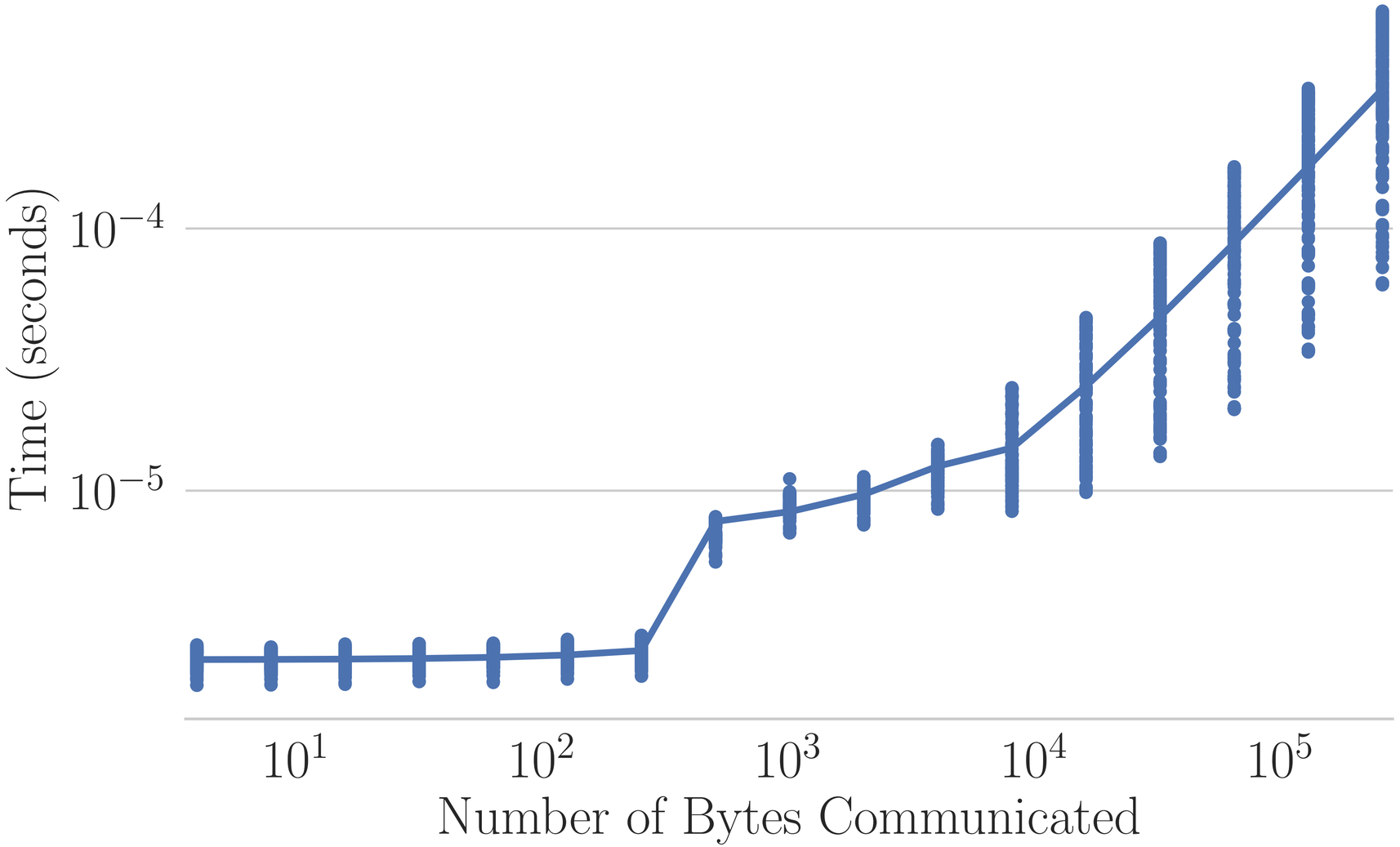
What Parallel Performance Really Looks Like

LANL HPC Summer School

Professor Amanda Bienz
University of New Mexico

Cost of Sending a Single Message



Message Passing

- To communicate a message, all data is split up into packets and the packets are sent through the network to the destination process
- Also, have an envelope that describes the message (size, tag, etc)
- Different protocols for sending messages:

Message Passing

- **Short** : All message data fits in envelope, sent directly to process

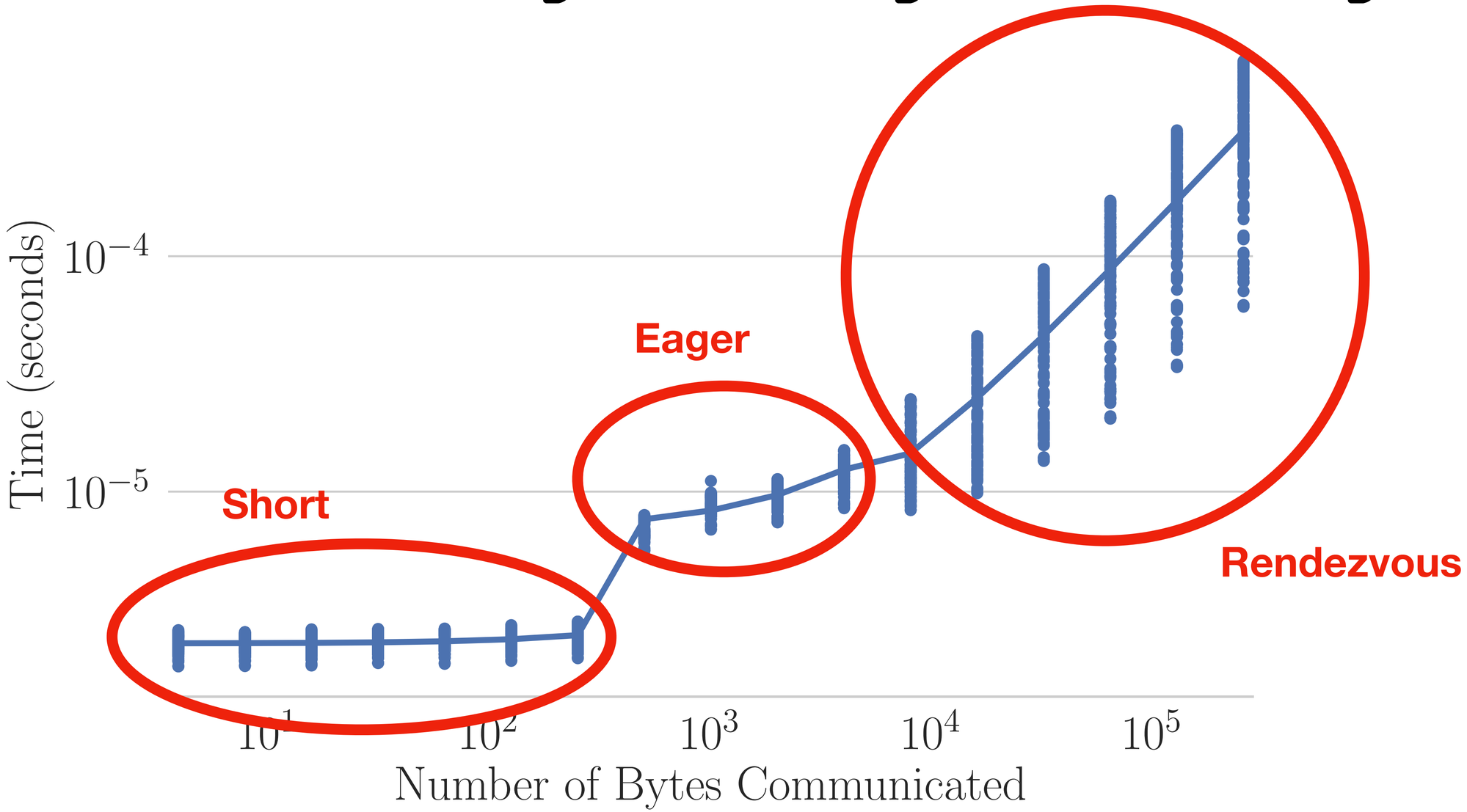
Message Passing

- **Eager :** Message does not fit in envelope, but still relatively small
  - Can assume the receiving process has buffer space available for this message
  - Pack up and send directly

Message Passing

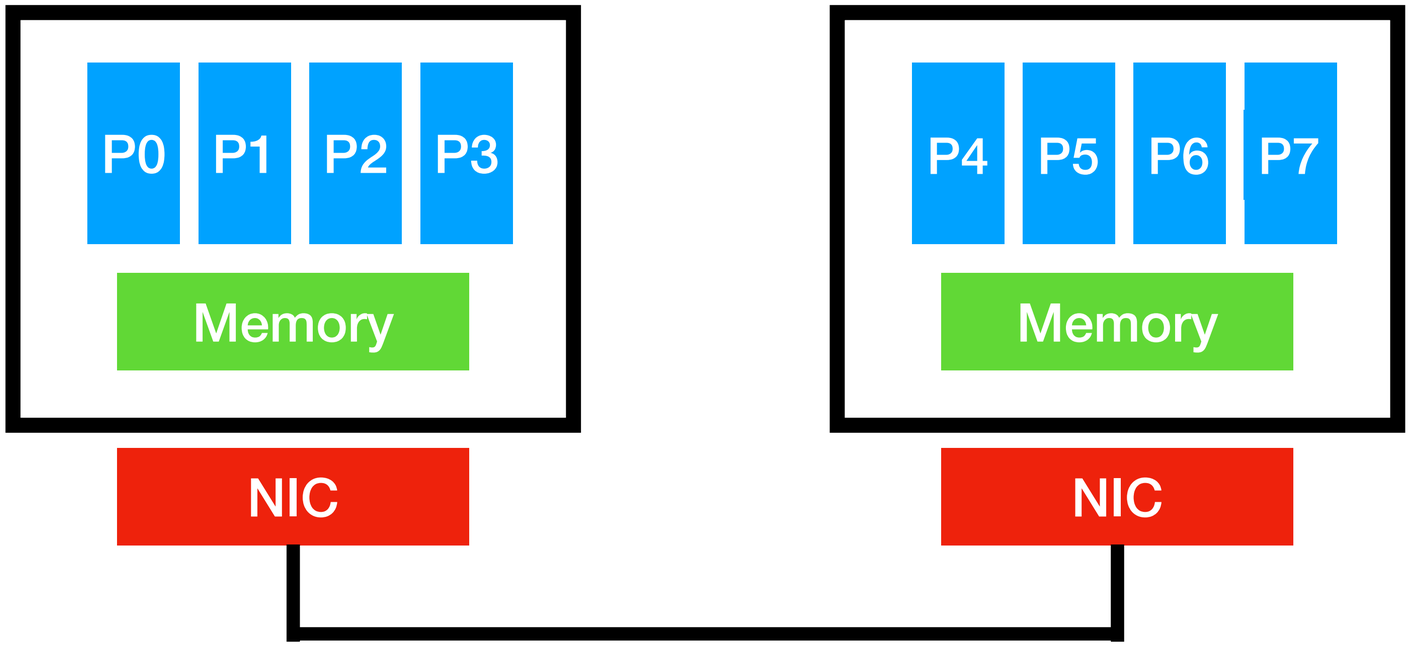
- **Rendezvous :** Largest messages
  - Cannot assume receiving process has buffer space for this message
  - Sending process sends a message to the receiving process, saying it wants to send a message of this size
  - Receiving process allocates the buffer space and sends back a message saying it is ready
  - Only then can sending process send the data

Cost of Sending a Single Message

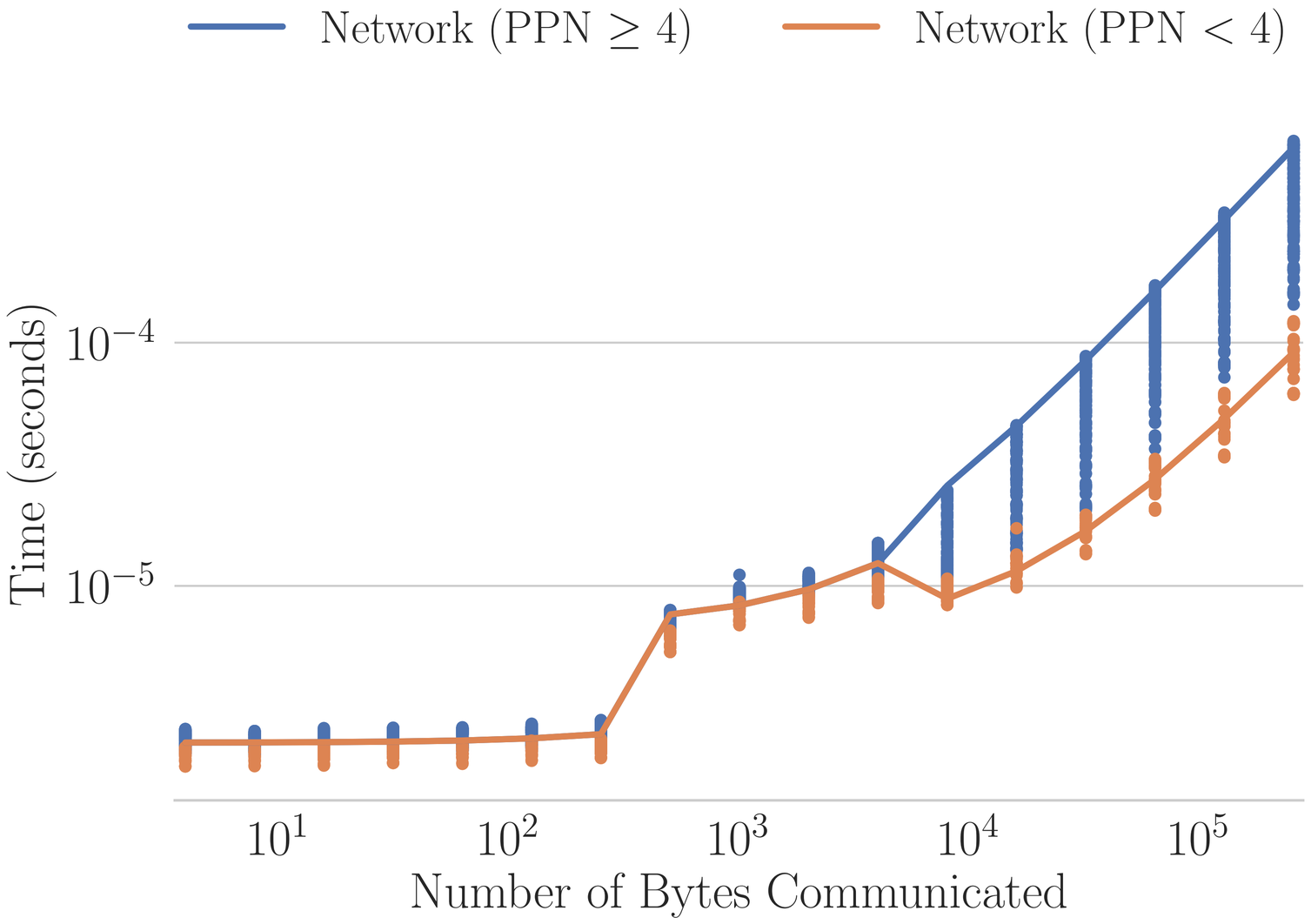


Supercomputer Architecture : Nodes

- Not actually processes connected to one another
- Supercomputers have symmetric multiprocessing (SMP) nodes
- Many processes per nod
- **Can have multiple processes communicating between nodes at once**

