OCP CHINA DAY | June 25th 2019 Beijing

The Co-Evolution of Data Center Hardware and Systems Software

Lintao Zhang | Principle Researcher

Microsoft Research Asia

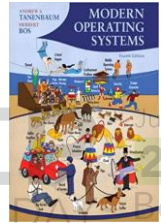
25th June, 2019

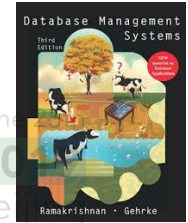
Introduction

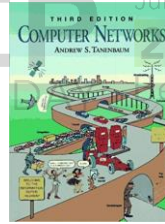
Many classic systems software are designed decades ago, with many common assumptions:

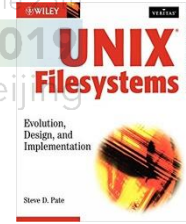
- CPU is the Central Processing Unit, it is fast, and is getting faster
- Disk is slow, especially for random accesses
- Network is unreliable
- Memory hierarchy is consisted of CPU cache, DRAM, and disk paging.
- Accessing hardware needs to go through OS kernel
- Locality is important, data should be stored close to compute
- ……

Many of the assumptions are no long true in modern data centers

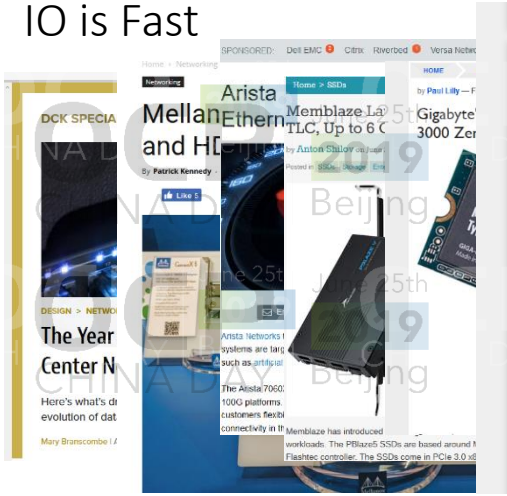








IO is Fast



tom's HARDWARE PRODUCT REVIEWS BUYING GUIDES E3 RASPBERRY PI DE

TRENDING Ryzen 9 3950X Radeon RX 5700 Zen 2 Microarchitecture DDR5 RAM

STORAGE › NEWS

Boot Your PC at 15GB/s With Gigabyte's New PCIe 4.0 SSD Adapter

2 COMMENTS

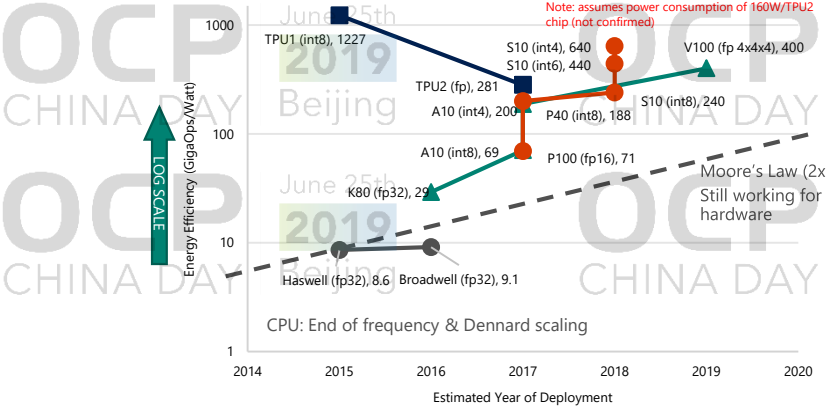
by Paul Alcorn May 27, 2019 at 2:42 PM

The debut of AMD's third-gen Ryzen chips hastened the arrival of blazingly-fast PCIe 4.0 SSDs, but we'll soon be jaded about SSDs with a "paltry" 5GB/s of throughput, especially with new models coming with even faster speeds. But what if you could boot your PC with an SSD adapter that hits 15 GB/s?

