



Collective Operations with Notified Communication in Shared Windows

**Muhammed Abdullah Al Ahad, Christian Simmendinger,
Roman Iakymchuk, Erwin Laure and Stefano Markidis**



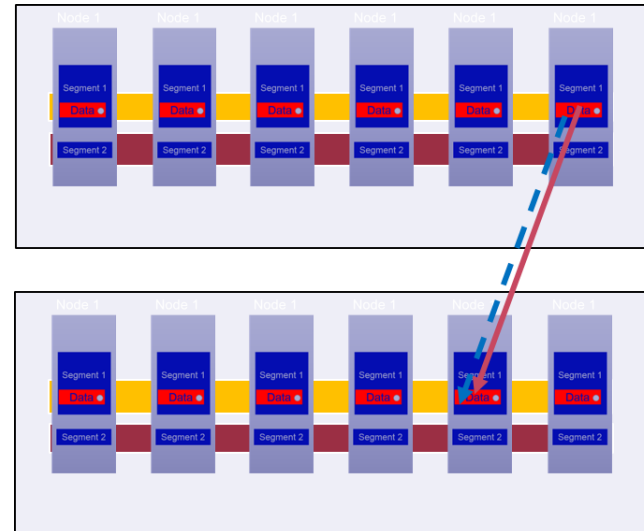
This project is funded from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 671602.

Notified communication in shared memory

Context: Extension of GASPI notified communication for shared memory.

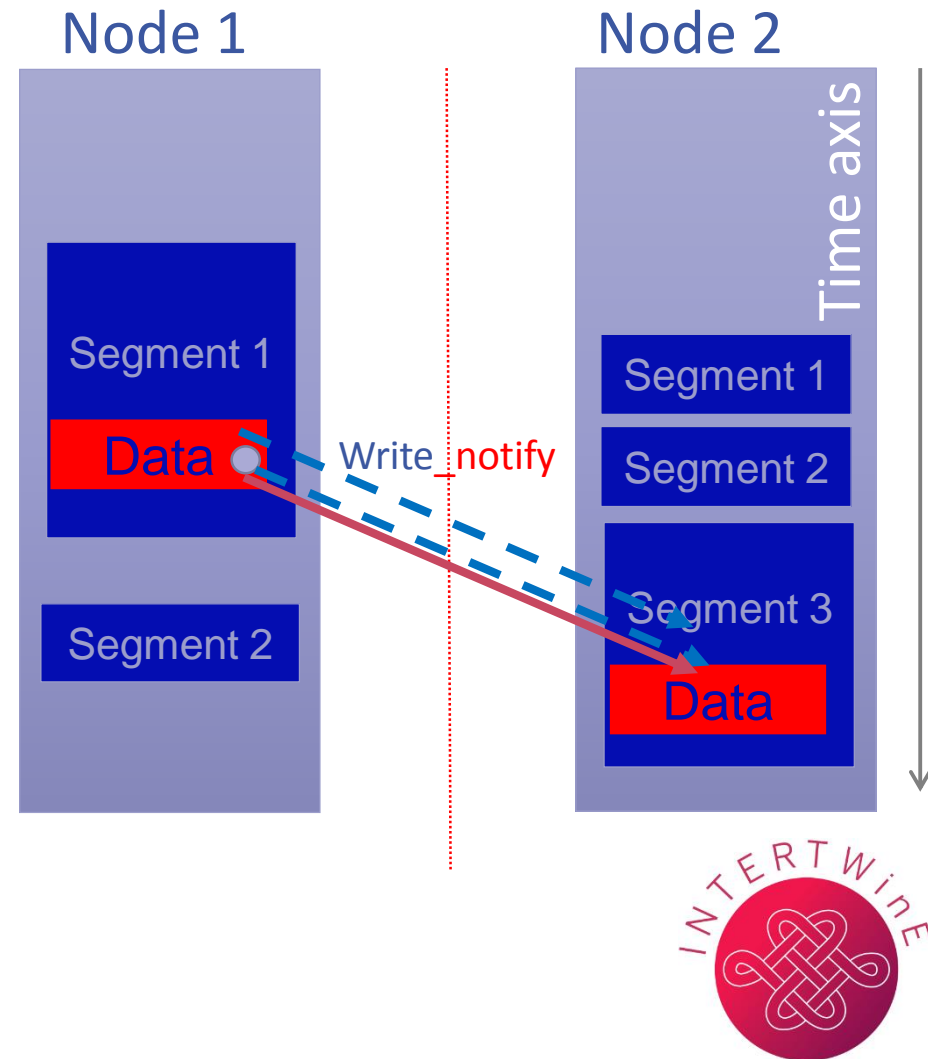
- Collectives as testbed for evaluating programmability and performance.

