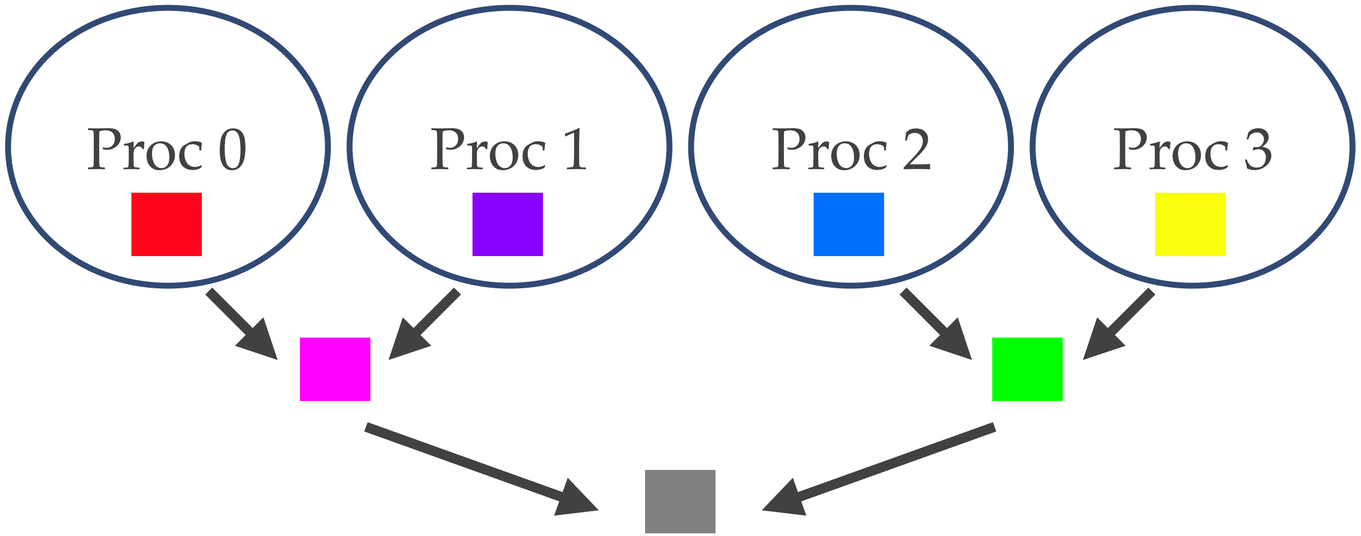
Introduction to Parallel Processing

Lecture 20 : Advanced Collective Operations

Professor Amanda Bienz

Allreduce Operation

- Reducing values across all processes in communicator
- For example, summing all values
  - Common in inner product / norm calculation in linear solvers



- **Focus of talk: reductions of small sizes over all processes**

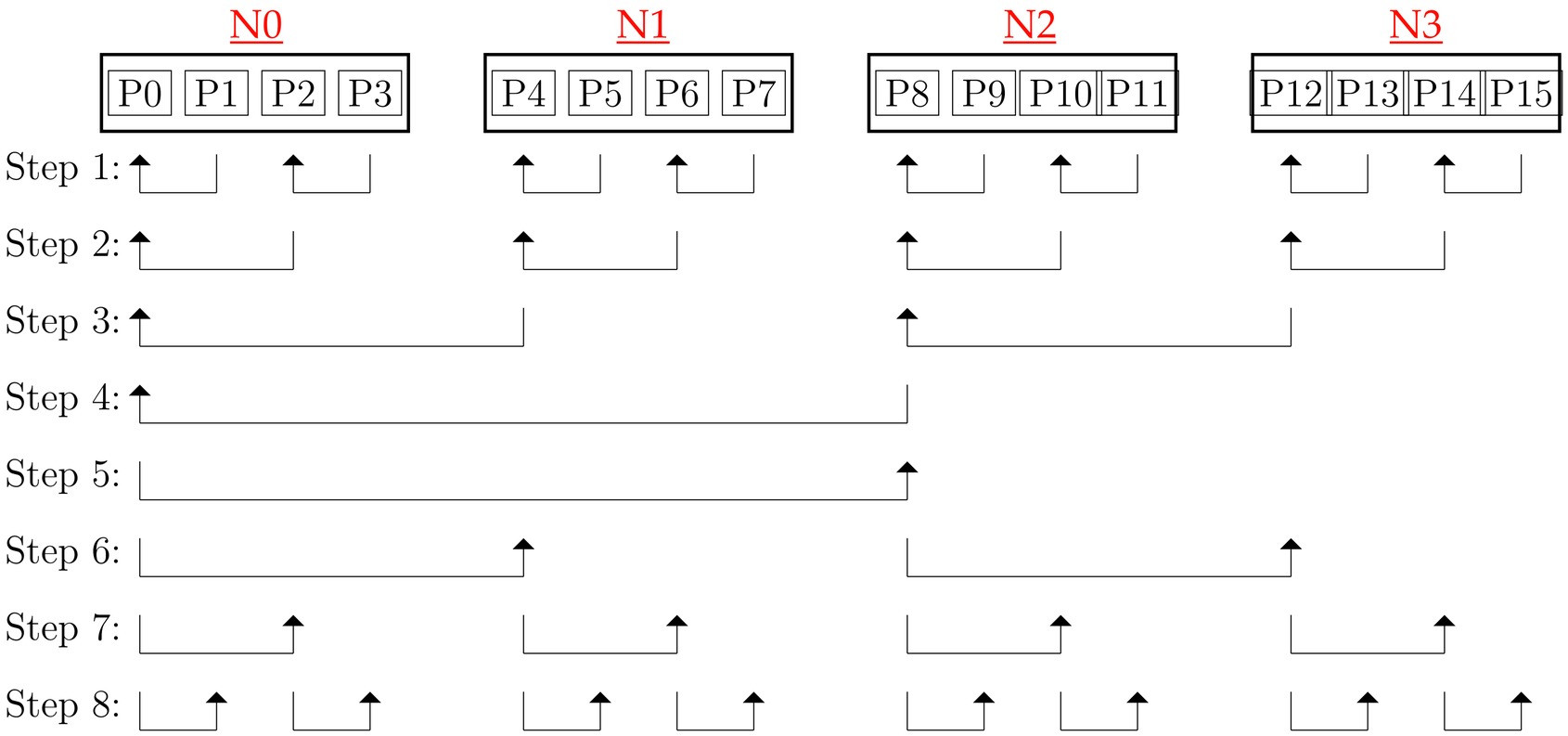
Cost (Bounds) of Allreduce Operations

- Allreduce of size s among p processes:
  - $(p-1) \cdot s$: floating point operations
  - $\frac{(p-1) \cdot s}{p}$: minimum floating point operations per process
  - $2 \cdot \frac{(p-1) \cdot s}{p}$: minimum number of floating point values to be transported
  - $\log_2(p)$: minimum messages to be communicated

Cost (Bounds) of Allreduce Operations

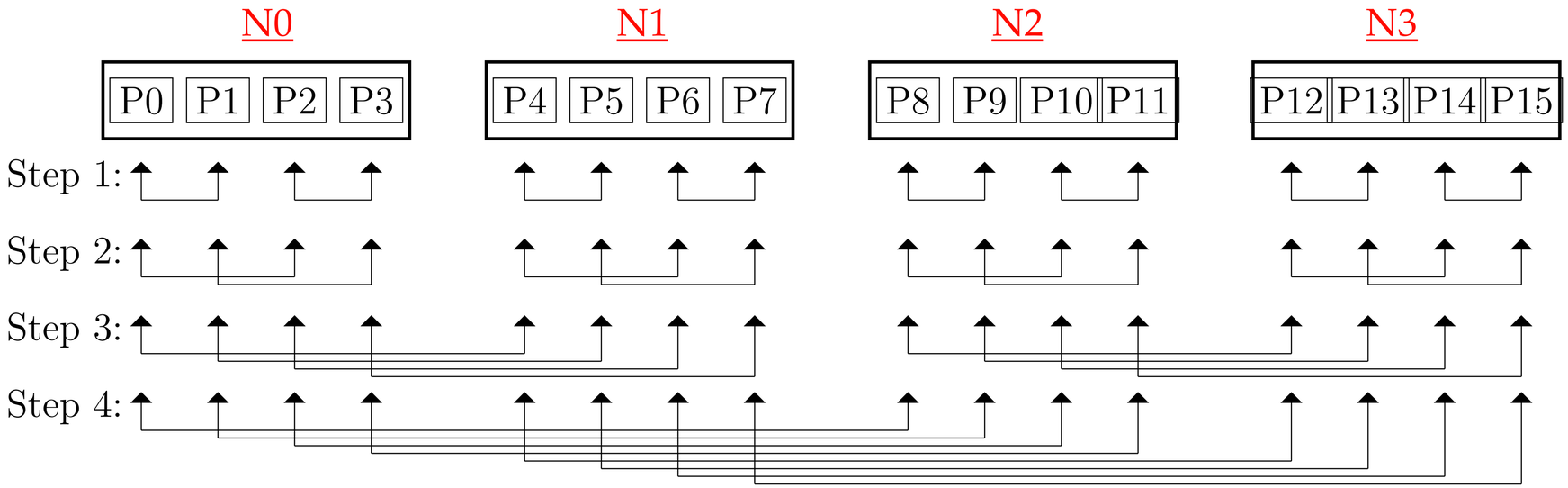
- Allreduce of size s among p processes:
  - $(p-1) \cdot s$: floating point operations
  - $\frac{(p-1) \cdot s}{p}$: minimum floating point operations per process
  - $2 \cdot \frac{(p-1) \cdot s}{p}$: minimum number of floating point values to be transported
  - $\log_2(p)$: minimum messages to be communicated

