VIENNA SCIENTIFIC CLUSTER

Parallelization with MPI

Beginner

28-29 April 2025
ONLINE

EURO AUSTRIA

MPI

Message Passing Interface

One of the most iconic tools in HPC.

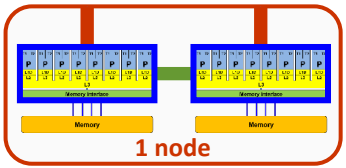
No matter if it‘s large-scale HPC simulations or training large AI models, the backbone that makes this possible is MPI or one of its derivatives.

MPI – communication is key

VSC-3 ping pong benchmark

to be shown
after exercises & homework
of “messages and point-to-point communication”

VSC-3 – fat-tree interconnect



-3- spine

Spine

Core

Leaf

-2- leaf

Edge

-1- edge

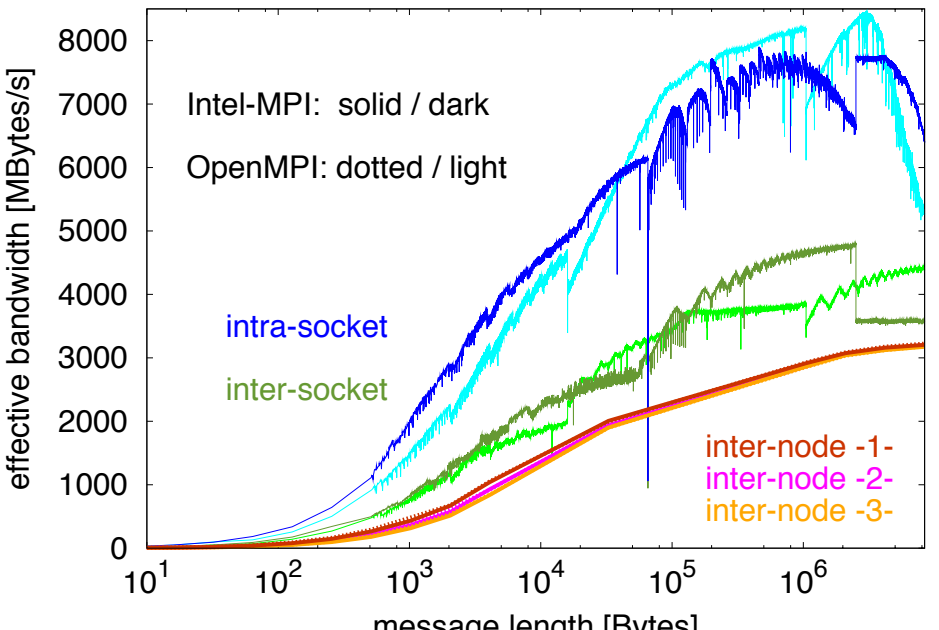
VSC-3 ping pong benchmark – latency

Latency [µs]	MPI_Send(…)			MPI_Ssend(…)	
	OpenMPI	Intel-MPI		Intel-MPI	OpenMPI
	C~Fortran	C <	Fortran	C~<F.	C~<F.
intra-socket	0.3	**0.3**	0.3	1.2	1.3
inter-socket	0.6	**0.7**	0.7	1.7	2.0
IB -1- edge	1.2	**1.4**	1.5	2.0	14.0
IB -2- leaf	1.6	**1.8**	1.9	2.5	15.0
IB -3- spine	2.1	**2.3**	2.4	3.0	18.0

