Simulation using MIC co-processor on Helios

Serhiy Mochalskyy, Roman Hatzky

PRACE PATC Course: Intel MIC Programming Workshop

High Level Support Team
Max-Planck-Institut für Plasmaphysik
Boltzmannstr. 2, D-85748 Garching, Germany

Outline

- MIC general architecture
- MIC network performance on the Robin cluster with one IB port
- MIC network performance on the Helios supercomputer with two IB ports
- Host, offload and native computation mode of the test N-Body code
- Micro OpenMP overhead benchmark

Outline

- MIC general architecture
- MIC network performance on the Robin cluster with one IB port
- MIC network performance on the Helios supercomputer with two IB ports
- Host, offload and native computation mode of the test N-Body code
- Micro OpenMP overhead benchmark

Helios cluster

Helios is a computer system dedicated to large-scale and high performance simulations in fusion science and engineering research.

CPU	**Intel Xeon E5 processor, Sandy-Bridge EP 2.7GHz**
Nodes	**4410**
Peak performance	**1.52 Pflops (70th in top 500, June 2016)**

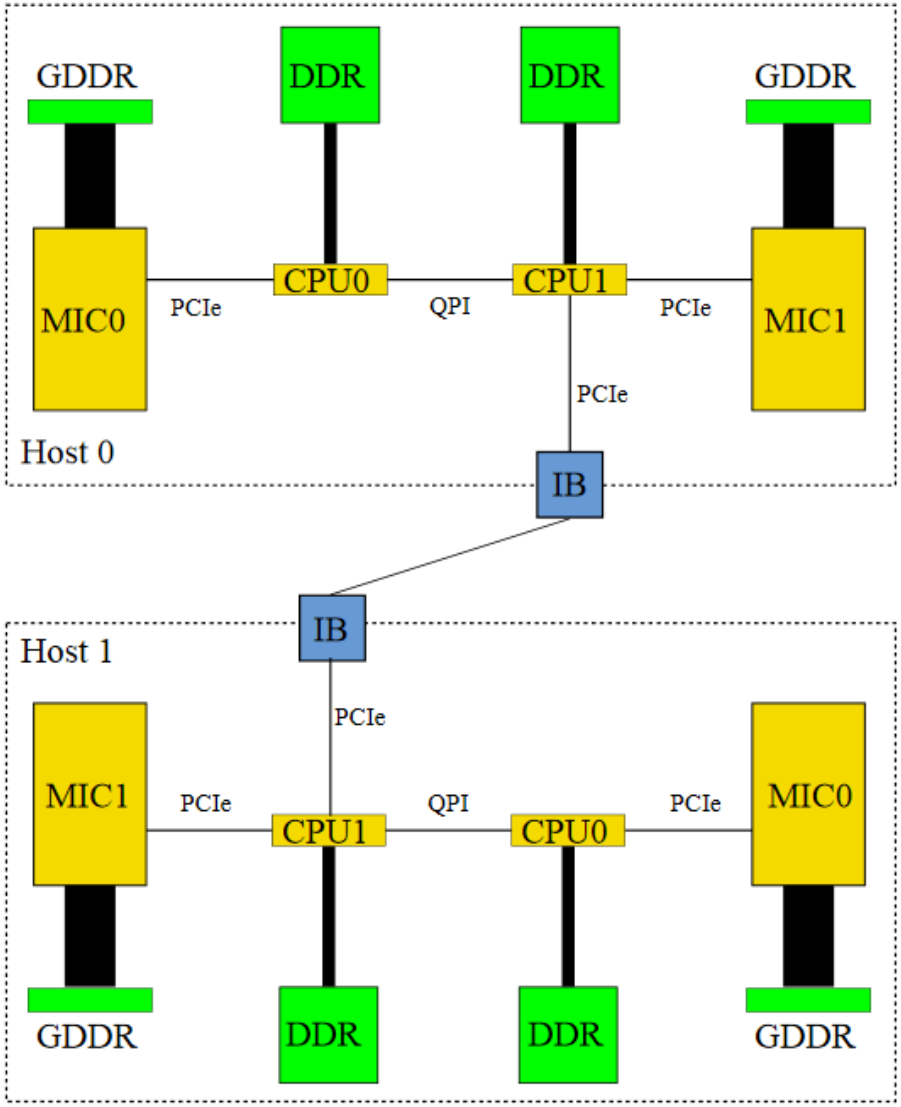
MIC	**Knights Corner**
Nodes	**180**

Intel Xeon Phi vs Intel Sandy Bridge

Processor	Sandy Bridge	Xeon Phi
Number of cores	8	60(1)
Memory	32 GB	8–16 GB
Peak performance	173 GFlops/s	1 TFlops/s
Memory bandwidth	40 GB/s	160 GB/s
Instruction execution	Out-of-order	In-order

- **~x4** increase in memory bandwidth
- **~x6** increase in peak performance
- **~x30** and **~x1.3** decrease in memory and performance per core
- In-order-execution requires 2–4 threads per core to fill the pipelines