Reduce and Broadcast



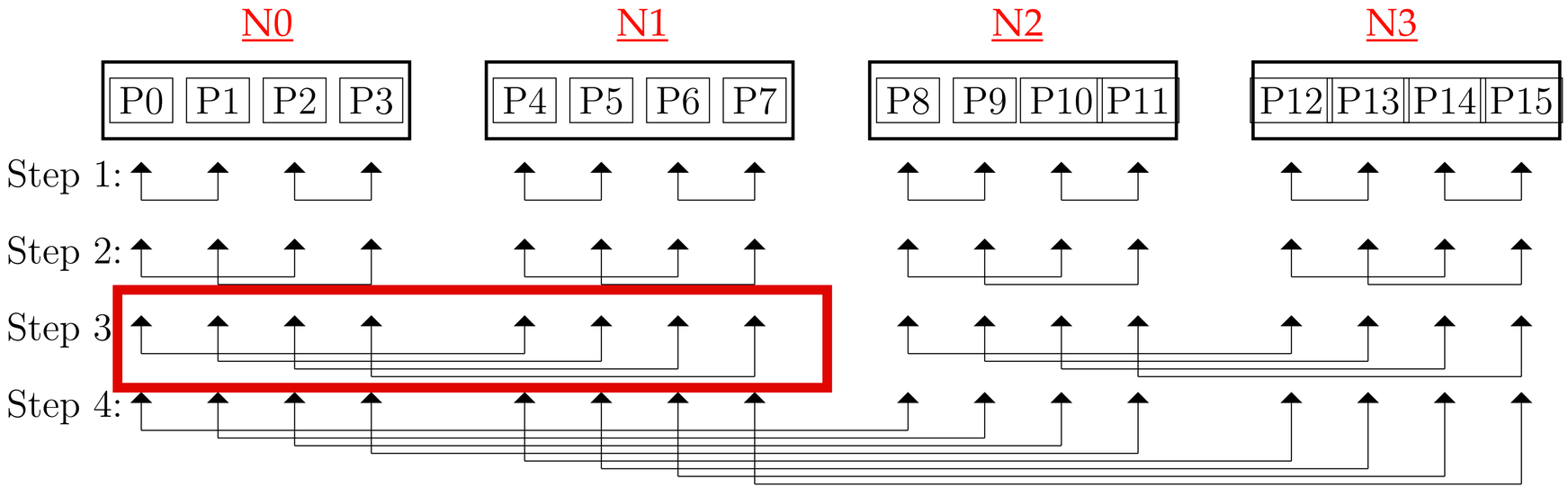
- First reduce to master process, then broadcast to all
- $2 \cdot \log_2(p)$ messages, idle processes

Recursive Doubling



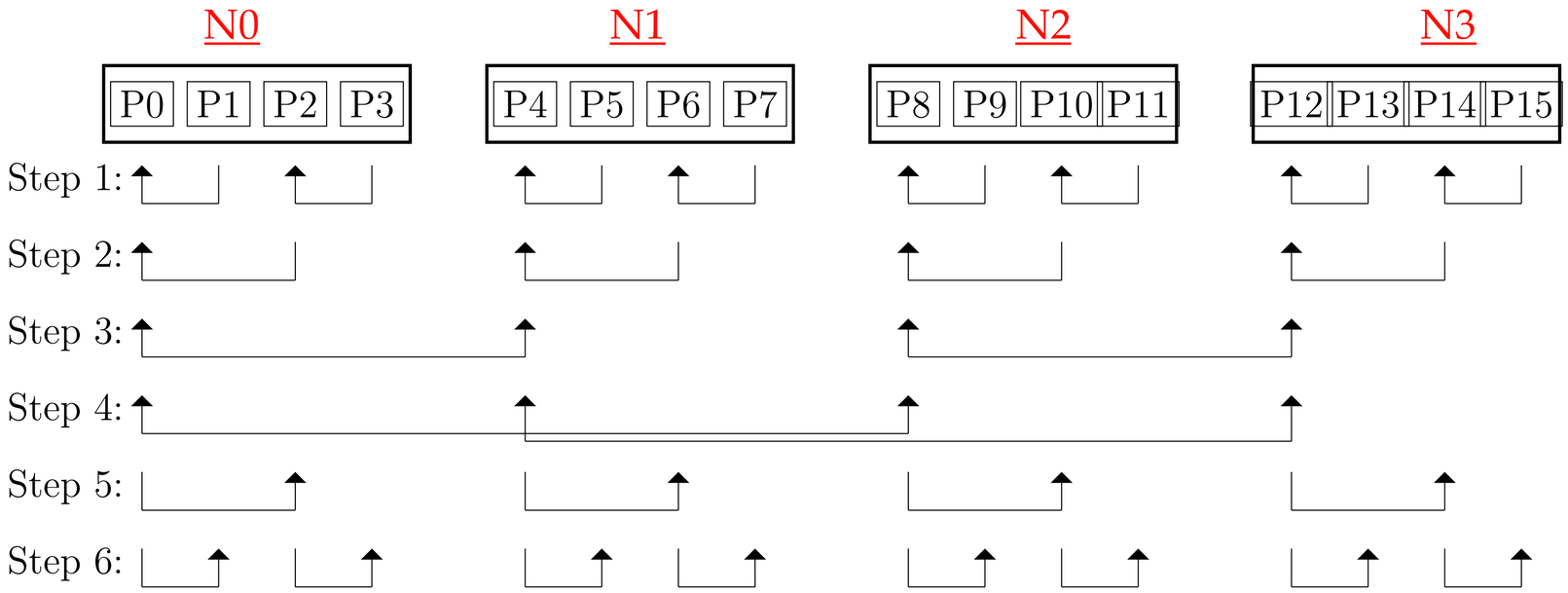
- Reduce among all processes at a time
- $\log_2(p)$ messages

Recursive Doubling



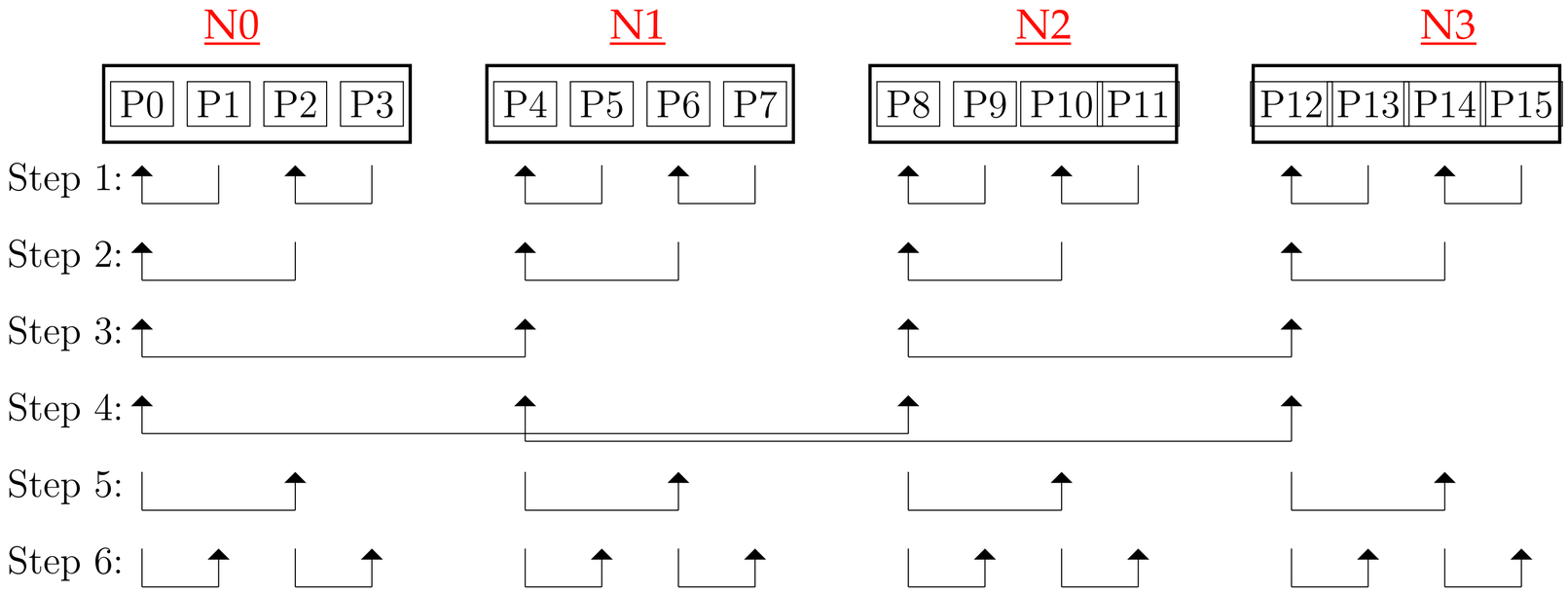
- Reduce among all processes at a time
- $\log_2(p)$ messages
- $\log_2(n)$ inter-node messages, where n is number of nodes
- Duplicate messages between nodes

Hierarchical Approach



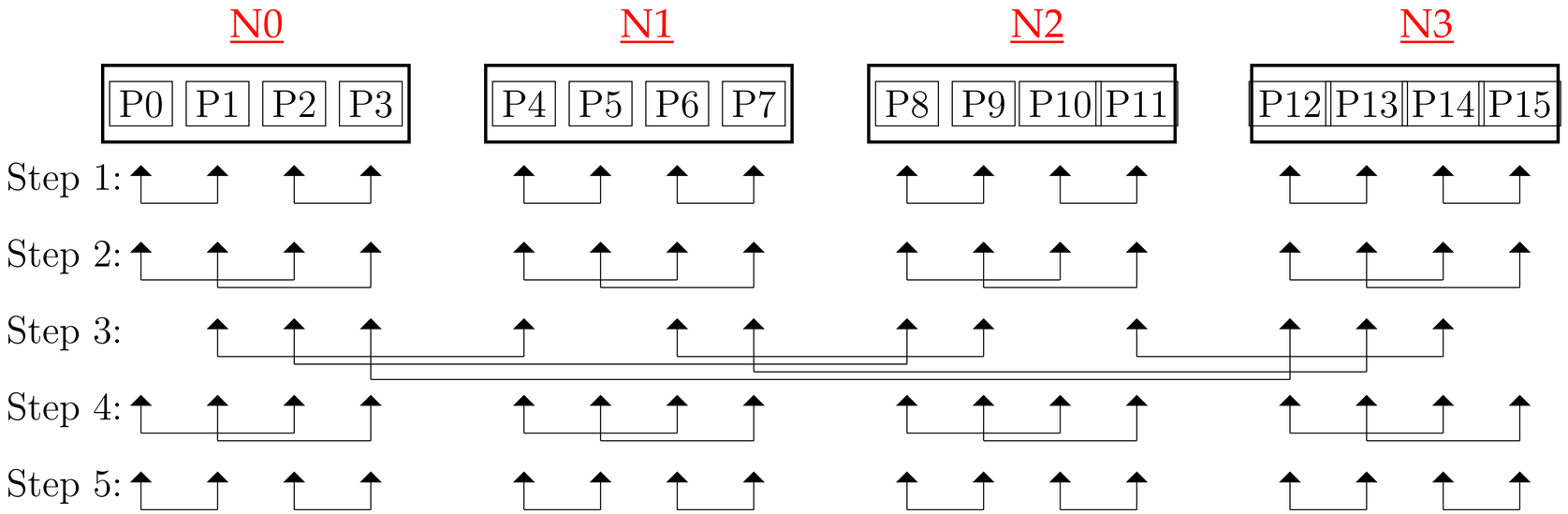
- Reduce to master process on each node
- Reduce among all nodes together
- Broadcast result on each node