- Allreduce
 - Pipelined ring
 - Dissemination
- AllgatherV



Synchronisation with Notifications

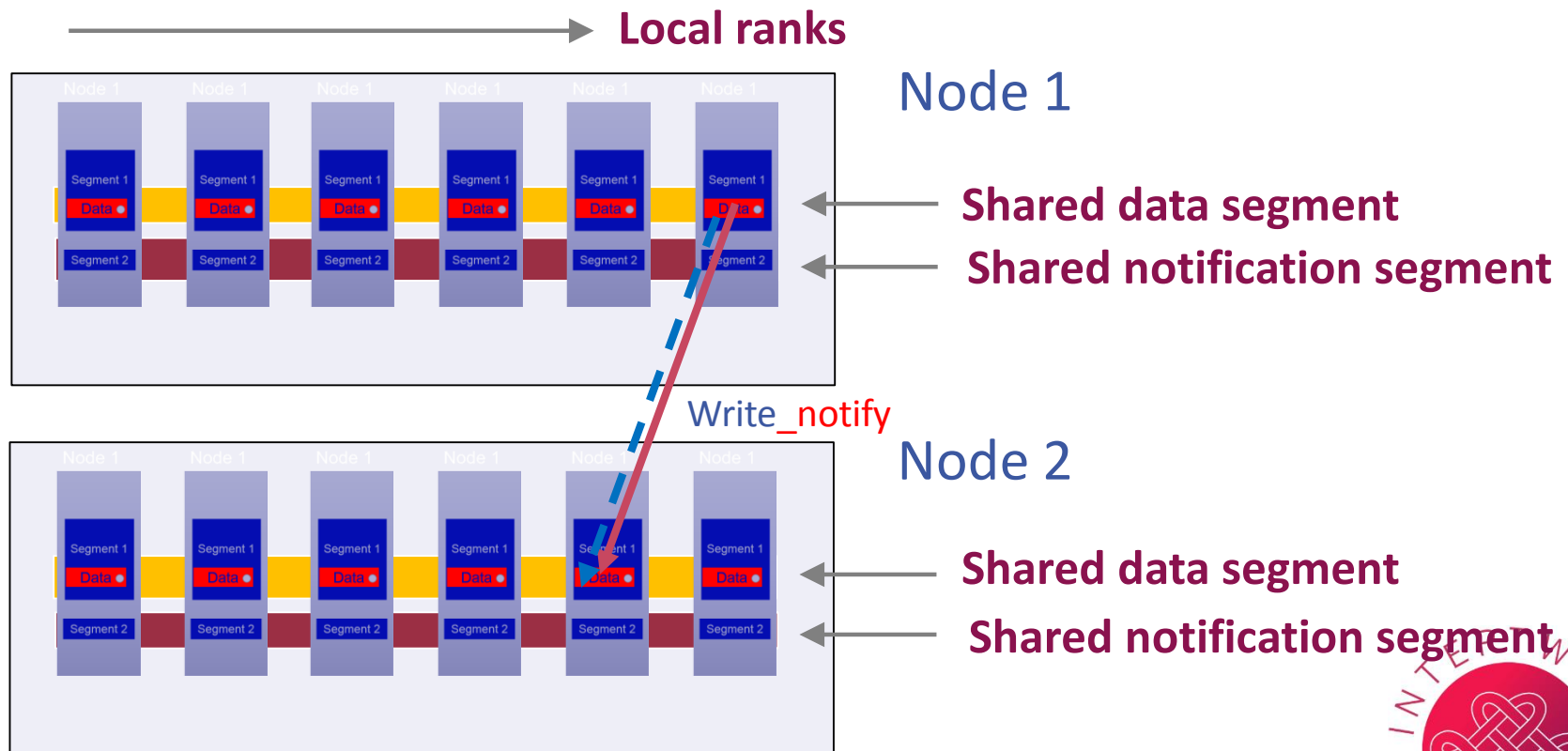
- Designated communication areas in GASPI are called memory segments.
- Segments are freely configurable.
- GASPI uses notified write and read (weak synchronization).
- GASPI also supports bundled write and notify as a single call: `write_notify` (and `read_notify`)
- GASPI is interoperable with MPI and allows incremental porting



Synchronisation with Notifications

- Q: How can we extend notified communication for shared segments ?

