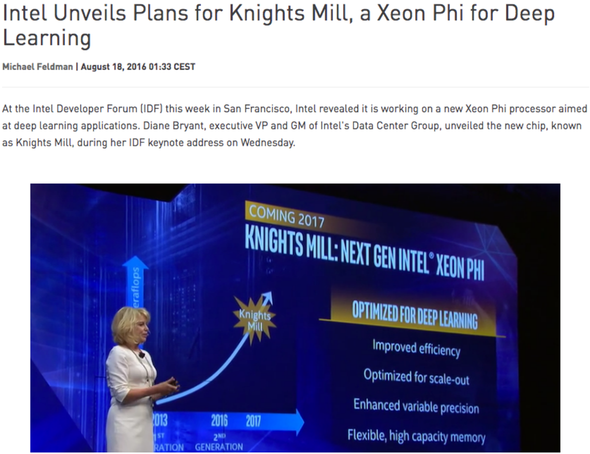
Solros: A Data-Centric Operating System Architecture for Heterogeneous Computing

Changwoo Min, Woonhak Kang, Mohan Kumar, Sanidhya Kashyap, Steffen Maass, Heeseung Jo, Taesoo Kim

Virginia Tech, eBay, Georgia Tech, Chonbuk National University

April 26, 2018

Cambrian Explosion of Processor Architecture



Intel Unveils Plans for Knights Mill, a Xeon Phi for Deep Learning

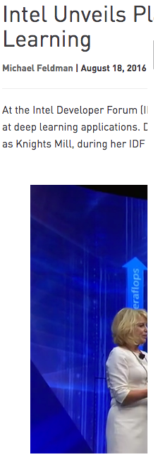
Michael Feldman | August 18, 2016 01:33 CEST

At the Intel Developer Forum (IDF) this week in San Francisco, Intel revealed it is working on a new Xeon Phi processor aimed at deep learning applications. Diane Bryant, executive VP and GM of Intel's Data Center Group, unveiled the new chip, known as Knights Mill, during her IDF keynote address on Wednesday.

COMING 2017
KNIGHTS MILL: NEXT GEN INTEL® XEON PHI
OPTIMIZED FOR DEEP LEARNING
Improved efficiency
Optimized for scale-out
Enhanced variable precision
Flexible, high capacity memory
eraflops
Knights Mill
013
2016
2017
1ST
RATION
2ND
GENERATION

Specialization of general-purpose processors

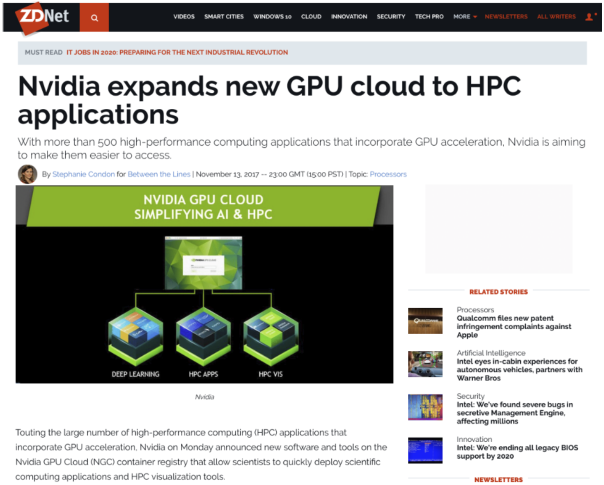
Cambrian Explosion of Processor Architecture



Intel Unveils Plans for Knights Mill, a Xeon Phi for Deep Learning

Michael Feldman | August 18, 2016

At the Intel Developer Forum (I
at deep learning applications. D
as Knights Mill, during her IDF



MUST READ IT JOBS IN 2020: PREPARING FOR THE NEXT INDUSTRIAL REVOLUTION

Nvidia expands new GPU cloud to HPC applications

With more than 500 high-performance computing applications that incorporate GPU acceleration, Nvidia is aiming to make them easier to access.

By Stephanie Condon for Between the Lines | November 13, 2017 -- 23:00 GMT (15:00 PST) | Topic: Processors

Nvidia

Touting the large number of high-performance computing (HPC) applications that incorporate GPU acceleration, Nvidia on Monday announced new software and tools on the Nvidia GPU Cloud (NGC) container registry that allow scientists to quickly deploy scientific computing applications and HPC visualization tools.