Hierarchical Approach



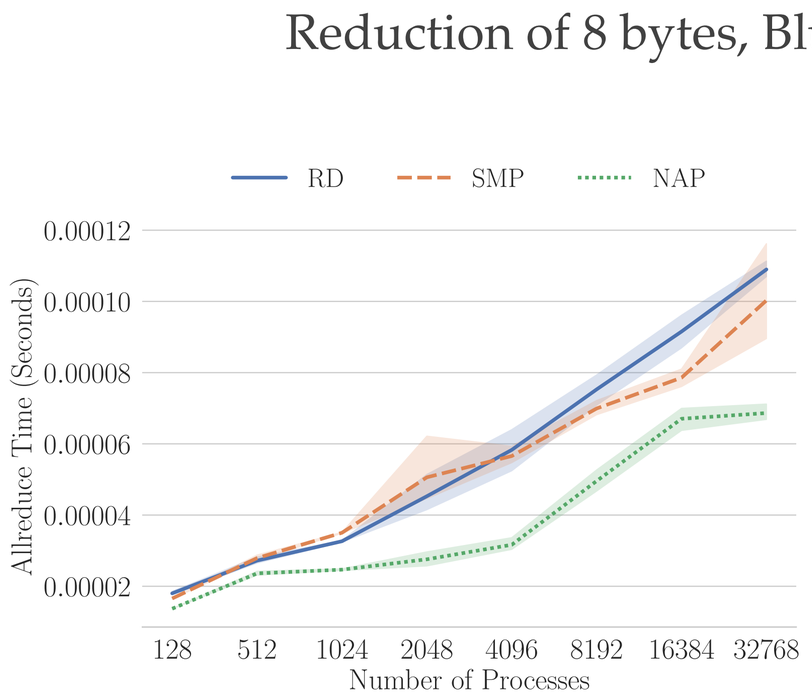
- Additional intra-node messages
- $\log_2(n)$ inter-node messages
- Idle processes

Locality-Aware Algorithm

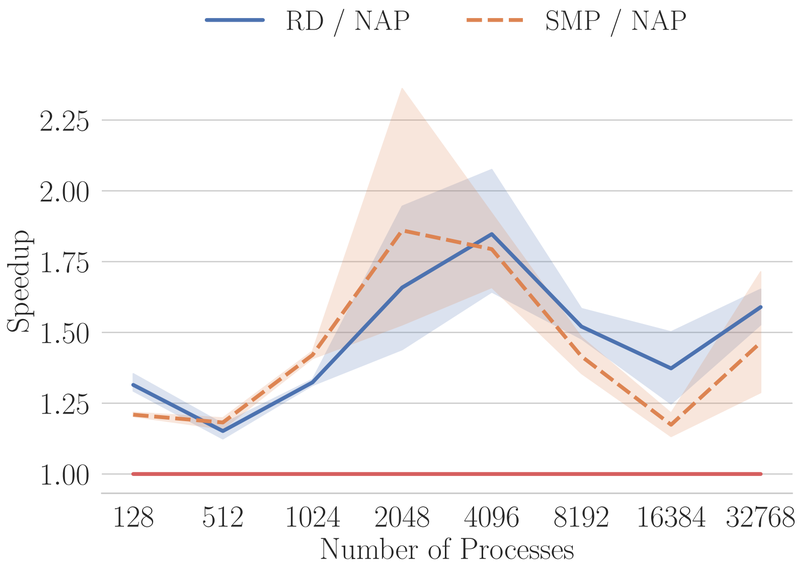


- Reduce among processes local to node
- Each local process reduces with a separate node
- Reduce inter-node results among processes local to node
- Local rank r on node n exchanges with local rank n on node r

Allreduce Times, Various Process Counts

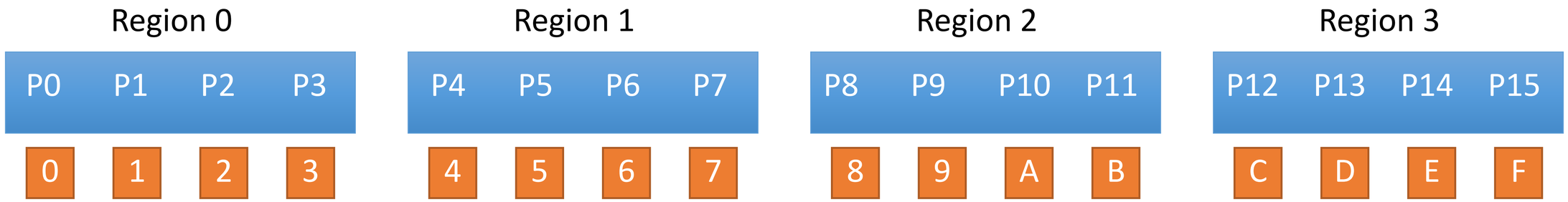


Allreduce Times



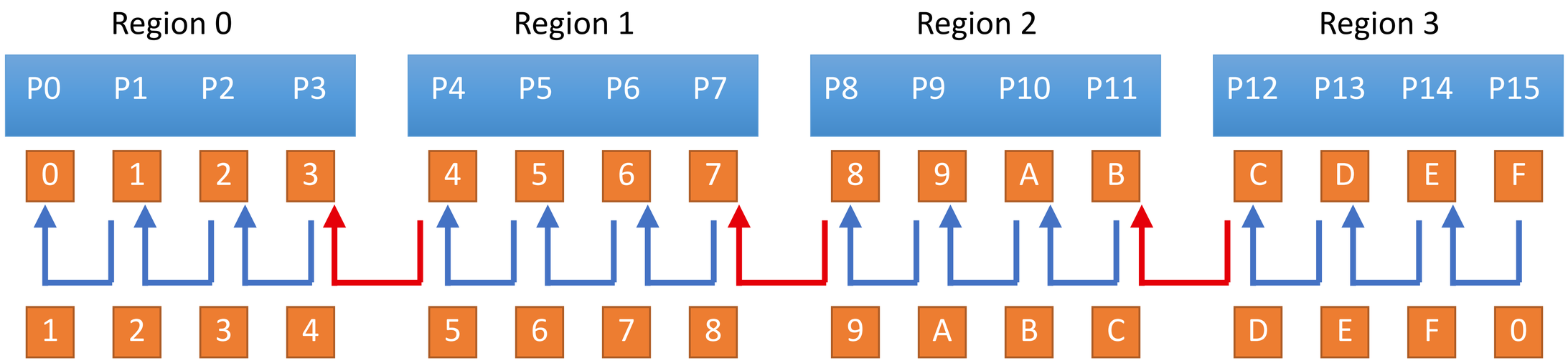
Speedup with NAP

Bruck Allgather : Initial Data

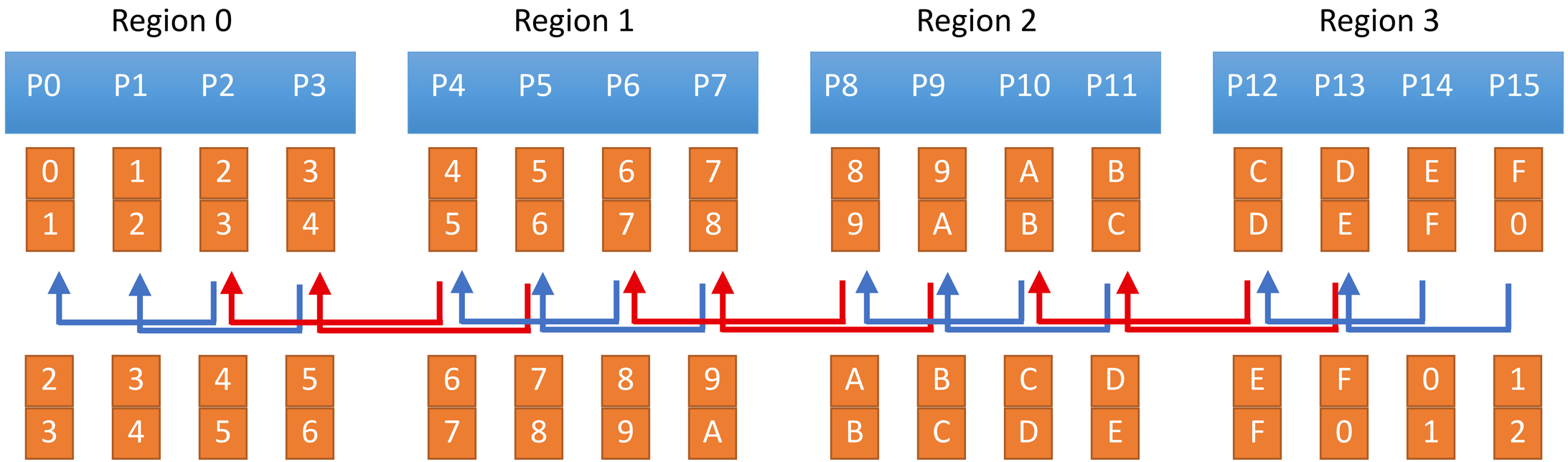


J. Bruck, Ching-Tien Ho, S. Kipnis, E. Upfal and D. Weathersby, "Efficient algorithms for all-to-all communications in multiport message-passing systems," in IEEE Transactions on Parallel and Distributed Systems, vol. 8, no. 11, pp. 1143-1156, Nov. 1997, doi: 10.1109/71.642949.

