Leibniz-Rechenzentrum
der Bayerischen Akademie der Wissenschaften

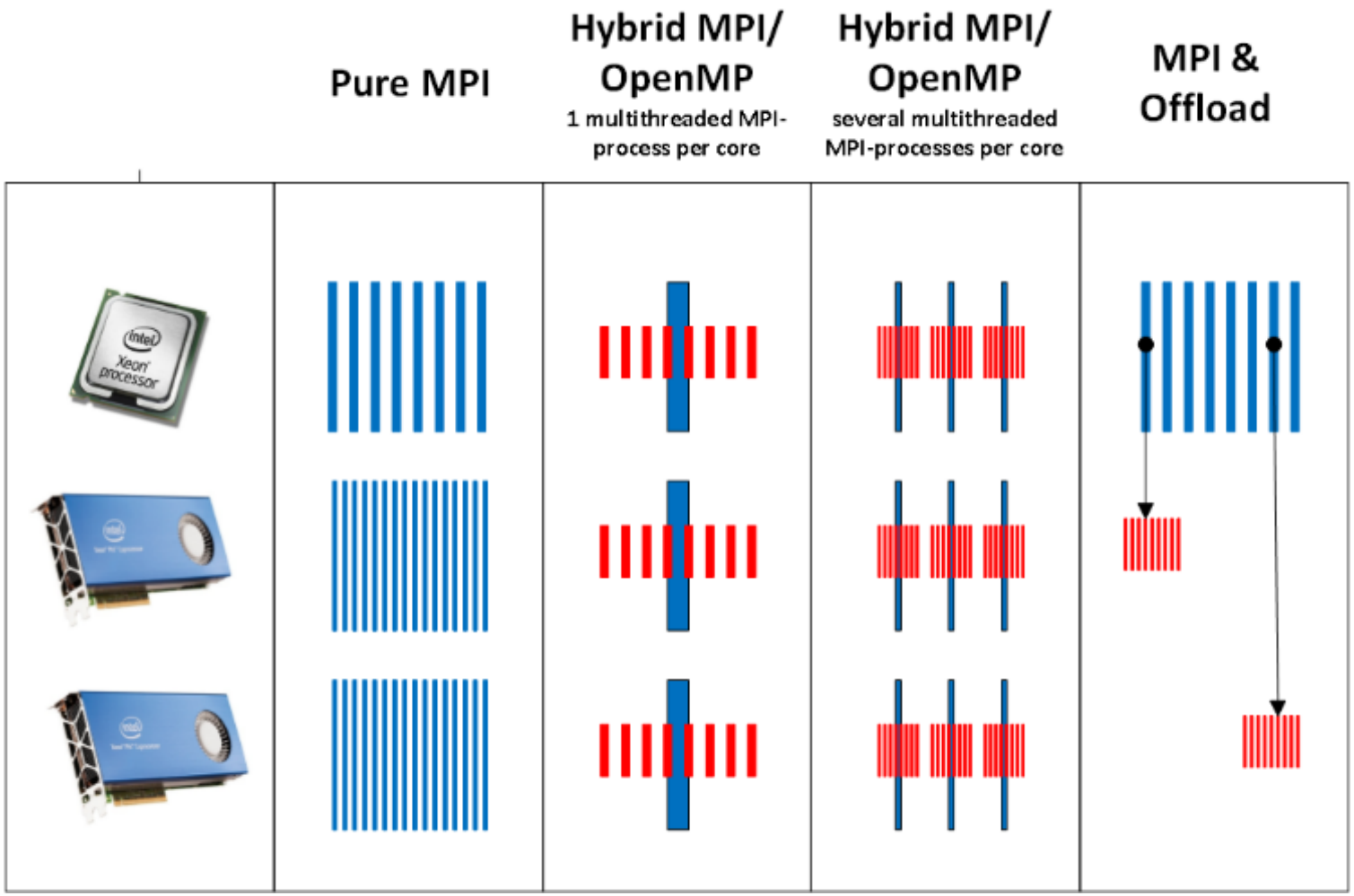
PRACE PATC Course:
Intel MIC Programming Workshop – MPI