RELATED STORIES

Processors
Qualcomm files new patent infringement complaints against Apple

Artificial Intelligence
Intel eyes in-cabin experiences for autonomous vehicles, partners with Warner Bros

Security
Intel: We've found severe bugs in secretive Management Engine, affecting millions

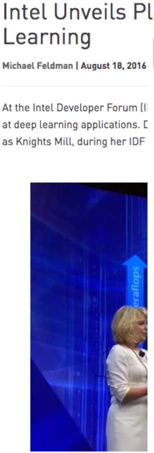
Innovation
Intel: We're ending all legacy BIOS support by 2020

NEWSLETTERS

Specializatio

Generalization of co-processors

Cambrian Explosion of Processor Architecture

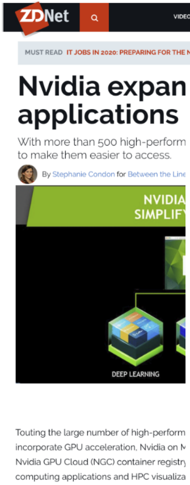


Intel Unveils Plans for Knights Mill, a Xeon Phi for Deep Learning

Michael Feldman | August 18, 2016

At the Intel Developer Forum (I
at deep learning applications. D
as Knights Mill, during her IDF

Specializatio



MUST READ IT JOBS IN 2020: PREPARING FOR THE NEXT INDUSTRIAL REVOLUTION

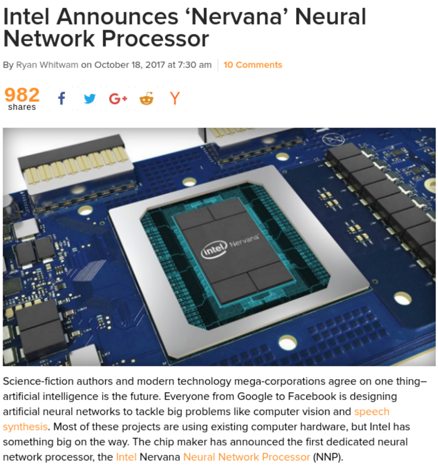
Nvidia expan
applications

With more than 500 high-perform
to make them easier to access.

By Stephanie Condon for Between the Line

Touting the large number of high-perform
incorporate GPU acceleration, Nvidia on M
Nvidia GPU Cloud (NGC) container registry
computing applications and HPC visualiza

Generalization c



Intel Announces 'Nervana' Neural Network Processor

By Ryan Whitwam on October 18, 2017 at 7:30 am | 10 Comments

982 shares

Science-fiction authors and modern technology mega-corporations agree on one thing–artificial intelligence is the future. Everyone from Google to Facebook is designing artificial neural networks to tackle big problems like computer vision and speech synthesis. Most of these projects are using existing computer hardware, but Intel has something big on the way. The chip maker has announced the first dedicated neural network processor, the Intel Nervana Neural Network Processor (NNP).

Specialization of co-processors

Blazingly fast IO Devices

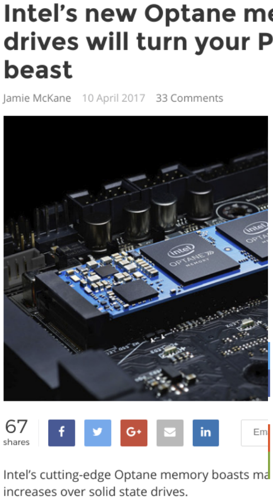
Intel's new Optane memory drives will turn your PC into a beast

Jamie McKane 10 April 2017 33 Comments

Intel's cutting-edge Optane memory boasts massive performance increases over solid state drives.

Blazingly fast storage/memory

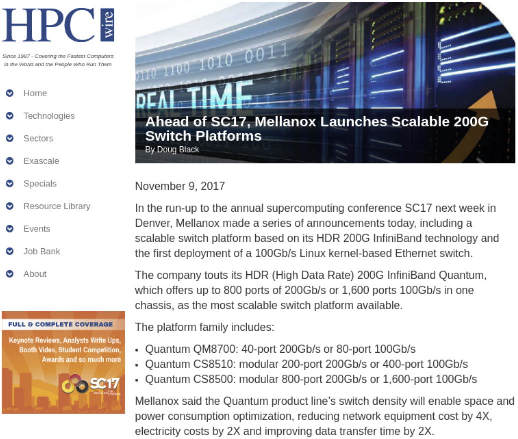
Blazingly fast IO Devices



Intel's new Optane memory drives will turn your PC into a beast

Jamie McKane 10 April 2017 33 Comments

Intel's cutting-edge Optane memory boasts ma
increases over solid state drives.