MIC nodes – general architecture

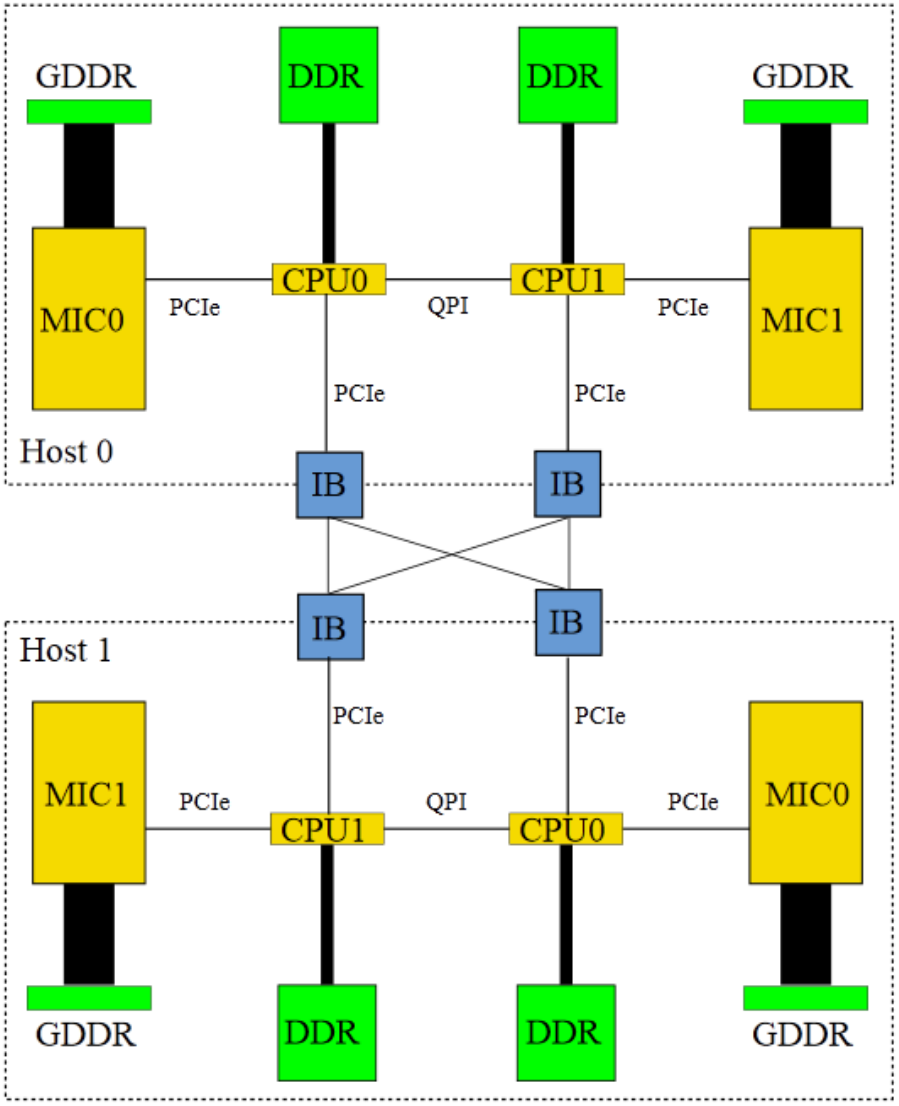


1 or 2 InfiniBand ports

Hydra and Robin → 1 IB port

Helios and Supermic → **2 IB** ports

MIC nodes – general architecture



1 or 2 InfiniBand ports

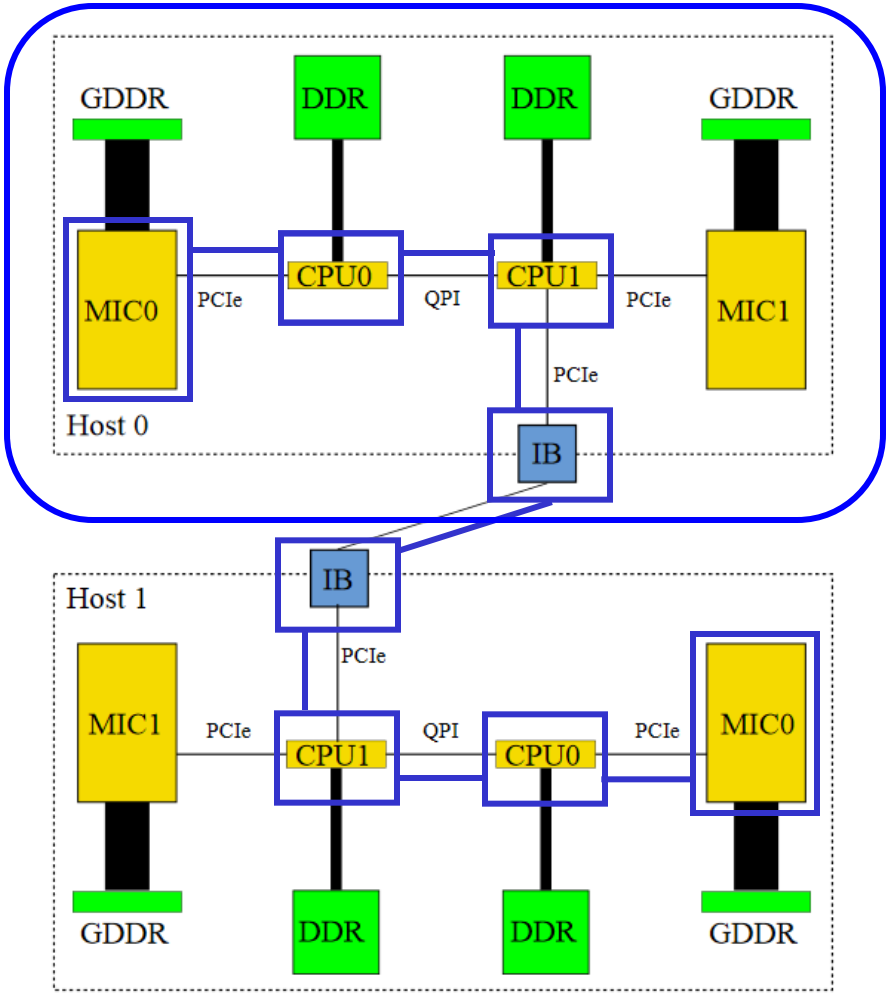
Hydra and Robin → **1 IB** port

Helios and Supermic → 2 IB ports

Outline

- MIC general architecture
- MIC network performance on the Robin cluster with one IB port
- MIC network performance on the Helios supercomputer with two IB ports
- Host, offload and native computation mode of the test N-Body code
- Micro OpenMP overhead benchmark

MIC network performance on the Robin cluster



PCIe+QPI+PCIe+IB+PCIe+QPI+PCIe

Intel MPI Benchmark suite: Ping-Pong test

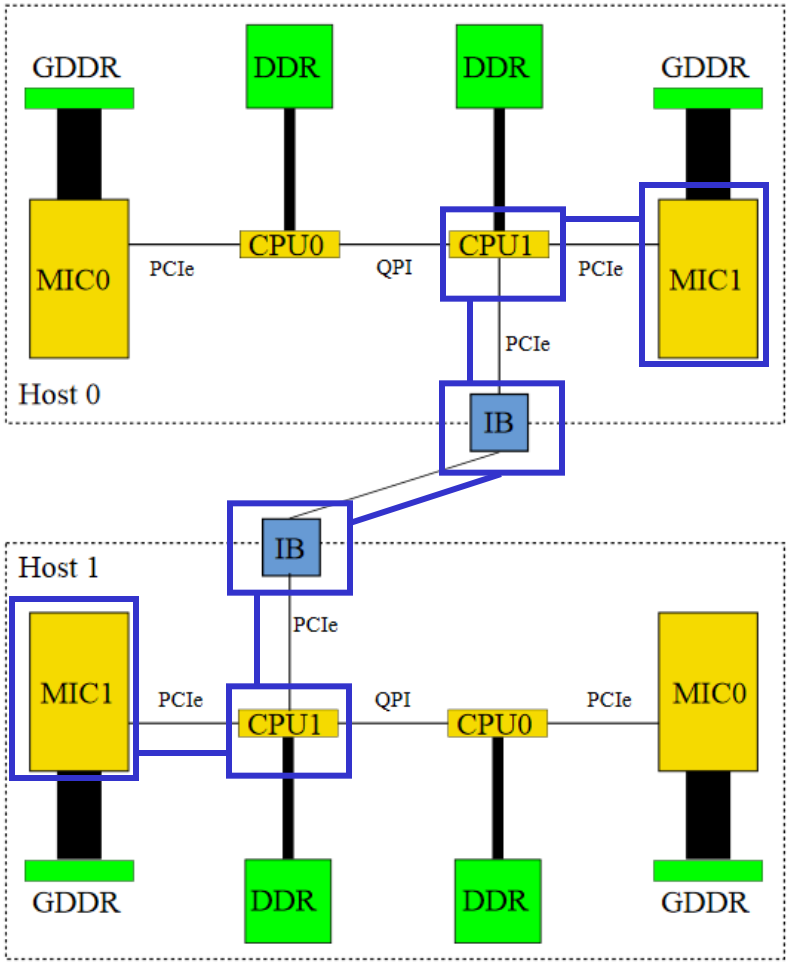
Intra-node

	CPU1			
host0	MIC0	4.90	2.73	
	MIC1	4.31	7.56	3.12
Latency (µs)		CPU1	MIC0	MIC1
		host0		

Inter-node

	CPU1	2.20		
host0	MIC0	4.71	9.04	
	MIC1	4.66	7.93	6.92
Latency (µs)		CPU1	MIC0	MIC1
		host1		

