Leibniz Supercomputing Centre
of the Bavarian Academy of Sciences and Humanities
ScalNEXT
Funded by the European Union
NextGenerationEU
SPONSORED BY THE
Federal Ministry of Education and Research

Modeling and Simulation of MPI Communication Dynamics in InfiniBand Networks Using SimGrid

Authors: Ehab Saleh, Amir Raoofy, Matthew Tovey, Josef Weidendorfer

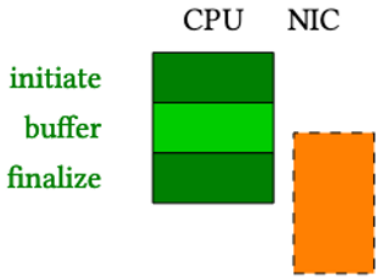
Presenter: Ehab Saleh - Leibniz Supercomputing Centre

Parallel-Algorithmen, Rechnerstrukturen und Systemsoftware (PARS)-Workshop
Technische Hochschule Ingolstadt
25th -26th Sep 2024

MPI communication modes

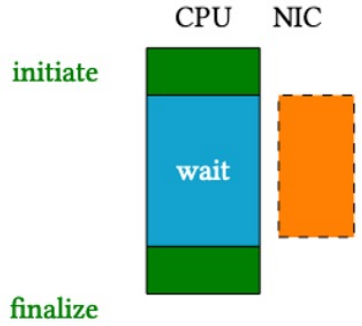
a) Buffered send:

- Corresponds to small messages.
- The message is copied immediately to the NIC.



b) Eager send:

- Corresponds to larger messages.
- The message buffer pointer is passed to the NIC.
- Sender must wait for a signal of completion from the NIC.



c) Rendezvous send:

- Even larger messages.
- Handshake protocol is used.

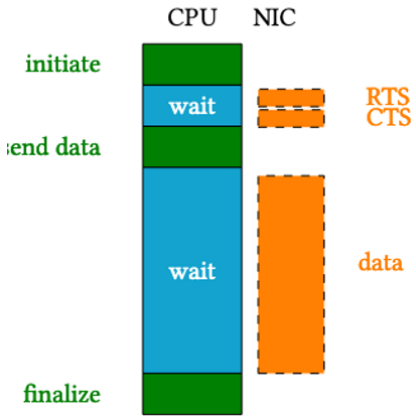
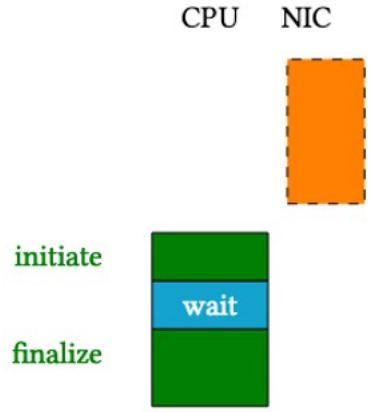


Figure is adapted from:Zhou et al. *MPI Progress for All*. https://arxiv.org/abs/2405.13807

MPI communication modes

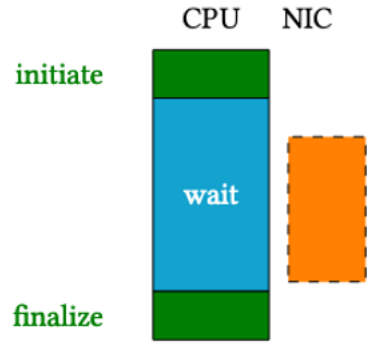
d) Eager unexpected receive:

- Receiving an eager message that arrived before posting the receive.



e) Eager expected receive:

- Receiving an eager message that arrived after posting the receive.



f) Rendezvous receive:

- Receiving a rendezvous message.