HPC wire
Since 1987 - Covering the Fastest Computers in the World and the People Who Run Them
Home
Technologies
Sectors
Exascale
Specials
Resource Library
Events
Job Bank
About
FULL & COMPLETE COVERAGE
Keynote Reviews, Analysts Write Ups, Booth Vides, Student Competition, Awards and so much more
SC17
REAL TIME

Ahead of SC17, Mellanox Launches Scalable 200G Switch Platforms

By Doug Black

November 9, 2017

In the run-up to the annual supercomputing conference SC17 next week in Denver, Mellanox made a series of announcements today, including a scalable switch platform based on its HDR 200G InfiniBand technology and the first deployment of a 100Gb/s Linux kernel-based Ethernet switch.

The company touts its HDR (High Data Rate) 200G InfiniBand Quantum, which offers up to 800 ports of 200Gb/s or 1,600 ports 100Gb/s in one chassis, as the most scalable switch platform available.

The platform family includes:

- Quantum QM8700: 40-port 200Gb/s or 80-port 100Gb/s
- Quantum CS8510: modular 200-port 200Gb/s or 400-port 100Gb/s
- Quantum CS8500: modular 800-port 200Gb/s or 1,600-port 100Gb/s

Mellanox said the Quantum product line's switch density will enable space and power consumption optimization, reducing network equipment cost by 4X, electricity costs by 2X and improving data transfer time by 2X.

Blazingly fast storage/

Blazingly fast network

Blazingly fast IO Devices

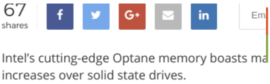
Intel's new Optane memory drives will turn your PC into a beast

Jamie McKane 10 April 2017 33 Comments

HPCwire
Since 1987 - Covering the Fastest Computers in the World and the People Who Run Them
0110 1100 1010 0011

How to exploit the full potential of such hardware devices without pain?

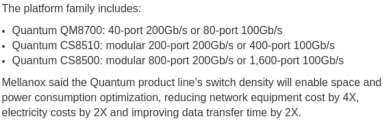
- System-wide performance
- Ease of programming



Intel's cutting-edge Optane memory boasts ma increases over solid state drives.

Blazingly fast storage/

FULL & COMPLETE COVERAGE
Keynote Reviews, Analysts Write Ups, Booth Vides, Student Competition, Awards and so much more
SC17



The platform family includes:

- Quantum QM8700: 40-port 200Gb/s or 80-port 100Gb/s
- Quantum CS8510: modular 200-port 200Gb/s or 400-port 100Gb/s
- Quantum CS8500: modular 800-port 200Gb/s or 1,600-port 100Gb/s

Mellanox said the Quantum product line's switch density will enable space and power consumption optimization, reducing network equipment cost by 4X, electricity costs by 2X and improving data transfer time by 2X.

Blazingly fast network

Outline

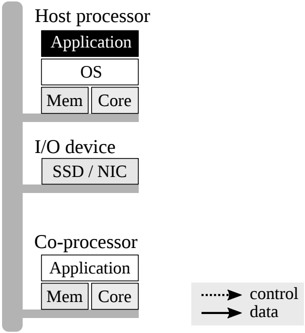
1. Heterogeneous Computing Architectures