MIC network performance on the Robin cluster



PCIe+ +PCIe+IB+PCIe+ +PCIe

Intel MPI Benchmark suit: Ping-Pong test

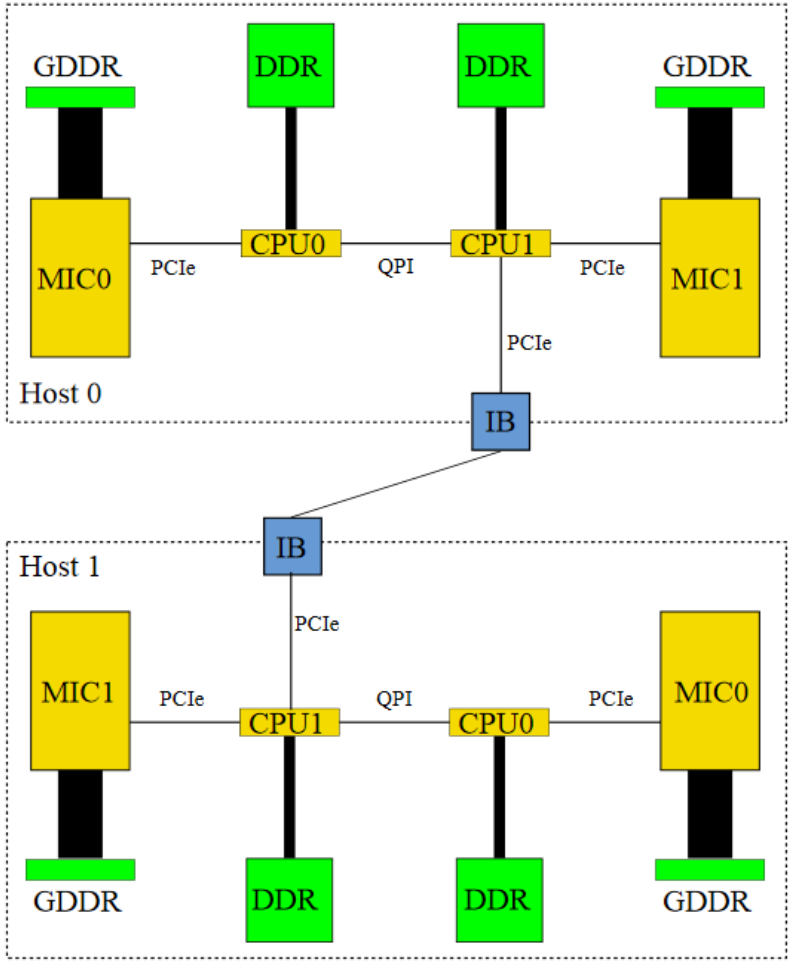
Intra-node

host0	**CPU1**			
	MIC0	**4.90**	**2.73**	
	MIC1	**4.31**	**7.56**	**3.12**
Latency (µs)		**CPU1**	**MIC0**	**MIC1**
		host0		

Inter-node

host0	**CPU1**	**2.20**		
	MIC0	**4.71**	**9.04**	
	MIC1	**4.66**	**7.93**	**6.92**
Latency (µs)		**CPU1**	**MIC0**	**MIC1**
		host1		

MIC network performance on the Robin cluster



Intra-node

host0	CPU1			
	MIC0	456	2016	
	MIC1	1609	416	2004
Bandwidth (MB/s)		CPU1	MIC0	MIC1
		host0		

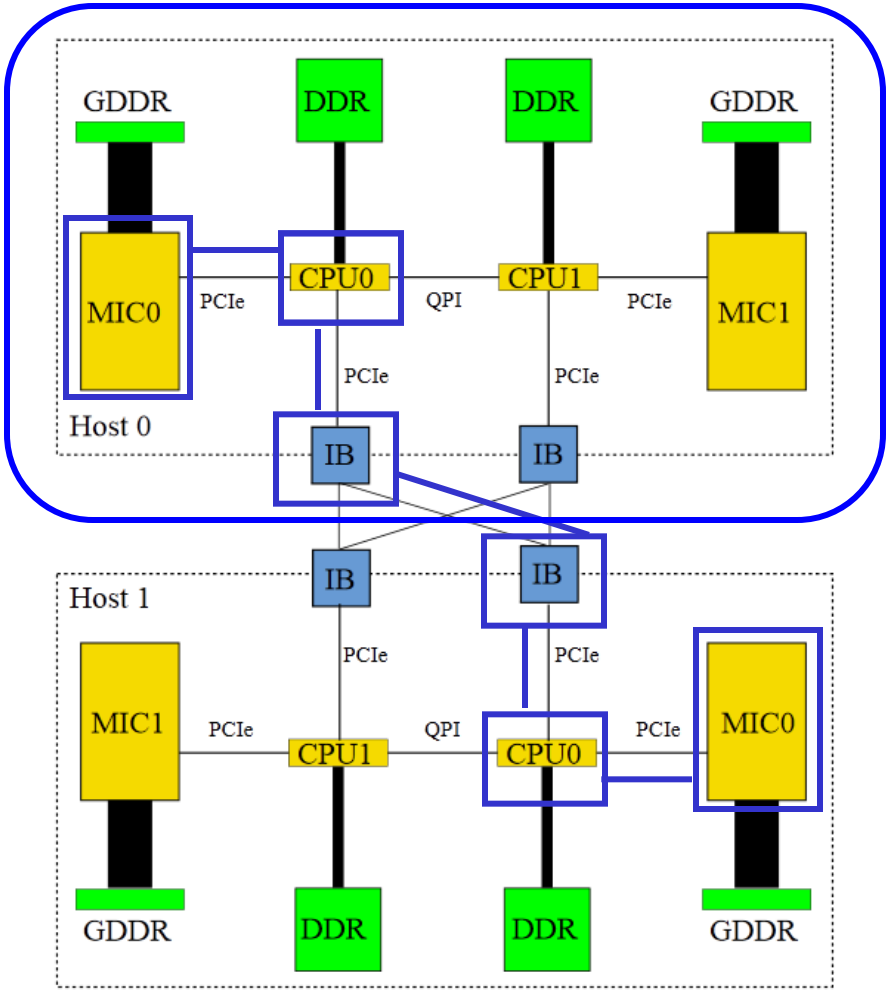
Inter-node

host0	CPU1	5729		
	MIC0	418	273	
	MIC1	1608	418	969
Bandwidth (MB/s)		CPU1	MIC0	MIC1
		host1		

Outline

- MIC general architecture
- MIC network performance on the Robin cluster with one IB port
- MIC network performance on the Helios supercomputer with two IB ports
- Host, offload and native computation mode of the test N-Body code
- Micro OpenMP overhead benchmark

MIC network performance on the Helios supercomputer



PCIe+ +PCIe+IB+PCIe+ +PCIe

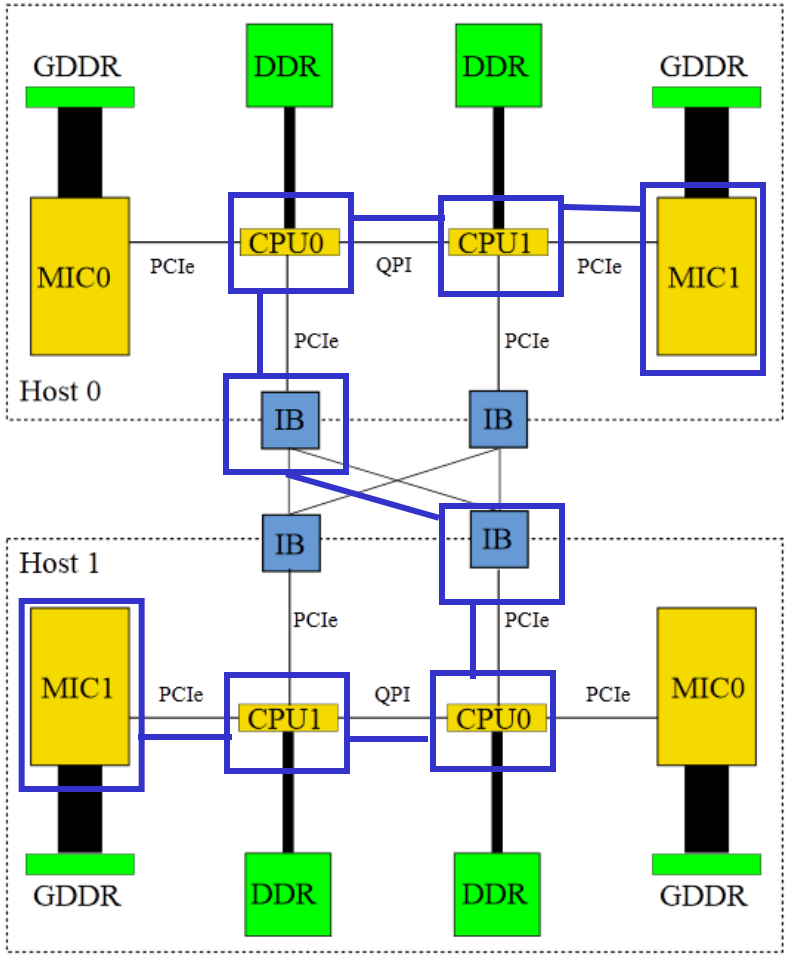
Intra-node

host0	CPU1	0.31		
	MIC0	3.29	2.70	
	MIC1	3.75	6.00	2.84
Latency (µs)		CPU1	MIC0	MIC1
		host0		