LRZ, 27.6.- 29.6.2016

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Intel Xeon Phi Programming Models: MPI

MPI on Hosts & MICs



MPI @ LRZ

- Default Module:
  - SuperMUC: mpi.ibm/1.4
  - SuperMIC: mpi.intel/5.1

- If you compile for MIC on SuperMUC login nodes use:
  - lu65fok@login07:~> module unload mpi.ibm
  - lu65fok@login07:~> module load mpi.intel

Important MPI environment variables

- Important Paths are already set by intel module, otherwise use:
  - . $ICC_BASE/bin/compilervars.sh intel64
  - . $MPI_BASE/bin64/mpivars.sh

- At LRZ the following MIC-specific environment variables are set per default on SuperMIC:
  - I_MPI_MIC=enable
  - I_MPI_HYDRA_BOOTSTRAP=ssh
  - I_MPI_FABRICS=shm:dapl
  - I_MPI_DAPL_PROVIDER_LIST=ofa-v2-mlx4_0-1,ofa-v2-scif0 (must be tuned)

Important MPI environment variables

- Important Paths are already set by intel module, otherwise use:
  - . $ICC_BASE/bin/compilervars.sh intel64
  - . $MPI_BASE/bin64/mpivars.sh
- Recommended environment on Salomon:

```
module load intel
export I_MPI_HYDRA_BOOTSTRAP=ssh
export I_MPI_MIC=enable
export I_MPI_FABRICS=shm:dapl
export I_MPI_DAPL_PROVIDER_LIST=ofa-v2-mlx4_0-1u,ofa-
v2-scif0,ofa-v2-mcm-1
export MIC_LD_LIBRARY_PATH =
$MIC_LD_LIBRARY_PATH:/apps/all/impi/5.1.2.150-iccifort-
2016.1.150-GCC-4.9.3-2.25/mic/lib/
```

Invocation of the Intel MPI compiler

Language	MPI Compiler	Compiler
C	mpiicc	icc
C++	mpiicpc	icpc
Fortran	mpiifort	ifort

I_MPI_FABRICS

- The following network fabrics are available for the Intel Xeon Phi coprocessor:

shm	Shared-memory
tcp	TCP/IP-capable network fabrics, such as Ethernet and InfiniBand (through IPoIB)
ofa	OFA-capable network fabric including InfiniBand (through OFED verbs)
dapl	DAPL–capable network fabrics, such as InfiniBand, iWarp, Dolphin, and XPMEM (through DAPL)

I_MPI_FABRICS

- The default can be changed by setting the I_MPI_FABRICS environment variable to I_MPI_FABRICS=<fabric> or I_MPI_FABRICS= <intra-node fabric>:<inter-nodes fabric>

- Intranode: Shared Memory, Internode: DAPL (Default on SuperMIC/MUC)
  - export I_MPI_FABRICS=shm:dapl

- Intranode: Shared Memory, Internode: TCP (Can be used in case of Infiniband problems)
  - export I_MPI_FABRICS=shm:tcp

MPI Usage in Native Mode: Best Practices

- When running MPI tasks on several hosts AND Xeon Phi coprocessors, several collective MPI functions like MPI Barriers do not return properly (cause deadlocks).
- In this case set i.e.
  - export I_MPI_DAPL_PROVIDER_LIST=ofa-v2-mlx4_0-1u
  - export I_MPI_ADJUST_BARRIER=1
  - export I_MPI_ADJUST_BCAST=1
- More details can be found under https://software.intel.com/en-us/articles/intel-mpi-library-collective-optimization-on-intel-xeon-phi
- To improve the performance of MPI_Put operations use:
  export I_MPI_SCALABLE_OPTIMIZATION=off