A: Use local notifications in shared data segments and share remote notifications across local ranks.



Testbed – Collectives in Shared Windows

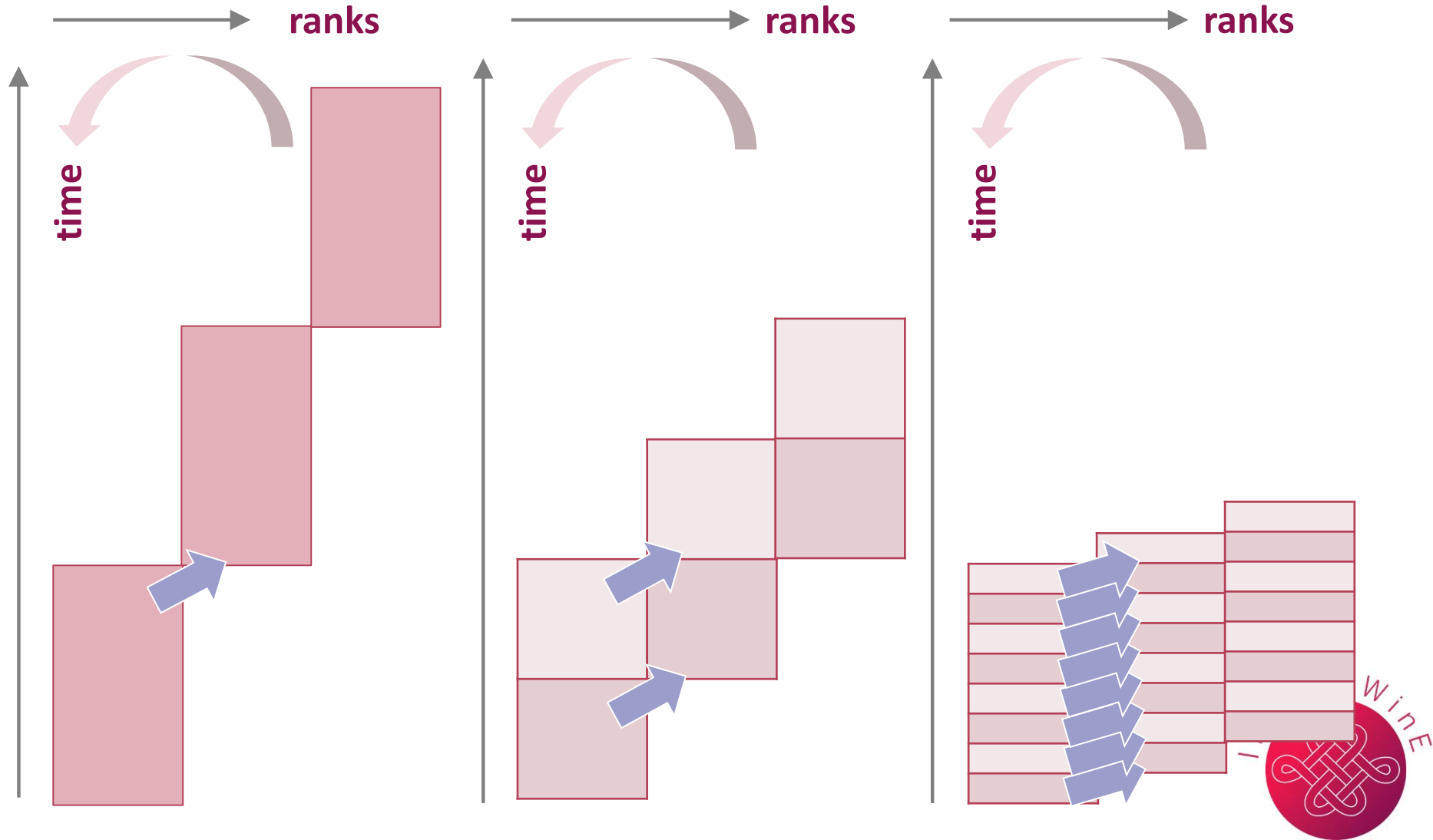
Testbed for notified communication - Collectives.