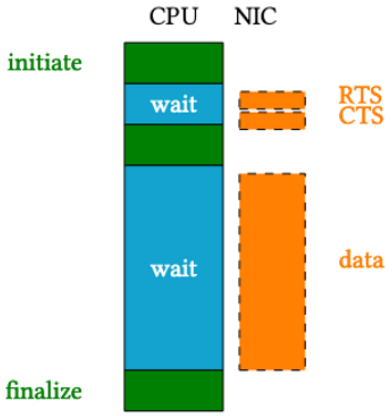


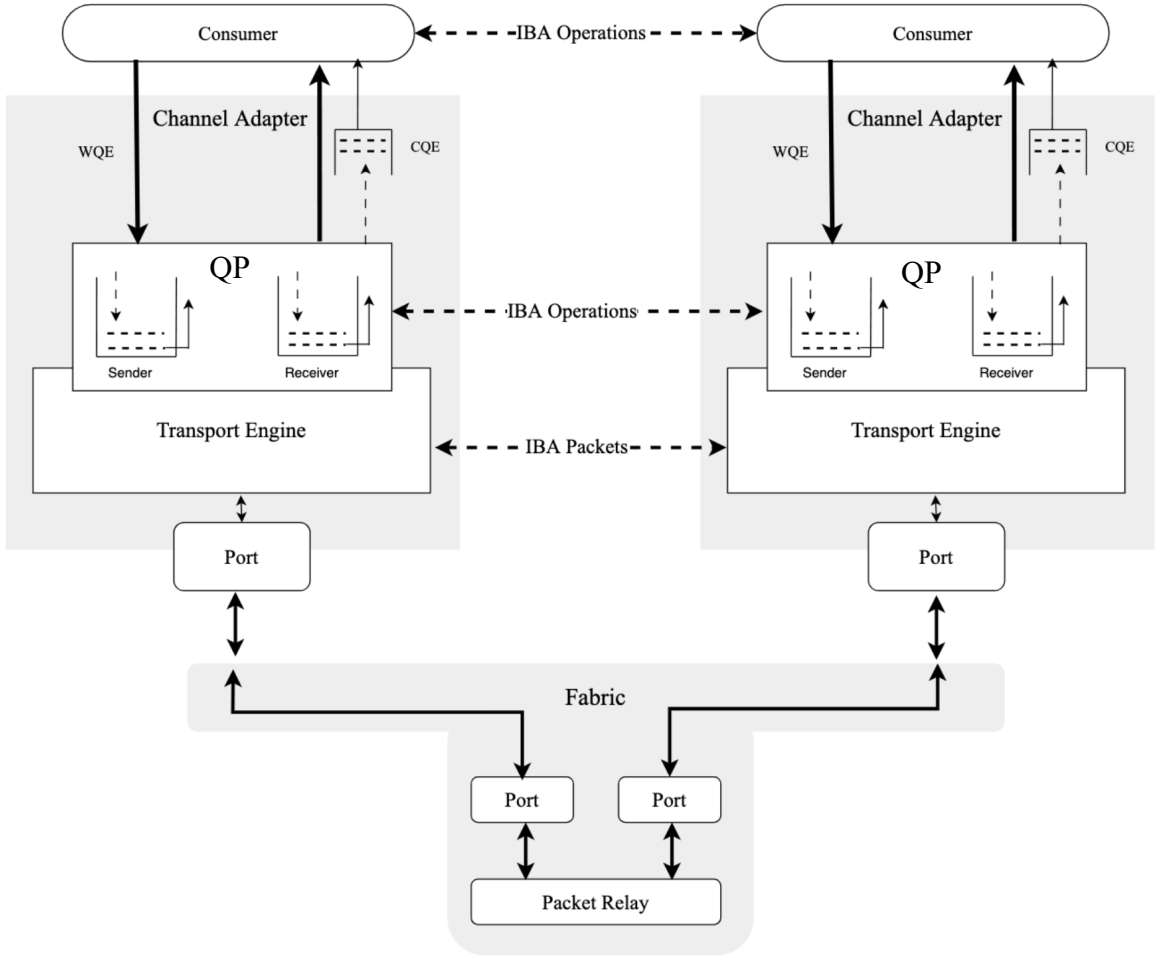
Figure is adapted from:Zhou et al. *MPI Progress for All*. https://arxiv.org/abs/2405.13807

InfiniBand architecture

- IB uses queue-based model:
  - Queue Pair (QP)
  - Work Queue Element (WQE)
  - Completion Queue Entry (CQE)

- IB supports two types of communication semantics:
  - Memory semantics
  - Send/Receive semantics

- MPI implementations use both adapter semantics when necessary.