Credit: Tom's Hardware

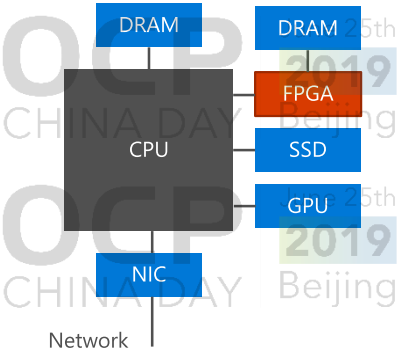
That's the purpose of Gigabyte's prototype quad-SSD adapter. This new add-in card snaps into your PCIe slot to get access to the full blazing 64 GB/s of theoretical throughput available through the PCIe 4.0 x16 connection. After you add in the annoying, yet unavoidable overhead of the PCIe interface and RAID, Gigabyte's new prototype has hit up to 15 GB/s when paired with four of the company's new Aorus PCIe 4.0 SSDs. Each of those SSDs hit up to 5GB/s, so the card could potentially push up to 20 GB/s with further tuning.

The rise of domain-specific computing

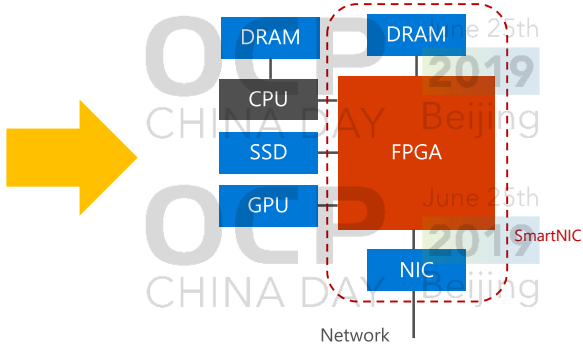


Credit: Doug Burger

Cloud Architecture built around Smart NIC



Traditional Thinking



SmartNIC Centric Thinking

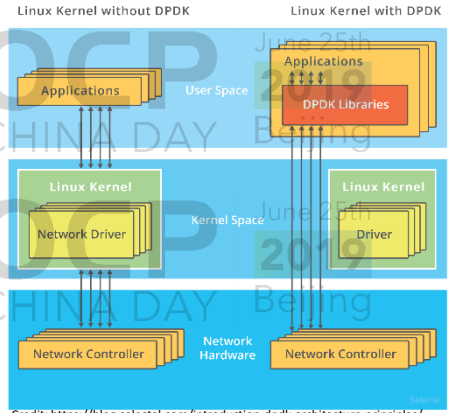
Kernel can be by-passed for High Speed IO

Traditional file system calls and network IO are usually in OS kernel

- May incur too much overhead

RDMA, DPDK and SPDK tries to move data plane to user space and by-pass kernel

- But they introduce new APIs



Credit: https://blog.selectel.com/introduction-dpdk-architecture-principles/

Agenda

1. Introduction
2. Two examples that take advantage of new hardware trends
3. Going forward and Conclusion

Agenda

1. Introduction
2. Two examples that take advantage of new hardware trends
3. Going forward and Conclusion

Example 1: Key-Value Store in Data Centers

Key Value Store is a key piece of systems infrastructure

- Traditionally used as a cache in the cloud
- Now often act as a share data structure
- Example usage: database, graph processing, parameter server,

MEMCACHED redis

Key		Value
User ID	→	Recent Tweets
SQL	→	Query result
Session	→	Browser cookies

Modern applications require:

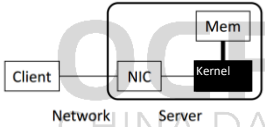
- High throughput
- Low tail latency
- Write intensive
- Support vector and atomic operations

GraphX

Key		Value
Node	→	List of edges
Feature	→	Weight
Shard	→	Sequence number

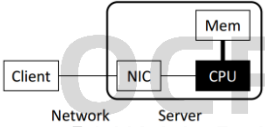
More demanding workload

Key-Value Store Architectures



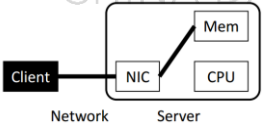
Kernel TCP/IP

Bottleneck: Network stack in OS
(~300 Kops per core)