Sample MPI Program

```
lu65fok@login12:~/tests> cat testmpi.c
#include <stdio.h>
#include <mpi.h>

int main (int argc, char* argv[]) {
  char hostname[100];
  int rank, size;

  MPI_Init (&argc, &argv);      /* starts MPI */
  MPI_Comm_rank (MPI_COMM_WORLD, &rank);       /* get current process id */
  MPI_Comm_size (MPI_COMM_WORLD, &size);       /* get number of processes */

  gethostname(hostname,100);
  printf( "Hello world from process %d of %d: host: %s\n", rank, size, hostname);
  MPI_Finalize();
  return 0;
}
```

MPI on hosts

- Compile for host using mpiicc / mpiifort:

  lu65fok@login12:~/tests> mpiicc testmpi.c -o testmpi-host

- Run 2 MPI tasks on host node i01r13a01

  lu65fok@login12:~/tests> mpiexec -n 2 -host i01r13a01 ./testmpi-host

  Hello world from process 0 of 2: host: i01r13a01

  Hello world from process 1 of 2: host: i01r13a01

MPI in native mode on 1 MIC

- Compile for MIC using mpiicc / mpiifort -mmic:

```
lu65fok@login12:~/tests> mpiicc  -mmic testmpi.c -o testmpi-mic
```

- Copy binary to MIC:

```
lu65fok@login12:~/tests> scp testmpi-mic  i01r13a01-mic0:
```

- Launch 2 MPI tasks from MIC node i01r13a01-mic0

```
lu65fok@i01r13a04:~/tests> ssh i01r13a01-mic0
[lu65fok@i01r13a01-mic0 ~]$ mpiexec -n 2 ./testmpi-mic
Hello world from process 1 of 2: host: i01r13a01-mic0
Hello world from process 0 of 2: host: i01r13a01-mic0
```

Do not mix up with mpicc and mpifort!!

```
lu65fok@login12:~/tests> mpicc  -mmic testmpi.c -o testmpi-mic
/usr/lib64/gcc/x86_64-suse-linux/4.3/../../../../x86_64-suse-linux/bin/ld: skipping incompatible
/lrz/sys/intel/mpi_41_3_048/mic/lib/libmpigf.so when searching for -lmpigf
/usr/lib64/gcc/x86_64-suse-linux/4.3/../../../../x86_64-suse-linux/bin/ld: skipping incompatible
/lrz/sys/intel/mpi_41_3_048/mic/lib/libmpigf.a when searching for -lmpigf
/usr/lib64/gcc/x86_64-suse-linux/4.3/../../../../x86_64-suse-linux/bin/ld: cannot find -lmpigf
/usr/lib64/gcc/x86_64-suse-linux/4.3/../../../../x86_64-suse-linux/bin/ld: skipping incompatible
/lrz/sys/intel/mpi_41_3_048/mic/lib/libmpi.so when searching for -lmpi
/usr/lib64/gcc/x86_64-suse-linux/4.3/../../../../x86_64-suse-linux/bin/ld: skipping incompatible
/lrz/sys/intel/mpi_41_3_048/mic/lib/libmpi.a when searching for -lmpi
/usr/lib64/gcc/x86_64-suse-linux/4.3/../../../../x86_64-suse-linux/bin/ld: cannot find -lmpi
/usr/lib64/gcc/x86_64-suse-linux/4.3/../../../../x86_64-suse-linux/bin/ld: skipping incompatible
/lrz/sys/intel/mpi_41_3_048/mic/lib/libmpigi.a when searching for -lmpigi
/usr/lib64/gcc/x86_64-suse-linux/4.3/../../../../x86_64-suse-linux/bin/ld: cannot find -lmpigi
collect2: ld returned 1 exit status
```

MPI on 1 MIC

- Compile for MIC using mpiicc / mpiifort -mmic:

  lu65fok@login12:~/tests> mpiicc -mmic testmpi.c -o testmpi-mic

- Copy binary to MIC

  (not necessary if home is mounted on MICs)

  lu65fok@login12:~/tests> scp testmpi-mic i01r13a01-mic0:

- Run 2 MPI tasks on MIC node i01r13a01-mic0

  lu65fok@i01r13a04:~/tests> mpiexec -n 2 -host i01r13a01-mic0 ./home/lu65fok/testmpi-mic

  Hello world from process 1 of 2: host: i01r13a01-mic0

  Hello world from process 0 of 2: host: i01r13a01-mic0

Full path needed!