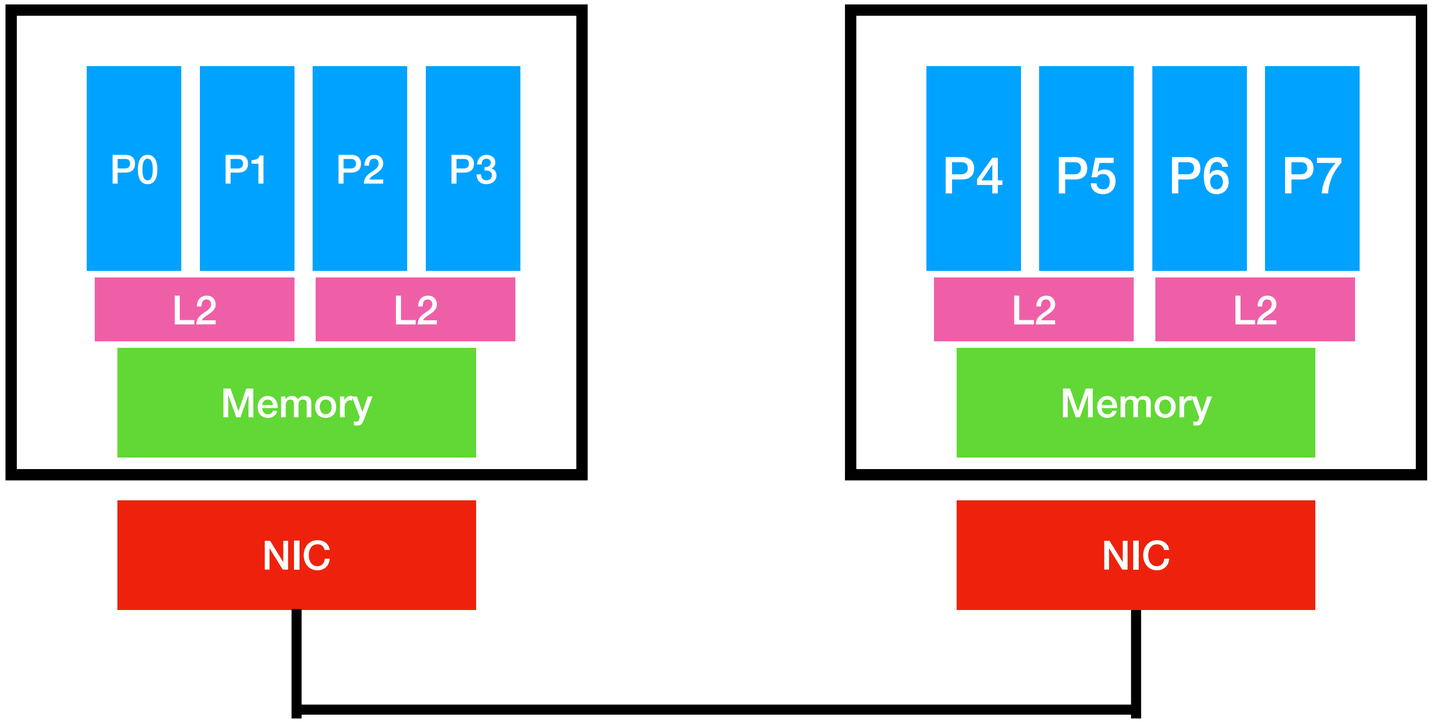


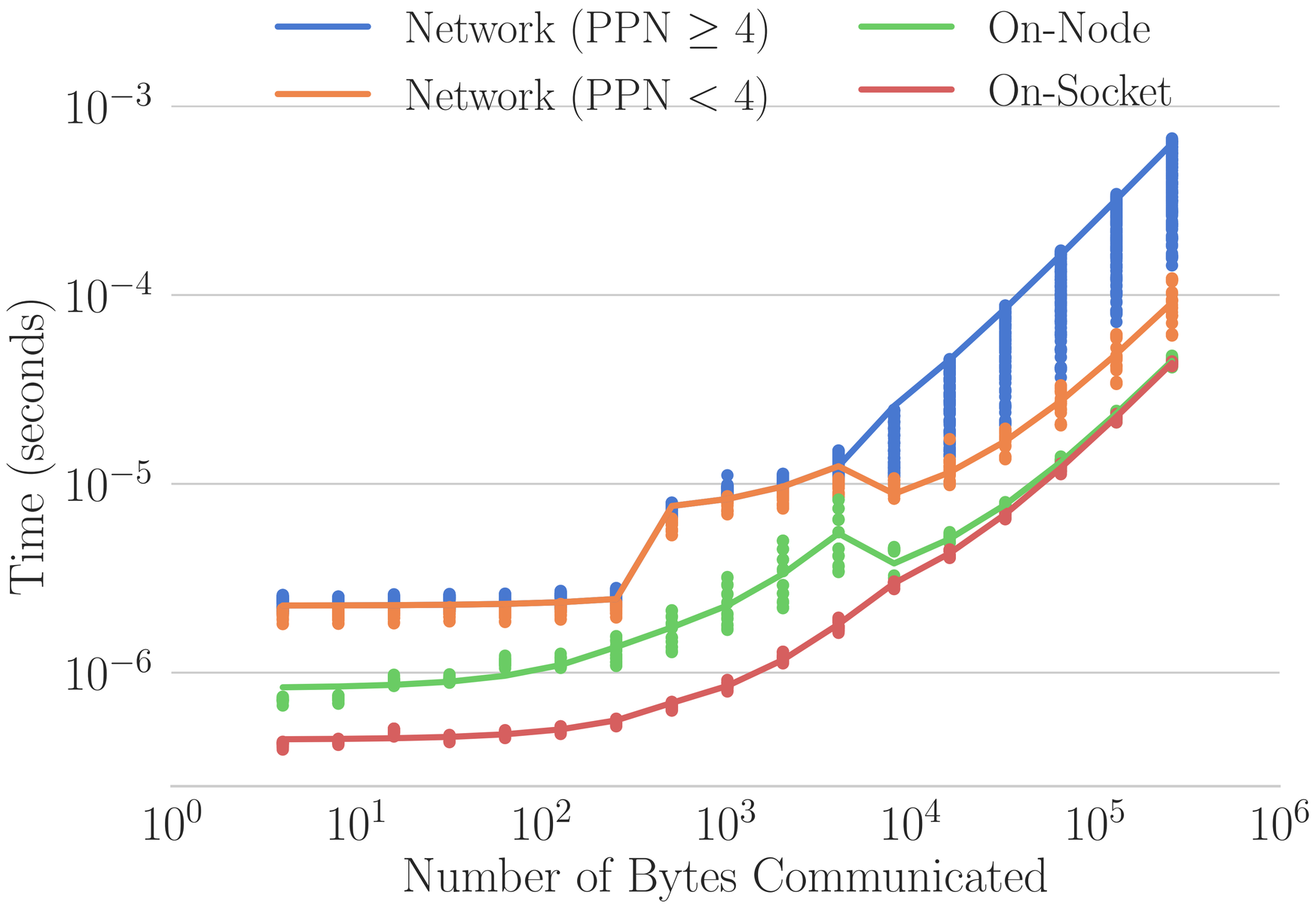
Inter-Node Communication



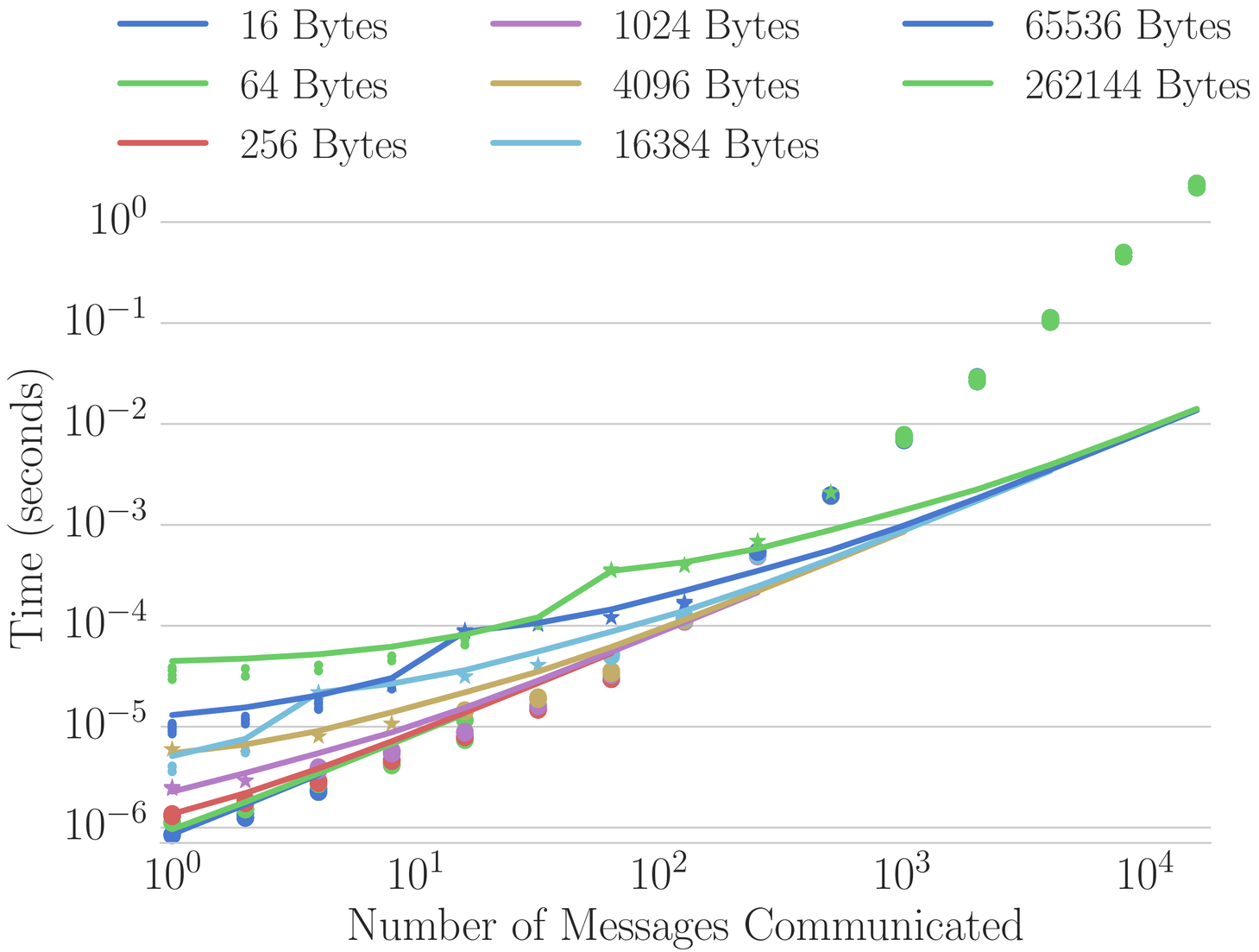
On-Node Communication

- Nodes usually have multiple sockets
- Processes on a socket share *cache*
- **Can have processes communication within a single node, on-socket and off-socket**

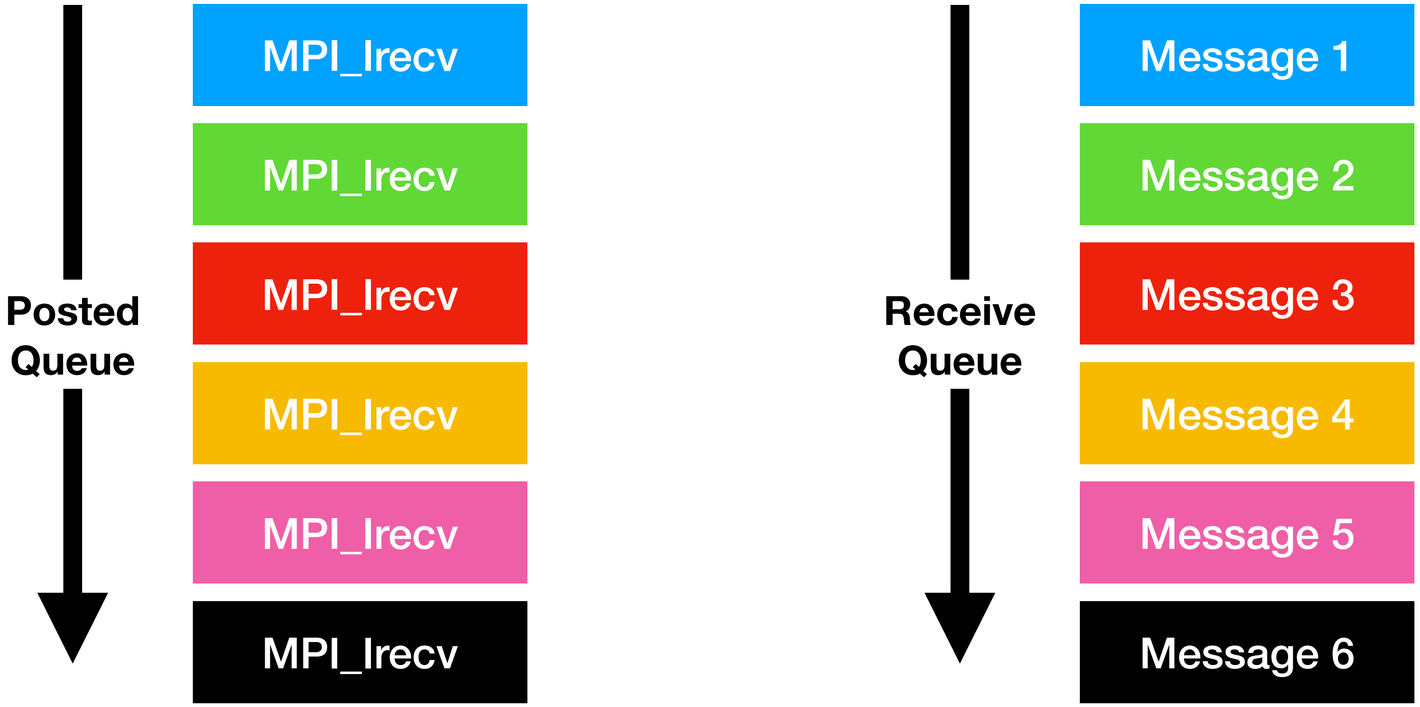



Network (PPN ≥ 4)
Network (PPN < 4)
On-Node
On-Socket
Time (seconds)
10^−3
10^−4
10^−5
10^−6
10^0
10^1
10^2
10^3
10^4
10^5
10^6
Number of Bytes Communicated

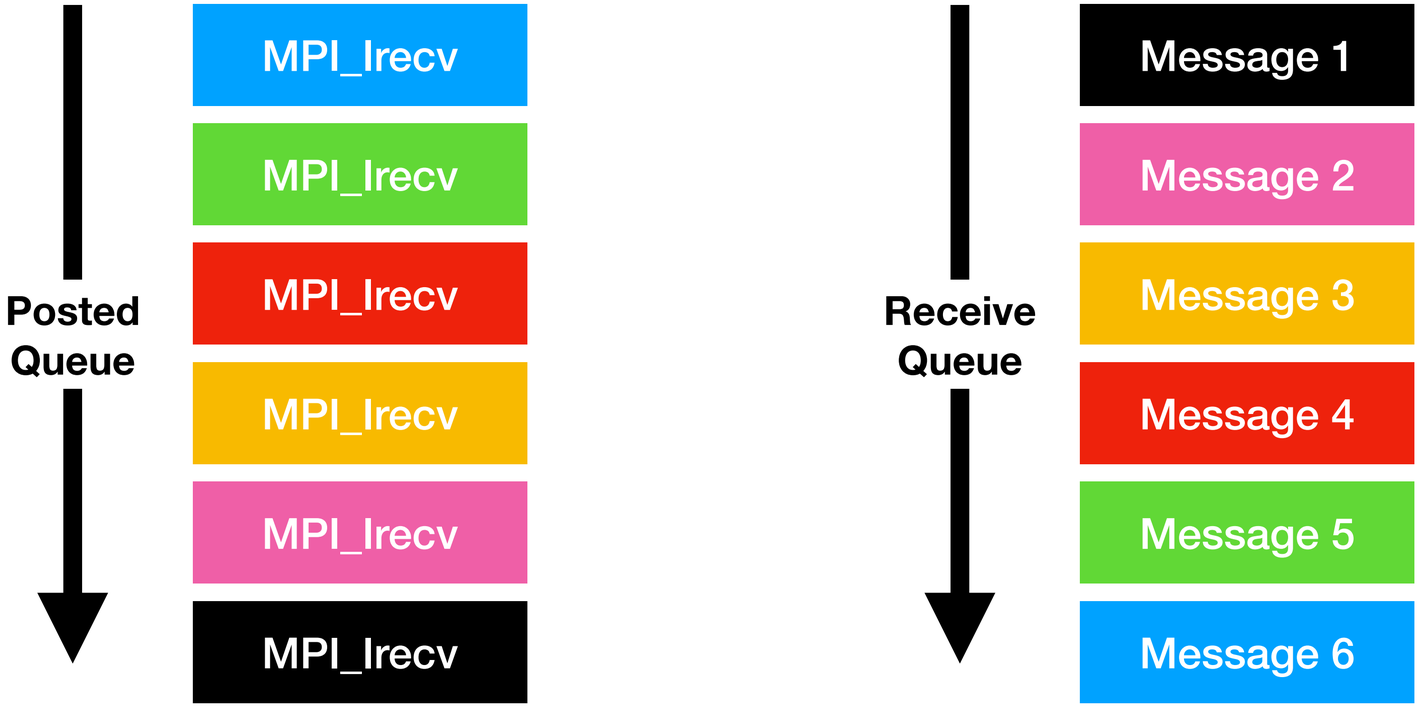
What about large numbers of messages?



