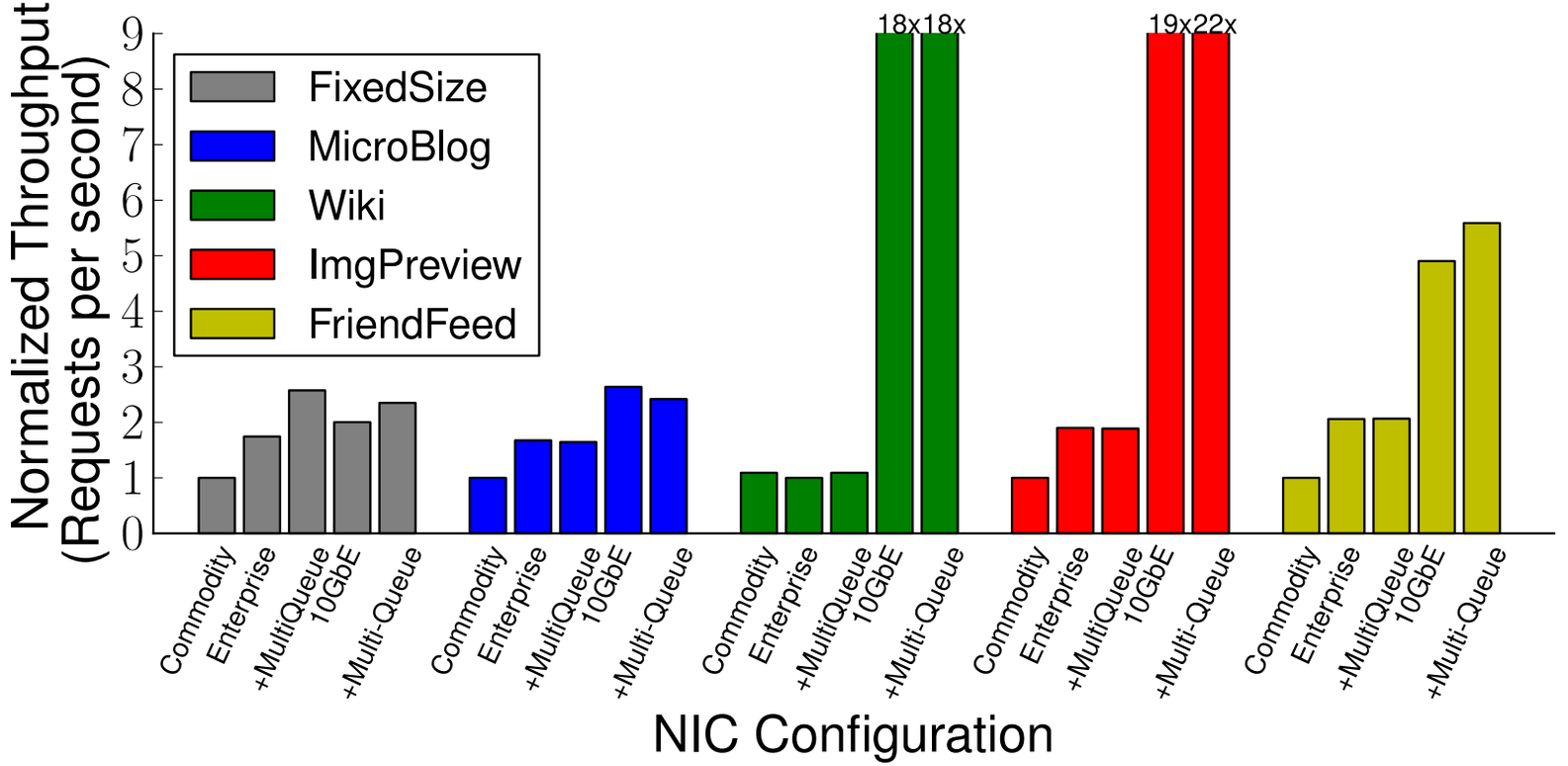
Inefficiency: Poor instruction supply



- Behavior consistent with other scale-out workloads [Ferdman '12]

High usage of kernel, TCP/IP and library code put tremendous pressure on instruction supply

Observation: NIC features matter



- Features often more important than raw bandwidth
  - E.g. MultiQueue gives up to 50% boost