2. Solros: Split-Kernel Approach
   - Solros Architecture
   - Operating System Services

3. Evaluation

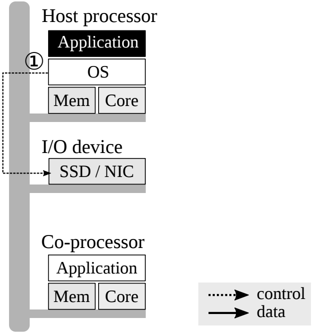
Host-Centric Approach

- Host OS controls co-processors and IO devices
- Examples: OpenCL, CUDA



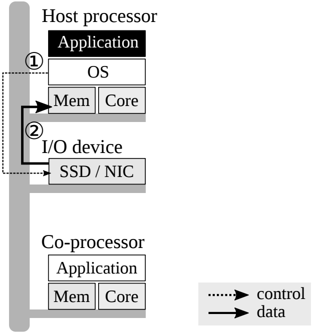
Host-Centric Approach

- Host OS controls co-processors and IO devices
- Examples: OpenCL, CUDA



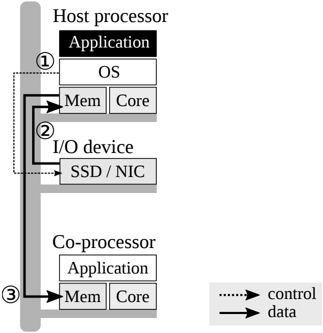
Host-Centric Approach

- Host OS controls co-processors and IO devices
- Examples: OpenCL, CUDA



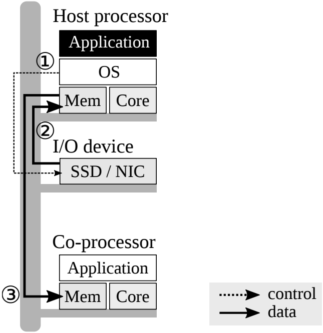
Host-Centric Approach

- Host OS controls co-processors and IO devices
- Examples: OpenCL, CUDA



Host-Centric Approach

- Host OS controls co-processors and IO devices
- Examples: OpenCL, CUDA



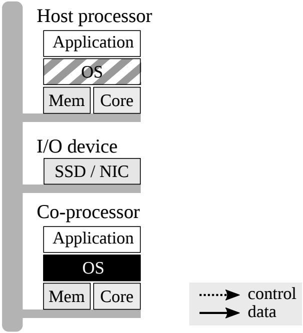
Problem

Redundant data communication

Complex to program and hard to optimize

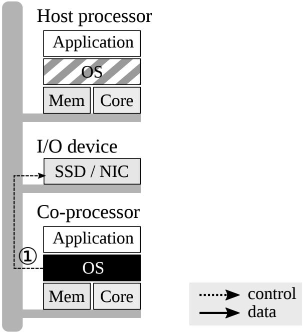
Coprocessor-Centric Architecture

- Co-processors control IO devices
- Examples: Xeon Phi (Linux), GPUfs [ASPLOS13], GPUNet [OSDI14]



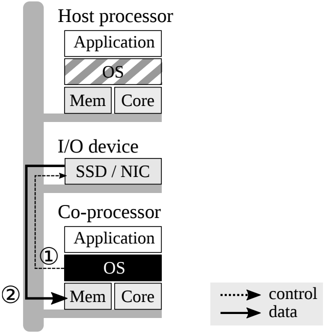
Coprocessor-Centric Architecture

- Co-processors control IO devices
- Examples: Xeon Phi (Linux), GPUfs [ASPLOS13], GPUNet [OSDI14]



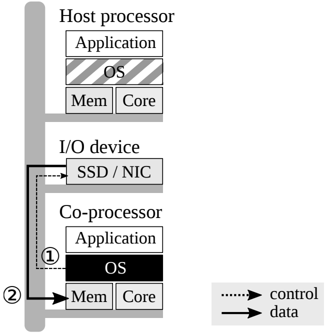
Coprocessor-Centric Architecture

- Co-processors control IO devices
- Examples: Xeon Phi (Linux), GPUfs [ASPLOS13], GPUNet [OSDI14]



Coprocessor-Centric Architecture

- Co-processors control IO devices
- Examples: Xeon Phi (Linux), GPUfs [ASPLOS13], GPUNet [OSDI14]



Problem

Significant effort required for porting IO stack to co-processor
Not completely exploiting powerful host processors

Outline

1. Heterogeneous Computing Architectures

2. Solros: Split-Kernel Approach
   - Solros Architecture
   - Operating System Services

3. Evaluation

Solros Goal

- Ease of programming
- Best use of processor architecture
- System-wide optimization

Solros Goal

- Ease of programming
- Best use of processor architecture
- System-wide optimization

Challenge

- Co-processor needs IO abstraction
- IO stacks is branch-divergent and difficult to parallelize
- It needs system-wide information

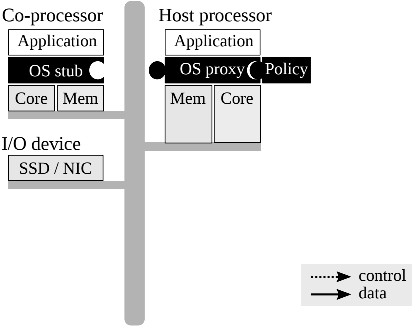
Solros Architecture

Split-Kernel Architecture

- Data-plane OS
  - Runs on a co-processor
  - Provides IO abstraction
  - Delegates actual IO operations to a control-plane OS
- Control-plane OS
  - Runs on a host processor
  - Runs actual IO stack
  - Performs system-wide coordination