module load intel/16.0.3 intel-mpi/5.1.3 | openmpi/1.10.2

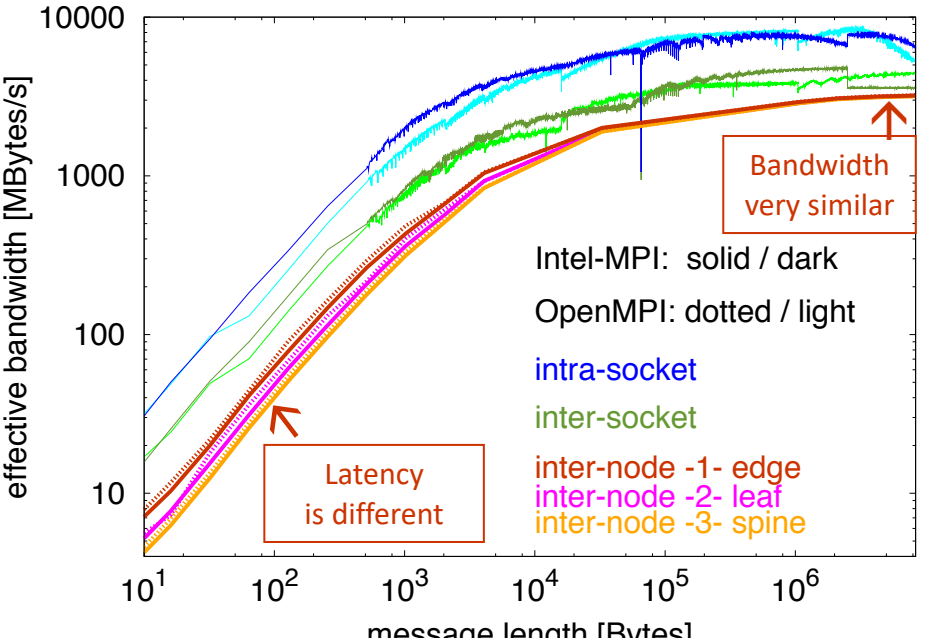
ping pong benchmark – bandwidth

VSC-3



ping pong benchmark – bandwidth

VSC-3



Optimizing MPI communication
a real-world example

to be shown

after exercises

of “collective communication”

WRONG (ORIG) version – a real world example

s_p_e:

loop over irank except self:

MPI_ISEND(snd_buf … irank …)

MPI_BARRIER(…)

loop over irank except self:

MPI_RECV(rcv_buf … irank …)

MPI_BARRIER(…)

DO SOME WORK (snd_buf is not used = okay)

MPI_BARRIER(…)

problem size	6x6		8x8	
timing	1_s_p_e	total	1_s_p_e	total
WRONG (ORIG)	1500 s	14 h	21000 s	90 h

REMOVE BARRIER – a real world example

s_p_e:

loop over irank except self:

MPI_ISEND(snd_buf … irank …)

~~MPI_BARRIER(…)~~

loop over irank except self:

MPI_RECV(rcv_buf … irank …)

~~MPI_BARRIER(…)~~

DO SOME WORK (snd_buf is not used = okay)

~~MPI_BARRIER(…)~~

problem size	6x6		8x8	
timing	1_s_p_e	total	1_s_p_e	total
WRONG (ORIG)	1500 s	14 h	21000 s	90 h
REMOVE BARRIER	960 s	12 h	3500 s	32 h

NEVER use MPI_BARRIER in production code !!!!!

CORRECT NB comm. – a real world example

s_p_e:

ASYNCHRONOUS :: snd_buf

loop over irank except self:
MPI_ISEND(snd_buf … irank …)

loop over irank except self:
MPI_RECV(rcv_buf … irank …)
MPI_WAIT(…)
if (.NOT. MPI_ASYNC_PROT…

DO SOME WORK (snd_buf is not used = okay)

problem size	6x6		8x8	
timing	1_s_p_e	total	1_s_p_e	total
WRONG (ORIG)	1500 s	14 h	21000 s	90 h
REMOVE BARRIER	960 s	12 h	3500 s	32 h
CORRECT NB comm.	425 s	10 h	960 s	18 h

BCAST version – a real world example

s_p_e:

~~ASYNCHRONOUS :: snd_buf~~

loop over irank except self:

~~MPI_ISEND(snd_buf … irank …)~~

loop over irank except self:

rcv_buf = snd_buf

MPI_BCAST(rcv_buf … irank …)

DO SOME WORK (snd_buf is not used = okay)

problem size	6x6		8x8	
timing	1_s_p_e	total	1_s_p_e	total
WRONG (ORIG)	1500 s	14 h	21000 s	90 h
REMOVE BARRIER	960 s	12 h	3500 s	32 h
CORRECT NB comm.	425 s	10 h	960 s	18 h
BCAST	420 s	10 h	1025 s	20 h