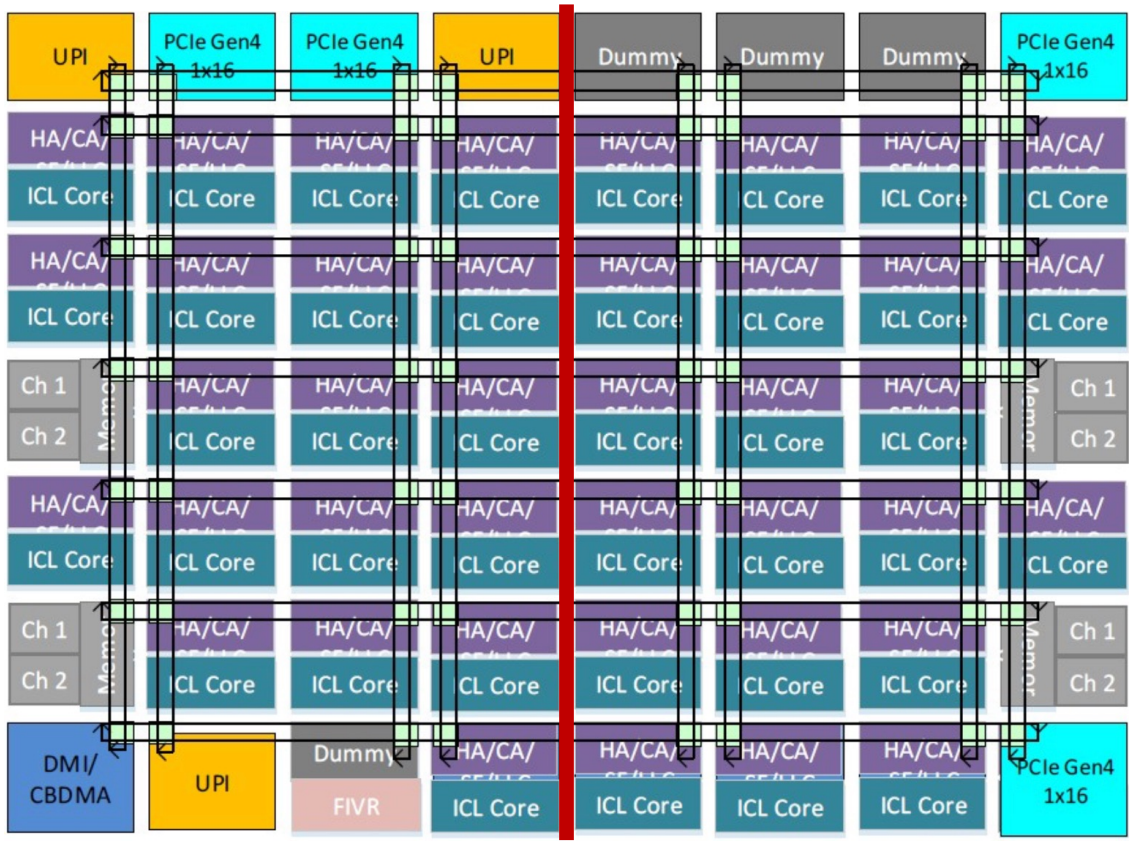


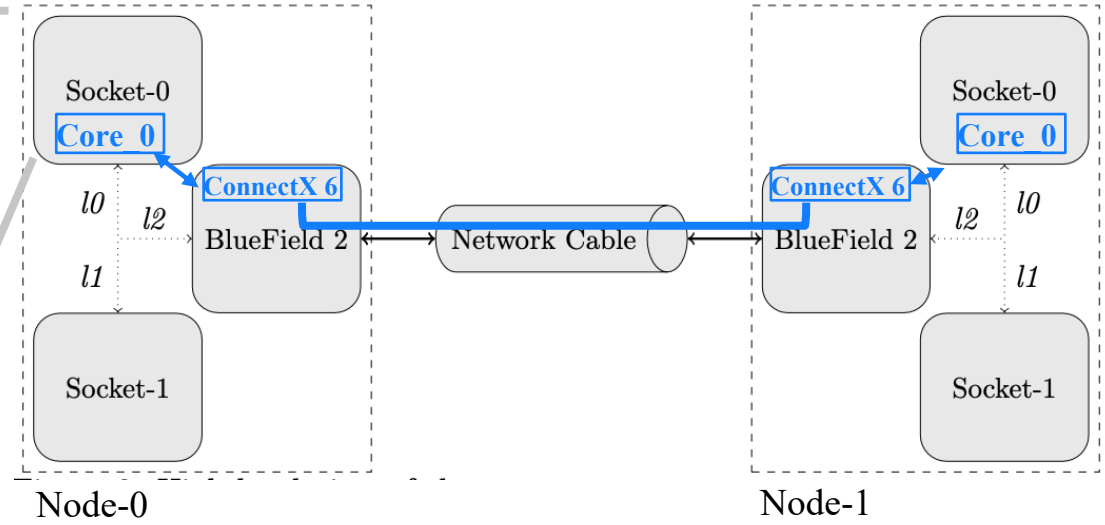
Modeling and simulation of MPI communication over IB

1- Choosing the real platform:

- Understanding the hardware affinities.

IceLake Socket Diagram





Modeling and simulation of MPI communication over IB

2- Execution of tests in the real platform and collection of duration data:

1- **MPI_Send**: Measures the time spent in blocking MPI_Send operation.

2- **MPI_Isend**: Measures the time spent MPI_Isend operation.

3- **MPI_Recv**: Time spent in MPI_Recv operation.

4- **Ping-Pong Benchmark**: Measures the elapsed time to perform a MPI_Send followed by a MPI_Recv.

5- **MPI_Test**: Measures the elapsed time to perform MPI_Test operation.

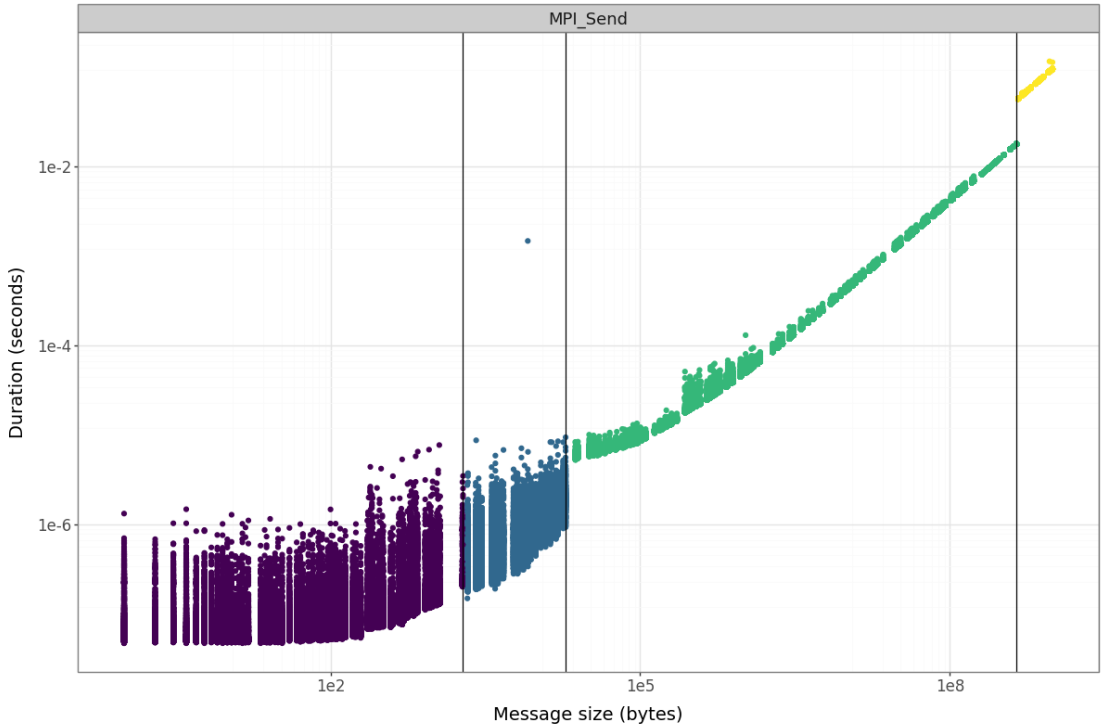
6- **MPI_Wtime**: Mesure the duration of MPI_Wtime operation.