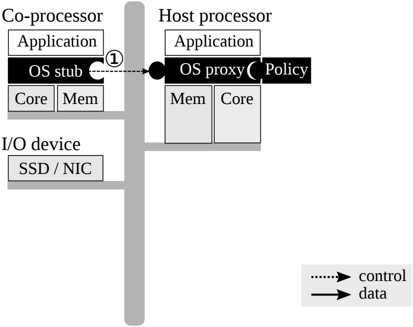
Solros Architecture

- Control-plane OS: actual OS service + system-wide coordination
- Data-plane OS: thin communication layer to host processor



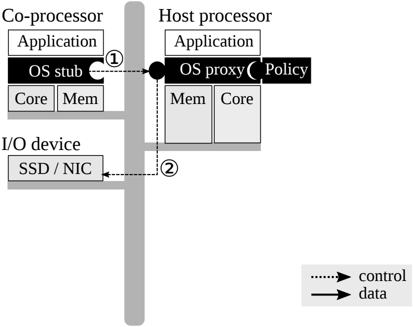
Solros Architecture

- Control-plane OS: actual OS service + system-wide coordination
- Data-plane OS: thin communication layer to host processor



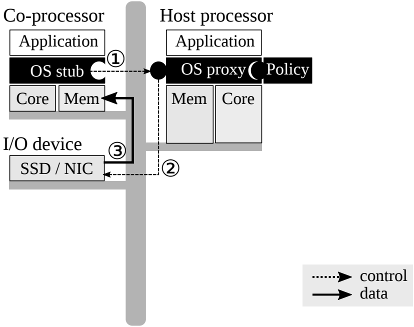
Solros Architecture

- Control-plane OS: actual OS service + system-wide coordination
- Data-plane OS: thin communication layer to host processor



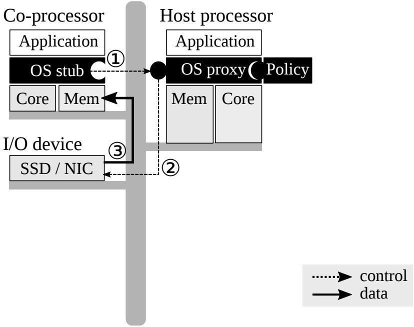
Solros Architecture

- Control-plane OS: actual OS service + system-wide coordination
- Data-plane OS: thin communication layer to host processor



Solros Architecture

- Control-plane OS: actual OS service + system-wide coordination
- Data-plane OS: thin communication layer to host processor



- Co-processor has OS abstraction with minimal effort
- Best use of each of the fat and lean processors
- Efficient global coordination among devices (policy)

Operating System Services

1. Transport service
2. Filesystem service
3. Network service

Operating System Services

1. Transport service
2. Filesystem service
3. Network service

Transport Service

High performance data transfer among devices are challenging:

- Uniform data transfer among devices
- High contention in massively-parallel co-processor
- Asymmetric performance between host processor and co-processor

Transport Service

High performance data transfer among devices are challenging:

- Uniform data transfer among devices
- High contention in massively-parallel co-processor
- Asymmetric performance between host processor and co-processor

Our approach

- Uniform data transfer ⇒ system-mapped PCIe window
- High contention ⇒ combining, replication, interleaving, etc.
- Asymmetric performance ⇒ flexibly configurable (host DMA engine vs. co-processor DMA engine)

Transport Service

High performance data transfer among devices are challenging:

- Uniform data transfer among devices
- High contention in massively-parallel co-processor
- Asymmetric performance between host processor and co-processor

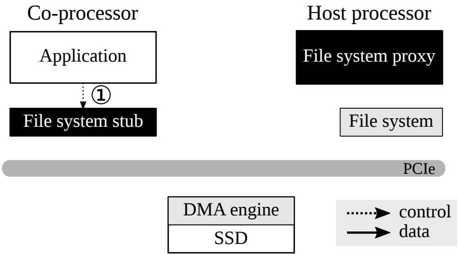
Our approach

- Uniform data transfer ⇒ system-mapped PCIe window
- High contention ⇒ combining, replication, interleaving, etc.
- Asymmetric performance ⇒ flexibly configurable (host DMA engine vs. co-processor DMA engine)

See details in the paper

Filesystem Service

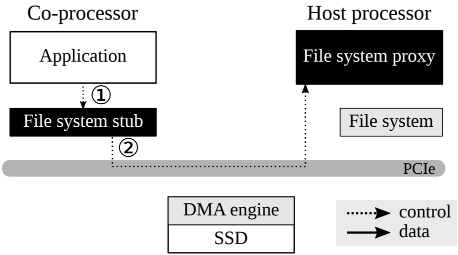
- **Peer-to-peer operation**
- Buffered operation