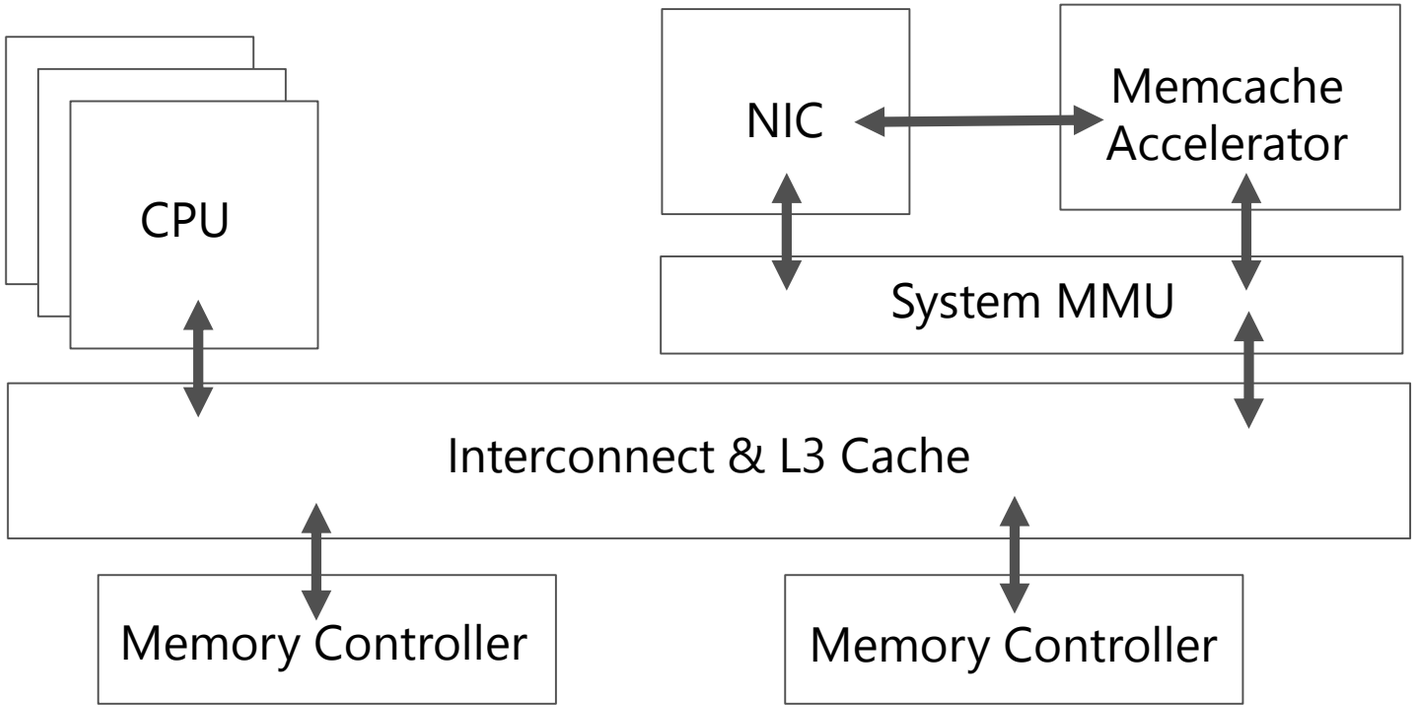
Takeaway: Integration of network and processing is key