MPI on 2 MICs

- Compile for MIC using mpiicc / mpiifort -mmic:

  lu65fok@login12:~/tests> mpiicc -mmic testmpi.c -o testmpi-mic

- Copy binary to MICs:

  (not necessary if home is mounted on MICs)

  lu65fok@login12:~/tests> scp testmpi-mic i01r13a01-mic0:

  lu65fok@login12:~/tests> scp testmpi-mic i01r13a01-mic1:

- Run 2 MPI tasks on MIC node i01r13a01-mic0

  lu65fok@login12:~/tests> mpiexec -n 2 -perhost 1 -host i01r13a01-mic0,i01r13a01-mic1 ./home/lu65fok/testmpi-mic

  Hello world from process 1 of 2: host: i01r13a01-mic1

  Hello world from process 0 of 2: host: i01r13a01-mic0

MPI on Host and 2 MICs attached to the host

lu65fok@login12:~/tests> mpirun -n 1 -host i01r13a01 ./testmpi-host : -n 1 -host i01r13a01-mic0 /home/lu65fok/testmpi-mic : -n 1 -host i01r13a01-mic1 /home/lu65fok/testmpi-mic

Hello world from process 0 of 3: host: i01r13a01

Hello world from process 2 of 3: host: i01r13a01-mic1

Hello world from process 1 of 3: host: i01r13a01-mic0

MPI on multiple Hosts & MICs

```
lu65fok@i01r13a01:~/tests> mpirun -n 1 -host i01r13a01 ./testmpi-host : -n 1 -host
i01r13a01-mic0 /home/lu65fok/testmpi-mic : -n 1 -host i01r13a01-mic1
/home/lu65fok/testmpi-mic : -n 1 -host i01r13a02 ./testmpi-host : -n 1 -host
i01r13a02-mic0 /home/lu65fok/testmpi-mic : -n 1 -host i01r13a02-mic1
/home/lu65fok/testmpi-mic
```

```
Hello world from process 3 of 6: host: i01r13a02
Hello world from process 0 of 6: host: i01r13a01
Hello world from process 2 of 6: host: i01r13a01-mic1
Hello world from process 5 of 6: host: i01r13a02-mic1
Hello world from process 1 of 6: host: i01r13a01-mic0
Hello world from process 4 of 6: host: i01r13a02-mic0
```

MPI Machine File

```
lu65fok@login12:~/tests> cat machinefile.txt
i01r13a01-mic0
i01r13a01-mic1
i01r13a02-mic0
i01r13a02-mic1

lu65fok@login12:~/tests> mpirun  -n 4 -machinefile machinefile.txt
/home/lu65fok/testmpi-mic
Hello world from process 3 of 4: host: i01r13a02-mic1
Hello world from process 2 of 4: host: i01r13a02-mic0
Hello world from process 1 of 4: host: i01r13a01-mic1
Hello world from process 0 of 4: host: i01r13a01-mic0
```

MPI Machine File

```
lu65fok@login12:~/tests> cat machinefile.txt
i01r13a01-mic0:2
i01r13a01-mic1
i01r13a02-mic0
i01r13a02-mic1

lu65fok@login12:~/tests> mpirun  -n 4 -machinefile machinefile.txt
/home/lu65fok/testmpi-mic
Hello world from process 3 of 4: host: i01r13a02-mic0
Hello world from process 0 of 4: host: i01r13a01-mic0
Hello world from process 2 of 4: host: i01r13a01-mic1
Hello world from process 1 of 4: host: i01r13a01-mic0
```