Zero-copy of data between co-processor memory and SSD
Minimal data transfer

Filesystem Service

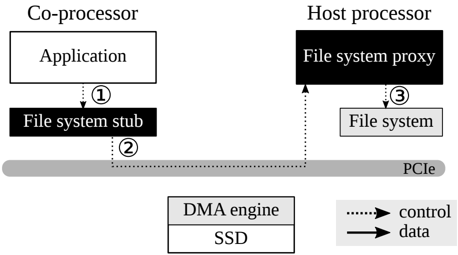
- **Peer-to-peer operation**
- Buffered operation



Zero-copy of data between co-processor memory and SSD
Minimal data transfer

Filesystem Service

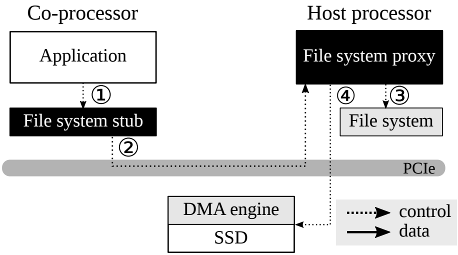
- **Peer-to-peer operation**
- Buffered operation



Zero-copy of data between co-processor memory and SSD
Minimal data transfer

Filesystem Service

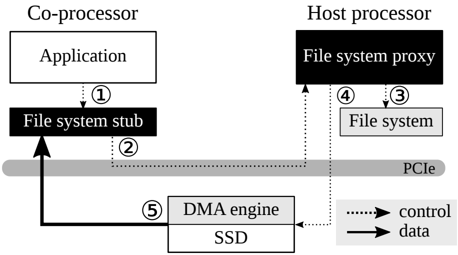
- **Peer-to-peer operation**
- Buffered operation



Zero-copy of data between co-processor memory and SSD
Minimal data transfer

Filesystem Service

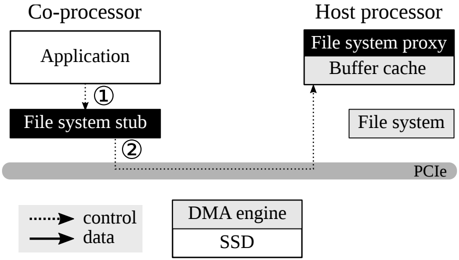
- **Peer-to-peer operation**
- Buffered operation



Zero-copy of data between co-processor memory and SSD
Minimal data transfer

Filesystem Service

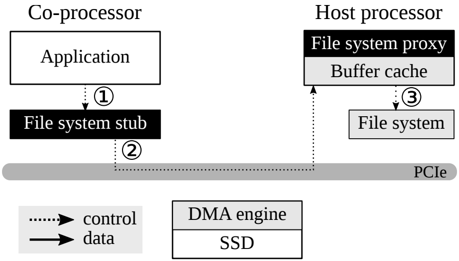
- Peer-to-peer operation
- **Buffered operation**



Reduce storage IO by leveraging shared buffer cache among co-processors
Avoid performance anomaly of peer-to-peer communication over PCIe bus

Filesystem Service

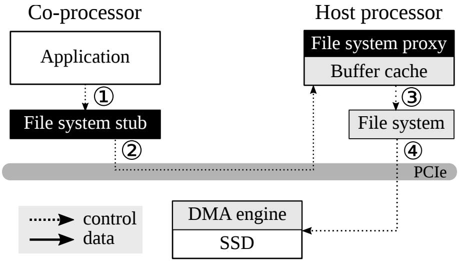
- Peer-to-peer operation
- **Buffered operation**



Reduce storage IO by leveraging shared buffer cache among co-processors
Avoid performance anomaly of peer-to-peer communication over PCIe bus

Filesystem Service

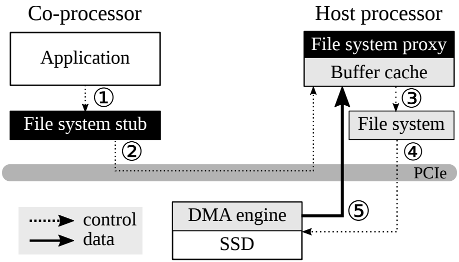
- Peer-to-peer operation
- **Buffered operation**



Reduce storage IO by leveraging shared buffer cache among co-processors
Avoid performance anomaly of peer-to-peer communication over PCIe bus