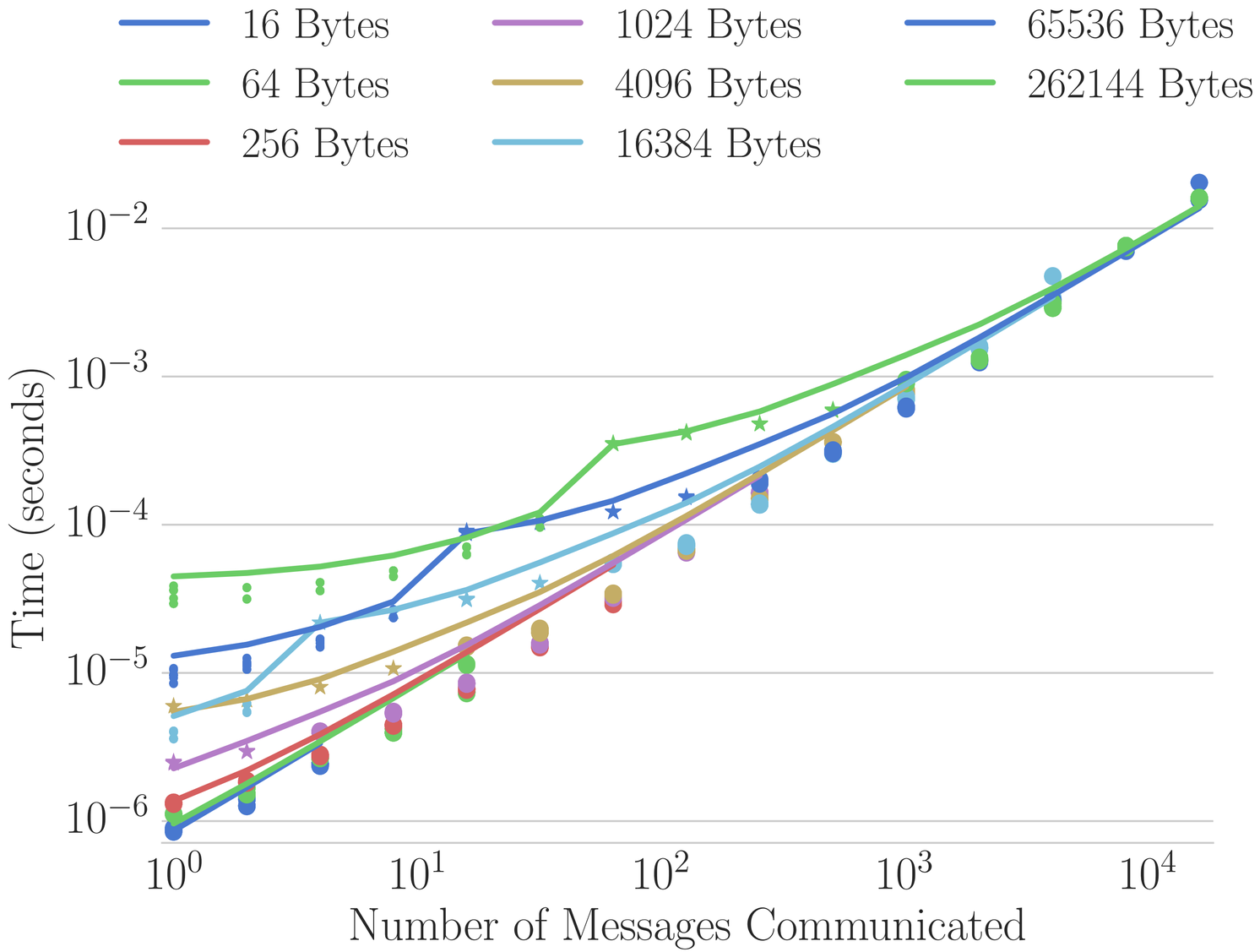
Matching in Receive Queues



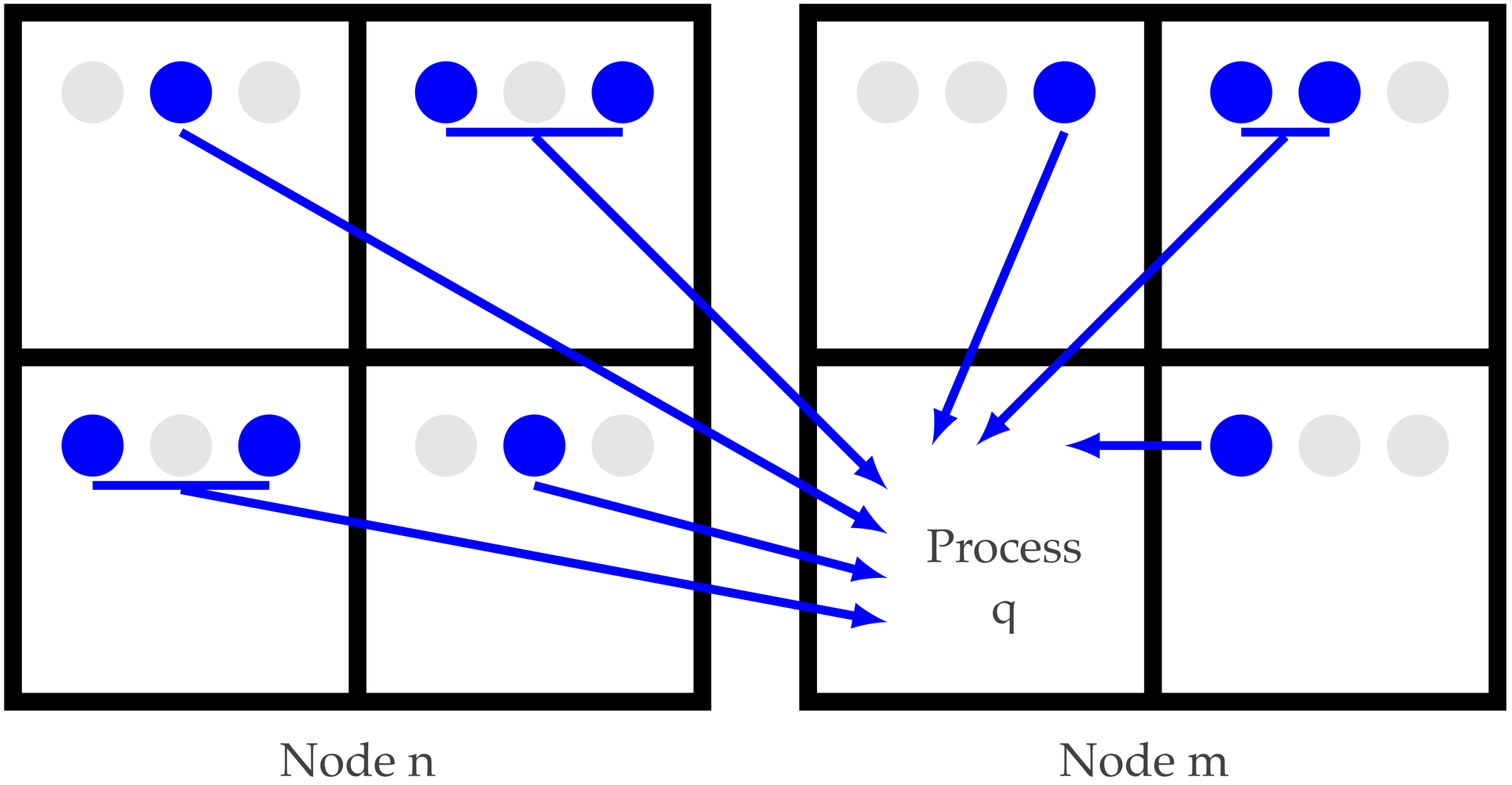
Matching in Receive Queues



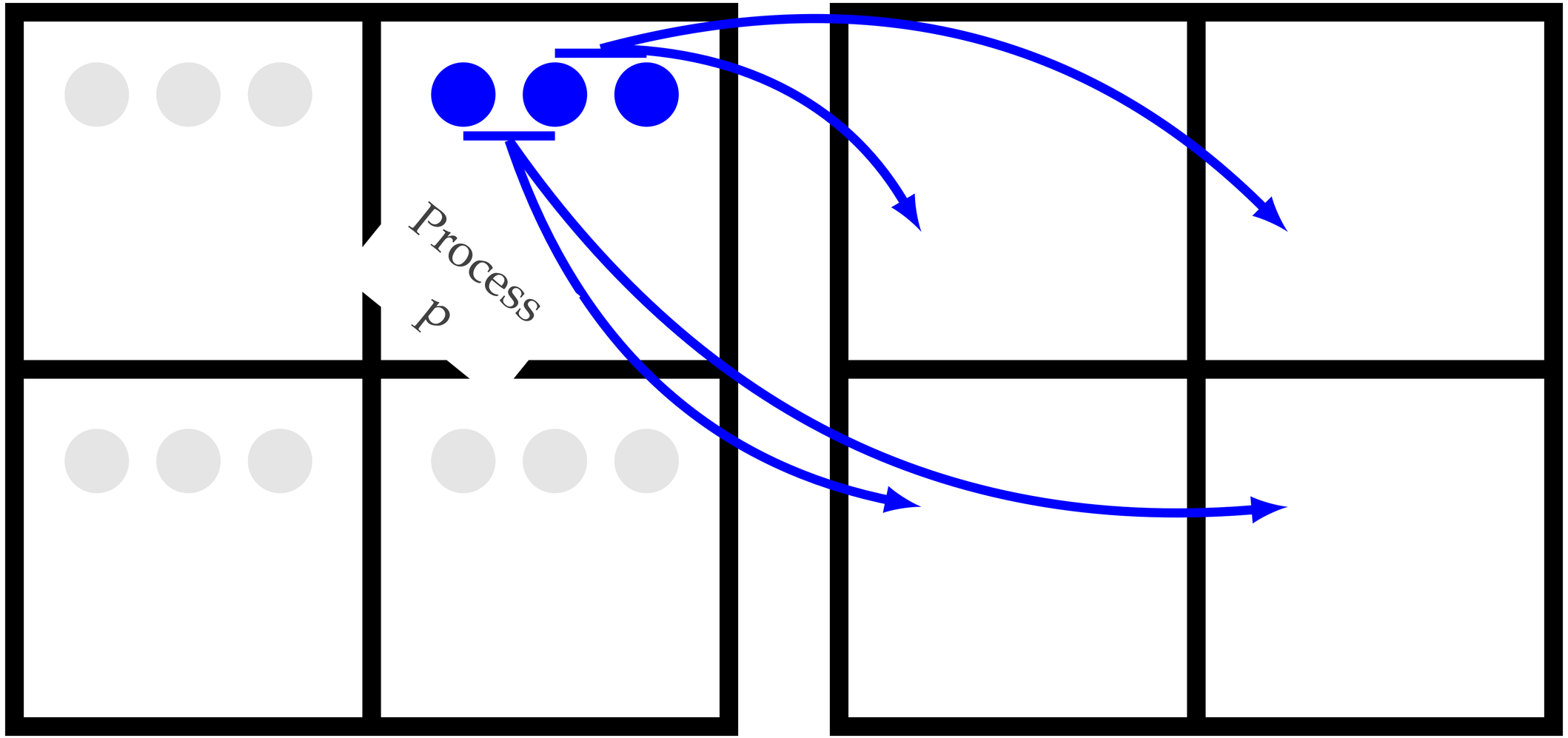
What about large numbers of messages?