Thin Servers with Smart Pipes

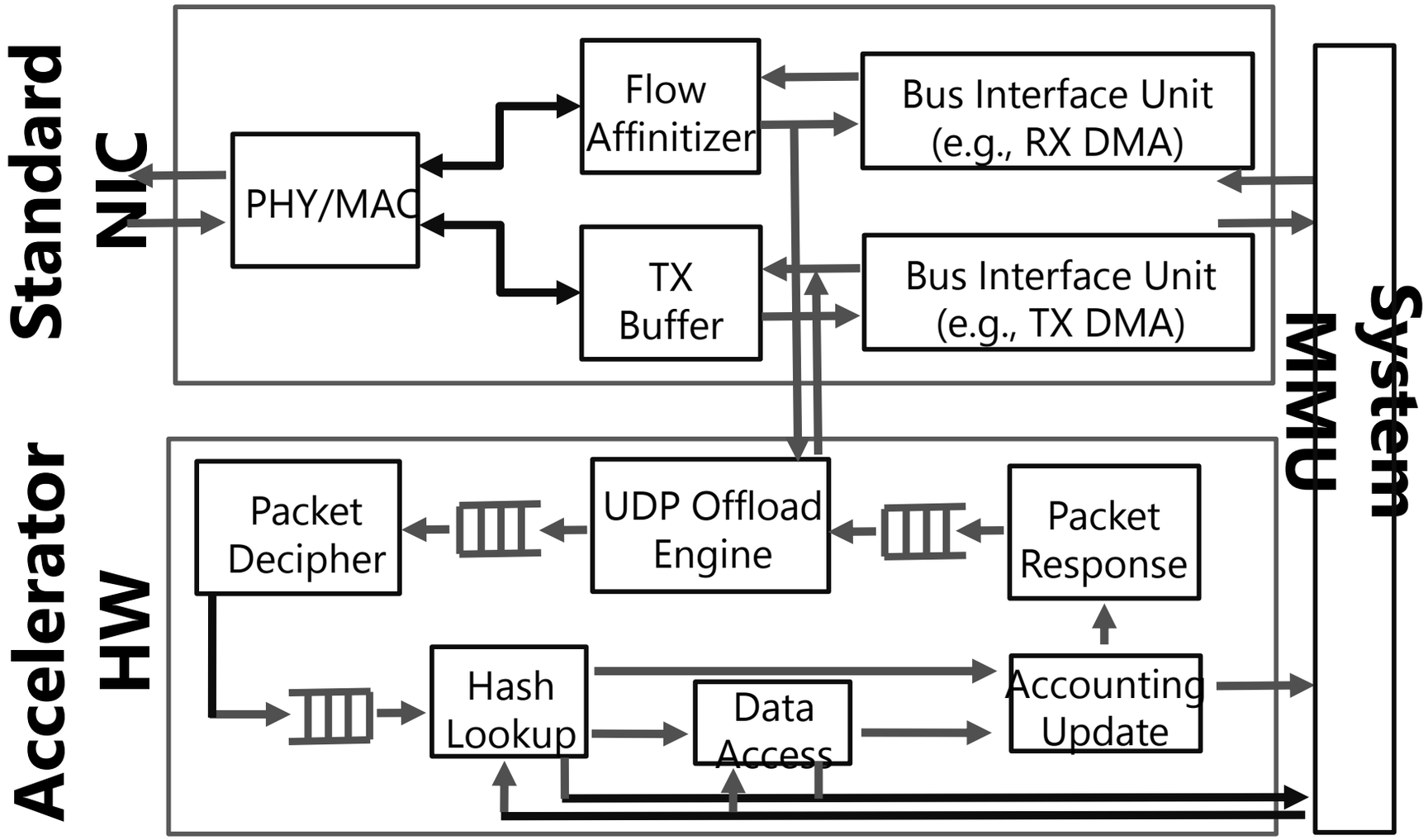
- Observation 1: Frequent network interactions
  - Creates significant inefficiencies on both architectures
- Observation 2: GET requests dominate traffic
  - 97% of operations in large-scale cluster [Atikoglu '12]
- SoC architecture integrating low-power cores with Memcached accelerator
  - Hardware GET processing and network stack near NIC
  - Wimpy CPU handles other operations

