An in-depth evaluation of the Intel Omni-Path network for LQCD applications

Peter Georg, Daniel Richtmann, and Tilo Wettig

Department of Physics, University of Regensburg

June 19, 2017

Overview: InfiniBand vs. Omni-Path

InfiniBand	Omni-Path
Connection-oriented (RC)	Connection-less (RDM)
Connection-less (UD)	

Overview: InfiniBand vs. Omni-Path

InfiniBand	Omni-Path
Connection-oriented (RC)	Connection-less (RDM)
Best for regular NN exchanges	Designed for MPI
Fabric Collective Accelerator (FCA)	Scalable All-To-All communication

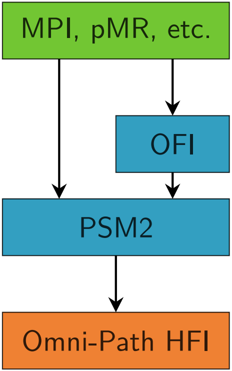
Overview: InfiniBand vs. Omni-Path

InfiniBand	Omni-Path
Connection-oriented (RC)	Connection-less (RDM)
Best for regular NN exchanges	Designed for MPI
Fabric Collective Accelerator (FCA)	Scalable All-To-All communication
Offloading	Onloading
Network co-processor	Network functions executed by CPU
Overlapping comm. & comp.	Shared resources

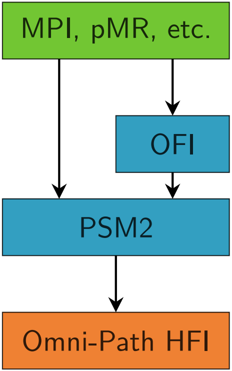
Overview: InfiniBand vs. Omni-Path

InfiniBand	Omni-Path
Connection-oriented (RC)	Connection-less (RDM)
Best for regular NN exchanges	Designed for MPI
Fabric Collective Accelerator (FCA)	Scalable All-To-All communication
Offloading	Onloading
Network co-processor	Network functions executed by CPU
Overlapping comm. & comp.	Shared resources
Reliable mature software stack	Quality of software stack?

Omni-Path Software



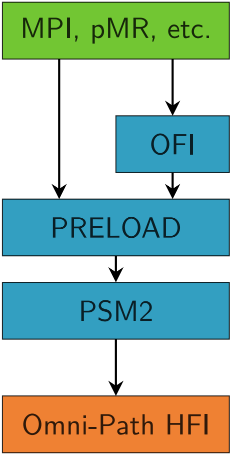
Omni-Path Software



Performance Scaled Messaging (PSM)

- Limited to one endpoint per process
- Can lookup already opened endpoint
- Access to endpoint needs to be coordinated
- Limited to one upper layer using PSM2 directly

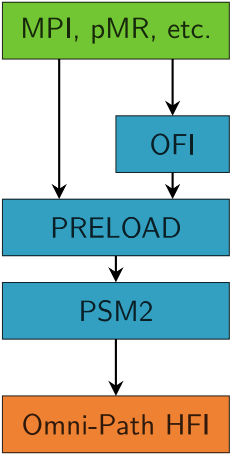
Omni-Path Software



Performance Scaled Messaging (PSM)

- Limited to one endpoint per process
- Can lookup already opened endpoint
- Access to endpoint needs to be coordinated
- Limited to one upper layer using PSM2 directly

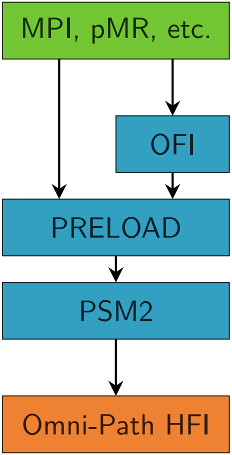
Omni-Path Software



Performance Scaled Messaging (PSM)

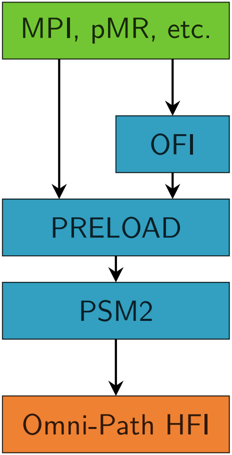
- Limited to one endpoint per process
- Can lookup already opened endpoint
- Access to endpoint needs to be coordinated
- Limited to one upper layer using PSM2 directly
- Supports three providers: self, shm, hfi
- Supports PIO and DMA send methods
- Supports TID and eager receive methods