Filesystem Service

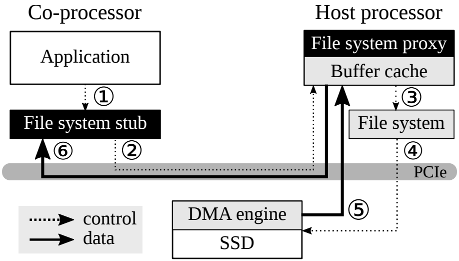
- Peer-to-peer operation
- **Buffered operation**



Reduce storage IO by leveraging shared buffer cache among co-processors
Avoid performance anomaly of peer-to-peer communication over PCIe bus

Filesystem Service

- Peer-to-peer operation
- **Buffered operation**



Reduce storage IO by leveraging shared buffer cache among co-processors
Avoid performance anomaly of peer-to-peer communication over PCIe bus

Implementation

- Host: 2-socket Xeon processor (12 cores each)
- Co-processor: 4 Xeon Phi (KNC, 61 cores, Linux, PCIe Gen 3x16)
- Storage device: 4 NVMe SSD
- NIC: 100 Gbps Ethernet

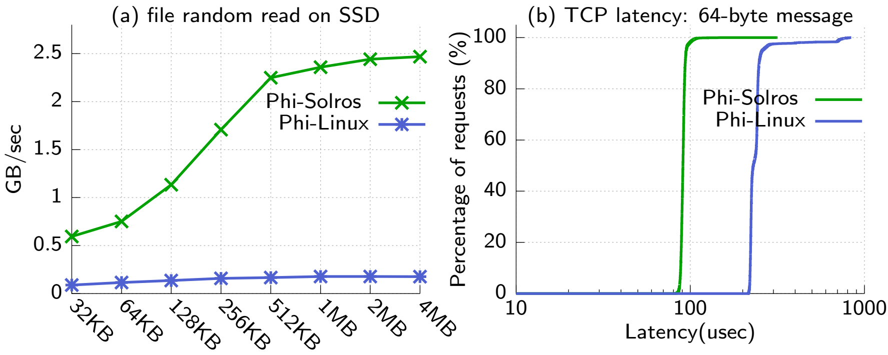
Module		**Lines of code**	
		Added lines	**Deleted lines**
Transport service		1,035	365
File system Service	Stub	5,957	2,073
	Proxy	2,338	124
Network Service	Stub	2,921	79
	Proxy	5,609	34
NVMe device driver		924	25
SCIF kernel module		60	14
Total		**18,844**	**2,714**

Evaluation

Questions:

- Performance of Solros services
- Impact on real-world applications

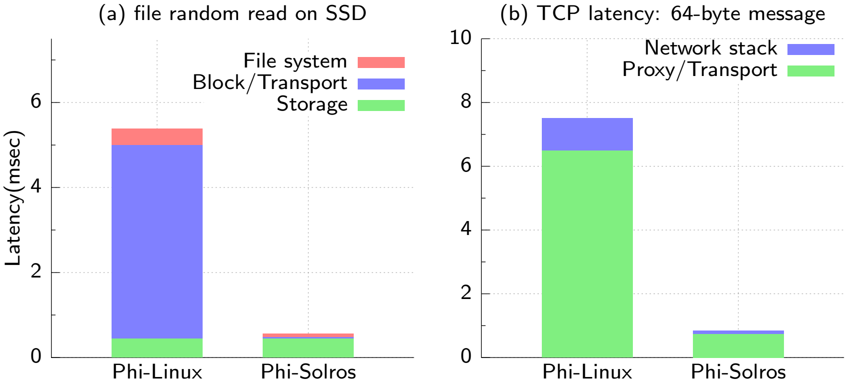
Performance of Solros Services



File IO performance: 19x faster than the stock Linux on Xeon Phi

TCP latency (99 percentile): 7x shorter than the stock Linux on Xeon Phi

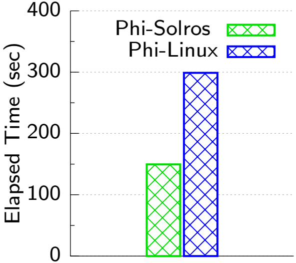
Performance of Solros Services

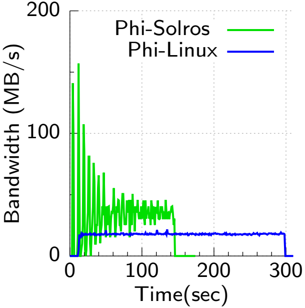


Significant performance gain in data transport
Running IO stack on co-processor is slower