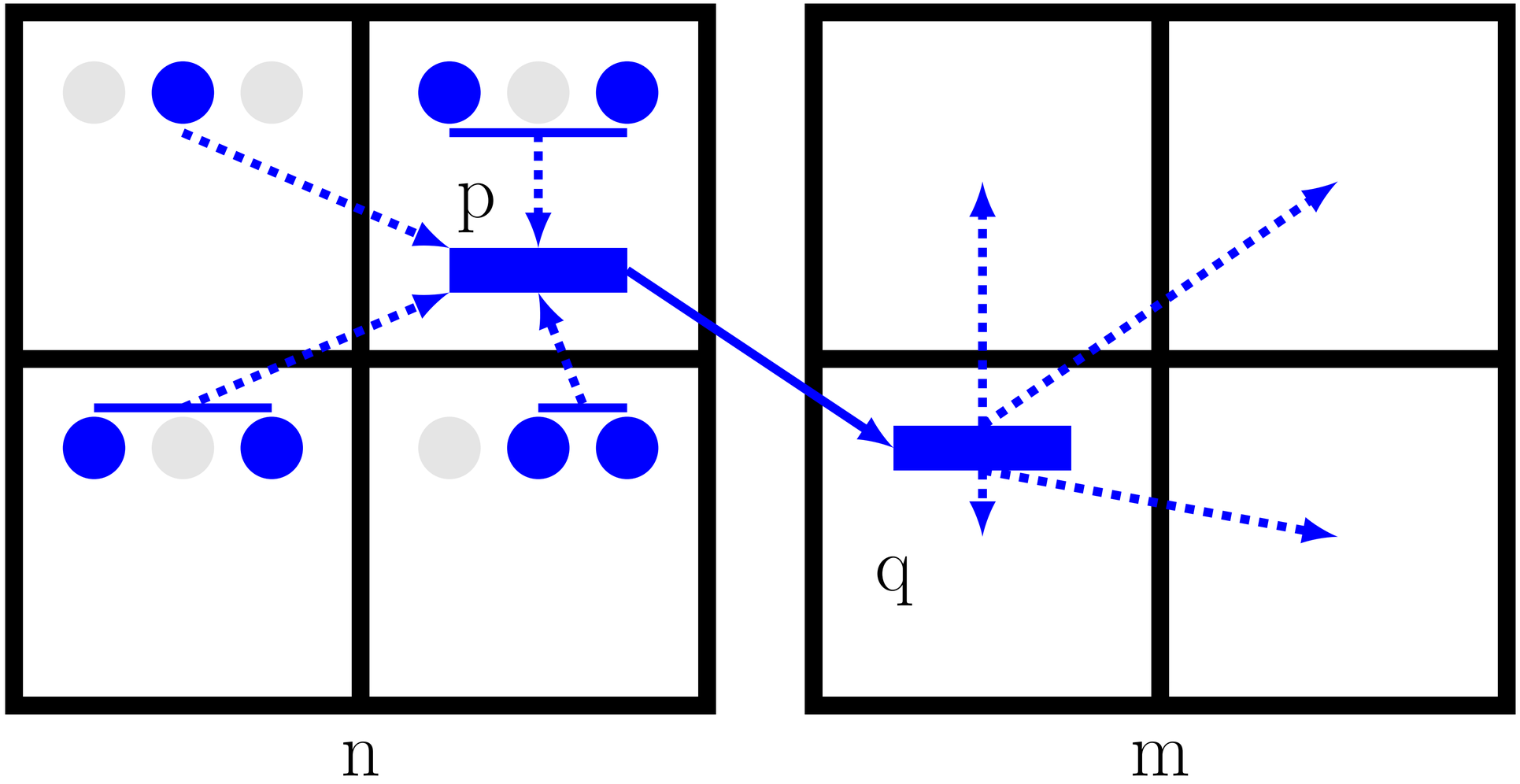
Standard Communication



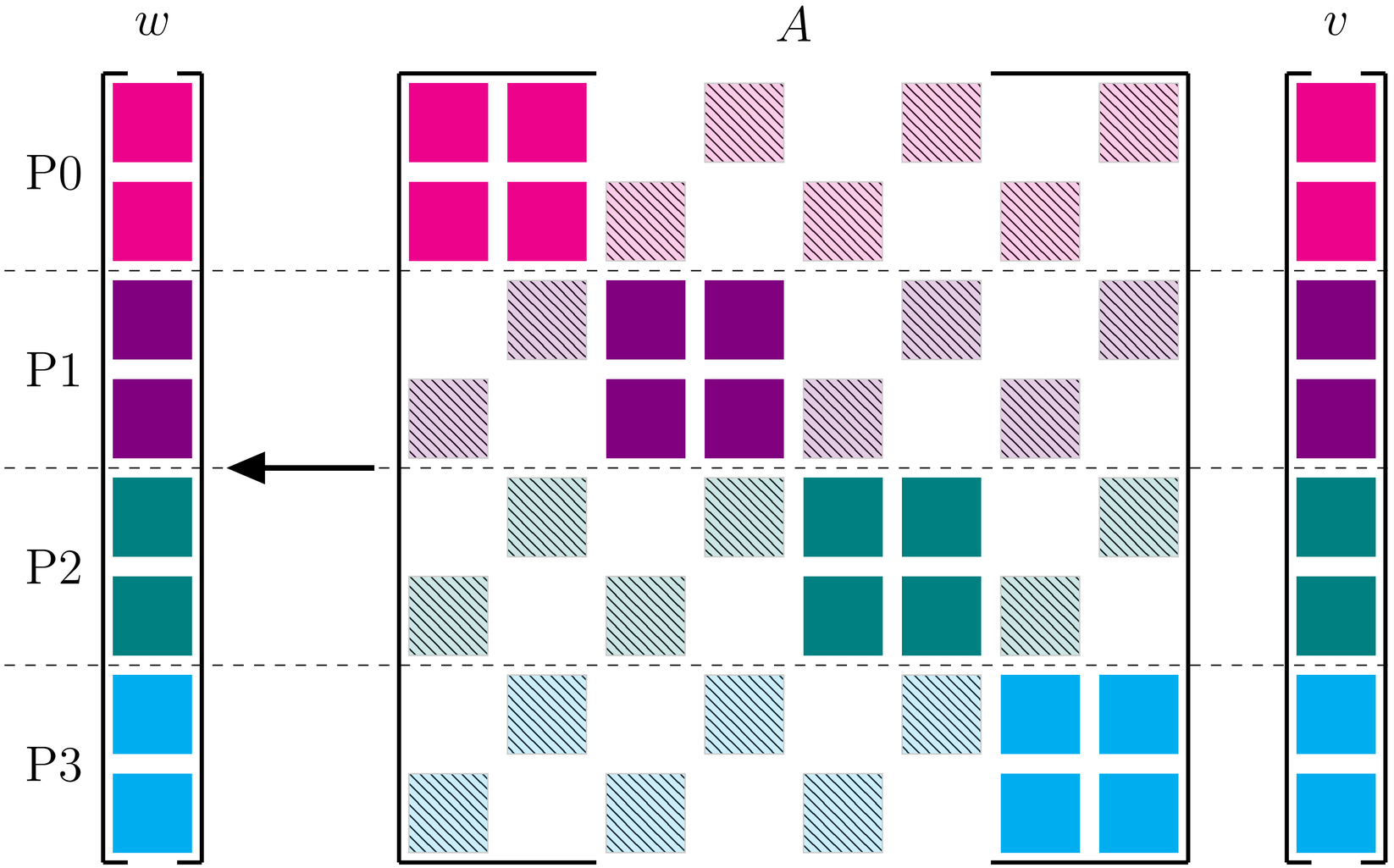
Standard Communication



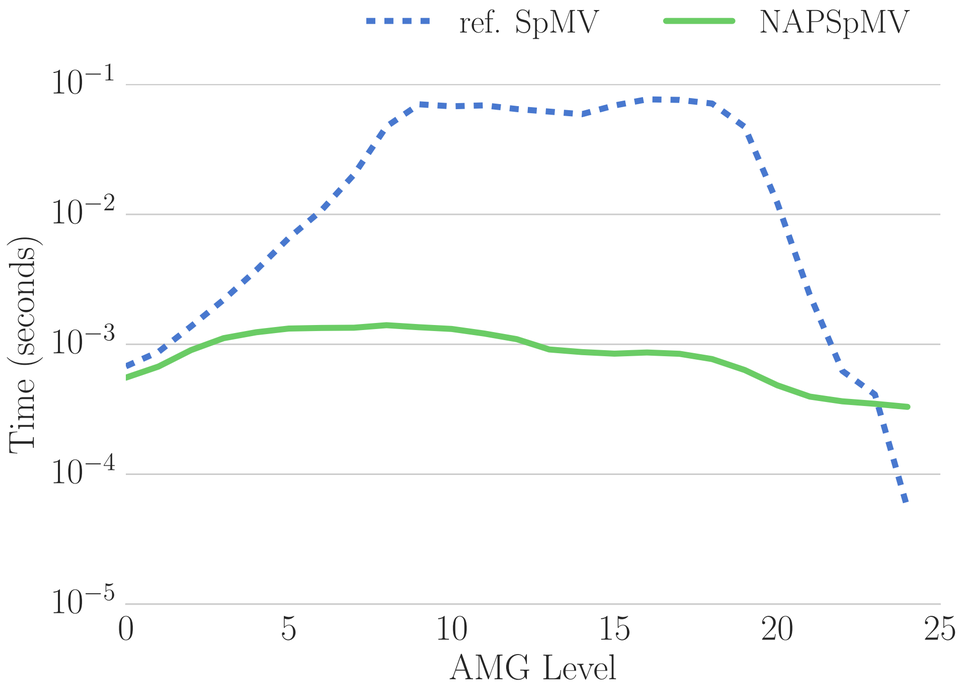
Node-Aware Communication



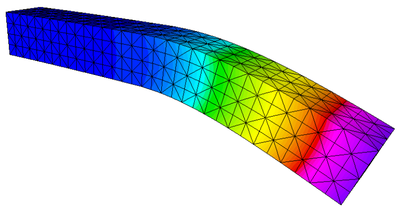
Parallel Sparse Matrix-Vector Multiplication (SpMV)

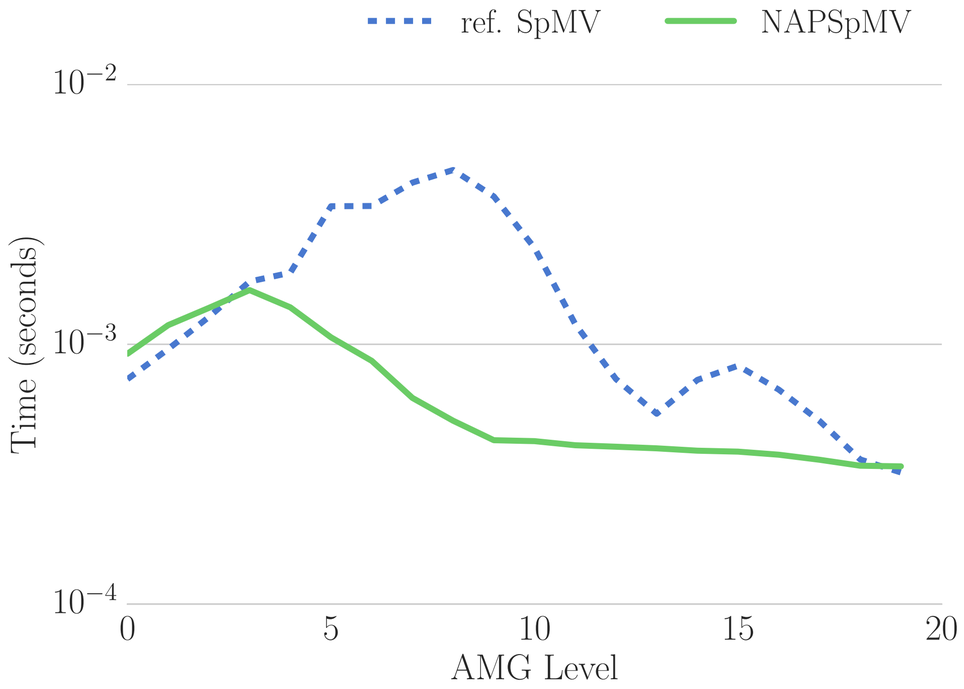


Node-Aware Sparse Matrix-Vector Multiplication

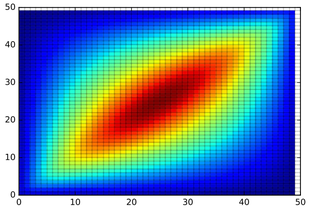


Linear Elasticity (MFEM)

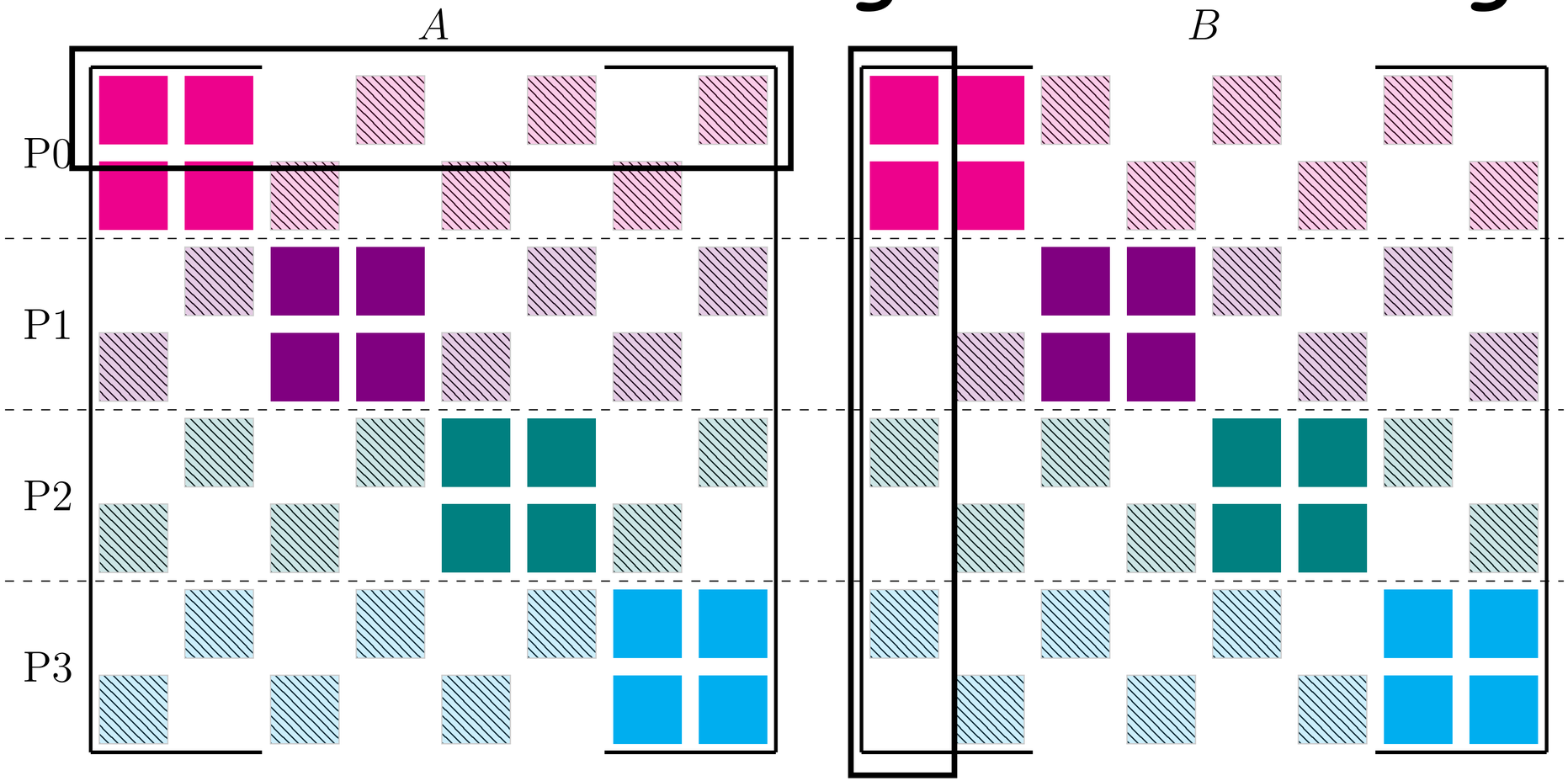




2D Rotated Anisotropic Diffusion



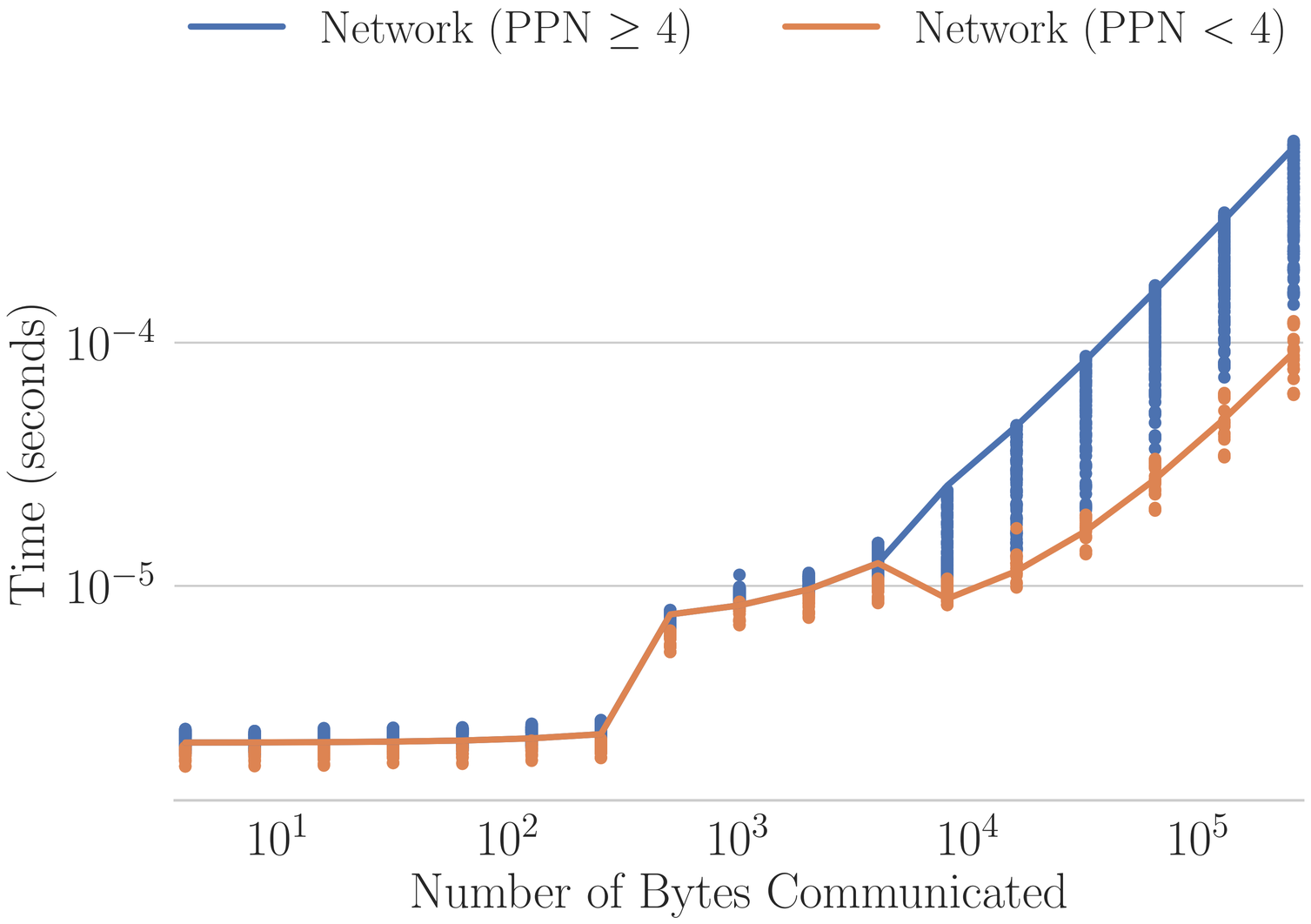
What About Larger Messages?