Allreduce with GASPI shared segments

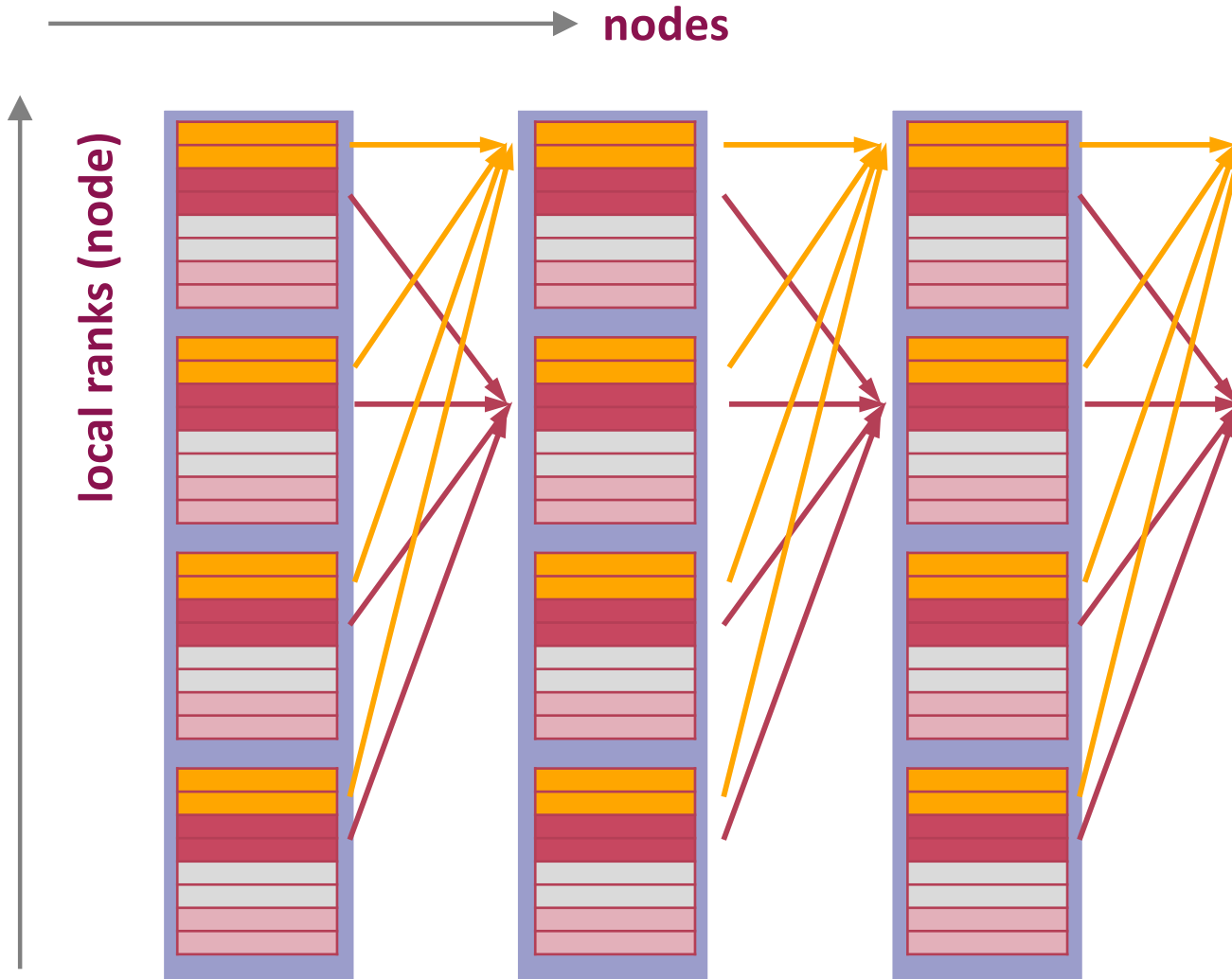
- Pipelined Ring for large message sizes
 - 2-stage pipelined process, reduce + broadcast
 - Highly parallel pipeline for fast ramp-up (latency $\sim 2(n-1)$)
 - Shared memory for node internal communication
 - Leverage available bandwidth / compute for reduce