Inter-node

host0	CPU1	1.24		
	MIC0	3.80	5.97	
	MIC1	4.15	6.47	6.95
Latency (µs)		CPU1	MIC0	MIC1
		host1		

MIC network performance on the Helios supercomputer



PCIe+QPI+PCIe+IB+PCIe+QPI+PCIe

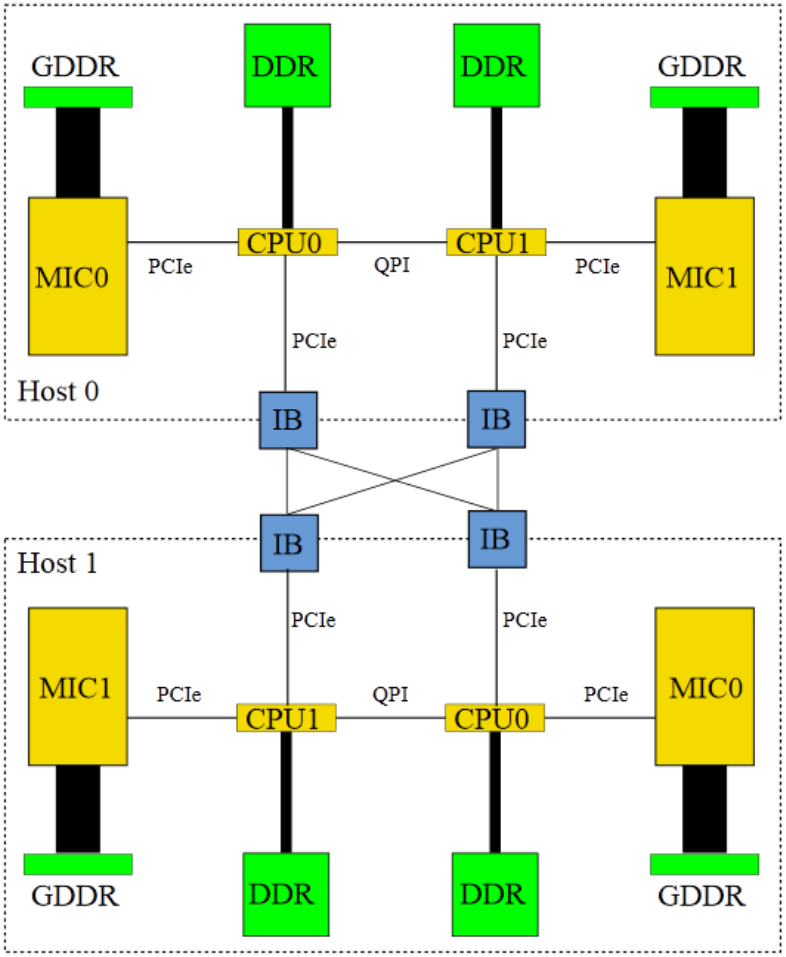
Intra-node

host0	CPU1	0.31		
	MIC0	3.29	2.70	
	MIC1	3.75	6.00	2.84
Latency (μs)		CPU1	MIC0	MIC1
		host0		

Inter-node

host0	CPU1	1.24		
	MIC0	3.80	5.97	
	MIC1	4.15	6.47	6.95
Latency (μs)		CPU1	MIC0	MIC1
		host1		

MIC network performance on the Helios supercomputer



Intra-node

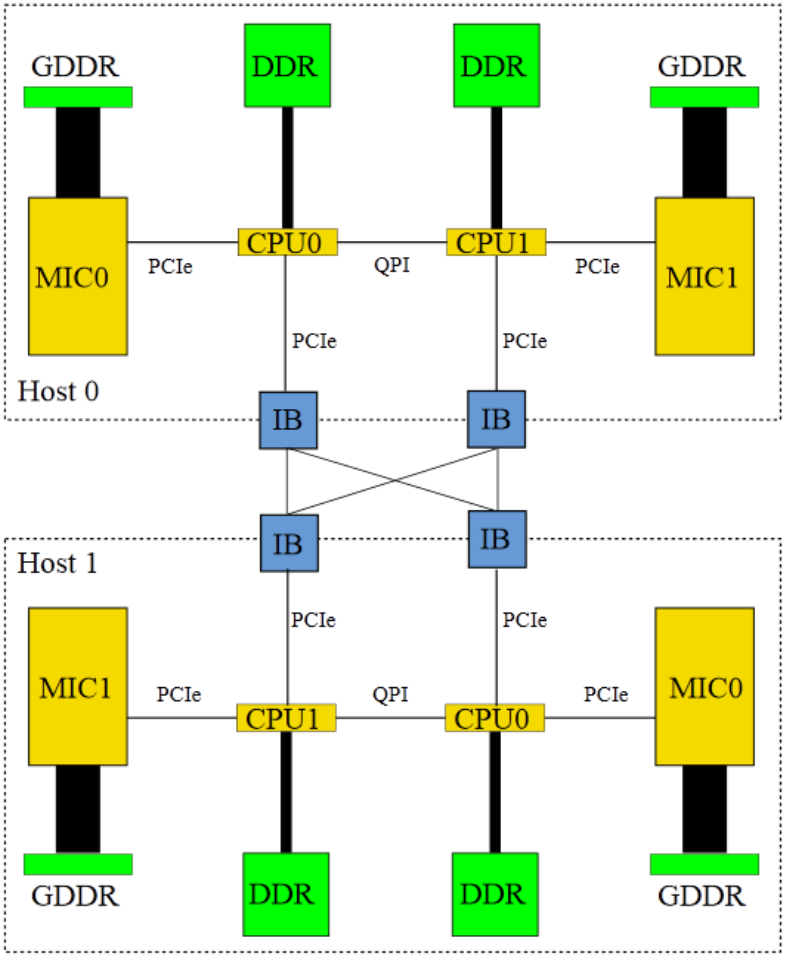
Bandwidth (MB/s)				
		CPU1	MIC0	MIC1
host0	CPU1	5061		
	MIC0	1611	1928	
	MIC1	480	413	1984

Inter-node

Bandwidth (MB/s)				
		CPU1	MIC0	MIC1
host0	CPU1	5943		
	MIC0	1640	984	
	MIC1	416	415	272

MIC network performance on Helios – optimized DAPL (Direct Acess Programming Library) provider