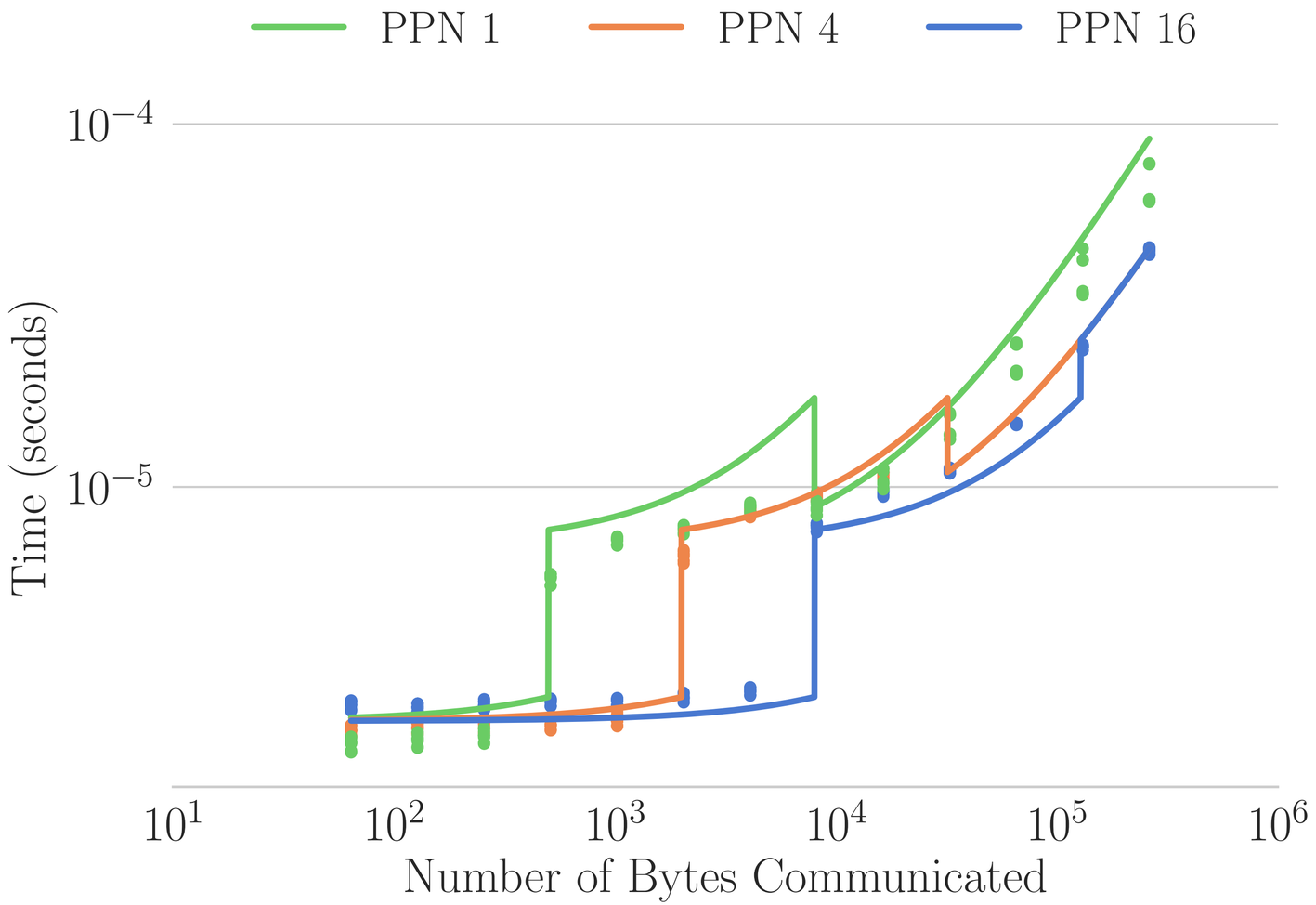
Must communicate entire rows of matrix for each patterned block

Much more data is communicated

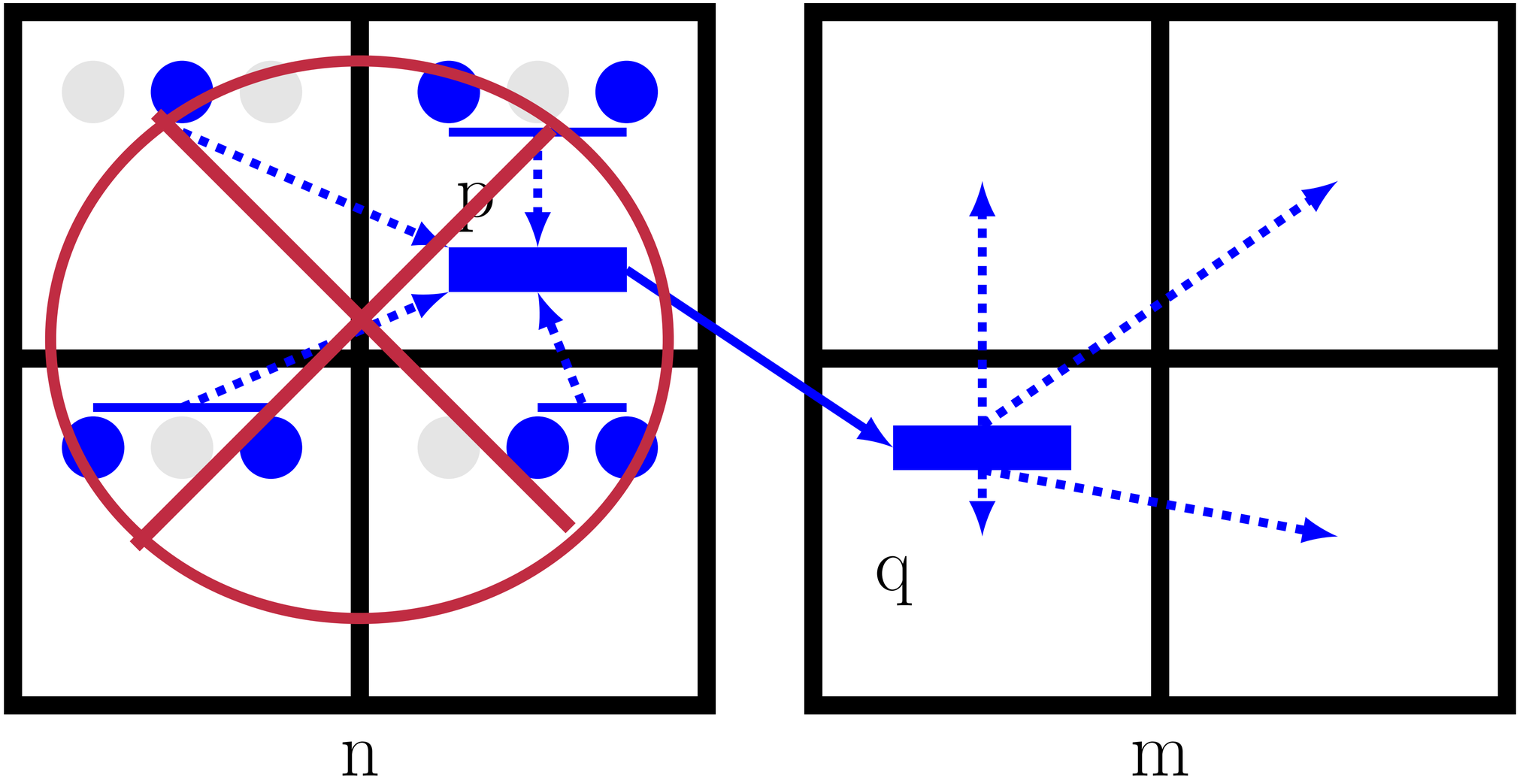
Inter-Node Communication



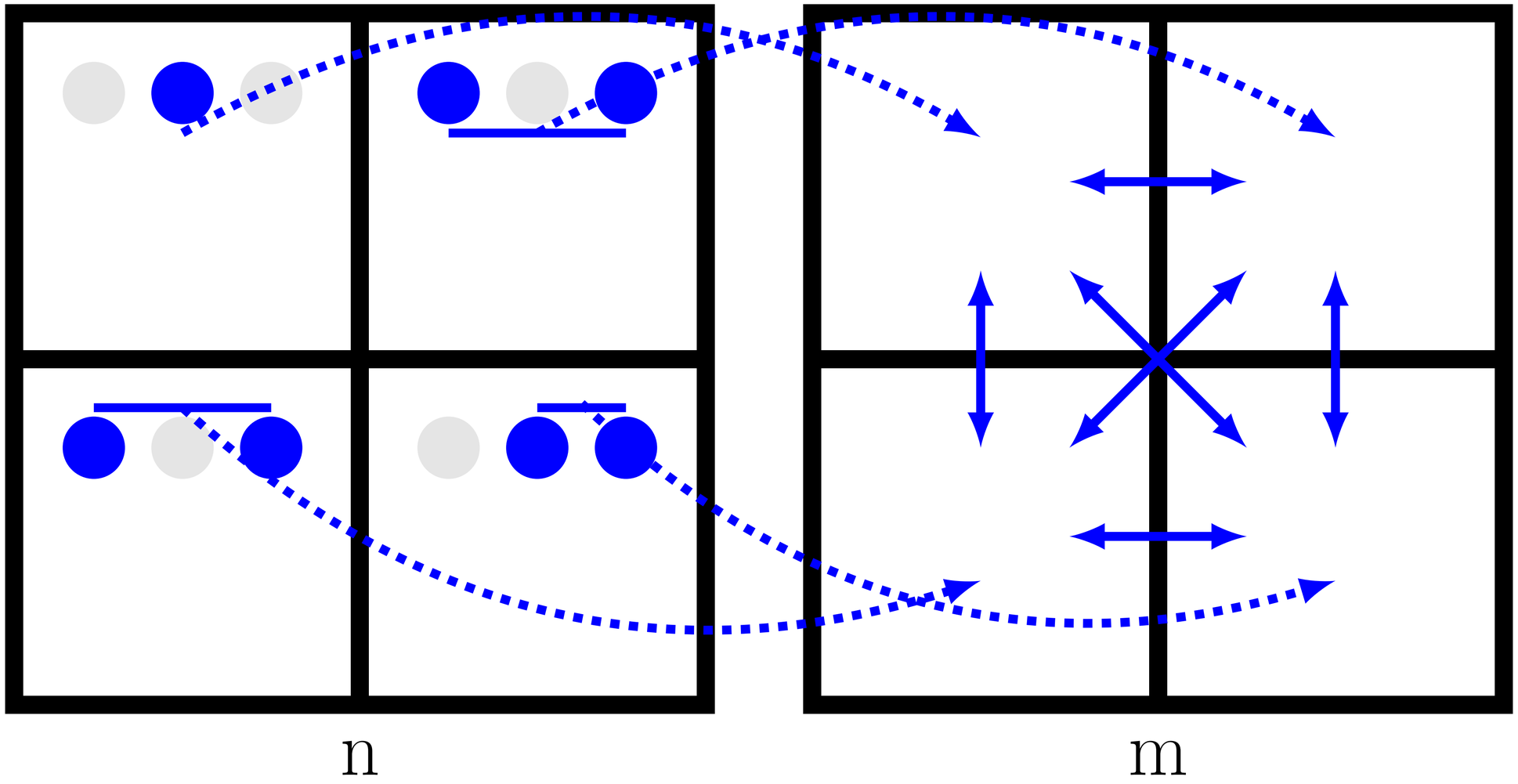
Inter-Node Communication



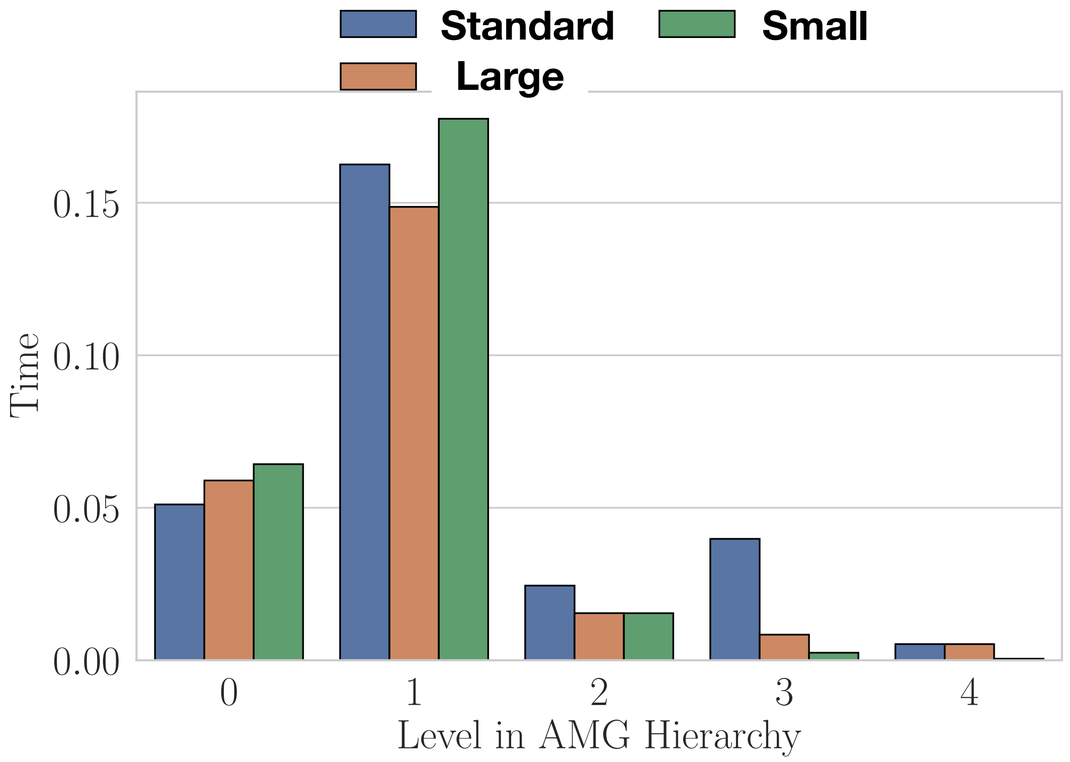
Node-Aware Communication for Large Messages