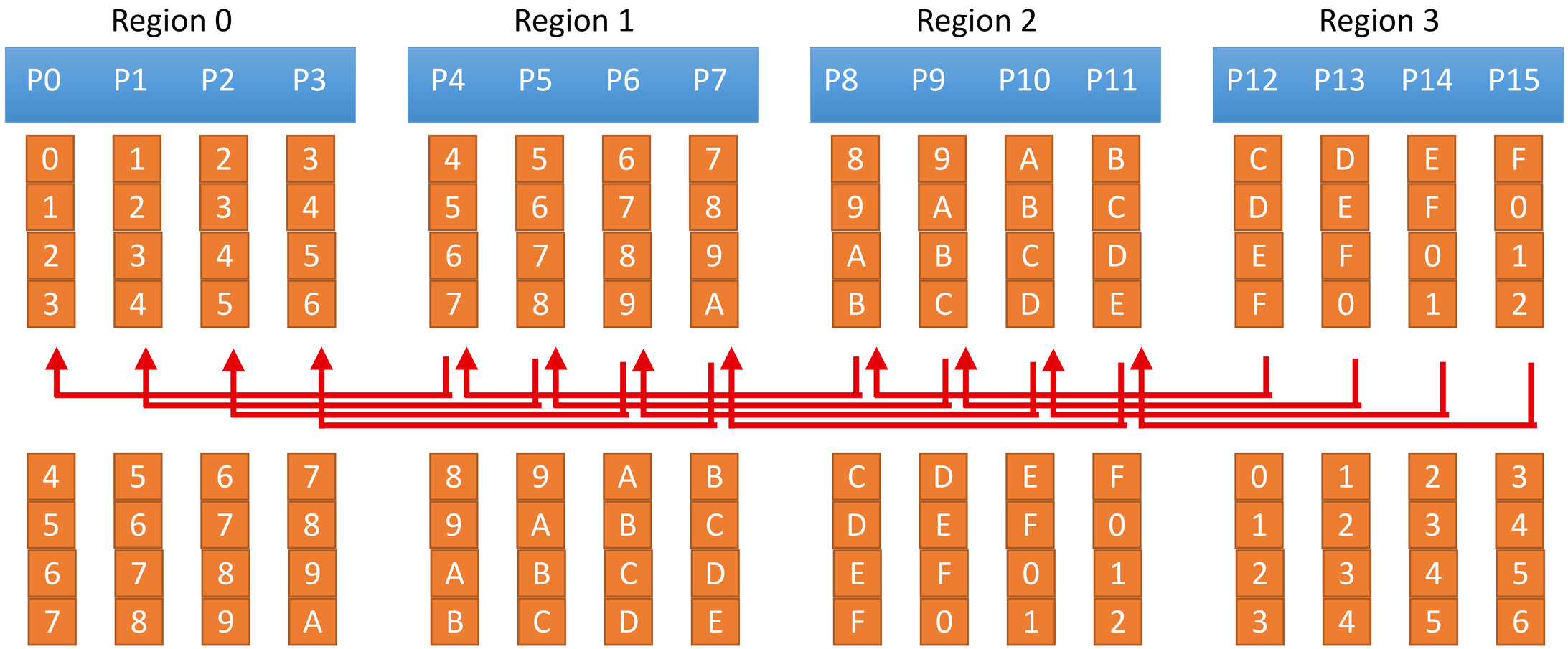
Bruck Allgather : Step 1



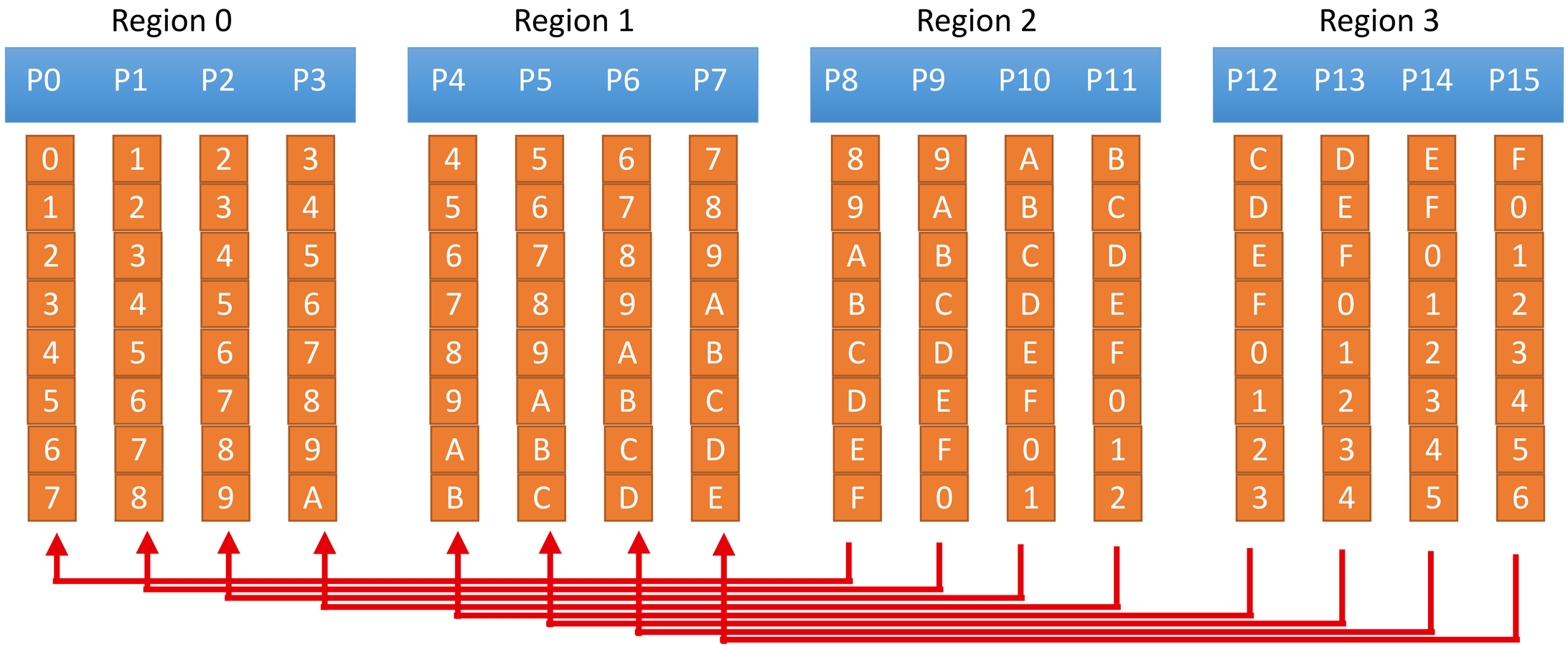
Bruck Allgather : Step 2



Bruck Allgather : Step 3



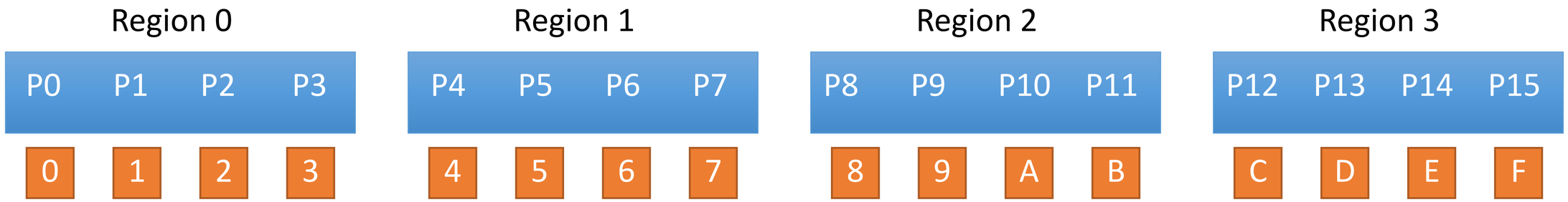
Bruck Allgather : Step 4



Bruck Allgather

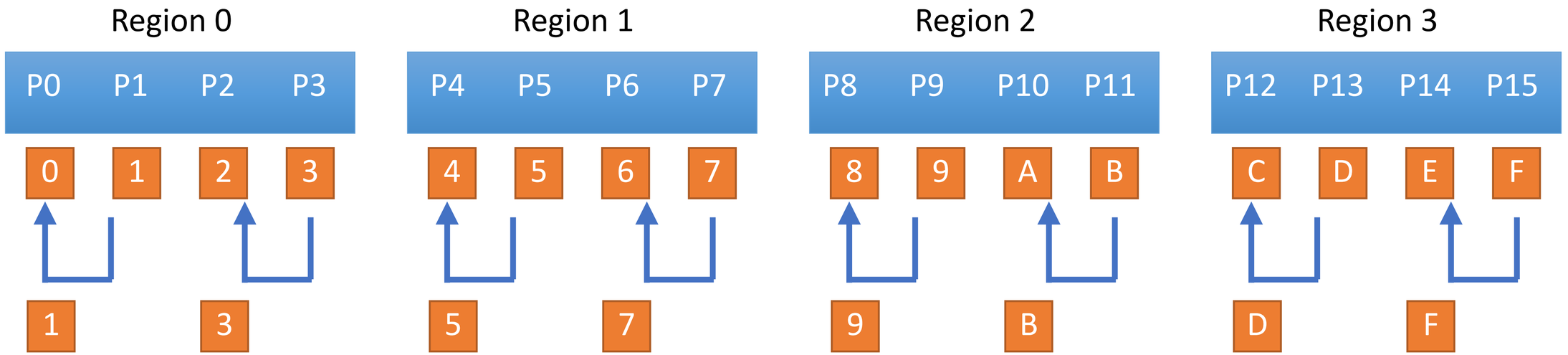
- Algorithmically minimizes number of messages
- **Multiple messages between non-local regions**
- **Data sent multiple times between a set of non-local regions**