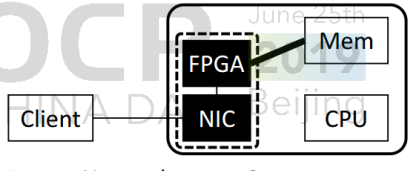
Kernel Bypass

Bottlenecks: CPU random memory access and KV operation computation
(~5 Mops per core)



One-sided RDMA

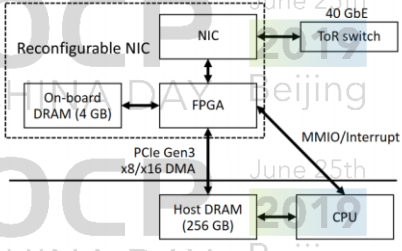
Communication overhead: multiple round-trips per KV operation (fetch index, data)
Synchronization overhead: write operations



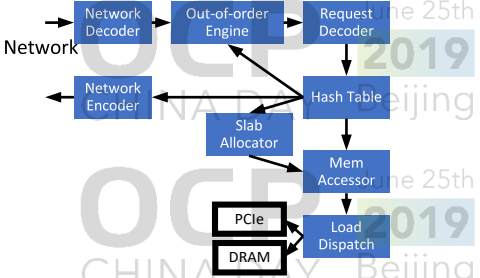
KV-Direct: Leverage SmartNIC

Offload KV processing on CPU to Programmable NIC

KV-Direct Architecture

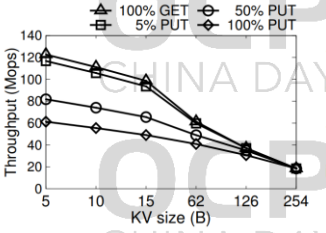


Hardware Configuration

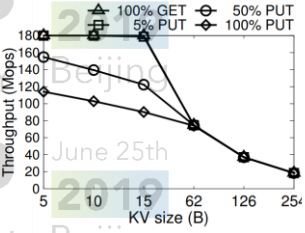


KV Processor

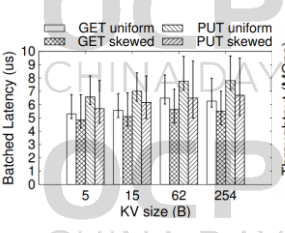
KV-Direct Performance Characteristics



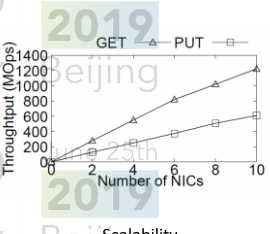
Uniform workload throughput



Skewed workload throughput



Latency

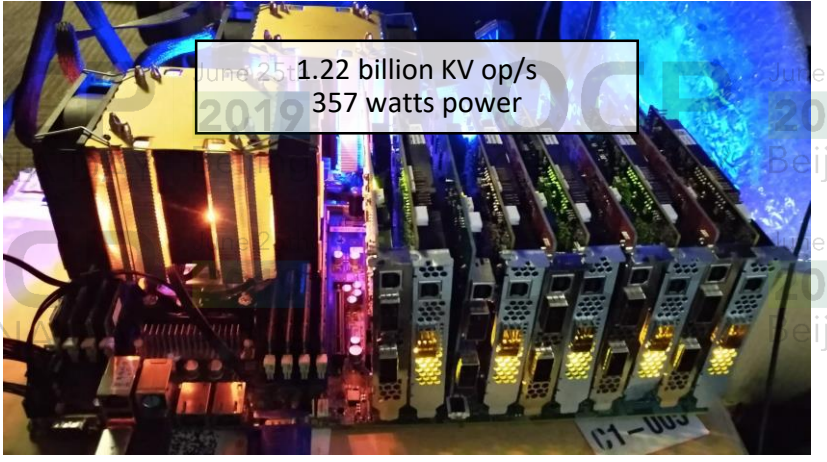


Scalability

KVS Performance Comparison

	Tput (Mops) (GET / PUT)	Power (Kops/W)	Architecture	Latency (us) (GET / PUT)
Memcached	1.5 / 1.5	5 / 5	TCP/IP	50 / 50
MemC3	4.3 / 4.3	14 / 14	TCP/IP	50 / 50
RAMCloud	6 / 1	20 / 3.3	Kernel bypass	5 / 14
MICA (12 NICs)	137 / 135	342 / 337	Kernel bypass	81 / 81
FARM	6 / 3	30 (261) / 15	One-side RDMA	4.5 / 10
DrTM-KV	115 / 14	500 (3972) / 60	One-side RDMA	3.4 / 6.3
HERD	35 / 25	490 / 300	Two-side RDMA	4 / 4
FPGA-Xilinx	14 / 14	106 / 106	Standalone FPGA	3.5 / 4.5