Allreduce / Allgather(V) - Pipelined Ring



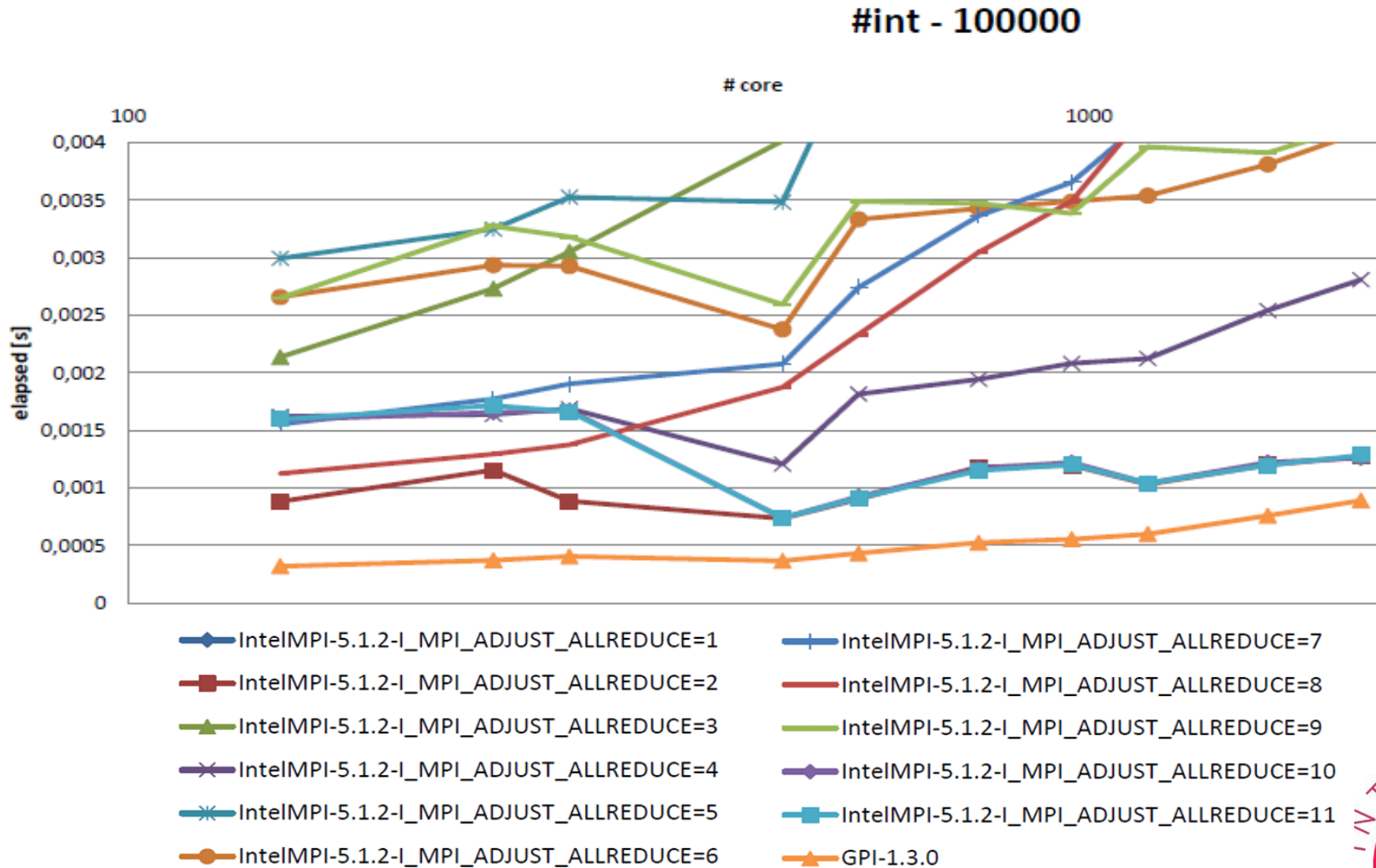
Allreduce / Allgather(V) - Load Distribution across Local Ranks