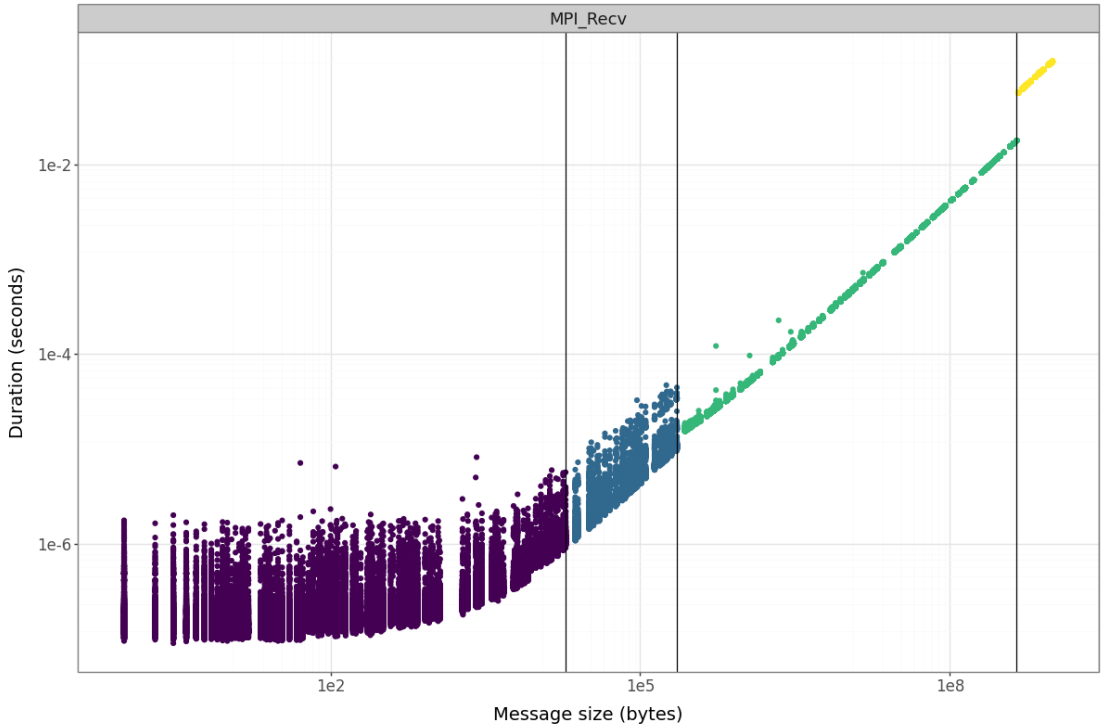
7- **MPI_Iprobe**: Meaure the duration of MPI_Iprobe operation.

Modeling and simulation of MPI communication over IB

3- Identifying thresholds:

- Identifying the threshold of the asynchronous and synchronous mode of MPI.
- The behaviour of communication cost cannot be represented with a single model.





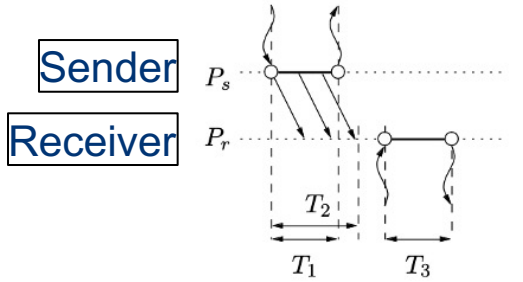
These experiments are conducted on two nodes of the BEAST systems using OpenMPI v5.0.3 on top of InfiniBand (Bluefield-2 with ConnectX-6 @200G HDR

Modeling and simulation of MPI communication over IB

3- Identifying thresholds:

- The hybrid network model of LogGPS is able to account for the full complexity of a real MPI implementation:

- Asynchronous: this mode corresponds to messages that fit in IB frame and are sent asynchronously by the kernel.

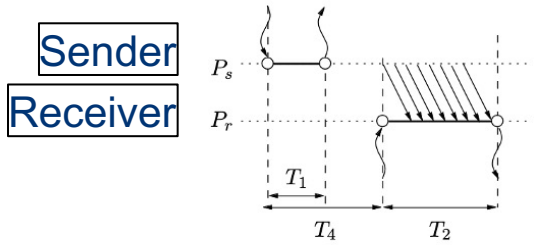


$$T_1 = o_s^{(i)} + kO_s^{(i)}$$
$$T_3 = o_r^{(i)} + kO_r^{(i)}$$

$(k \leq S_a)$

(a) Asynchronous mode

- Detached: this mode corresponds to messages that do not block the sender but require the receiver to post the reception before the communication actually takes place.

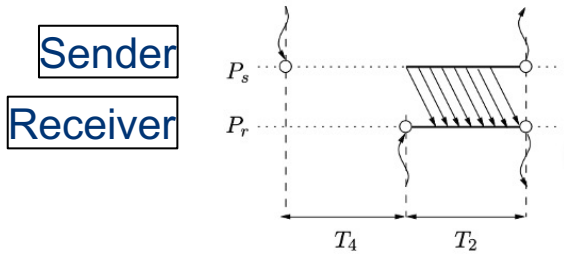


$$T_1 = o_s^{(i)} + kO_s^{(i)}$$
$$T_2 = L^{(i)} + k/B^{(i)}$$
$$T_4 = t_r - t_s$$

$(S_a < k \leq S_d)$

(b) Detached mode

- Large(Synchronous):both sender and receiver synchronize using a rendezvous protocol before the data is sent



$$T_2 = L^{(i)} + k/B^{(i)}$$
$$T_4 = t_r - t_s$$

$(k > S_d)$

(c) Synchronous mode

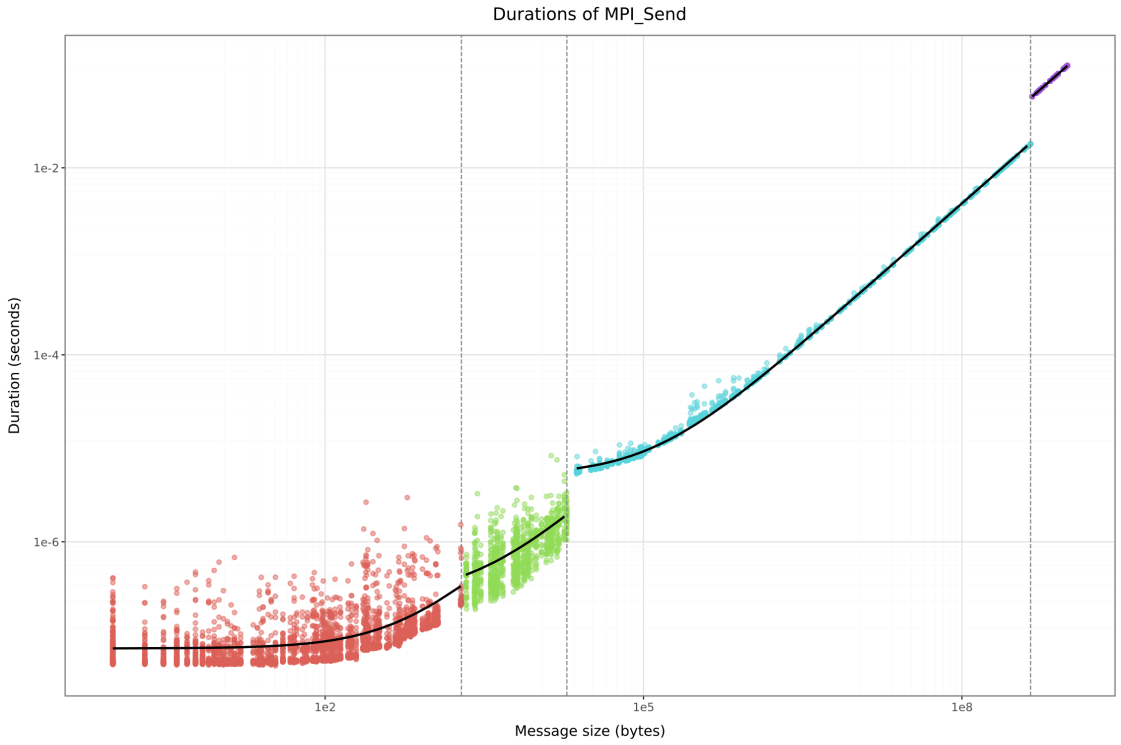
Modeling and simulation of MPI communication over IB