Mega-KV	166 / 80	330 / 160	GPU	280 / 280
KV-Direct (1 NIC)	180 / 114	1487 (5454) / 942 (3454)	Programmable NIC	4.3 / 5.4
KV-Direct (10 NICs)	1220 / 610	3417 (4518) / 1708 (2259)	Programmable NIC	4.3 / 5.4

New Milestone for KVS Performance



Takeaway for KV-Direct

KV-Direct leverage SmartNICs hardware to implement KV Store

- Solve the semantic mismatch of RDMA and KV access
- CPU is used only for control
- Push performance to the limit of hardware
- Achieve orders of magnitude power efficiency improvement
- Published in SOSP17

Many other workloads can be improved in a similar manner

- Such as DNN training, storage, NFV ……
- This is an emerging paradigm change in data center

Example 2: Socket for Communication

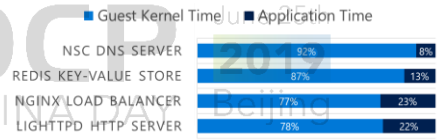
Socket is the universal communication primitive

- Designed half a century ago
- Complicated semantic
- Low performance, high overhead

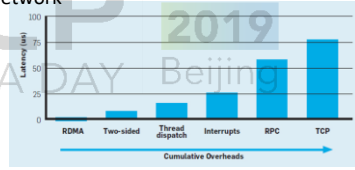
Challenges:

- How to be fully compatible with native socket
- How to leverage modern hardware such as RDMA and multicore
- How to maintain isolation and security

Socket wastes CPU



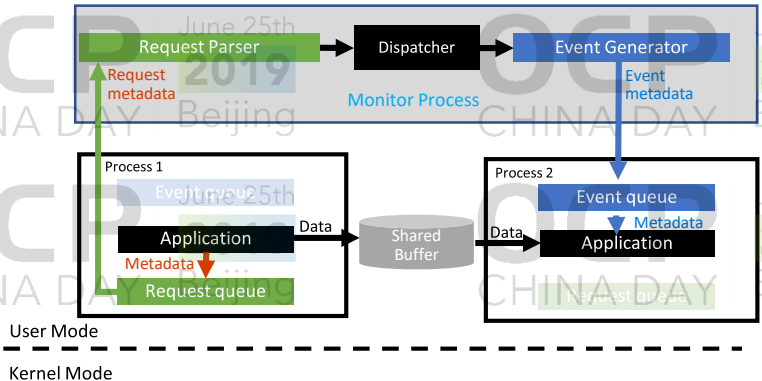
Socket wastes low latency DC network



Source: Luiz Barroso, Mike Marty, David Patterson, Parthasarathy Ranganathan.