$ export I_MPI_DAPL_PROVIDER_LIST=ofa-v2-mlx4_0-1u,ofa-v2-mcm-1



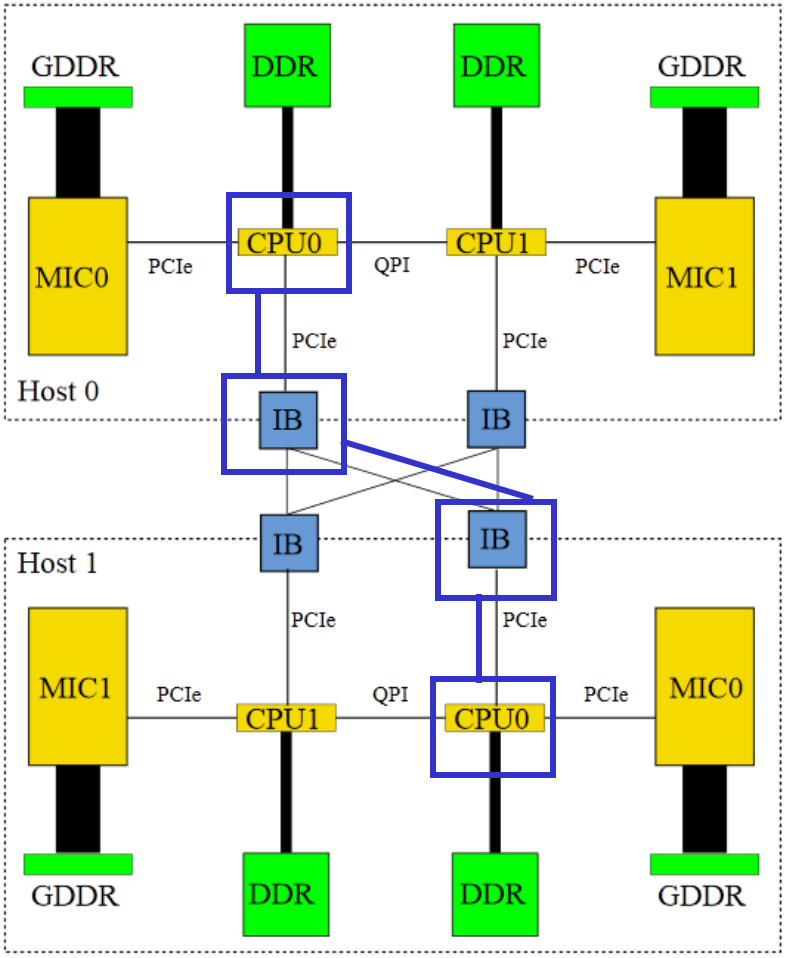
Inter-node

host0	CPU1	5943		
	MIC0	1640	984	
	MIC1	416	415	272
Bandwidth (MB/s)		CPU1	MIC0	MIC1
		host1		

Inter-node optimized DAPL

host0	CPU1	5836		
	MIC0	4135	3447	
	MIC1	1780	2235	1349
Bandwidth (MB/s)		CPU1	MIC0	MIC1
		host1		

