Parallel Programming

Prof. **Paolo Bientinesi**

pauldj@aices.rwth-aachen.de

WS 16/17

Collective Communication

```
          Barrier
Broadcast   ↔   Reduce
  Scatter   ↔   Gather
Allgather   ↔   Reduce-scatter
        Allreduce
         Alltoall
            ⋮
```

References

- "Collective Communication: Theory, Practice, and Experience", Chan, Heimlich, Purkayastha, van de Geijn. (FLAME working note #22)
- Collective Communications in MPI

  `http://www.mcs.anl.gov/research/projects/mpi/tutorial/gropp/node72.html`

Collective Communication

- Synchronization
  `Barrier` ← **Almost never needed!**
- Data Movement
  `Broadcast, Scatter, Gather, Allgather, Alltoall`
- Reductions
  `Reduce, Reduce-scatter, Allreduce, Scan, ...`

For all collectives: no tags; blocking.

`int MPI_BCast(...)`

Before:

Node_0	Node_1	Node_2	Node_3
		α	

After:

Node_0	Node_1	Node_2	Node_3
α	α	α	α

- How would you implement the broadcast in terms of inividual sends and receives?
- How many steps does it take to broadcast to np processes?

`int MPI_Reduce(...)`

Before:

Node_0	Node_1	Node_2	Node_3
δ_0	δ_1	δ_2	δ_3

After:

Node_0	Node_1	Node_2	Node_3
		$\text{Op}_{i=0}^{p-1} \delta_i$	

`MPI_Op`:
`MPI_MAX, MPI_MIN, MPI_SUM, MPI_PROD, MPI_LAND, MPI_BAND, ...,`

`Op`: Associative. Why is this needed?

`MPI_Datatype`:
`MPI_CHAR, MPI_INT, MPI_UNSIGNED, MPI_FLOAT, MPI_DOUBLE, ...`