Hierarchical Bruck Allgather : Initial Data

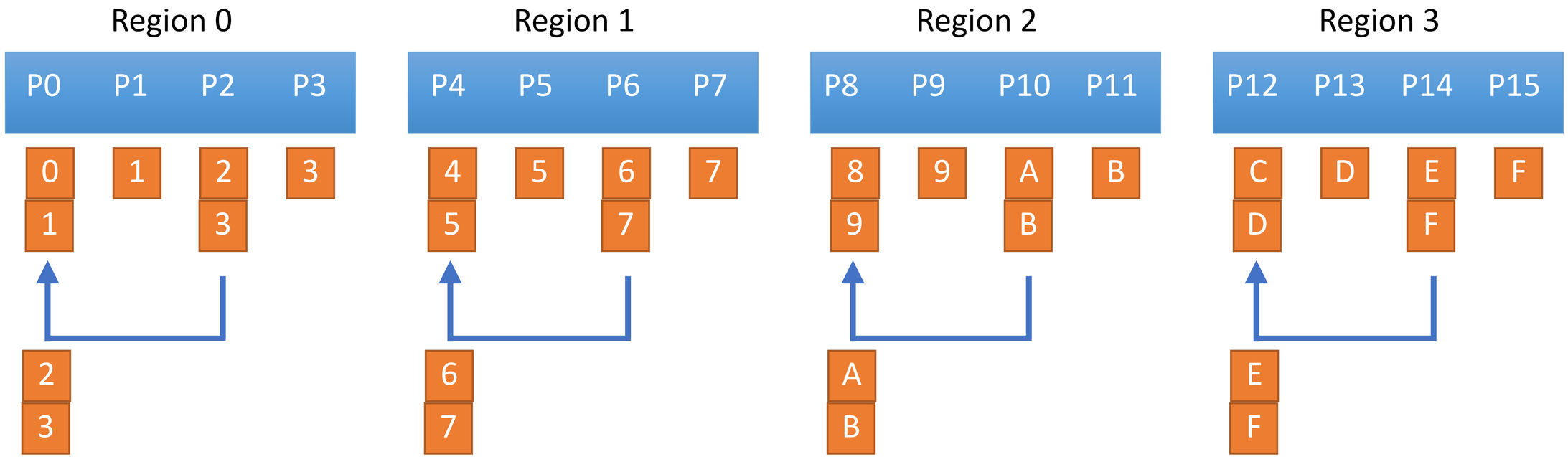


- Jesper Larsson Träff and Antoine Rougier. 2014. MPI Collectives and Datatypes for Hierarchical All-to-all Communication. In Proceedings of the 21st European MPI Users' Group Meeting (EuroMPI/ASIA '14). Association for Computing Machinery, New York, NY, USA, 27–32.
- X. Luo et al., "HAN: a Hierarchical AutotuNed Collective Communication Framework," 2020 IEEE International Conference on Cluster Computing (CLUSTER), 2020, pp. 23-34, doi: 10.1109/CLUSTER49012.2020.00013.
- R. Graham et al., "Cheetah: A Framework for Scalable Hierarchical Collective Operations," 2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, 2011, pp. 73-83, doi: 10.1109/CCGrid.2011.42.
- Zhu, Hao, et al. “Hierarchical collectives in MPICH2.” *European Parallel Virtual Machine/Message Passing Interface Users’ Group Meeting.* Springer. Berling, Heidelberg, 2009.