Achieves both high performance and low-power

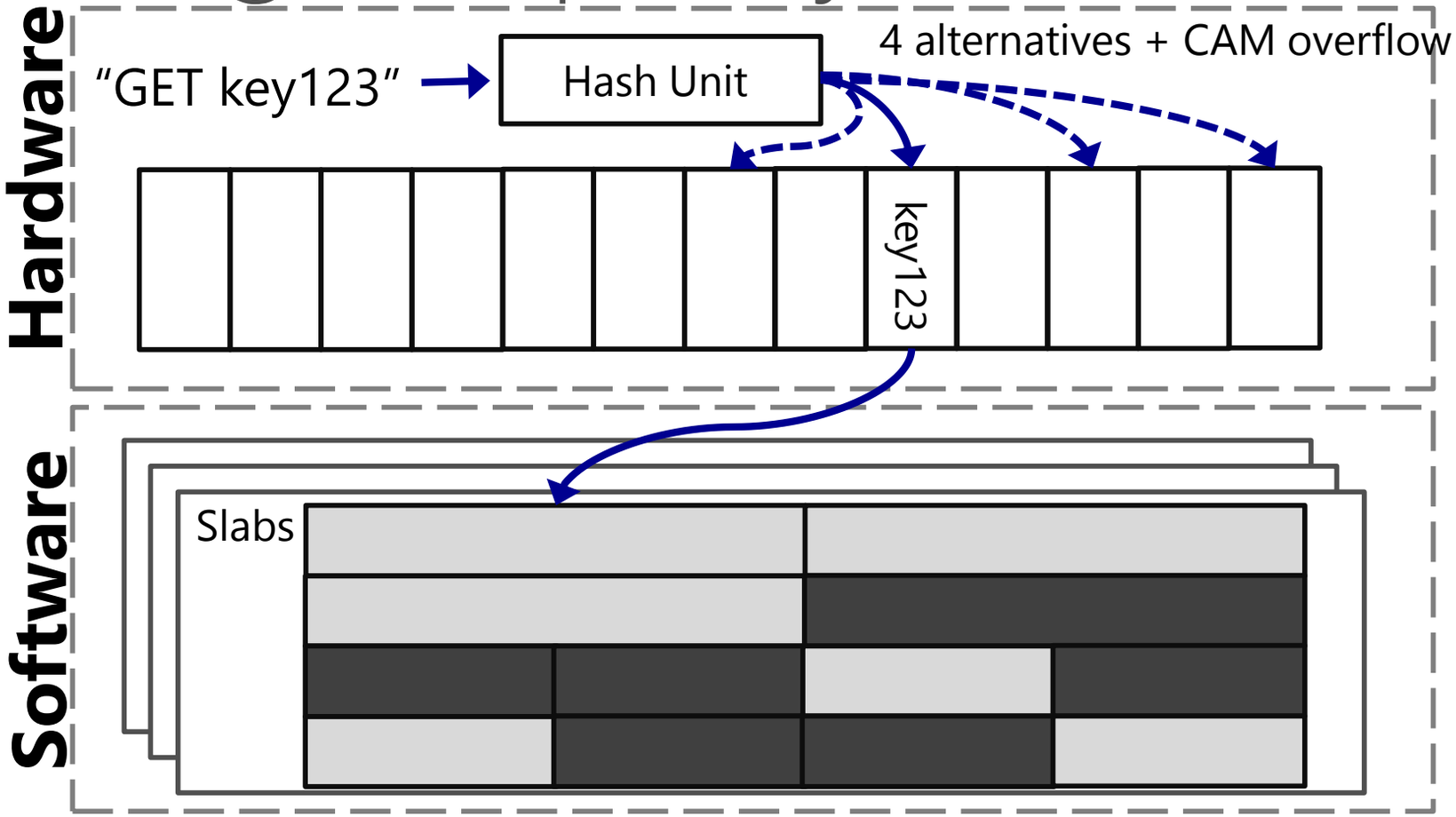
TSSP SoC architecture



Memcache accelerator



Partitioning complexity across SW/HW



HW: GETs, hash table (incl. SW-directed updates)

SW: memory, replacement, SETs, etc.

TSSP evaluation

- TSSP Evaluation based on FPGA hardware prototype [Chalamalasetti '13]
  - Use empirically measured latency and throughput
  - Estimated power from existing SoC design
  - Xeon/Atom empirically measured on commodity HW

System	Power (W)	Perf. (k qps)	Perf./W
TSSP	16	282	17.63
Xeon-TCP	143	410	2.87
Xeon-UDP	143	372	2.60
Atom-UDP	35	58	1.66
Atom-TCP	35	38	1.09

Conclusions

- Deploying efficient Memcached systems is difficult
  - No current CPUs good fit for performance, power, cost
  - Frequent network interactions causes key bottlenecks
- Insights drive Thin Servers with Smart Pipes SoC design
  - Low-power cores integrated with network controller
  - Memcached accelerator for GET requests
  - 6-16X performance-per-watt improvement
- Further opportunities for accelerators in the data center
  - Better I/O, media processing, machine learning & inference

Save the planet and return
your name badge before you
leave (on Tuesday)

Microsoft Privacy Policy statement applies to all information collected. Read at research.microsoft.com

© 2014 Microsoft Corporation. All rights reserved. Microsoft, Windows, Windows Vista and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries.
The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation. MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.