4- Segmentation:

- Using piece-wise modelling to group MPI communication modes into segments.
- Calculating the communication cost after determining the average duration of each segment.

$$Communication\_cost = coefficient \times msg\_size + intercept$$

size_min	size_max	Intercept	Coefficient
0	1920	7.2255e-08	1.3864e-10
1920	18947	2.5349e-07	8.9679e-11
18947	443618990	5.1178e-06	4.1525e-11
443618990	inf	0.0003746	1.2494e-10



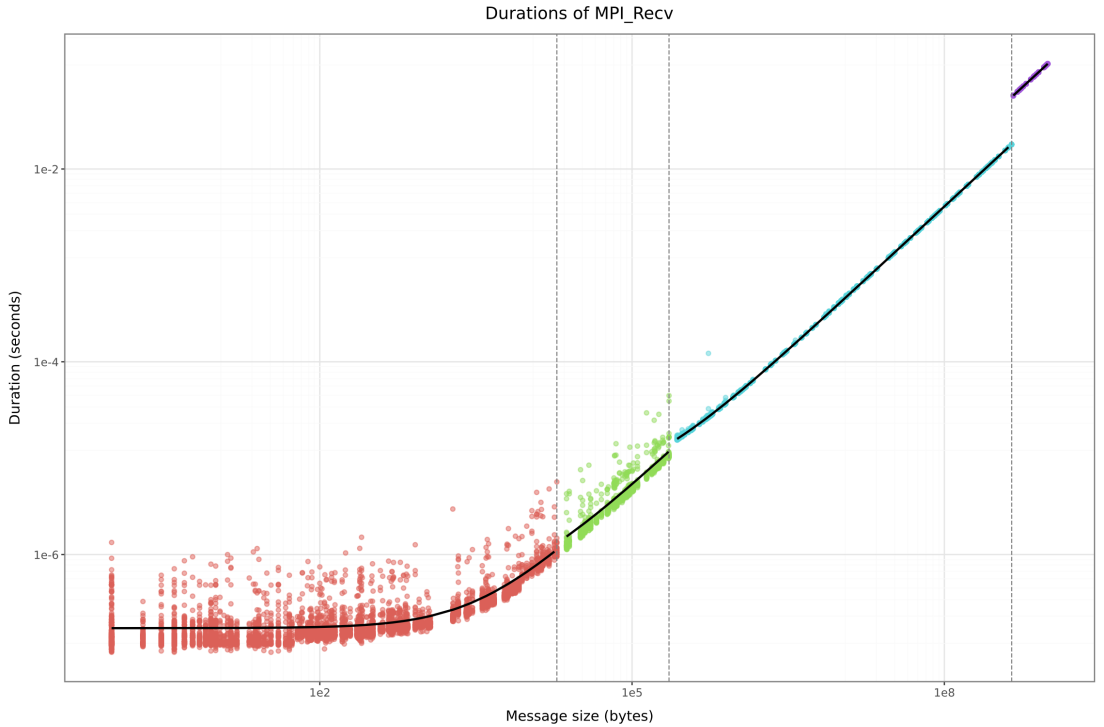
Modeling and simulation of MPI communication over IB

4- Segmentation:

- Using piece-wise modelling to group MPI communication modes into segments.
- Calculating the communication cost after determining the average duration of each segment.

$$Communication\_cost = coefficient \times msg\_size + intercept$$

size_min	size_max	Intercept	Coefficient
0	18947	1.7229e-07	5.0202e-11
18947	226629	1.55671e-07	5.0351e-11
226629	443618990	4.8031e-06	4.0736e-11
443618990	inf	0.00034669	1.2489e-10