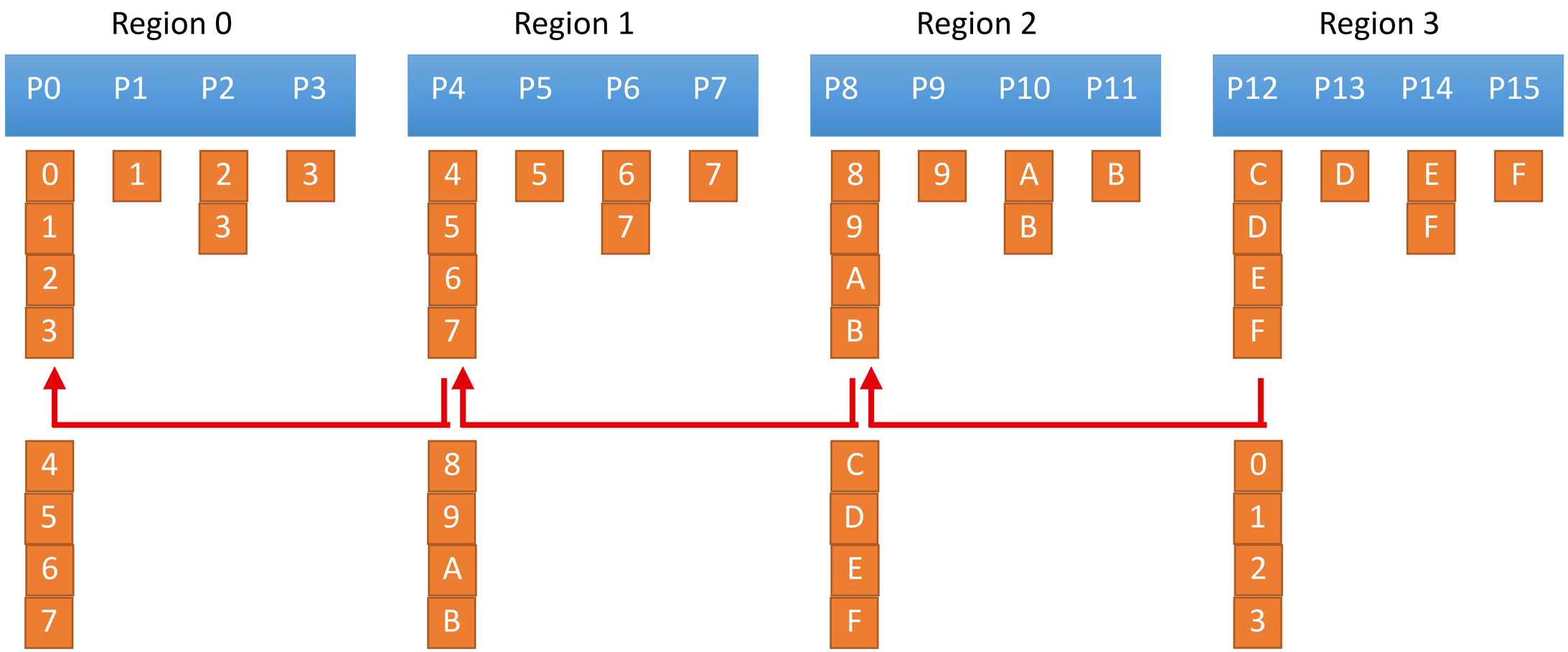
Hierarchical Bruck Allgather : Step 1



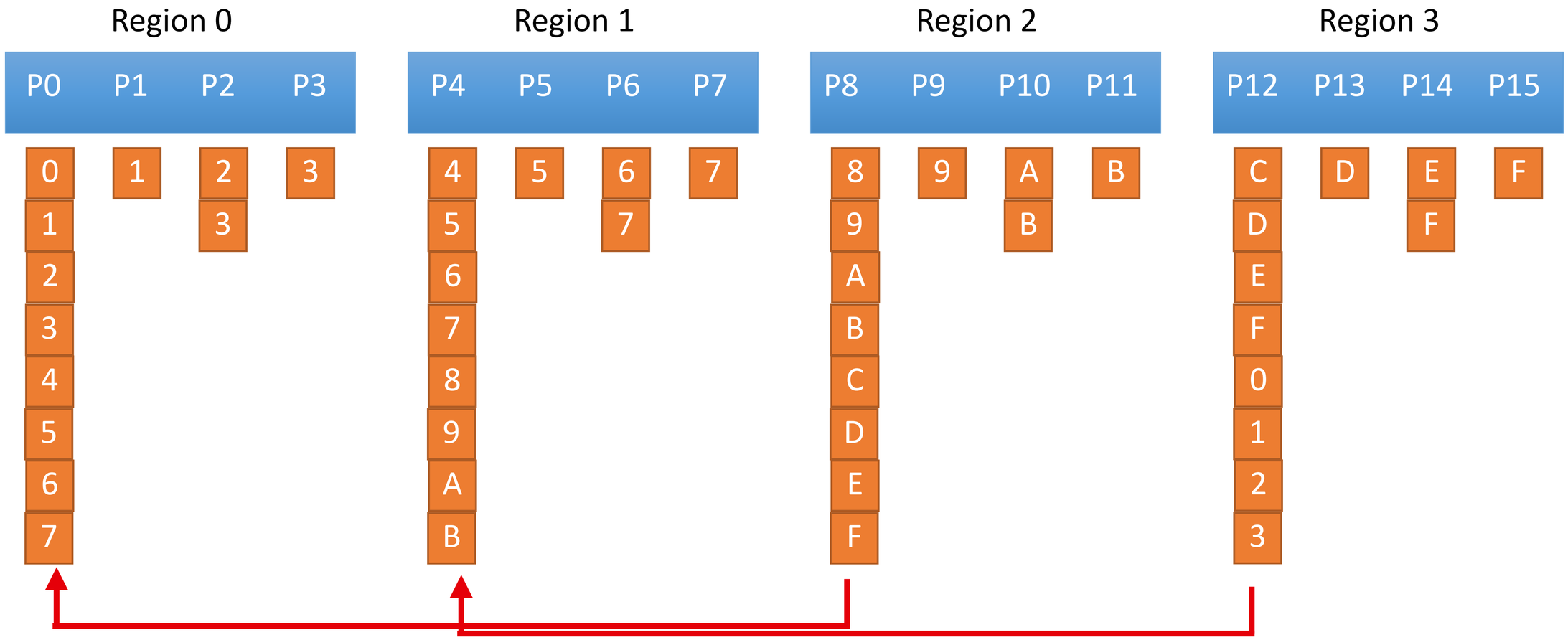
Hierarchical Bruck Allgather : Step 2



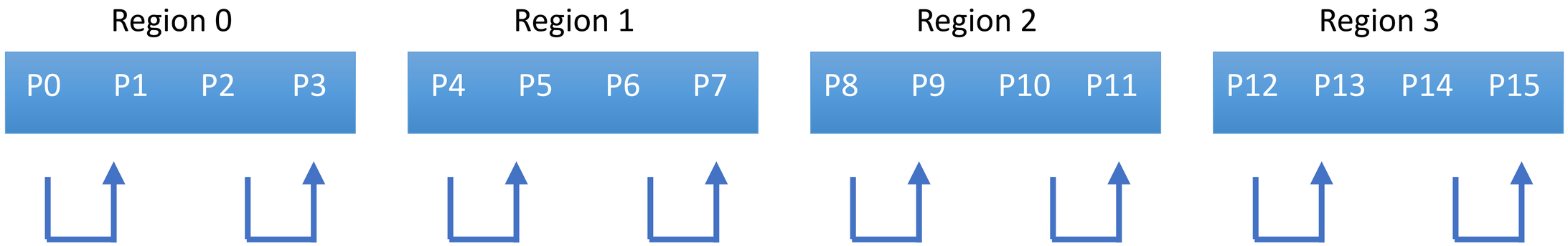
Hierarchical Bruck Allgather : Step 3



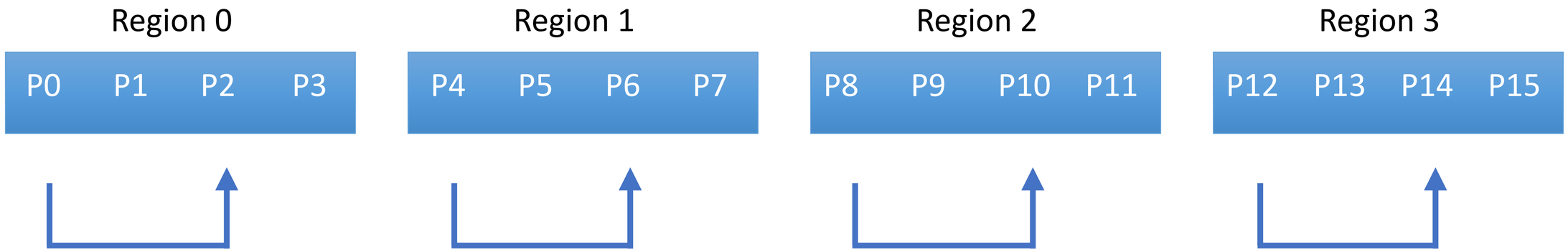
Hierarchical Bruck Allgather : Step 4



Hierarchical Bruck Allgather : Step 5



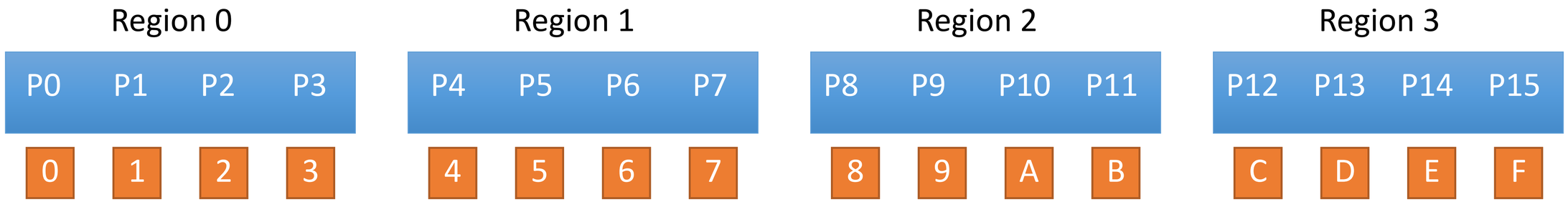
Hierarchical Bruck Allgather : Step 6



Hierarchical Bruck Allgather

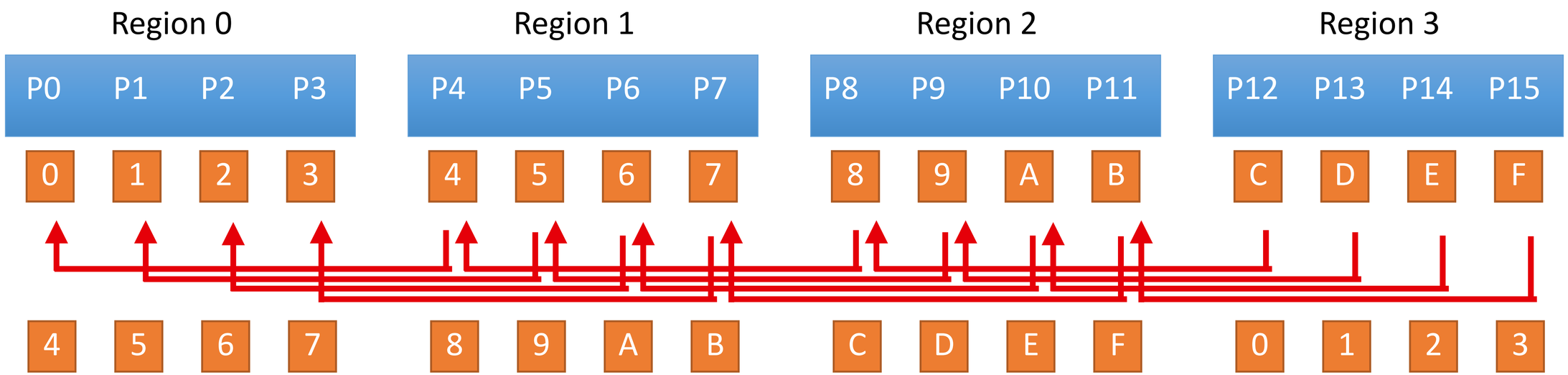
- Extra local messages within a region
- Only one message between any set of non-local regions
- Data only sent once between any set of non-local regions
- **Many idle processes**

MultiLane Bruck Allgather : Initial Data

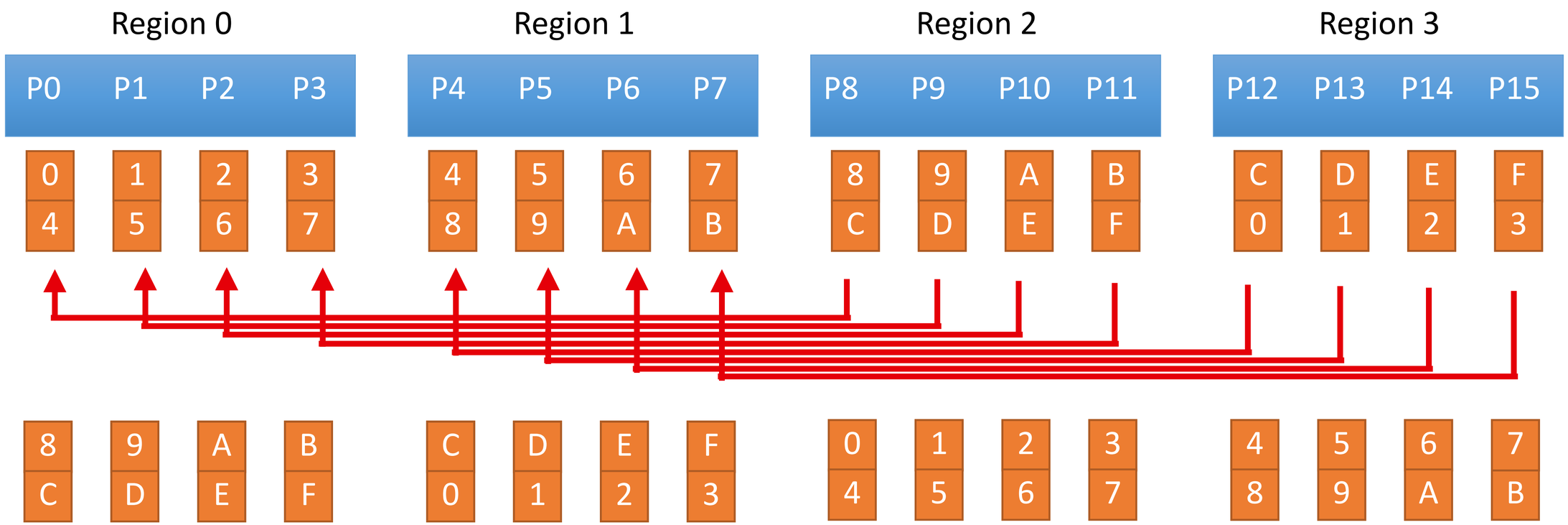


J. L. Träff and S. Hunold, "Decomposing MPI Collectives for Exploiting Multi-lane Communication," 2020 IEEE International Conference on Cluster Computing (CLUSTER), 2020, pp. 270-280, doi: 10.1109/CLUSTER49012.2020.00037.