Real-world Application - Image Search

- Image search engine is running on Xeon Phi
- Image database is on NVMe SSD (shared read-only)
- Image search queries are from network





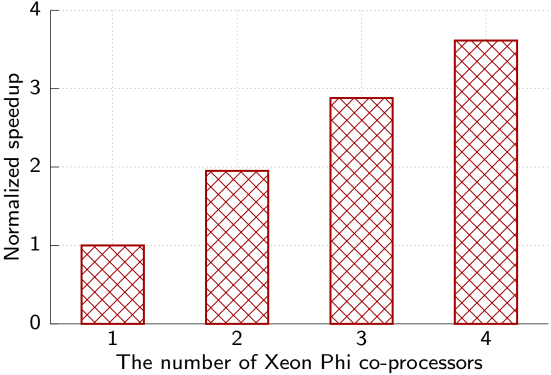
Solros performs 2x faster than stock Linux on Xeon Phi

Conclusion

- Solros, a new operating system architecture for co-processors and fast IO devices
- Control-plane, data-plane architecture allow:
  - Supporting high-level OS abstraction on co-processor
  - Efficient global coordination among devices
  - Near ideal IO performance from co-processor
- We will release source code soon

Image Search - Scalability

- Increase the number of Xeon Phi to 4

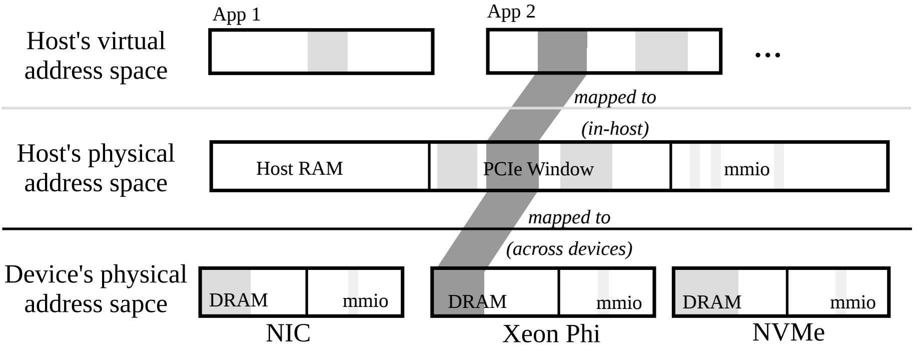


Solros load balancing mechanism achieves near linear scaling

Related work

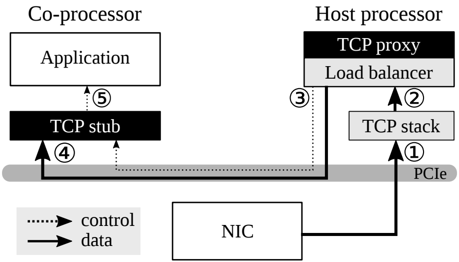
- Control-plane/data-plane OS: Arrakis [OSDI'14], IX [OSDI'14]
- OS for heterogeneous systems: Helios [SOSP]09], M3 [ASPLOS'16], Hydra [ASPLOS'08]
- IO support for GPU: PTask [SOSP'11], GPUfs [ASPLOS'13], GPUnet [OSDI'14]

Transport service



Network Service (TCP)

- Outbound operation
- **Inbound operation**



A load balancer on a host distributes incoming TCP connections to one of least-loaded co-processors.

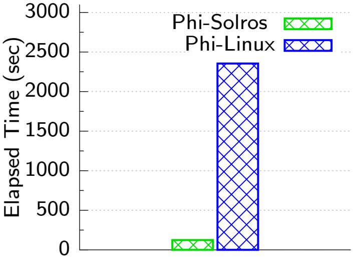
See details in the paper.

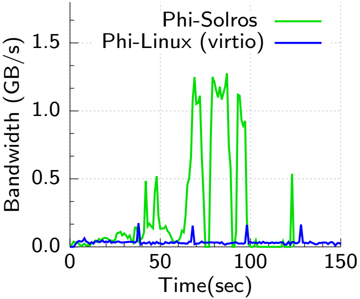
Discussion

- Hardware support other than Xeon Phi
  - Two atomic instructions: transport service
  - MMU: isolation among co-processor applications
- Scalability of control-plane OS
  - Limited by scalability of OS service, PCIe interconnect, and performance of IO devices

Real-world Application - Text Search

- CLucene text indexing engine running on Xeon Phi
- Text data is on NVMe SSD





Solros performs 19x faster than stock Linux (ext4/virtio) on Xeon Phi