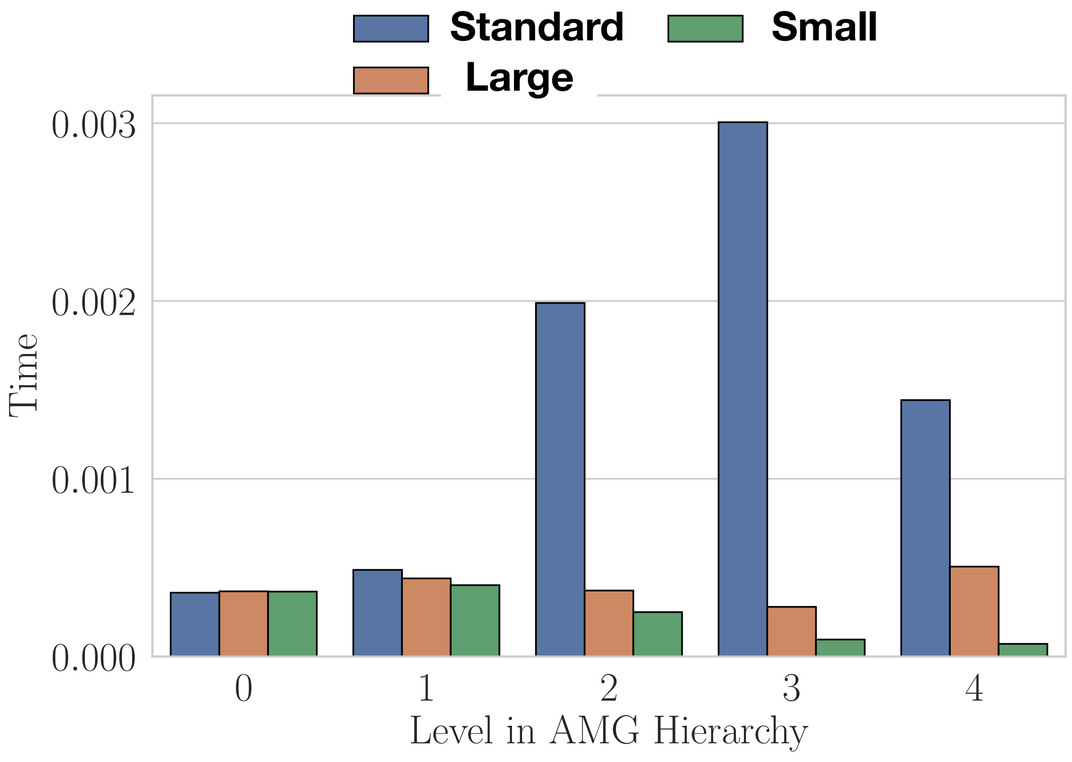
Node-Aware Communication for Large Messages