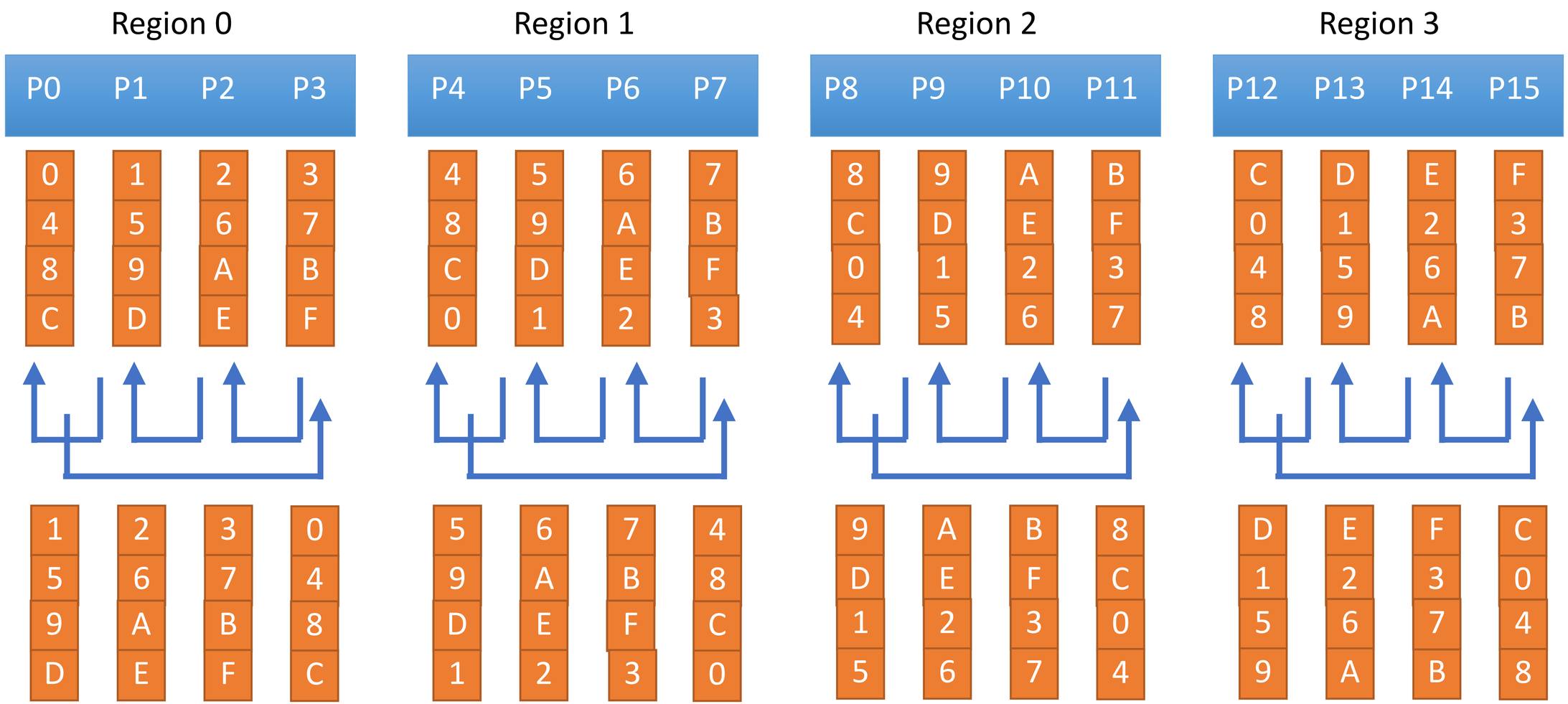
MultiLane Bruck Allgather : Step 1



MultiLane Bruck Allgather : Step 2



MultiLane Bruck Allgather : Step 3

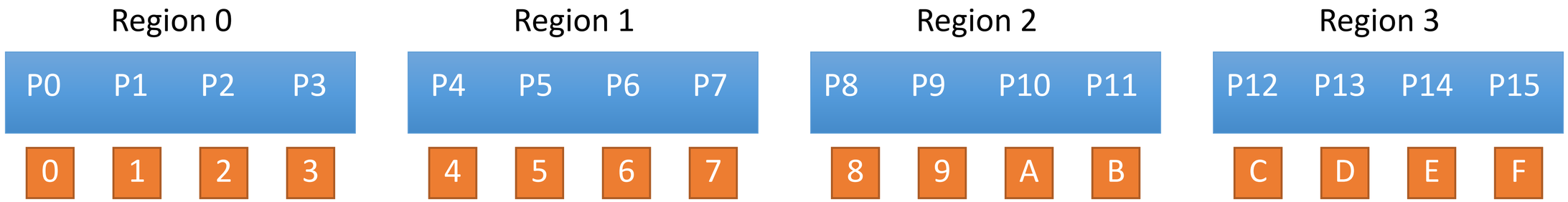


MultiLane Bruck Allgather : Step 4

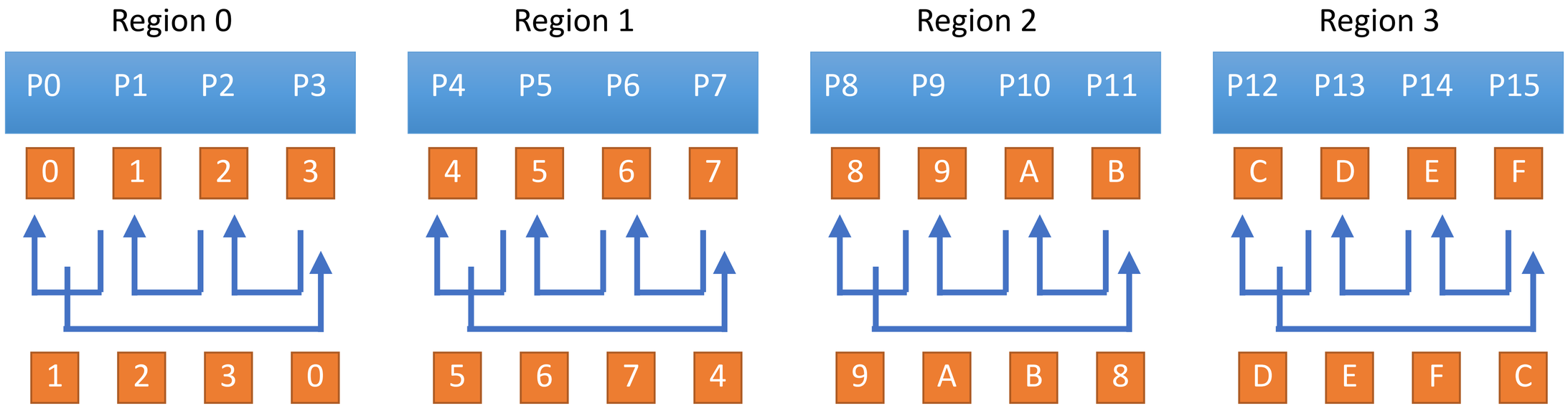
MultiLane Bruck Allgather

- Uses all processes
- Algorithmically minimizes messages
- Data only sent once between any set of non-local regions
- **Multiple messages between sets of nodes**
- **Ideal for larger message counts**

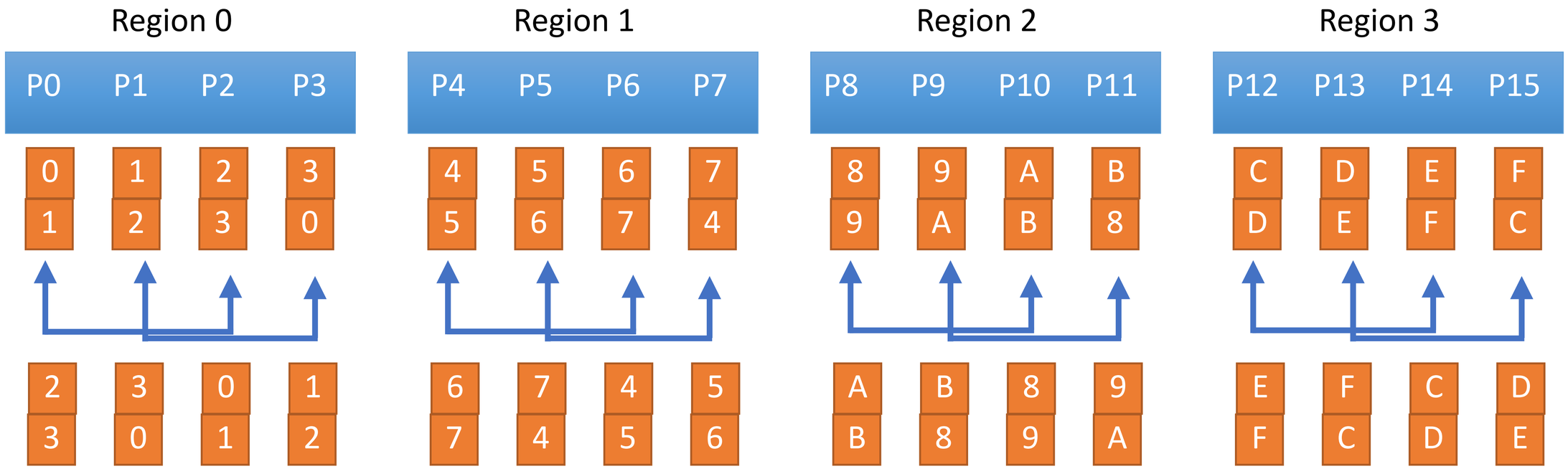
Locality-Aware Bruck Allgather : Initial Data