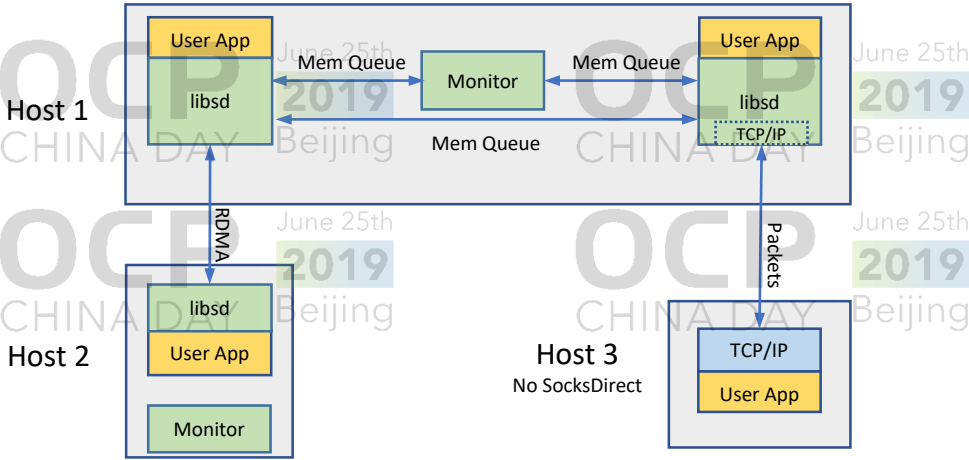
Attack of the Killer Microseconds, Communications of the ACM, April 2017

SocksDirect: Fast and Compatible User Space Socket

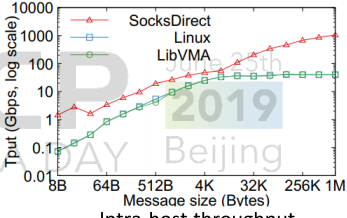


Kernel Bypass with Dedicated Monitor Process

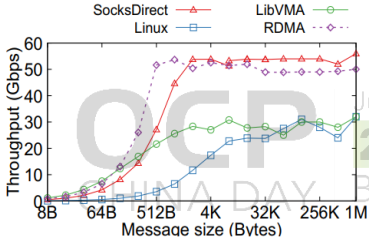
SocksDirect Supports Different Transports for Data



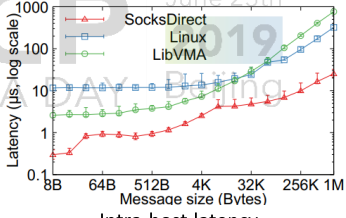
SocksDirect Performance



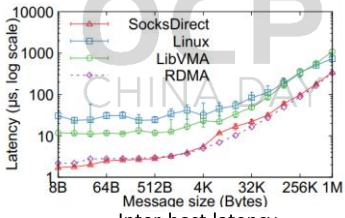
Intra-host throughput



Inter-host throughput

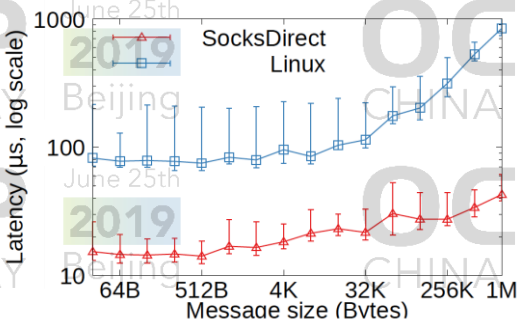


Intra-host latency



Inter-host latency

Application End-to-End Performance with SocksDirect



Nginx HTTP End-to-End Latency

Takeaway for SocksDirect

SocksDirect implements a high-performance Socket system in user space

- Bypass OS kernel to reduce overhead
- Dedicates a core for monitor process (take advantage of multi-core CPU)
- Can act as a drop-in replacement for classic socket without app modification
- Achieve orders of magnitude performance improvements
- Published in SIGCOMM19

Other traditional OS systems functions can be improved in a similar manner

- Such as file systems

How is this related to OCP

Microsoft is contributing hardware innovations

- SmartNIC: FPGA enabled programable NIC
- Zipline : Compression Engine based on FPGA
- Cerberus: Hardware root of trust
- Denali: Cloud optimized Flash storage

New hardware in Data Centers require us to rethink systems software design

- Which abstraction to use for non-volatile Storage Class Memory?
- Will CPU be relegated to become a second-class citizen in data center?
- How can we evolve the systems abstraction while keep software modification to a minimum?

Hardware and Systems software needs to evolve together

- We demonstrate two cases that can greatly improve systems performance by taking advantage of new hardware

OCP CHINA DAY June 25th 2019 Beijing

Thank you