Node-Aware Communication

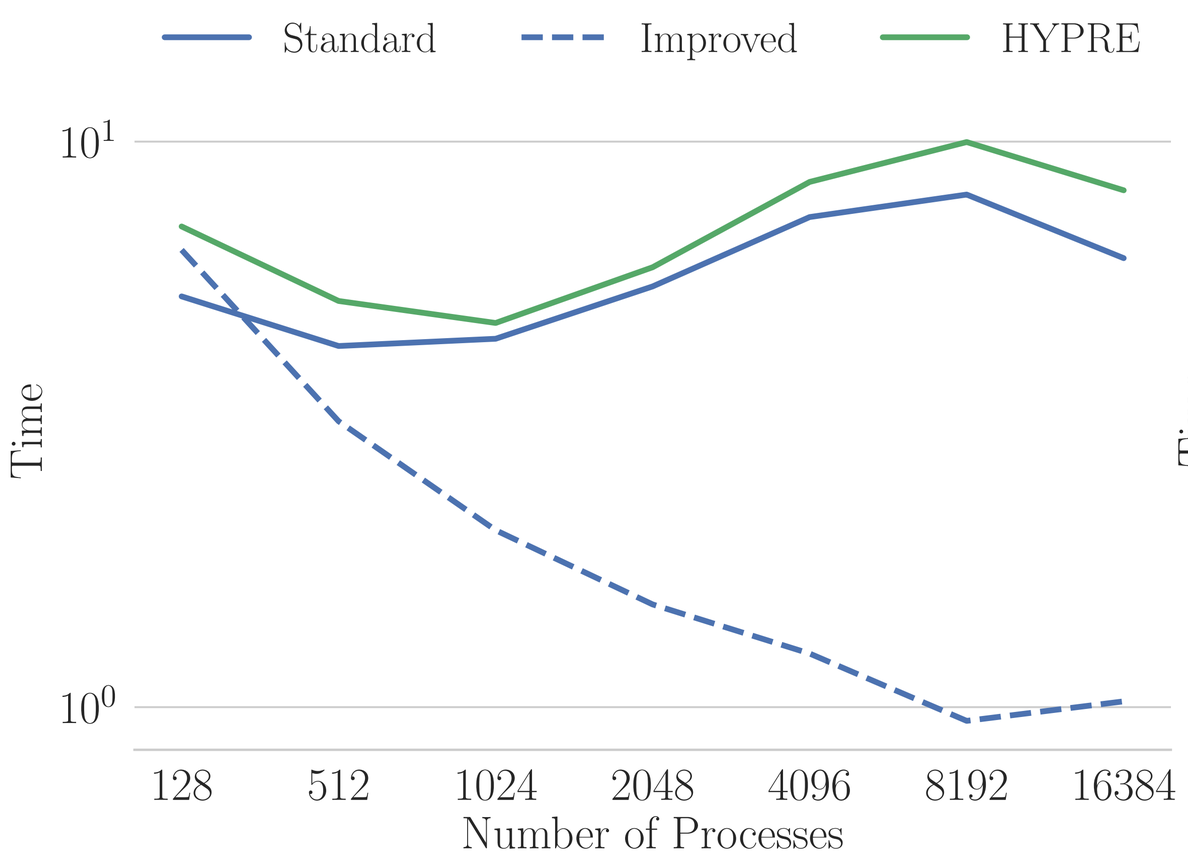


SpGEMM

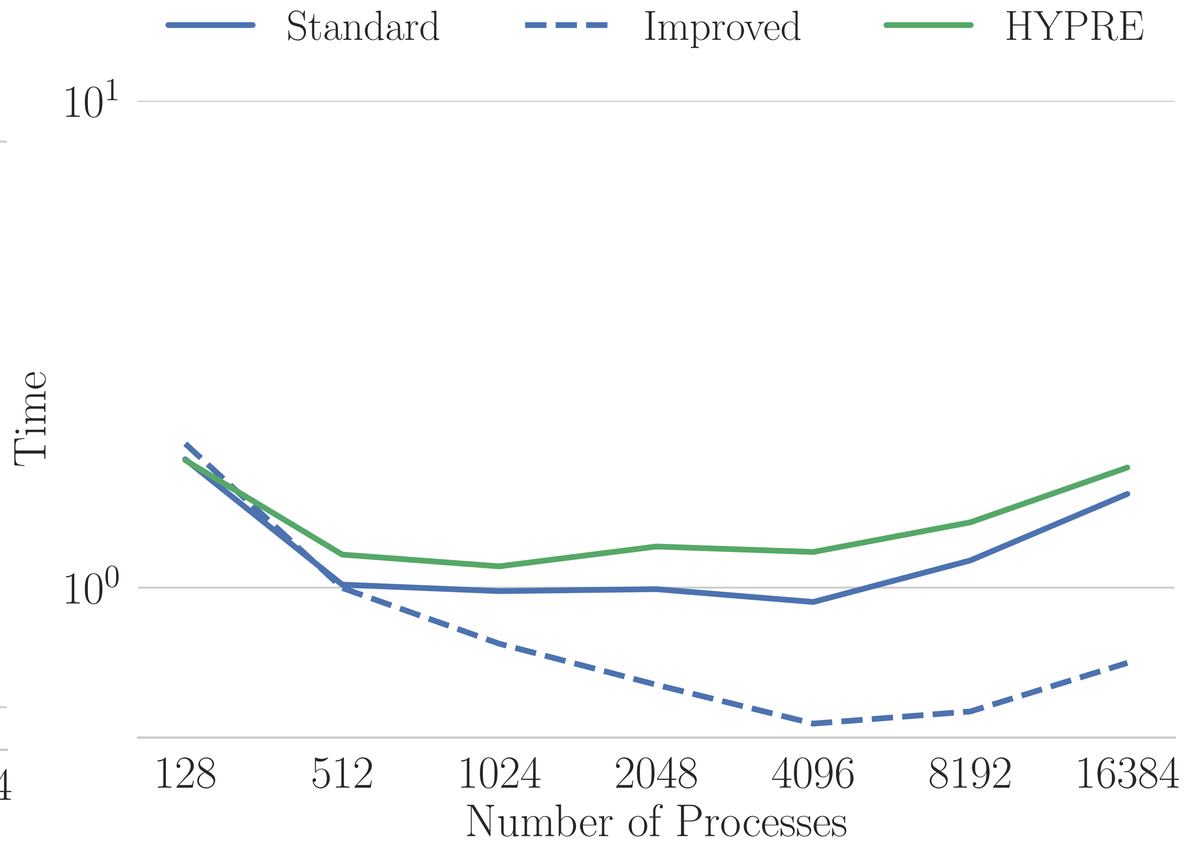


SpMV

Node-Aware Performance

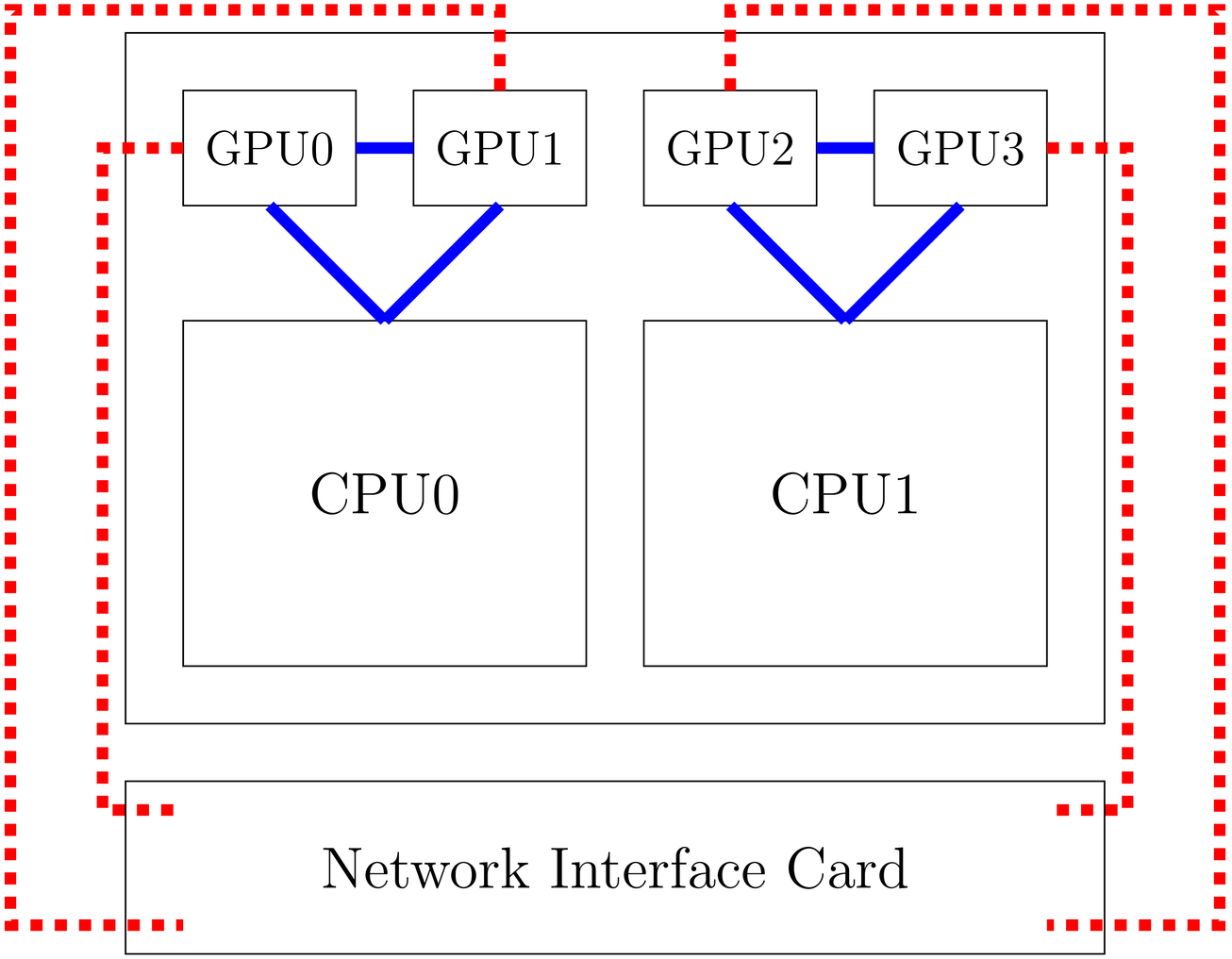


MFEM Laplacian

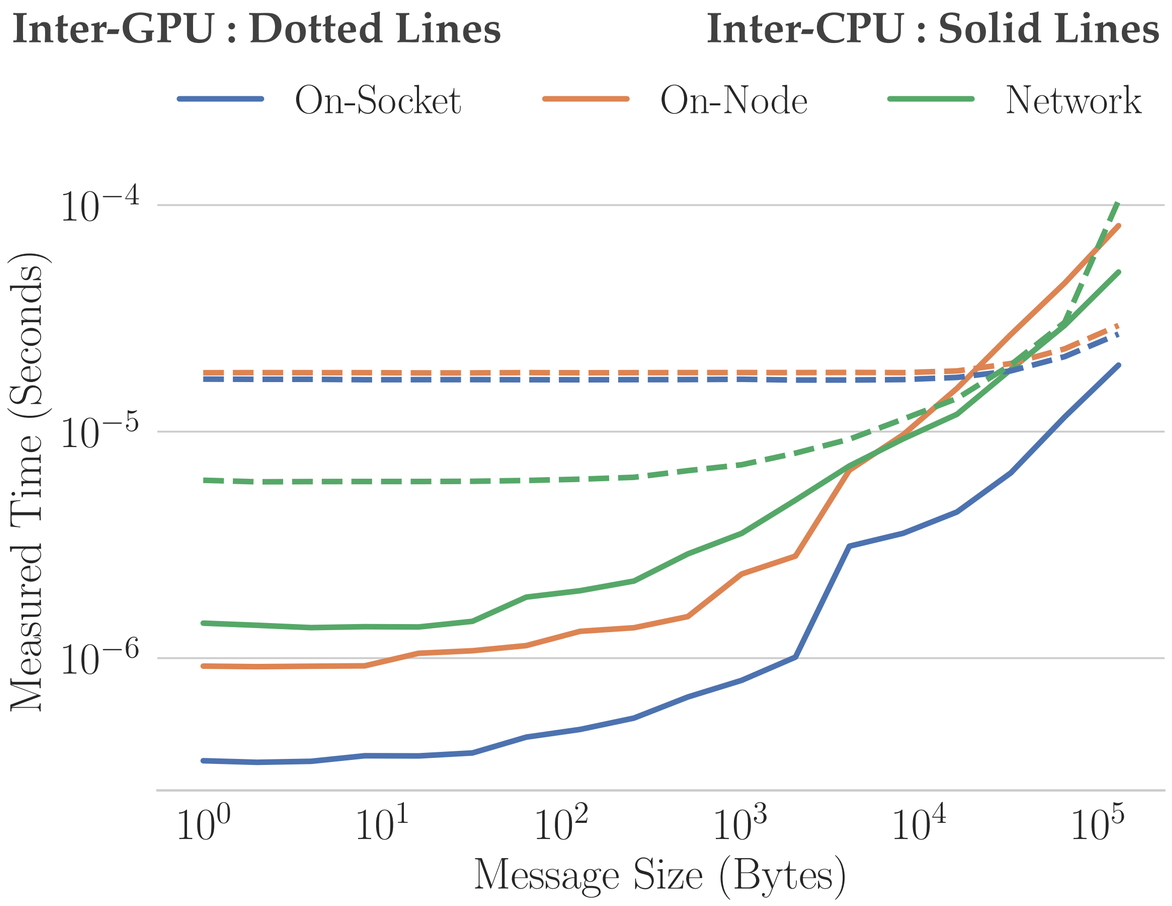


MFEM Grad-Div

Heterogeneous Systems

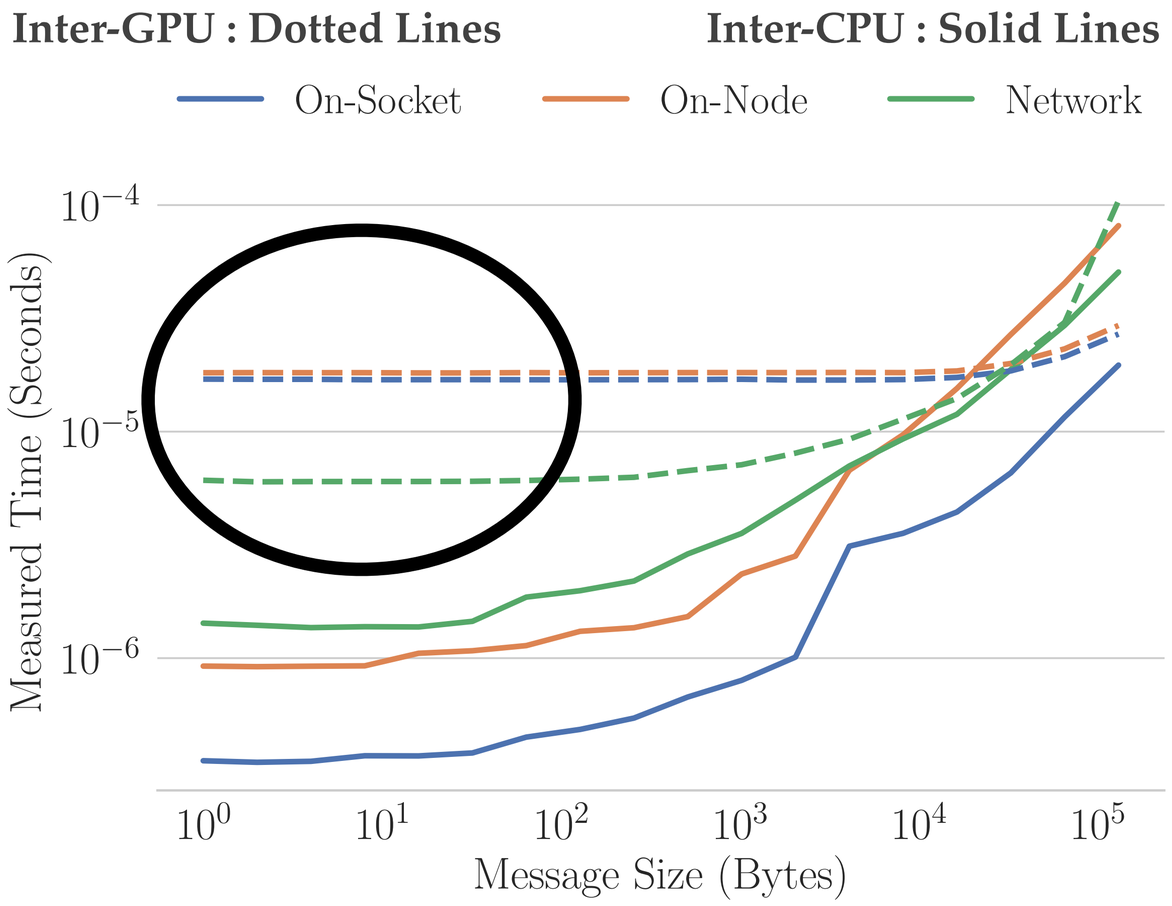


Performance Measurements



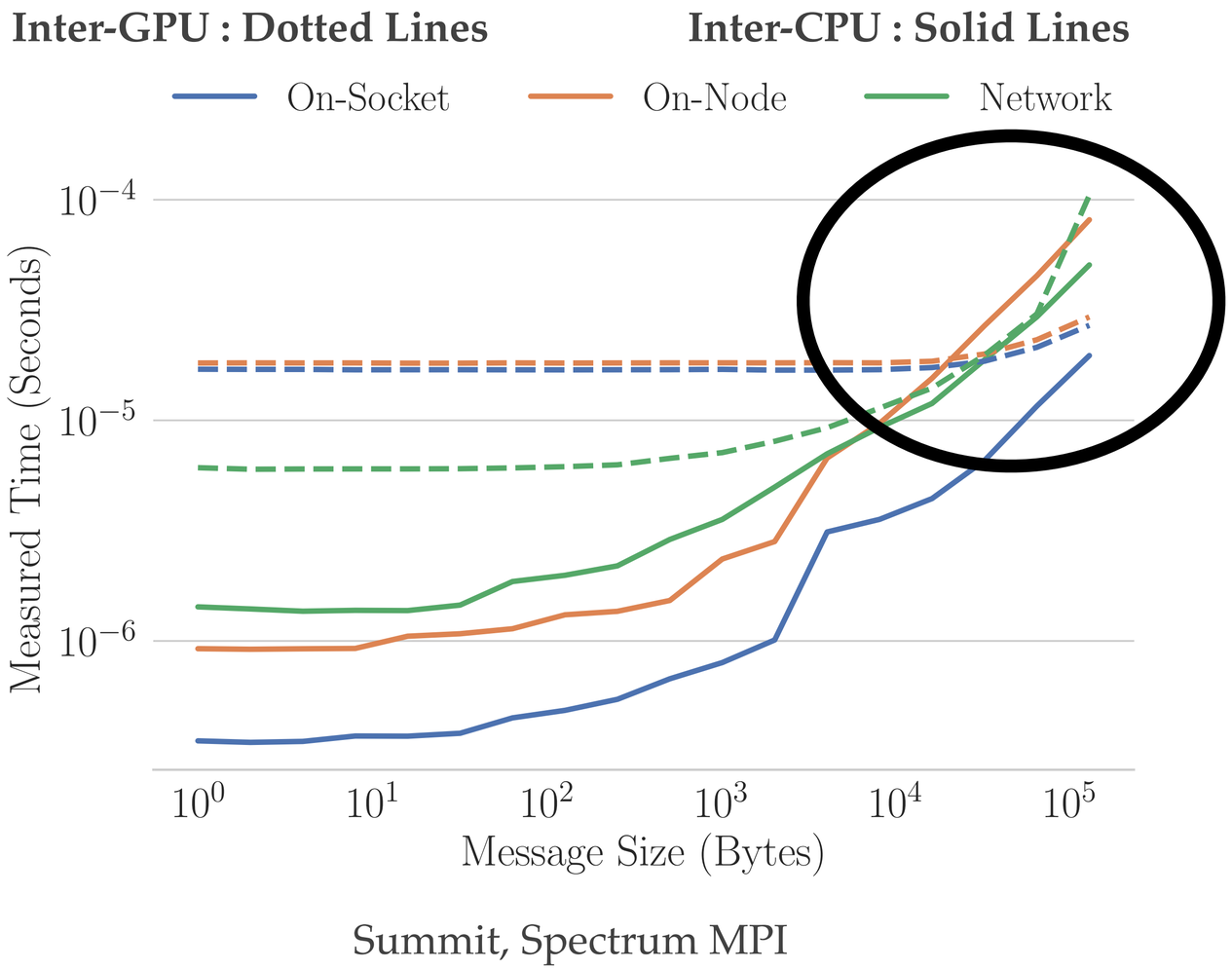
Summit, Spectrum MPI

Performance Measurements



Summit, Spectrum MPI

Performance Measurements

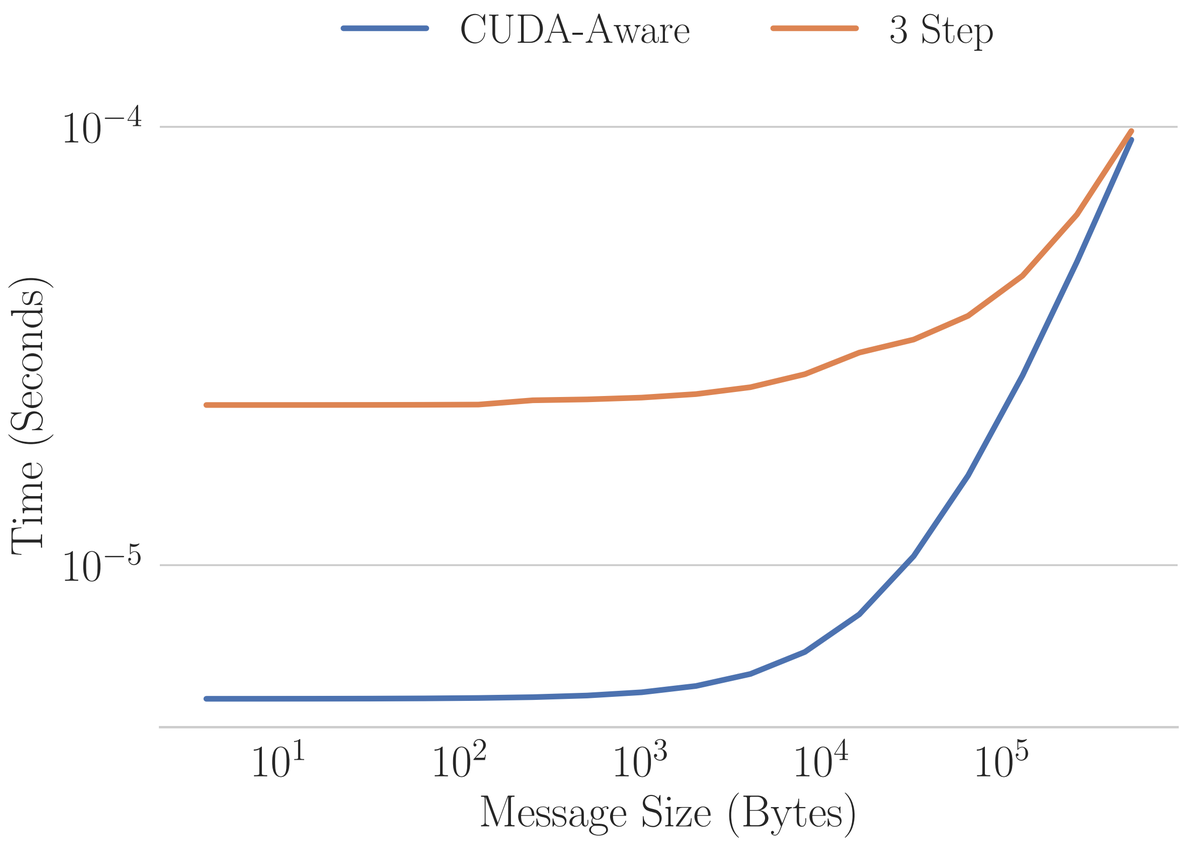


Summit, Spectrum MPI

How to Communicate Data Between GPUs?

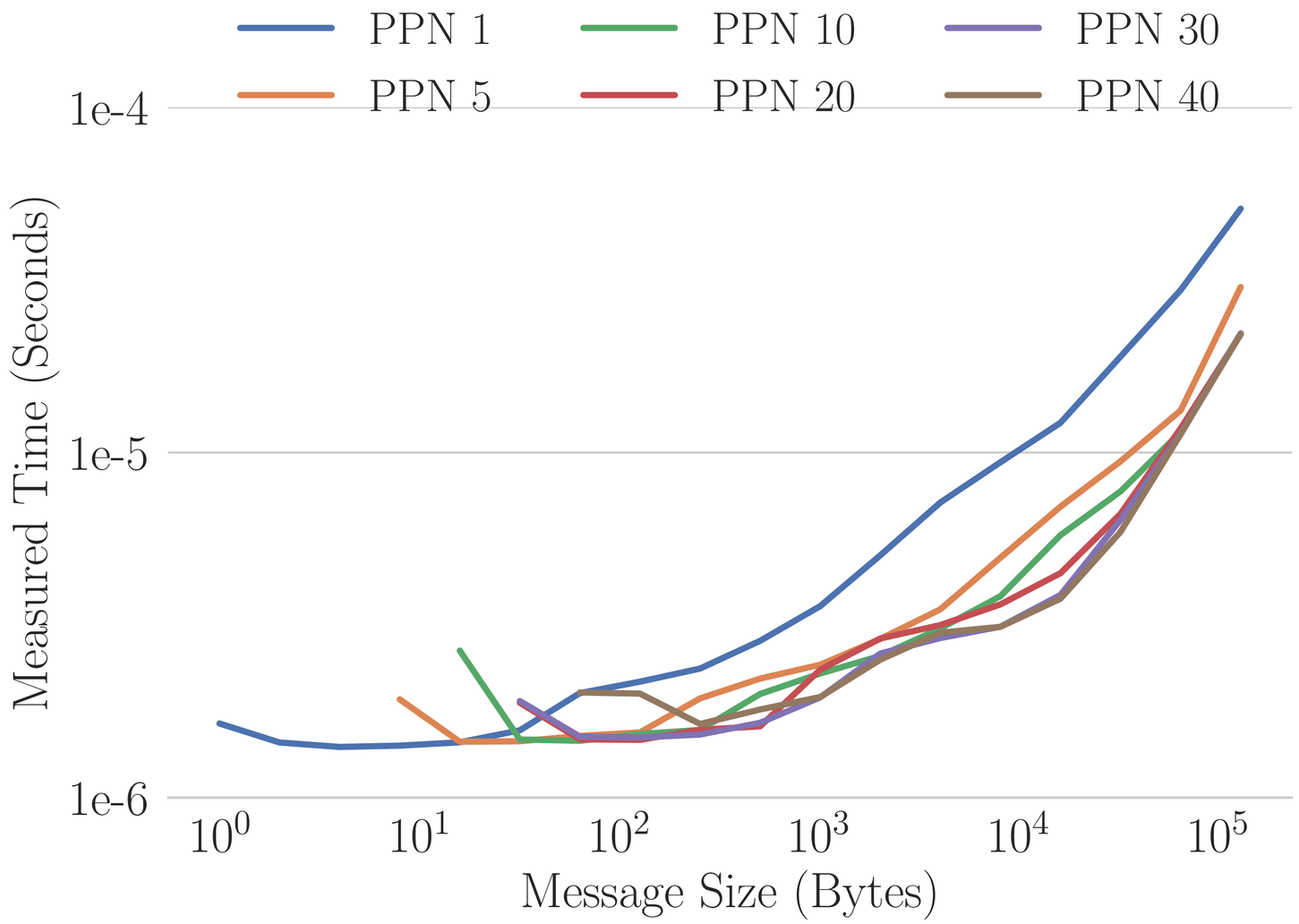
- Inter-GPU communication is more costly than inter-CPU communication
- However, we can assume data needs to be moved between GPUs
- Two approaches :
  - **GPUDirect :** communicate directly between GPUs
  - **3Step :** copy to CPU, communicate between CPUs, copy received data to GPU

How to Communicate Data Between GPUs?