Omni-Path Software



Low-level counterpart to InfiniBand Verbs

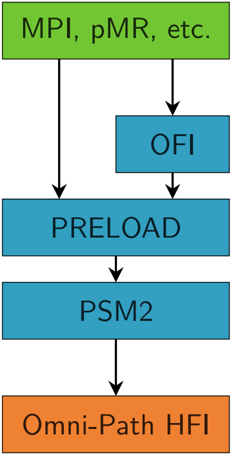
Omni-Path Software



Low-level counterpart to InfiniBand Verbs

- Does not exist
- Is not required

Omni-Path Software



Low-level counterpart to InfiniBand Verbs

- Does not exist
- Is not required

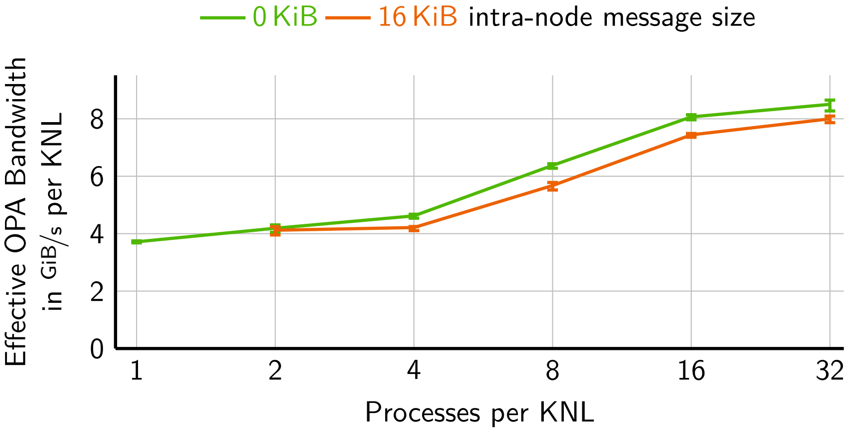
Everything done in software anyway

What does that mean for me as a software developer?

What does that mean for me as a software developer?

> *While there are future plans to extend PSM2 to support multi-threaded applications, PSM2 is currently a single-threaded library. This means that the user cannot make any concurrent PSM2 library calls. While threads may be a valid execution model for the wider set of potential PSM2 clients, applications should currently expect better effective use of OPA resources (and hence better performance) by dedicating a single PSM2 communication endpoint to every CPU core.*
>
> — Intel, *psm2.h*

Omni-Path Bandwidth vs. Processes per KNL (16 KNLs)

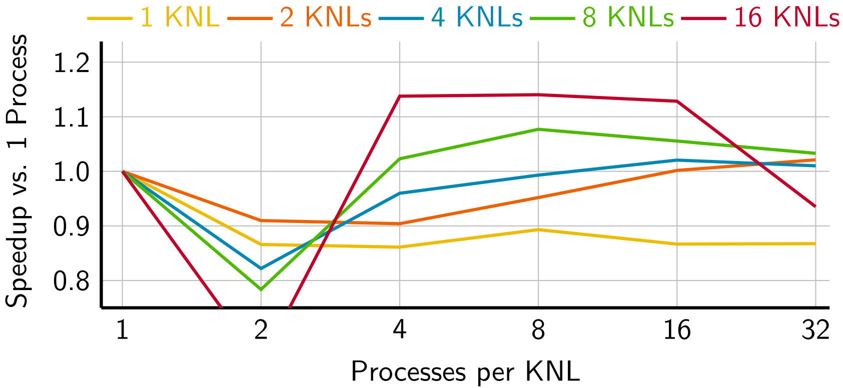


- Inter-node message size: 512 KiB / number of processes per KNL
- Bandwidth increases with number of processes
- Threaded communication can further reduce intra-node overhead