Offload from MPI Tasks

```
#include <unistd.h>
#include <stdio.h>
#include <mpi.h>

int main (int argc, char* argv[]) {
  char hostname[100];
  int rank, size;
  MPI_Init (&argc, &argv);      /* starts MPI */
  MPI_Comm_rank (MPI_COMM_WORLD, &rank);        /* get current process id */
  MPI_Comm_size (MPI_COMM_WORLD, &size);        /* get number of processes */

  gethostname(hostname,100);

  #pragma offload target(mic)
{
    char michostname[50];
    gethostname(michostname, 50);
    printf("MIC: I am %s and I have %ld logical cores. I was called by  process %d of %d: host: %s \n", michostname,
        sysconf(_SC_NPROCESSORS_ONLN), rank, size, hostname);
  }
  printf( "Hello world from process %d of %d: host: %s\n", rank, size, hostname);
  MPI_Finalize();
  return 0;

}
```

Offload from MPI Tasks using 1 host

```
lu65fok@login12:~/tests> mpiicc testmpioffload.c -o testmpioffload
lu65fok@login12:~/tests> mpirun -n 4  -host i01r13a01 ./testmpioffload
Hello world from process 3 of 4: host: i01r13a01
Hello world from process 1 of 4: host: i01r13a01
Hello world from process 0 of 4: host: i01r13a01
Hello world from process 2 of 4: host: i01r13a01
MIC: I am i01r13a01-mic0 and I have 240 logical cores. I was called by process 3 of 4: host: i01r13a01
MIC: I am i01r13a01-mic0 and I have 240 logical cores. I was called by process 0 of 4: host: i01r13a01
MIC: I am i01r13a01-mic0 and I have 240 logical cores. I was called by process 1 of 4: host: i01r13a01
MIC: I am i01r13a01-mic0 and I have 240 logical cores. I was called by process 2 of 4: host: i01r13a01
```

Offload from MPI Tasks using multiple hosts

```
lu65fok@login12:~/tests> mpirun -n 4 -perhost 2   -host
i01r13a01,i01r13a02 ./testmpioffload
Hello world from process 2 of 4: host: i01r13a02
Hello world from process 0 of 4: host: i01r13a01
Hello world from process 3 of 4: host: i01r13a02
Hello world from process 1 of 4: host: i01r13a01
MIC: I am i01r13a02-mic0 and I have 240 logical cores. I was called by
process 2 of 4: host: i01r13a02
MIC: I am i01r13a01-mic0 and I have 240 logical cores. I was called by
process 1 of 4: host: i01r13a01
MIC: I am i01r13a01-mic0 and I have 240 logical cores. I was called by
process 0 of 4: host: i01r13a01
MIC: I am i01r13a02-mic0 and I have 240 logical cores. I was called by
process 3 of 4: host: i01r13a02
```