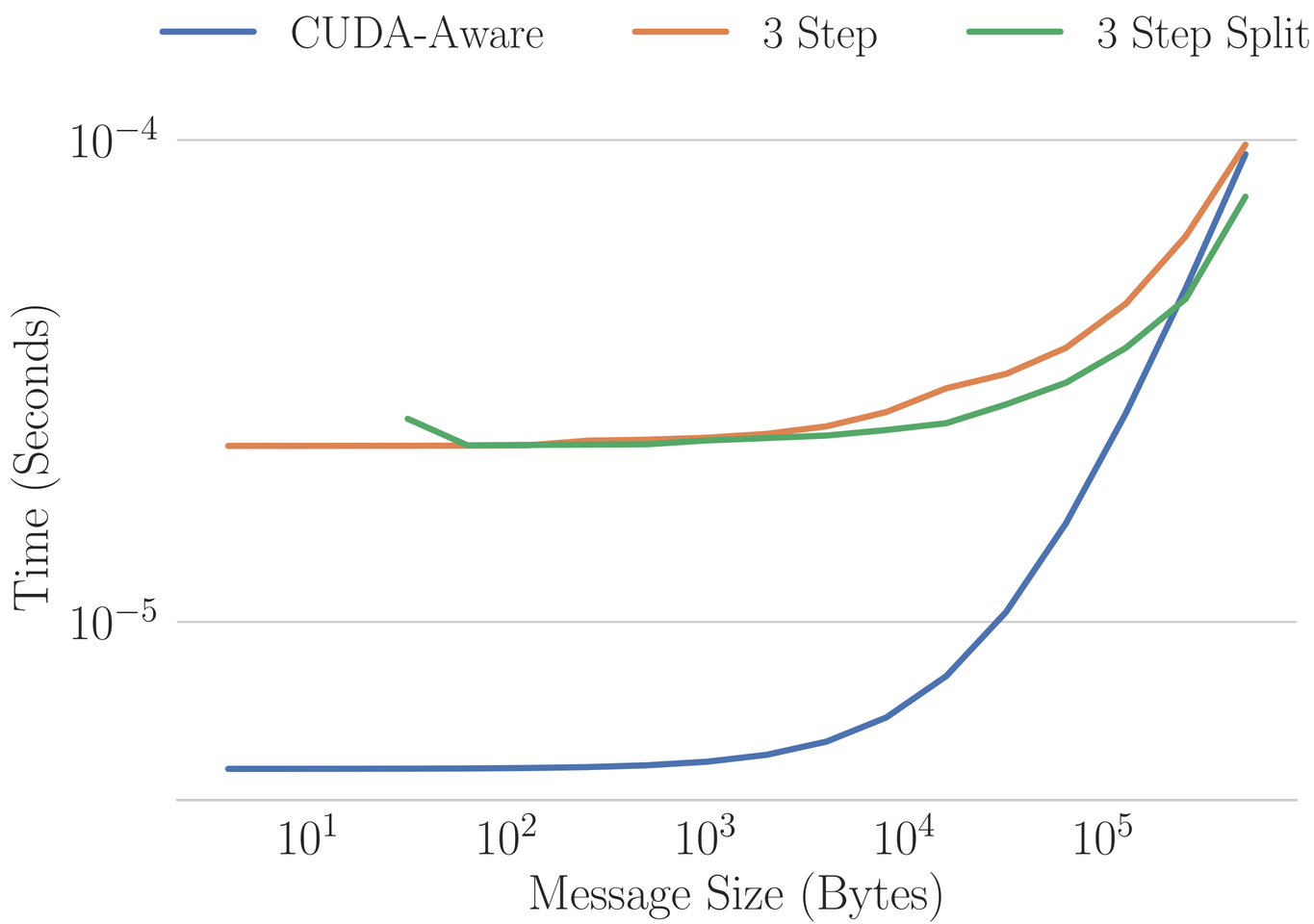


Summit, Spectrum MPI

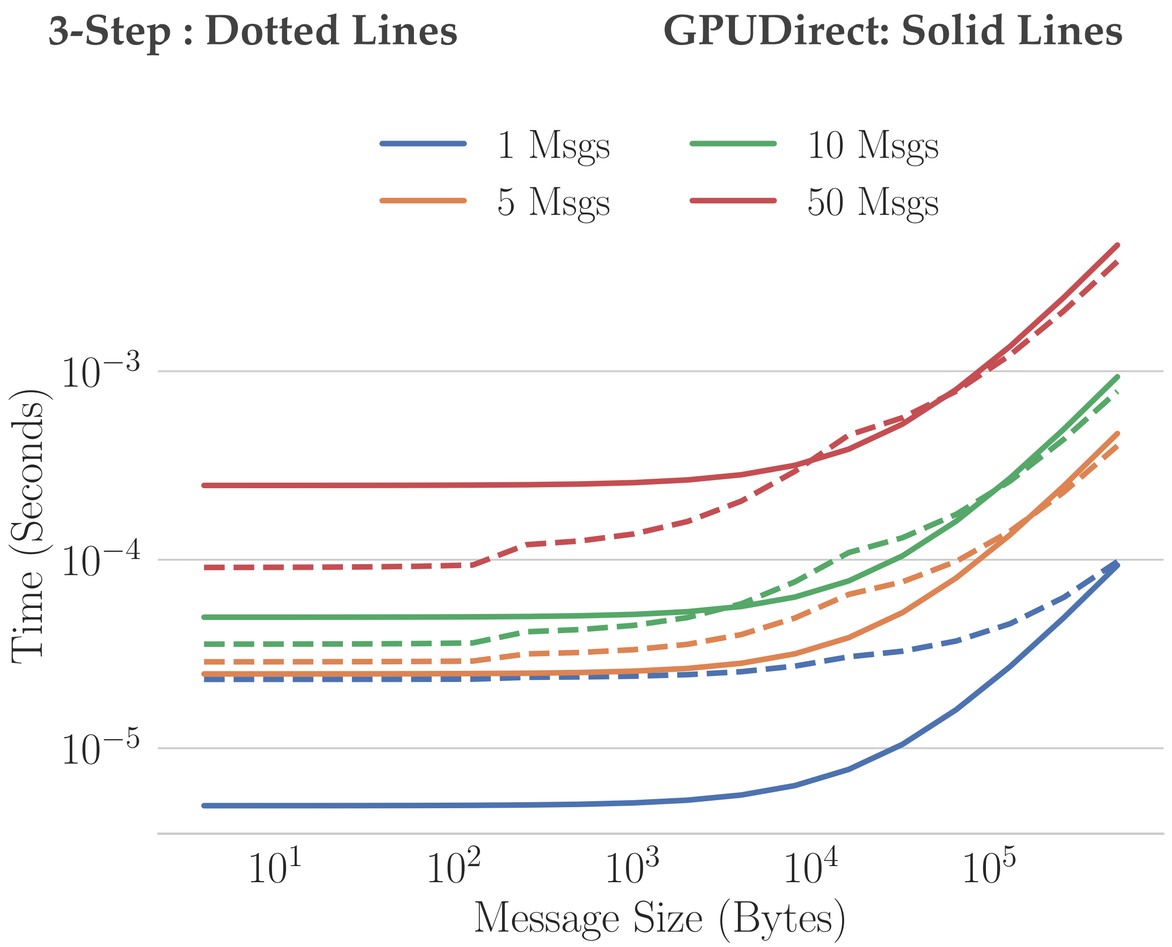
Remember, Many CPU Cores Per Node



Remember, Many CPU Cores Per Node



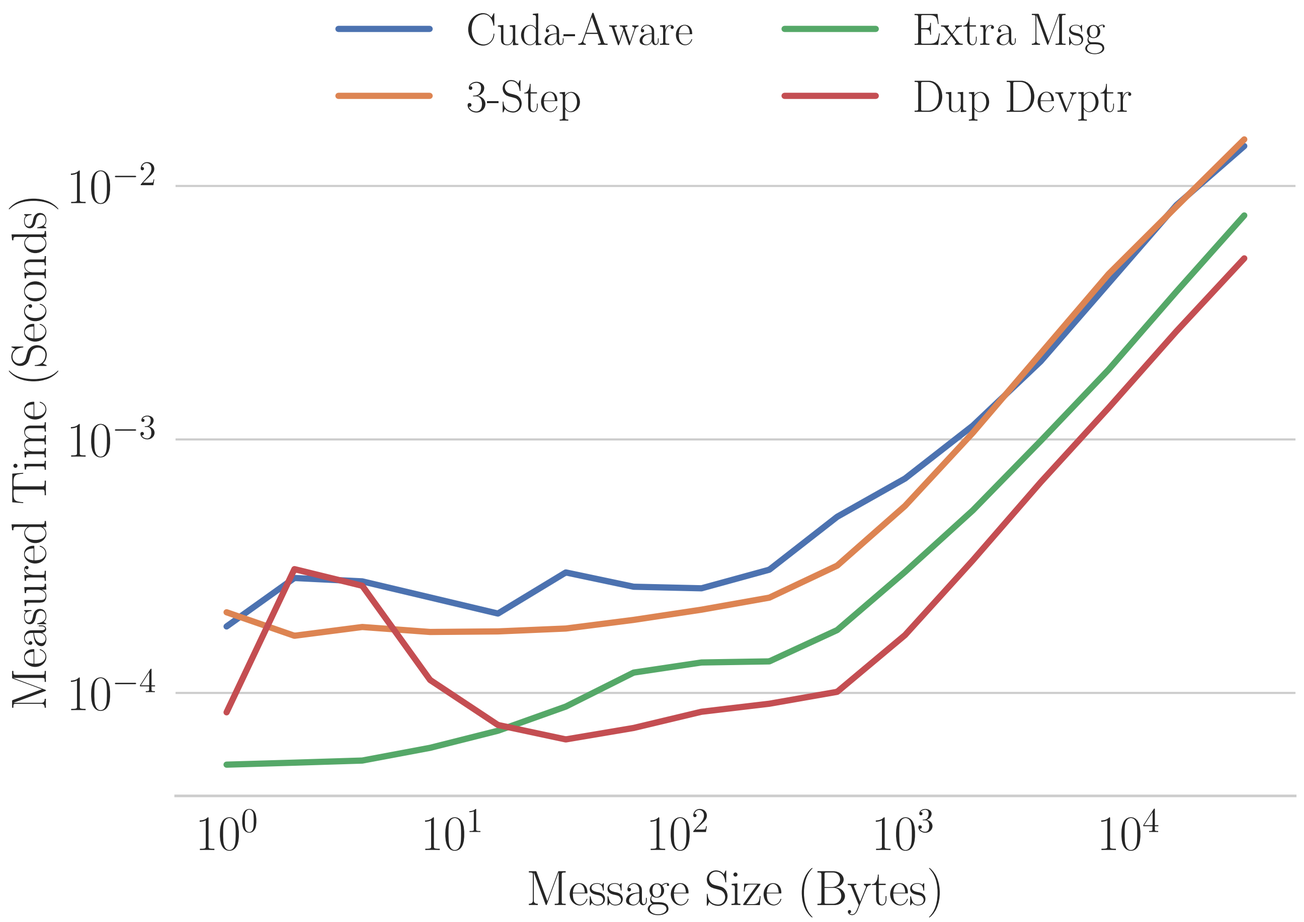
Sending Multiple Messages



Speeding Up Communication

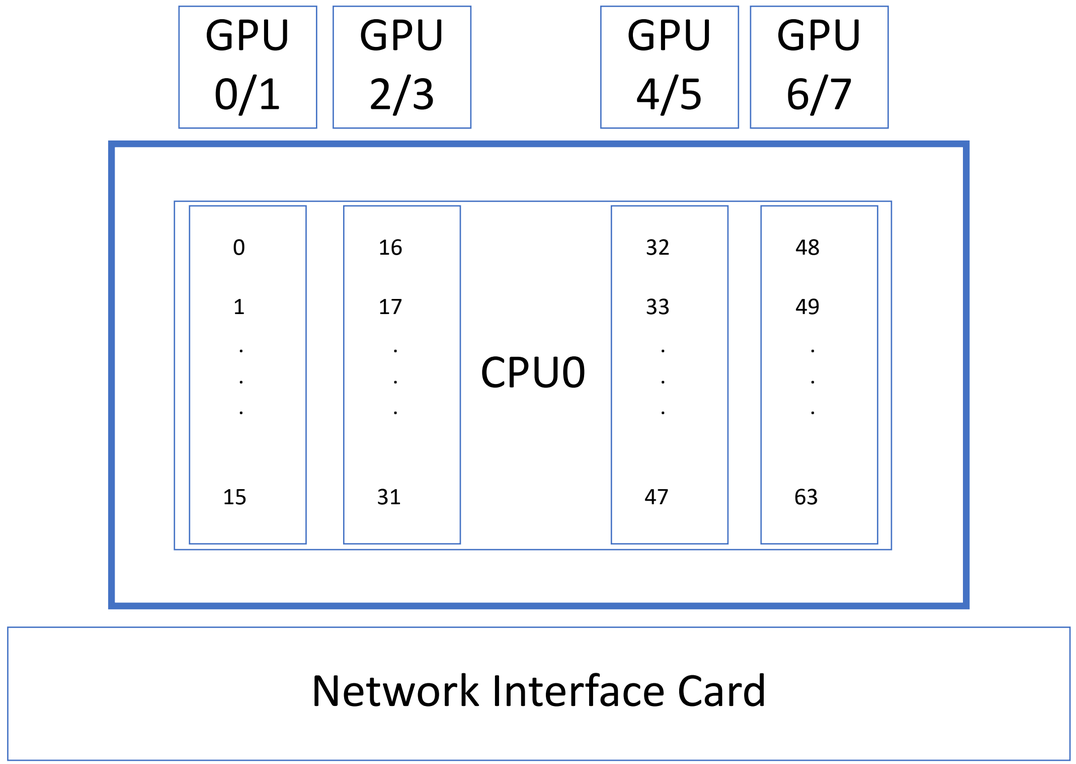
- **GPUDirect** : communicate messages directly between GPUs, avoiding cudaMemcpy entirely
- **3-Step** : copy all data to a single CPU, inter-CPU communication, and copy received data back to GPU
- **Extra Msg** : copy all data to single CPU, this CPU redistributes data between all available CPU cores per GPU, inter-CPU communication among all 40 CPU cores, gather received data to single CPU per GPU, and copy this data to GPU
- **Dup Devptr** : each available CPU core per GPU calls it's own cudaMemcpy on a duplicated device pointer

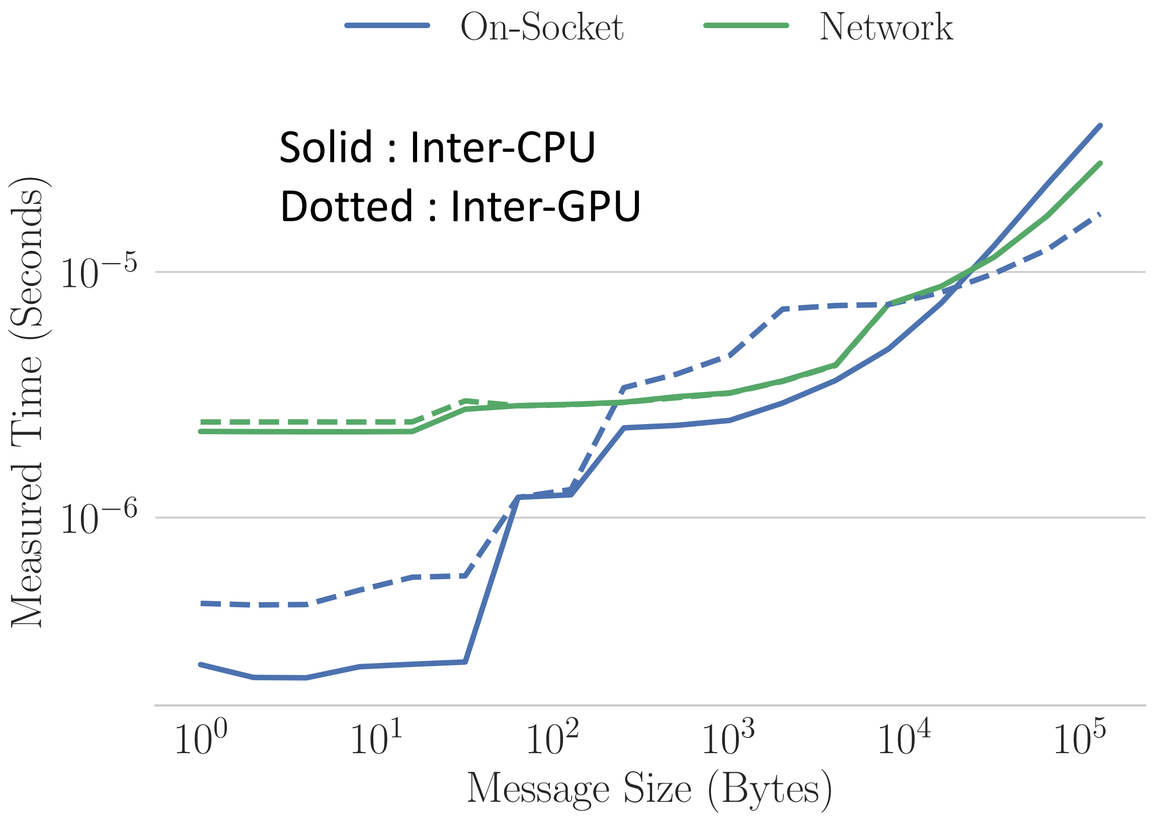
Speeding Up Communication



Summit, Spectrum MPI

Emerging Systems : Frontier





Thanks!

- I’m happy to answer questions, or chat over lunch :)

- www.amandabienz.com
- bienz@unm.edu