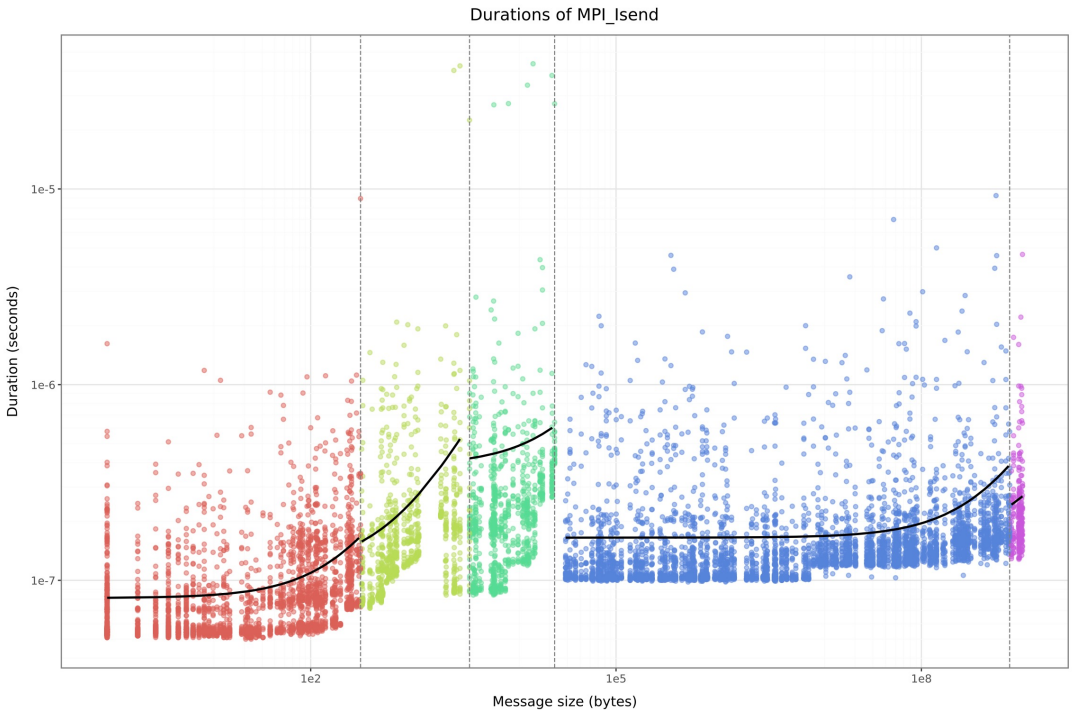
Modeling and simulation of MPI communication over IB

4- Segmentation:

- Using piece-wise modelling to group MPI communication modes into segments.
- Calculating the communication cost after determining the average duration of each segment.

$$Communication\_cost = coefficient \times msg\_size + intercept$$

size_min	size_max	Intercept	Coefficient
0	309	8.1115e-08	2.8054e-10
309	3636	1.1293e-07	1.4141e-10
3636	24933	3.8661e-07	9.1019e-12
24933	737529836	1.6517e-07	3.0364e-16
737529836	inf	1.6517e-07	3.0364e-16



Modeling and simulation of MPI communication over IB

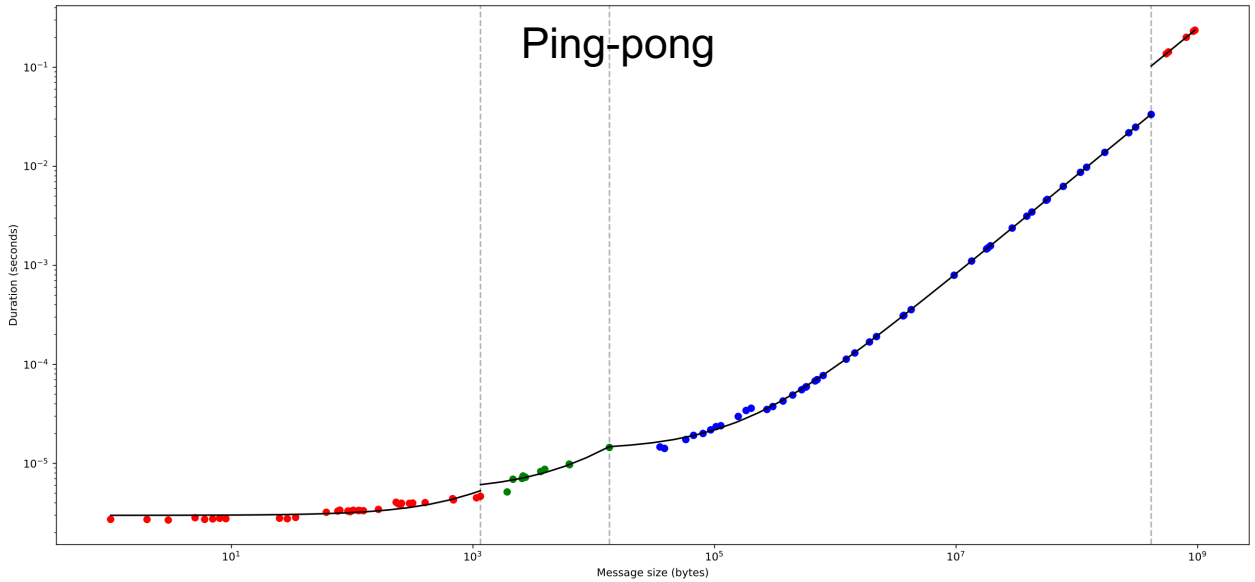
4- Segmentation:

- Calculating the physical latency and bandwidth using ping-pong benchmark:
- **Recall:**
  - For small message sizes:

    Communication_cost=Coefficient × *msg_size* + *Intercept*

  - For larger message sizes:

    Communication_cost= msg_size/Bandwidth + Latency



One-way Latency= intercept/2

One-way Bandwidth=1/ (coefficient/2) = 2/ coefficient

Bandwidth: 2.251149e+10 Bs

Latency: 2.364632e-06 s

Modeling and simulation of MPI communication over IB

4- Segmentation:

- Using Ping-Pong Benchmark to calculate the latency and bandwidth factors:
  - Multipliers to apply to the physical latency&bandwidth (i.e., the one described in the platform) of the set of links involved in a communication.
  - Expected_Bandwidth and Expected_Latency are the optimal values which are estimated by the user.