MIC network performance on the Helios supercomputer – CPUs



PCIe+IB+PCIe

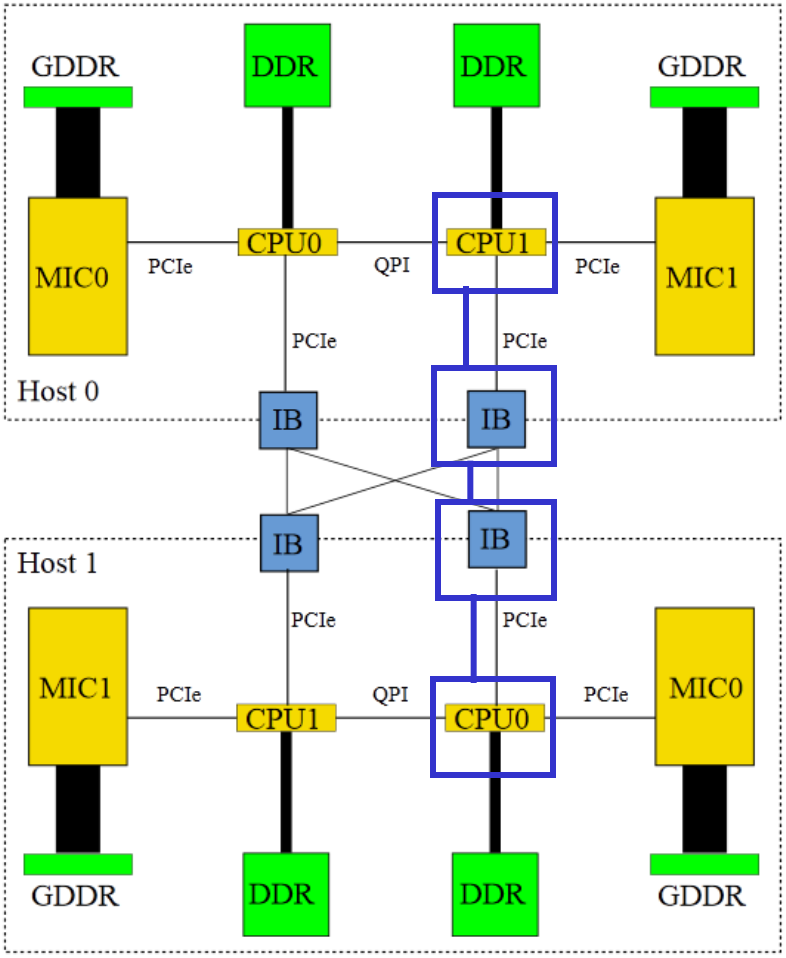
Inter-node

host0	CPU0	1.27	1.23
	CPU1	1.28	1.25
Latency (μs)		CPU0	CPU1
		host1	

Inter-node

host0	CPU0	4987	5029
	CPU1	5075	5058
Bandwidth (MB/s)		CPU0	CPU1
		host1	

MIC network performance on the Helios supercomputer – CPUs



PCIe+IB+PCIe

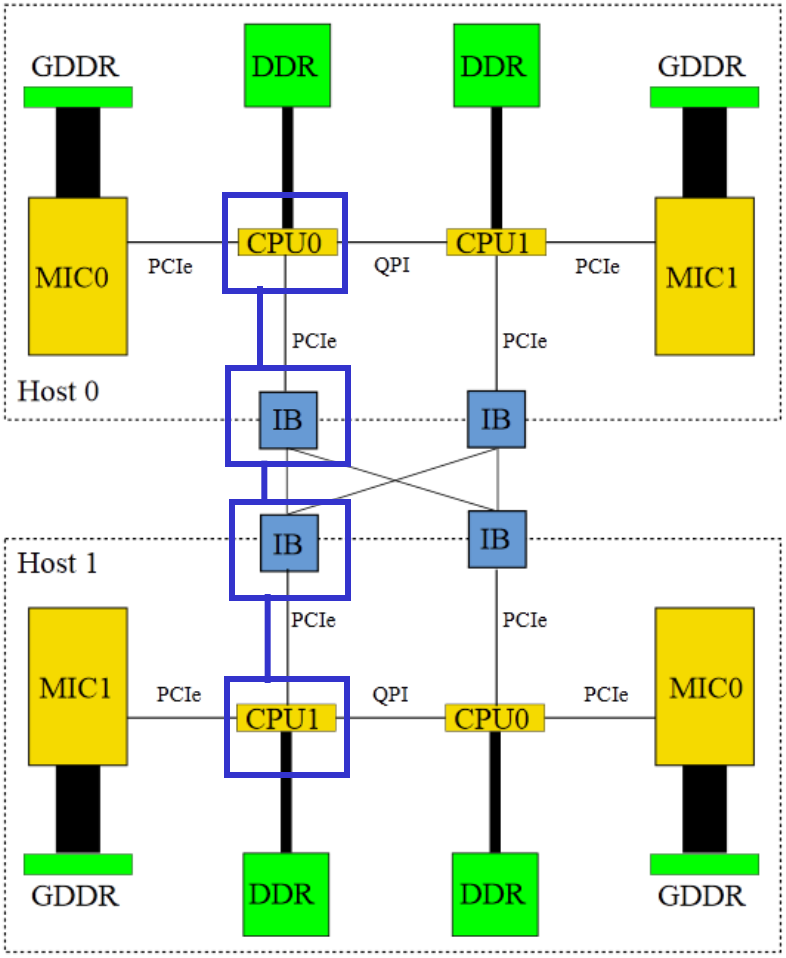
Inter-node

host0	CPU0	1.27	1.23
	CPU1	1.28	1.25
Latency (μs)		CPU0	CPU1
		host1	

Inter-node

host0	CPU0	4987	5029
	CPU1	5075	5058
Bandwidth (MB/s)		CPU0	CPU1
		host1	

MIC network performance on the Helios supercomputer – CPUs



PCIe+IB+PCIe

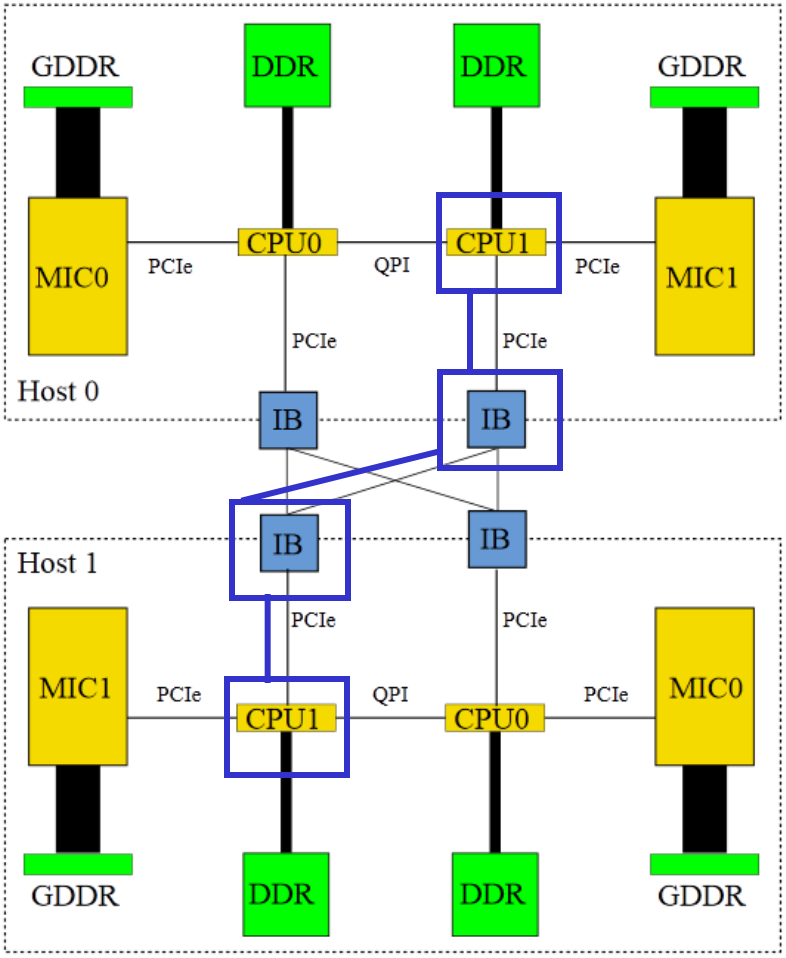
Inter-node

host0	CPU0	1.27	1.23
	CPU1	1.28	1.25
Latency (μs)		CPU0	CPU1
		host1	

Inter-node

host0	CPU0	4987	5029
	CPU1	5075	5058
Bandwidth (MB/s)		CPU0	CPU1
		host1	

MIC network performance on the Helios supercomputer – CPUs



PCIe+IB+PCIe

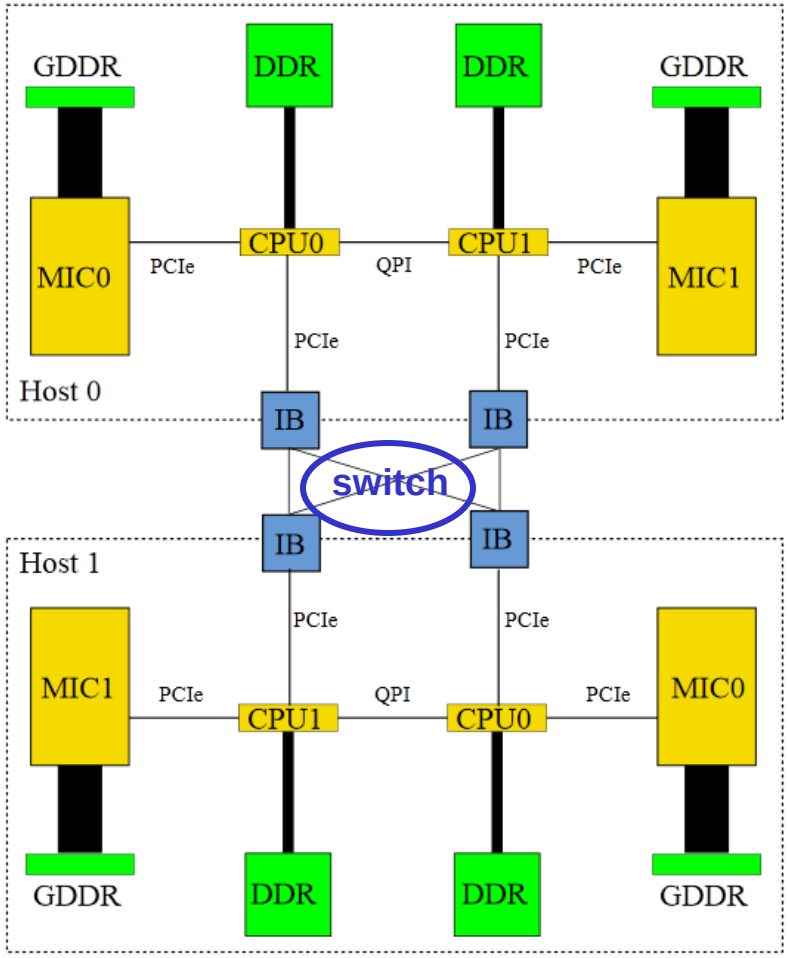
Inter-node

host0	CPU0	1.27	1.23
	CPU1	1.28	1.25
Latency (μs)		CPU0	CPU1
		host1	

Inter-node

host0	CPU0	4987	5029
	CPU1	5075	5058
Bandwidth (MB/s)		CPU0	CPU1
		host1	

MIC network performance on the Helios supercomputer – new DAPL provider in dat.conf



dat – direct access transport

/etc/dat.conf for mic0

ofa-v2-mcm-1 u2.0 nonthreadsafe default libdaplomcm.so.2 dapl.2.0 "mlx4_0 1" ""
ofa-v2-mlx4_0-1u u2.0 nonthreadsafe default libdaploucm.so.2 dapl.2.0 "mlx4_0 1" ""

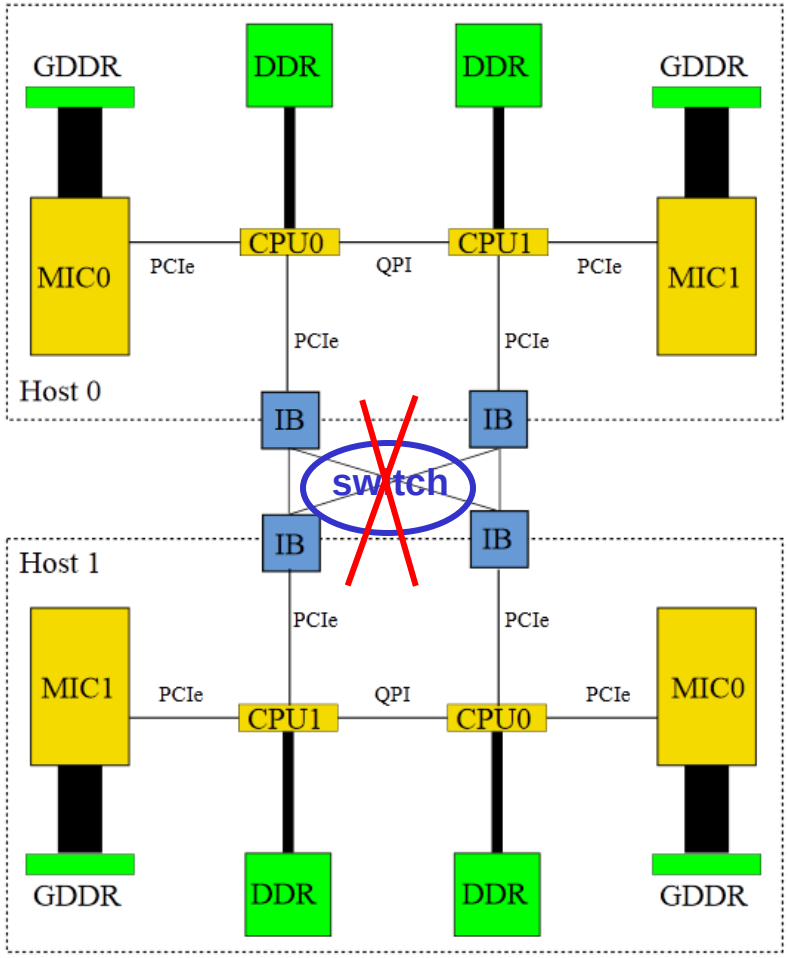
/etc/dat.conf for mic1

ofa-v2-mcm-1 u2.0 nonthreadsafe default libdaplomcm.so.2 dapl.2.0 "mlx4_1 1" ""
ofa-v2-mlx4_1-1u u2.0 nonthreadsafe default libdaploucm.so.2 dapl.2.0 "mlx4_1 1" ""