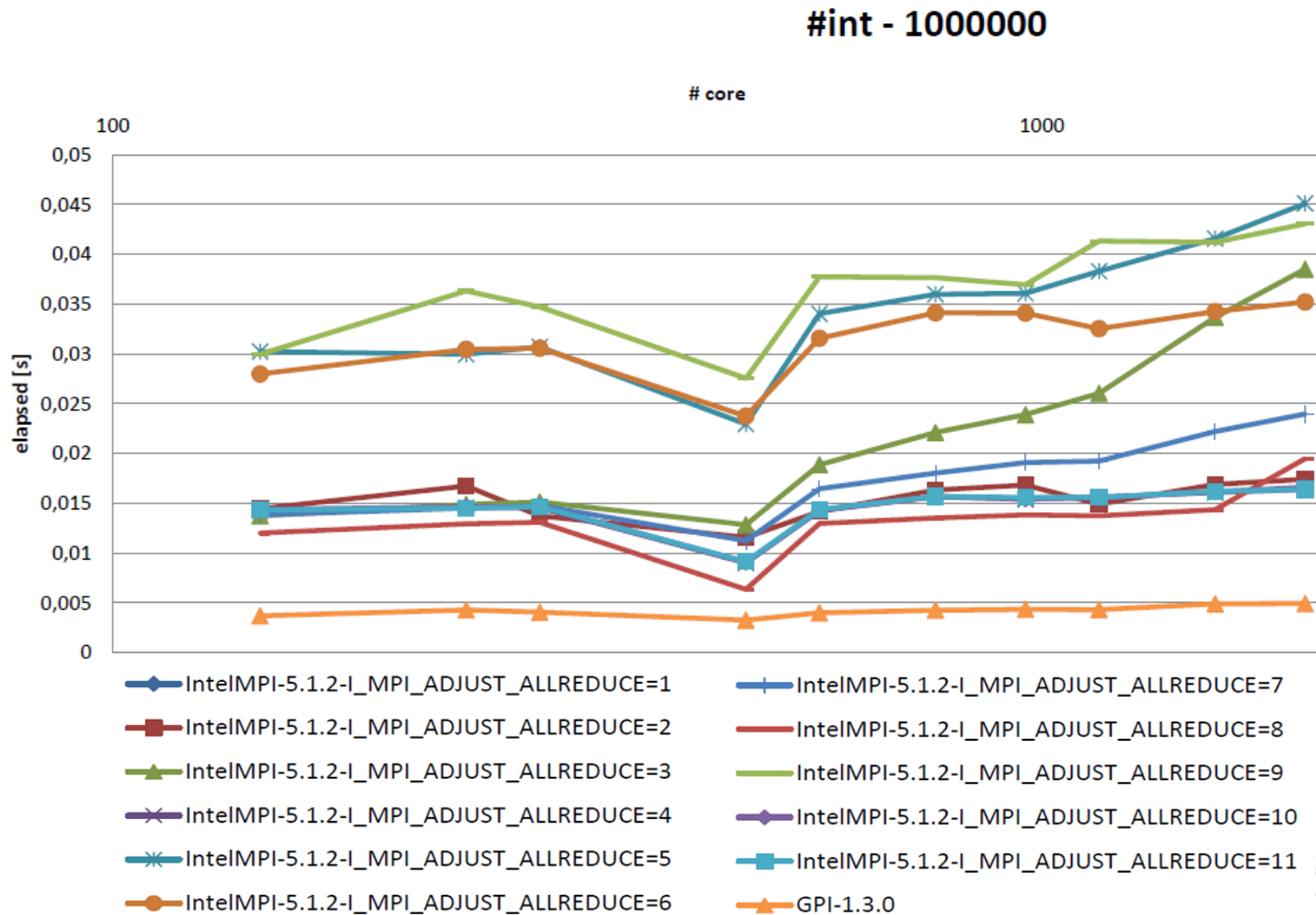
**All ranks
participate in
reducing and
forwarding data
at any point in
time**



Allreduce – Pipelined Ring



Allreduce – Pipelined Ring



Testbed – Collectives in Shared Windows

Testbed for notified communication - Collectives.

Allreduce with GASPI shared segments

- Dissemination algorithm
 - $\log_2(m)$ dissemination steps, where $m \leq 2^n$.
 - Hide communication of additionally required dissemination steps for $m \neq 2^n$
 - Decompose number of ranks as

$$n = \sum_{k=l}^0 c_k 2^k, \text{ with } c_k \in \{0, 1\}$$

- In step j , rank i writes additional out of band msg to target t where

$$t = \left(i + \sum_{k=l}^{j+1} c_k 2^k \right)$$



Allreduce – Async Dissemination Algorithm

Step 0

Diagram illustrating the initial state of the Async Dissemination Algorithm (Step 0). The diagram shows a grid of 15 ranks (0 to 14) and 15 states (0 to 14). The rank 0 is highlighted in green, indicating it is the source of the data. The state 0 is highlighted in green, indicating it is the initial state. The grid is labeled "ranks" horizontally and "state" vertically.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	0	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
3	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
4	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
5	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
6	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
7	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
8	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
9	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
10	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
11	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
12	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
13	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
14	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1



Allreduce - Async Dissemination Algorithm

Step 0

Diagram illustrating the initial state of the Allreduce - Async Dissemination Algorithm at Step 0. The diagram shows a 15x15 grid of values, where the horizontal axis represents ranks (0 to 14) and the vertical axis represents states (0 to 14). The grid is labeled "ranks" and "state".

state \ ranks	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	14	0	1	2	3	4	5	6	7	8	9	10	11	12	13
2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
3	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
4	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
5	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
6	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
7	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
8	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
9	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
10	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
11	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
12	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
13	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
14	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1



Allreduce - Async Dissemination Algorithm

Step 1

Diagram illustrating Step 1 of the Allreduce - Async Dissemination Algorithm. The diagram shows a 15x15 grid representing the state of the algorithm. The horizontal axis is labeled "ranks" and the vertical axis is labeled "state".