$$\text{Bandwidth\_Factor} = \frac{1}{Coefficient \times Expected\_Bandwidth}$$

$$\text{Latency\_Factor} = \frac{Intercept}{Expected\_Latency}$$

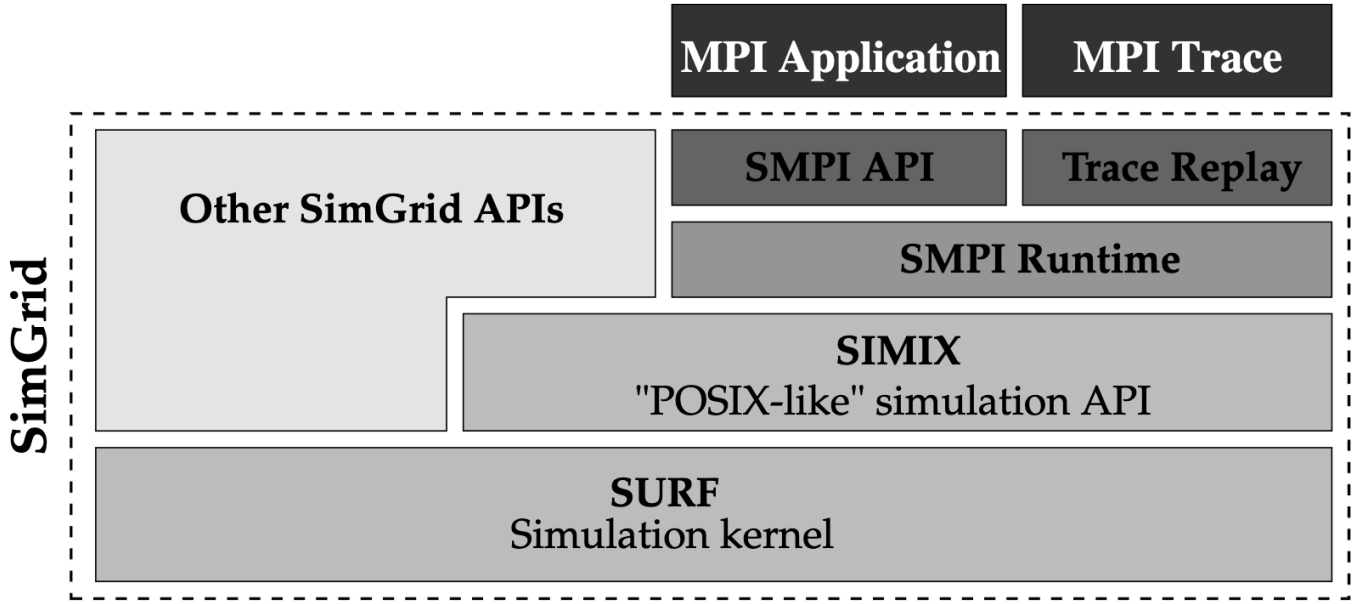
Modeling and simulation of MPI communication over IB

5- Simulation and validation:

- The SMPI interface facilitates the examination of MPI applications by emulating them in the top of SimGrid simulator.

- SMPI support a large fraction of the MPI interface:
  - **ompi:** default selection logic of OpenMPI (version 4.1.2)
  - **mpich**: default selection logic of MPICH (version 3.3b)
  - **mvapich2**: selection logic of MVAPICH2 (version 1.9) tuned on the Stampede cluster
  - **impi**: preliminary version of an Intel MPI selector (version 4.1.3, also tuned for the Stampede cluster).

- For our experiments we used OMPI 5.0.3 to collect the durations of MPI calls in the real system



Modeling and simulation of MPI communication over IB

5- Simulation and validation:

```
> smpicc tests.c -o model.out

> smpirun -platform model.xml -hostfile hostfile  \
    --cfg=smpi/os:"0:7.2255e-08:1.3864e-10;1920:2.5349e-07 :8.9679e-11;18947:5.1178e-06:4.1525e-11;443618990:0.0:0.0" \
    --cfg=smpi/or:"0:1.7229e-07:5.0202e-11; 18947:0:0; 226629:0:0;443618990:0.0:0.0" \
    --cfg=smpi/ois:"0:8.1115e-08 :2.8054e-10;309:1.1293e-07 :1.4141e-10;3636:3.8661e-07:9.1019e-12;24933:1.6517e-07 :3.0364e-16" \
    --cfg=smpi/async-small-thresh:18947 \
    --cfg=smpi/wtime:2.4601346799999998e-08 \
    --cfg=smpi/iprobe:1.4022049861649997e-08\
    --cfg=smpi/test:4.892816343545e-08 \
    --cfg=smpi/send-is-detached-thresh:443618990 \
    --cfg=smpi/bw-factor:"18947:0.97;443618990:0.319" \
    --cfg=smpi/lat-factor:"18947:67.31 ;443618990:1112.83" \
    --cfg=network/model:SMPI \
    ./model.out
```