Offload from MPI Tasks: Using both MICs

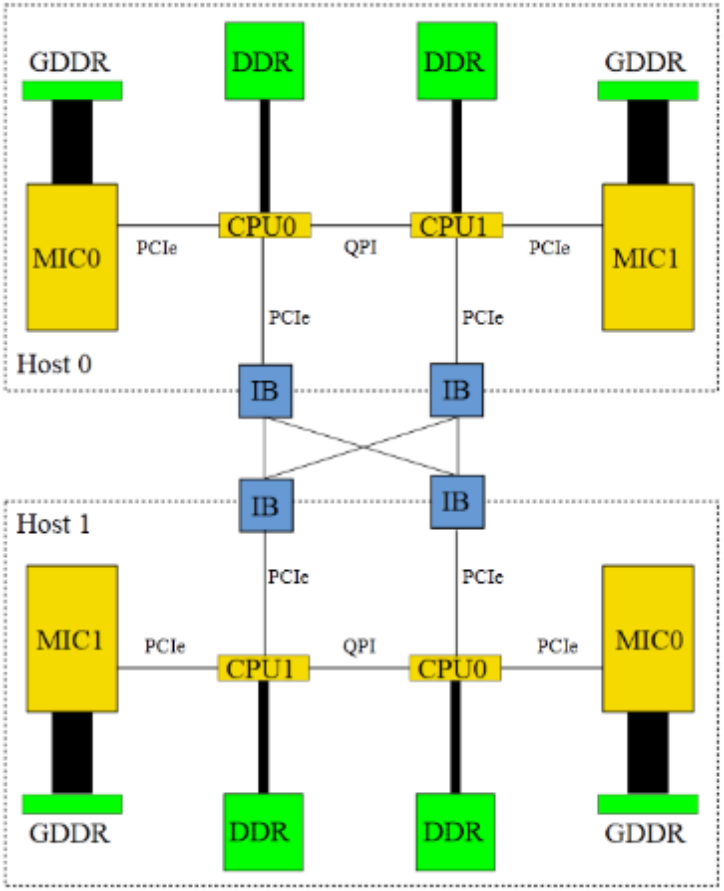
```
#pragma offload target(mic:rank%2)
 {
   char michostname[50];
   gethostname(michostname, sizeof(michostname));
   printf("MIC: I am %s and I have %ld logical cores. I was called by
              process %d of %d: host: %s \n", michostname,
              sysconf(_SC_NPROCESSORS_ONLN), rank, size,
hostname);
 }
```

Offload from MPI Tasks: Using both MICs

```
lu65fok@login12:~/tests> mpirun -n 4 -perhost 2   -host
i01r13a01,i01r13a02 ./testmpioffload
Hello world from process 0 of 4: host: i01r13a01
Hello world from process 2 of 4: host: i01r13a02
Hello world from process 3 of 4: host: i01r13a02
Hello world from process 1 of 4: host: i01r13a01
MIC: I am i01r13a02-mic1 and I have 240 logical cores. I was called
by  process 3 of 4: host: i01r13a02
MIC: I am i01r13a01-mic1 and I have 240 logical cores. I was called
by  process 1 of 4: host: i01r13a01
MIC: I am i01r13a01-mic0 and I have 240 logical cores. I was called
by  process 0 of 4: host: i01r13a01
MIC: I am i01r13a02-mic0 and I have 240 logical cores. I was called
by  process 2 of 4: host: i01r13a02
```

SuperMIC architecture

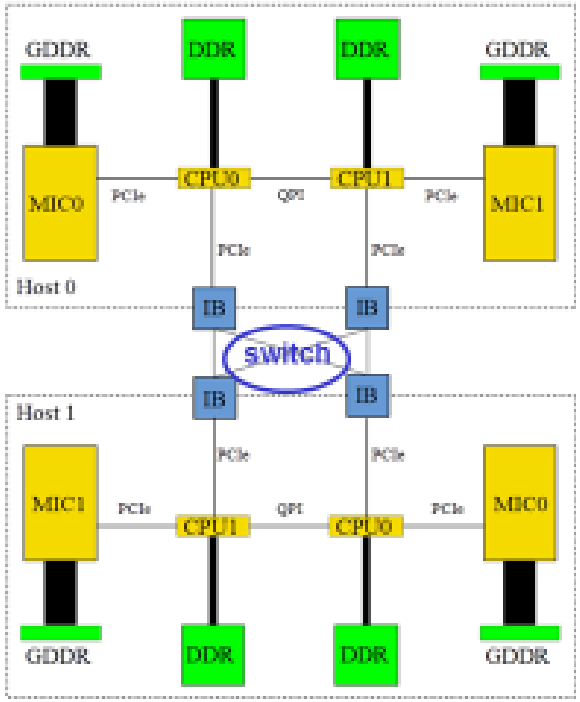
SuperMIC / Helios:
two IB ports



(Figure from M. Haefele)

Improved MPI performance

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 mic 1

```
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		

(Slide from Serhiy Mochalskyy)

Lab: MPI

Thank you!