Inter-node new dat.conf

host0	**CPU1**			
	MIC0		**3340**	**3338**
	MIC1		**3345**	**3330**
Bandwidth (MB/s)		**CPU1**	**MIC0**	**MIC1**
		host1		

MIC network performance on the Helios supercomputer



Host0mic0–Host1mic0

1069.35
3383.44
3393.16

Memory bandwidth (MB/s)

Host0mic1–Host1mic1

1685.87
3346.75
3355.48

Mixed

Host0mic0–Host1mic0

1186.84

Host0mic1–Host1mic1

1632.26

Intel Manycore Platform Software Stack (IMPSS) v 3.6.1 Open Fabrics Enterprise Distribution (OFED) v 3.18

~3550 MB/s

Outline

- MIC general architecture
- MIC network performance on the Robin cluster with one IB port
- MIC network performance on the Helios supercomputer with two IB ports
- Host, offload and native computation mode of the test N-Body code
- Micro OpenMP overhead benchmark

Host, offload and native computation mode test using N-Body code

Execution time in (s)

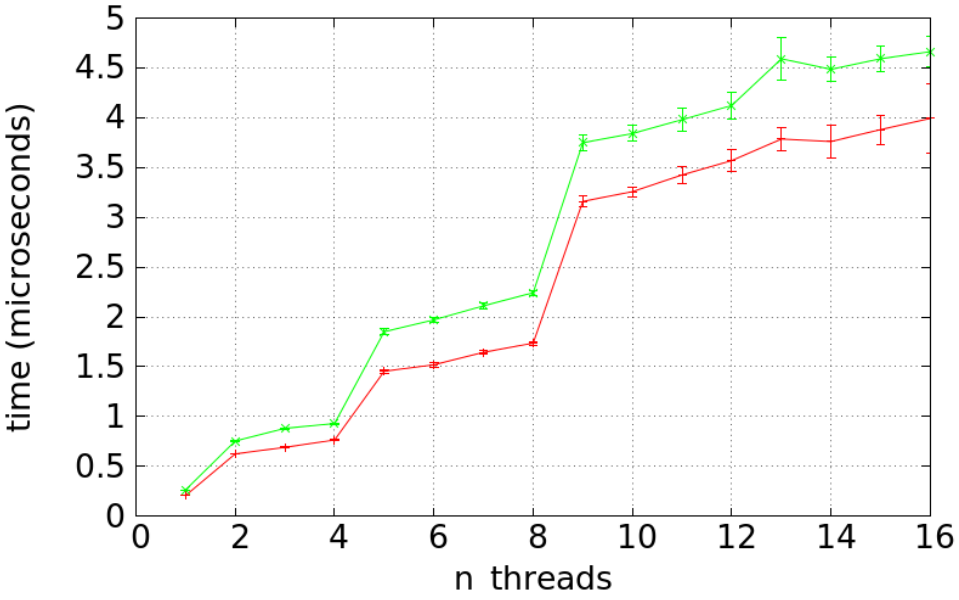
Number of cores	Intel Sandy Bridge	Intel Xeon Phi (offload)	Intel Xeon Phi (native)
1	55	130.61	126.60
2	28	66.47	62.69
4	14	33.78	30.75
8	7	18.78	15.86
16	3.5	12.02	9.97
32		7.44	4.72
64		6.19	3.46
128		4.09	1.59
236		3.96	1.39

Outline

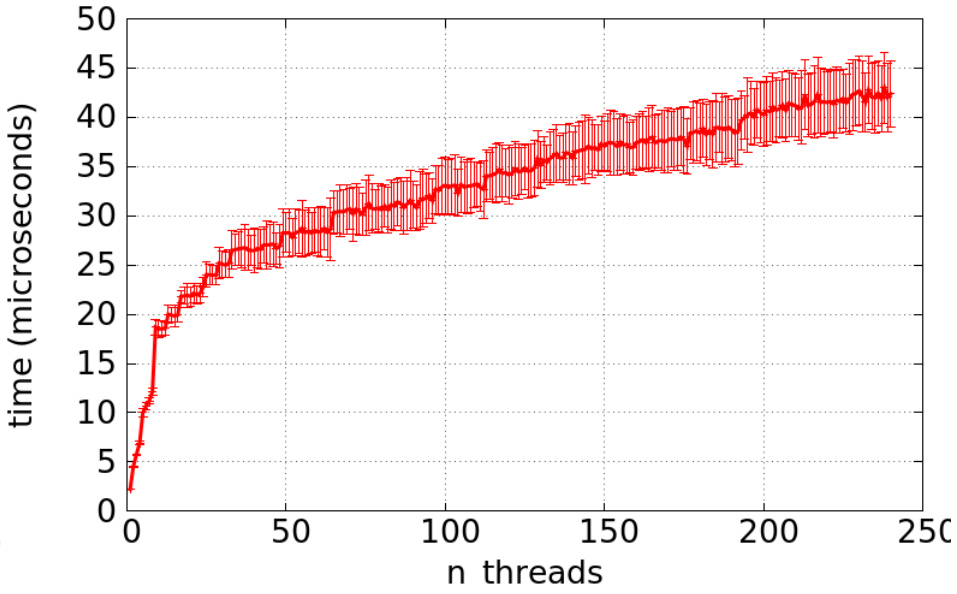
- MIC general architecture
- MIC network performance on the Robin cluster (made by M. Haefele) with one IB port
- MIC network performance on the Helios supercomputer with two IB ports
- Host, offload and native computation mode of the test N-Body code
- Micro OpenMP overhead benchmark