MPI_Send

MPI_Recv

MPI_Isend

Asynchronous mode threshold

Constants for MPI_Wtime, MPI_Iprobe and MPI_Test

Detached mode threshold

Bandwidth factors

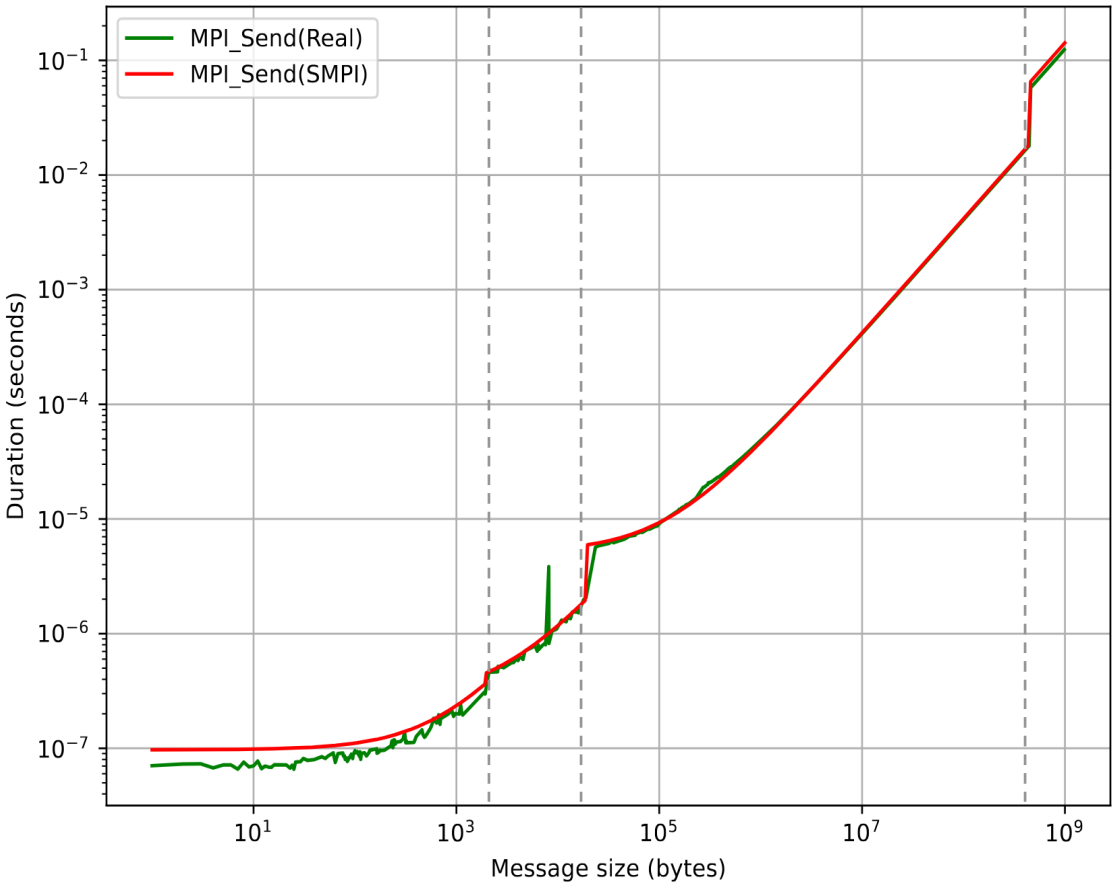
Latency factors

Modeling and simulation of MPI communication over IB

5- Simulation and validation:

- For MPI_Send, the model aligns with the reality for all communication modes.

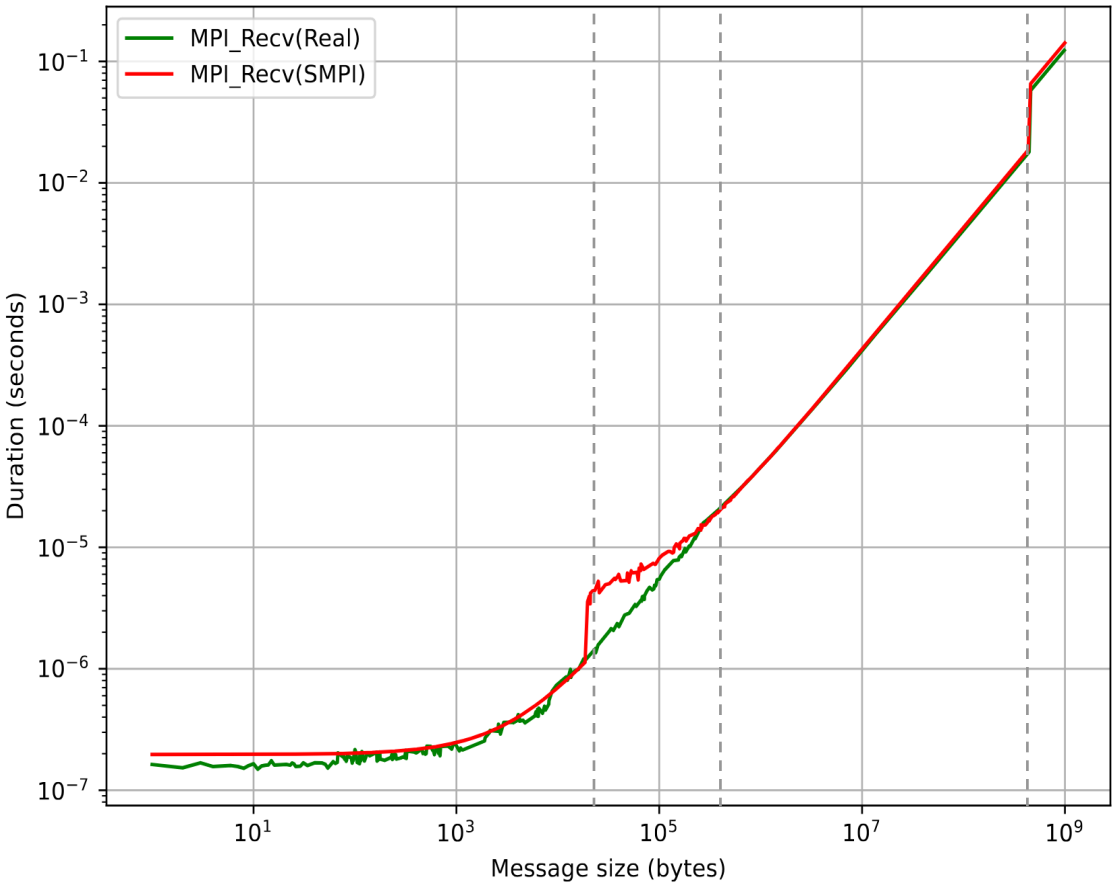
- For small message sizes, the model deviates from the reality as the MPI_Send relies totally on the CPU overhead (Duration may vary at each iteration even for the same message size).



Modeling and simulation of MPI communication over IB

5- Simulation and validation:

- For MPI_Recv, the model was consistent with the real experiment, except for the start of the detached mode.
- In real experiment, MPI_Send is still waiting for the receiver to post MPI_Recv(Unexpected eager receive).
- In simulation, SMPI already injected with the proper threshold of the detached mode (Expected eager receive).



Contributions and conclusions

- We introduced a piece-wise linear model to simulate MPI P2P communication over InfiniBand network.
- We used real measurements on systems with BlueField-2 (ConnectX-6) NICs to verify the presented model and validate SimGrid simulations.
- The LogGPS hybrid model can be used to represent the communication behavior of MPI over IB.
- The SMPI interface shows exactly how MPI communications would behave optimally.

Thank You!

Leibniz Supercomputing Centre
of the Bavarian Academy of Sciences and Humanities