The grid contains values for each rank (0 to 14) and state (0 to 14). The values are as follows:

state \ ranks	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	14	0	1	2	3	4	5	6	7	8	9	10	11	12	13
2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
3	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
4	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
5	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
6	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
7	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
8	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
9	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
10	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
11	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
12	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
13	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
14	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1

Key features of the grid:

- The top row (state 0) contains the rank numbers 0 to 14.
- The leftmost column (rank 0) contains the state numbers 0 to 14.
- The diagonal elements (where rank equals state) are highlighted in orange (e.g., 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14).
- The off-diagonal elements are -1.
- The bottom-left cell (rank 0, state 14) is highlighted in green and contains the value 1.



Allreduce - Async Dissemination Algorithm

Step 1

Diagram illustrating Step 1 of the Allreduce - Async Dissemination Algorithm. The diagram shows a 15x15 grid of data, with rows labeled 'state' and columns labeled 'ranks'. The grid is divided into four quadrants by a vertical line between rank 3 and rank 4, and a horizontal line between state 12 and state 13.

The top-left quadrant (ranks 0-3, states 0-3) contains a 4x4 grid of values (0-14) in a specific pattern. The top-right quadrant (ranks 4-14, states 0-3) contains a 4x11 grid of values (0-14) in a specific pattern. The bottom-left quadrant (ranks 0-3, states 13-14) contains a 4x4 grid of values (0-14) in a specific pattern. The bottom-right quadrant (ranks 4-14, states 13-14) contains a 4x11 grid of values (0-14) in a specific pattern.

The grid is divided into four quadrants by a vertical line between rank 3 and rank 4, and a horizontal line between state 12 and state 13.

state \ ranks	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	14	0	1	2	3	4	5	6	7	8	9	10	11	12	13
2	13	14	0	1	2	3	4	5	6	7	8	9	10	11	12
3	12	13	14	0	1	2	3	4	5	6	7	8	9	10	11
4	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
5	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
6	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
7	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
8	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
9	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
10	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
11	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
12	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
13	3	4	5	6	7	8	9	10	11	12	13	14	0	1	2
14	2	3	4	5	6	7	8	9	10	11	12	13	14	0	1
15	1	2	3	4	5	6	7	8	9	10	11	12	13	14	0



Allreduce - Async Dissemination Algorithm

Step 2

Diagram illustrating Step 2 of the Allreduce - Async Dissemination Algorithm. The diagram shows a 15x15 grid of data, with rows labeled 'state' and columns labeled 'ranks'. The grid is divided into four quadrants, each containing a 4x4 sub-grid of data. The top-left quadrant (ranks 0-3, states 0-3) is highlighted in orange. The top-right quadrant (ranks 4-7, states 0-3) is highlighted in blue. The bottom-left quadrant (ranks 0-3, states 10-13) is highlighted in orange. The bottom-right quadrant (ranks 4-7, states 10-13) is highlighted in blue. The middle two rows (states 4-7) are filled with -1. The bottom row (state 14) is filled with -1.

state \ ranks	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	14	0	1	2	3	4	5	6	7	8	9	10	11	12	13
2	13	14	0	1	2	3	4	5	6	7	8	9	10	11	12
3	12	13	14	0	1	2	3	4	5	6	7	8	9	10	11
4	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
5	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
6	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
7	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
8	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
9	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
10	3	4	5	6	7	8	9	10	11	12	13	14	0	1	2
11	2	3	4	5	6	7	8	9	10	11	12	13	14	0	1
12	1	2	3	4	5	6	7	8	9	10	11	12	13	14	0