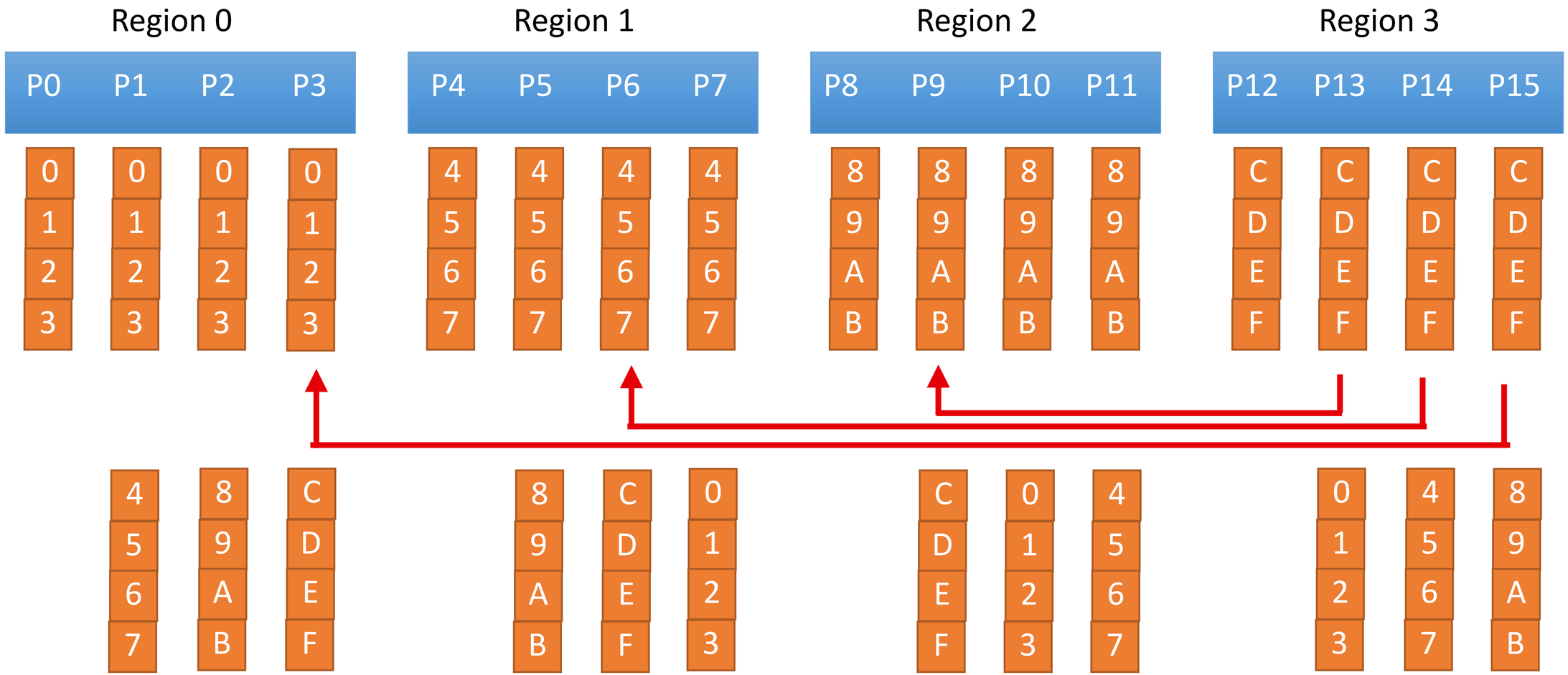
Locality-Aware Bruck Allgather : Step 1



Locality-Aware Bruck Allgather : Step 2

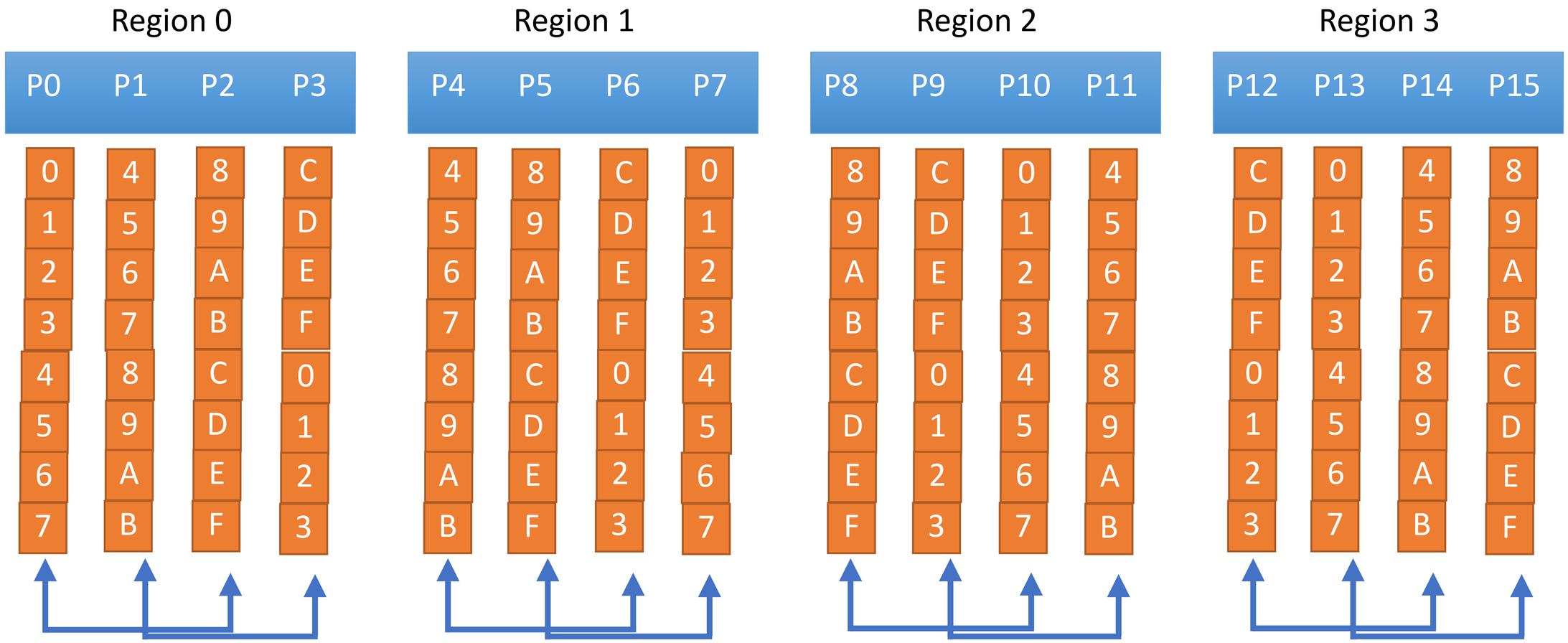


Locality-Aware Bruck Allgather : Step 3



Locality-Aware Bruck Allgather : Step 4

Locality-Aware Bruck Allgather : Step 5



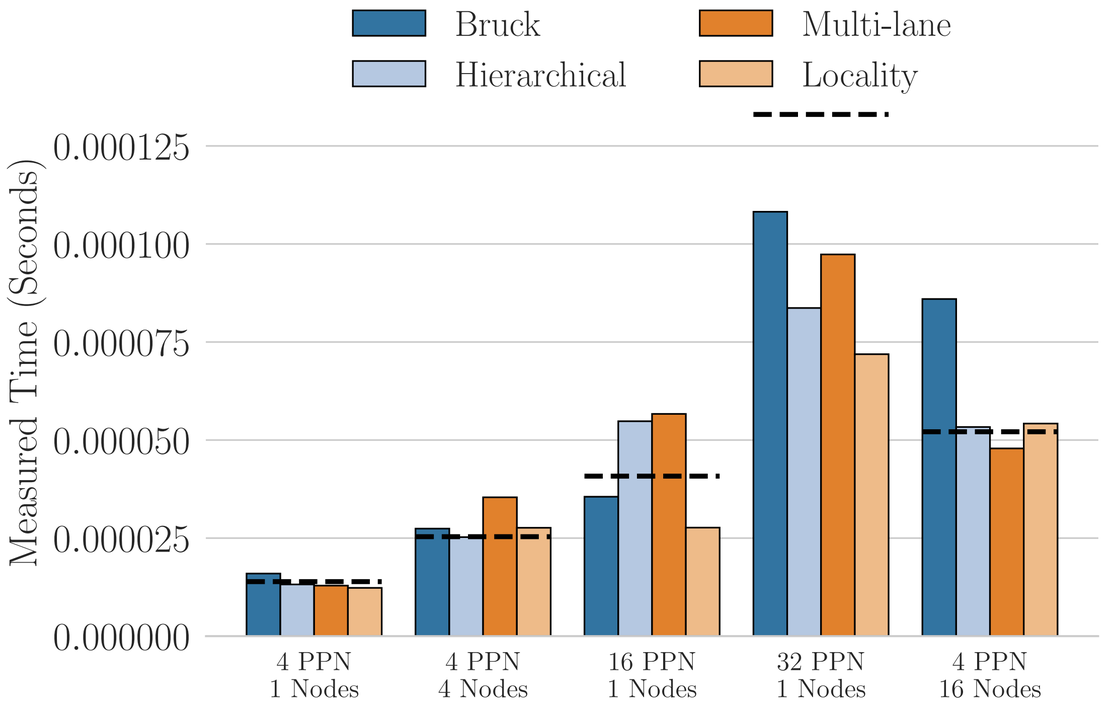
Locality-Aware Bruck Allgather

- Extra local messages within a region
- Only one message between any set of non-local regions
- Data only sent once between any set of non-local regions
- Uses all processes per node
  - Each sends to / receives from a set of unique non-local regions

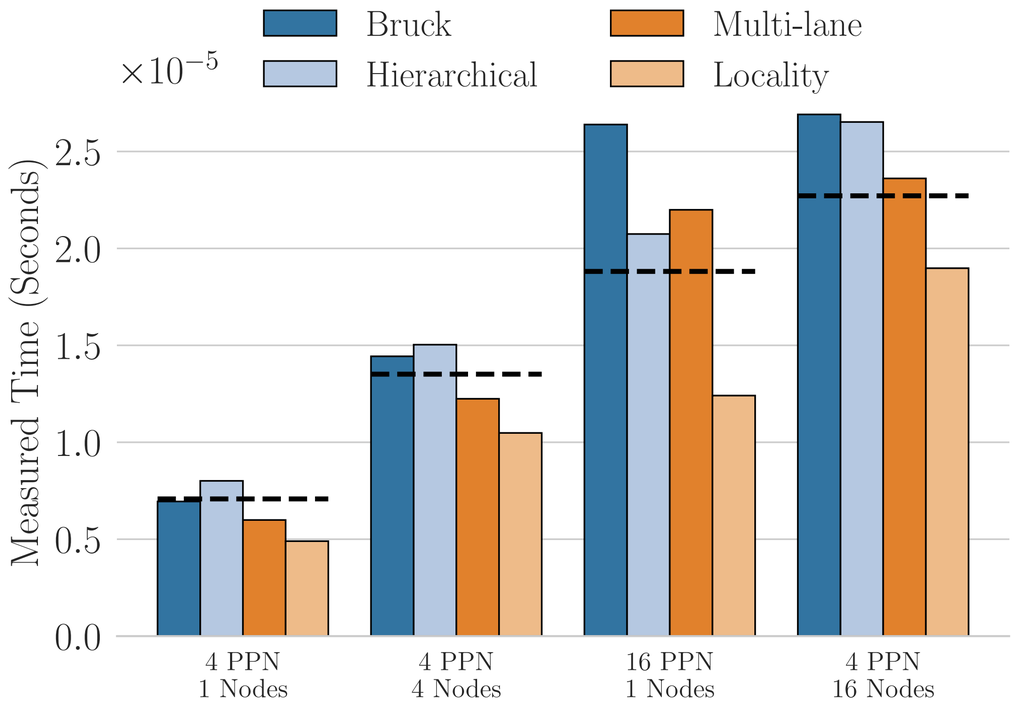
- Ideal for small messages (such as Bruck Allgather)

Measured Times

Dotted line : System MPI



Quartz Supercomputer



Lassen Supercomputer (CPUs only)