Solver Benchmarks

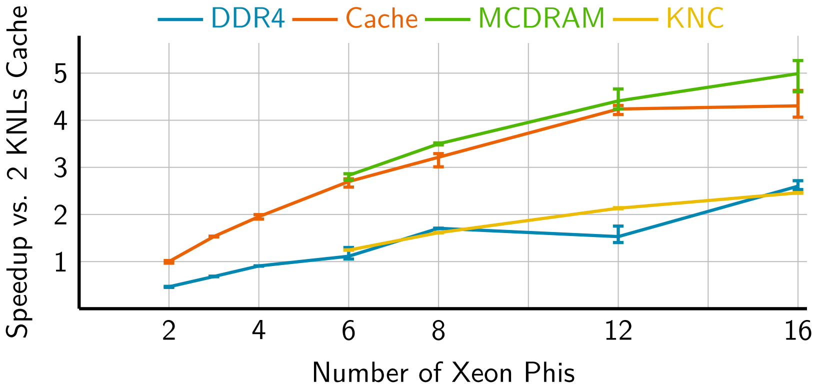
Cluster	QPACE 2	QPACE 3
Solver	DD-αAMG (two-level multigrid)	
Lattice size	$32\times32\times32\times96$ ($>$1 node) $16\times16\times16\times32$ ($=$1 node)	
Lattice splitting	Tuned	
CPU	Xeon Phi 7120	Xeon Phi 7210
Memory	16 GB GDDR5	16 GB MCDRAM 48 GB DDR4
Fabric	Connect IB	Omni-Path
Compiler	Intel 2015 U3	Intel 2017 U2
MPI	Intel MPI 5.1 U3 (DAPL)	Intel MPI 2017 U2 (OFI)
OS	CentOS 7.1	CentOS 7.3
Fabric Suite	OFED 3.18-2	IFS 10.3.1
CPU stack	Intel MPSS 3.7.1	Intel XPPSL 1.5.0

Solver runtime vs. Processes per KNL



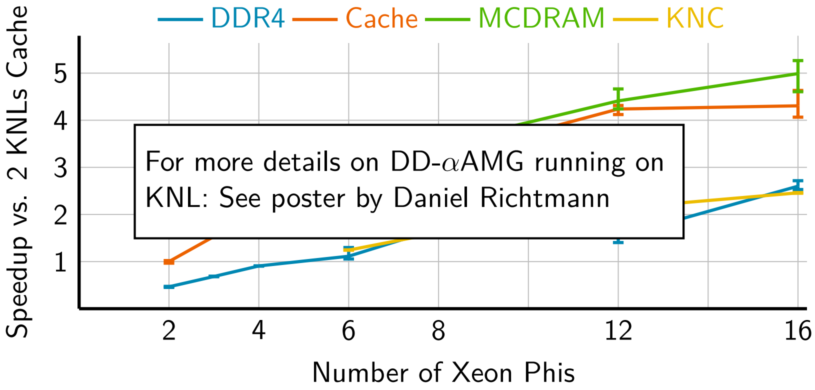
- Single-node performance decreases with number of processes
- Multi-node performance increases with number of processes

Off-chip strong scaling



- DDR4 with 100 Gbit/s OPA ≈ KNC with 28 Gbit/s IB FDR
- Cache does not scale beyond 12 KNLs
- Using MCDRAM we see expected speedup (2.2x) vs. KNC

Off-chip strong scaling



For more details on DD-αAMG running on KNL: See poster by Daniel Richtmann

- DDR4 with 100 $^{\text{Gbit}}/_{\text{s}}$ OPA ≈ KNC with 28 $^{\text{Gbit}}/_{\text{s}}$ IB FDR
- Cache does not scale beyond 12 KNLs
- Using MCDRAM we see expected speedup (2.2x) vs. KNC

Conclusion

KNL and Omni-Path prefer different programming models

- Multi-threading vs. multi-processing
- There is still hope as it is (mainly) a software problem
- Intel supposedly working on major software update
- Issue can be alleviated by software modifications
- Omni-Path software stack is **WIP**

Omni-Path does not solve our scalability problems

- Bandwidth is not the main bottleneck
- Latency and message rate are key performance indicators
- A second HFI most likely won't help much (yet)
- Main problem is getting data from CPU to HFI (via PCIe)
  - KNL-F does not solve this issue

Take-home message

Tune your applications

- Tune number of processes per KNL
- Tune lattice splitting, i.e., geom for Chroma
- Consider threaded communication (not using MPI)

Take-home message

Tune your applications

- Tune number of processes per KNL
- Tune lattice splitting, i.e., geom for Chroma
- Consider threaded communication (not using MPI)

Hope everything will be better in 2018...