“OpenMP reduction” overhead

Helios **Sandy bridge** and **MIC host**

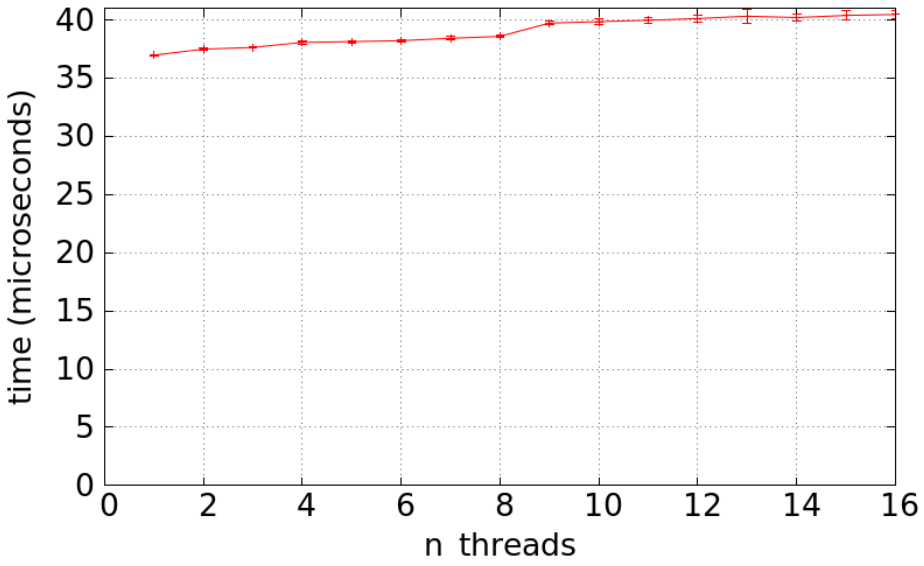


Helios MIC native mode

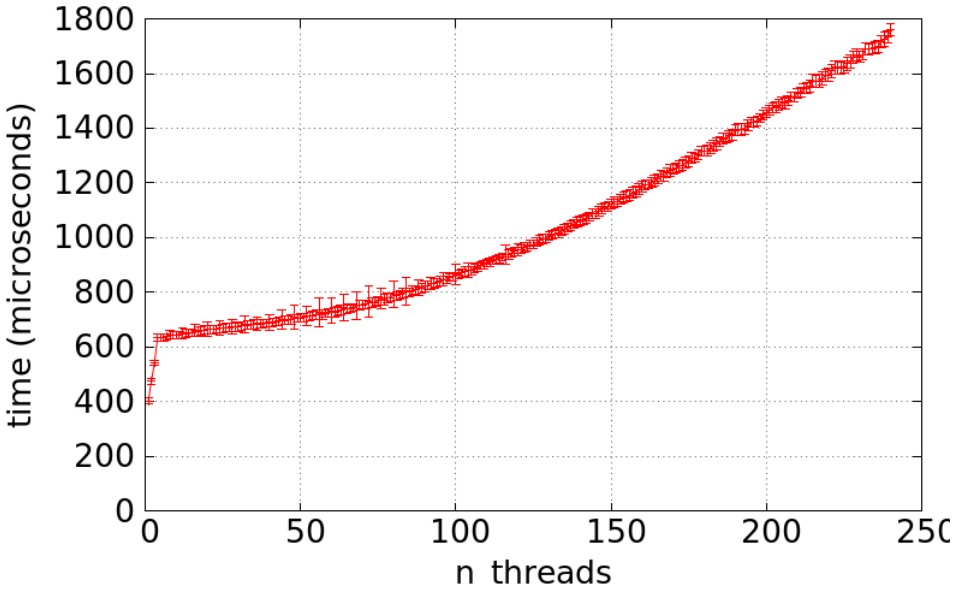


"OpenMP firstprivate" overhead using arraysize 59,049 bytes

MIC host



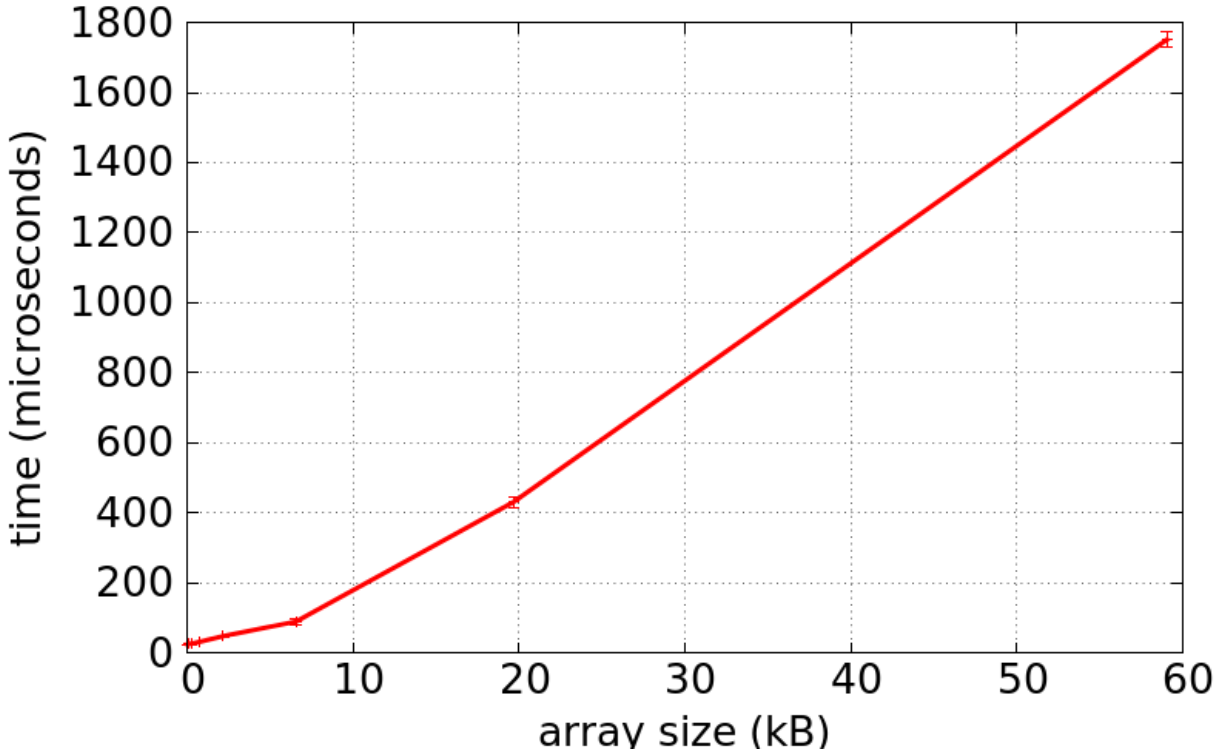
Helios MIC native mode



In real simulation the overhead time can be equal to the computational kernel time

"OpenMP firstprivate" overhead using different array size with 240 threads

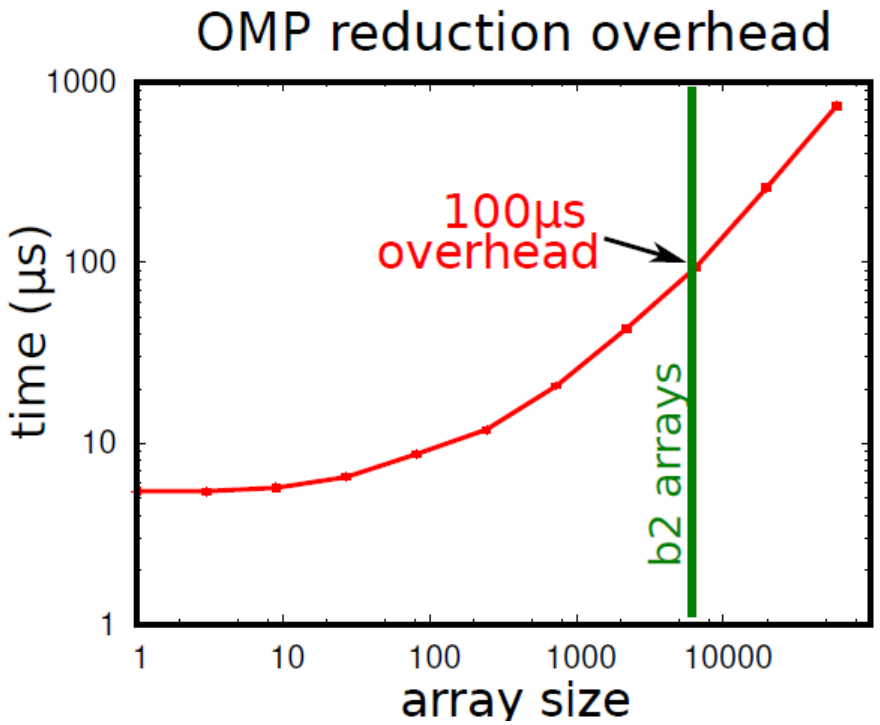
Helios MIC native mode



OMP overhead in real simulation – b2 code

20 cores on Ivy-Bridge (Hydra)

```
!$omp parallel do
private(is,ix,iy) &
!$omp reduction(+: s)
do is = 0, ns-1
  do iy = -1, ny
    do ix = -1, nx
      s(ix,iy) = s(ix,iy) + &
      a(ix,iy,is)* r(ix,iy,is)&
      * vol(ix,iy) * ne(ix,iy)
    enddo
  enddo
enddo
```



Thank you for your attention