Allreduce - Async Dissemination Algorithm

Step 2

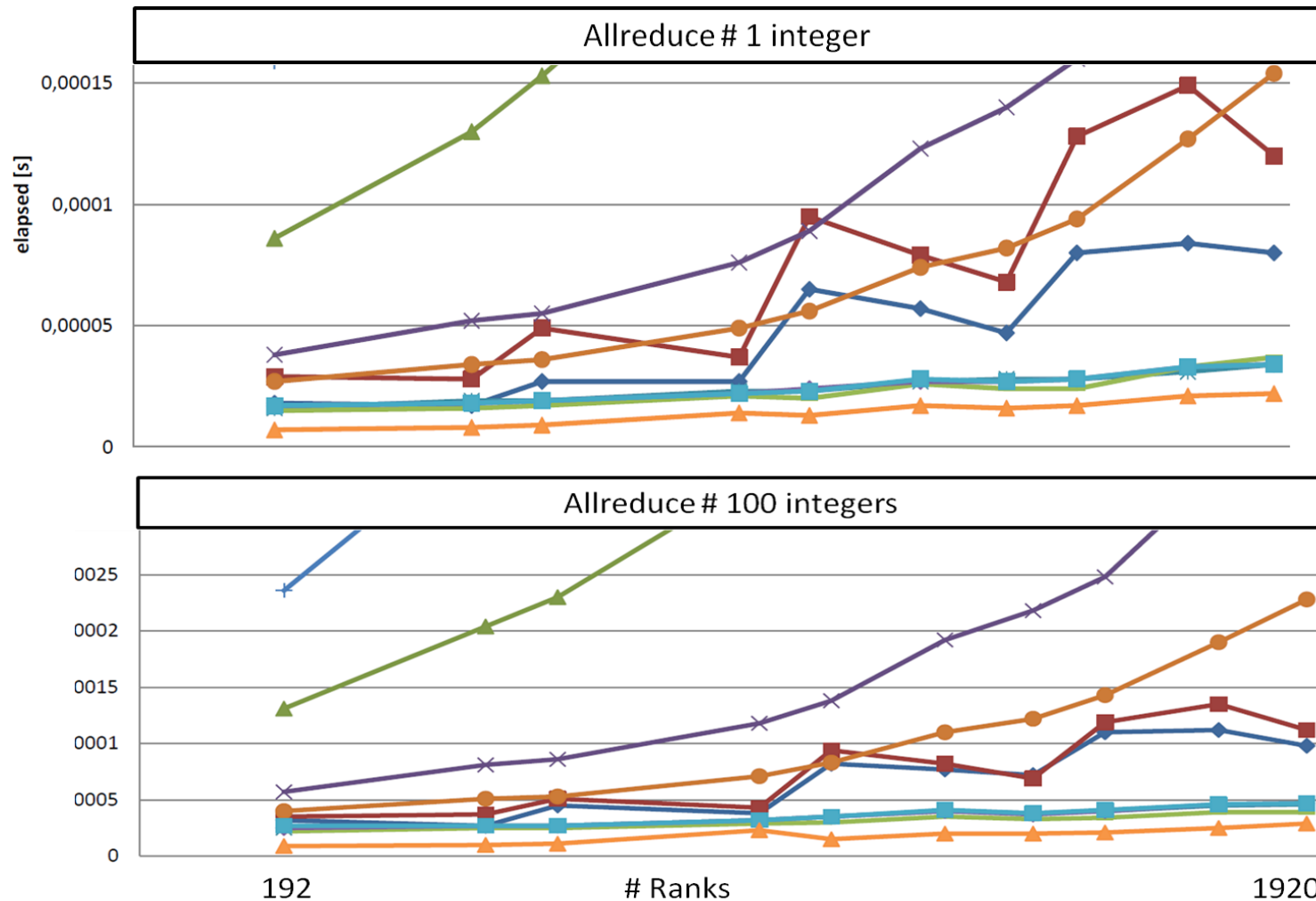
Diagram illustrating the Async Dissemination Algorithm (Step 2) using a 15x15 grid. The horizontal axis is labeled **ranks** and the vertical axis is labeled **state**.

The grid shows the state of each rank (0-14) across 15 states (0-14). The values represent the rank of the source of the data being disseminated. The grid is divided into four colored regions: yellow (bottom-left), green (bottom-middle), blue (middle-left), and purple (top-right).

state \ ranks	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
14	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
13	14	0	1	2	3	4	5	6	7	8	9	10	11	12	13
12	13	14	0	1	2	3	4	5	6	7	8	9	10	11	12
11	12	13	14	0	1	2	3	4	5	6	7	8	9	10	11
10	11	12	13	14	0	1	2	3	4	5	6	7	8	9	10
9	10	11	12	13	14	0	1	2	3	4	5	6	7	8	9
8	9	10	11	12	13	14	0	1	2	3	4	5	6	7	8
7	8	9	10	11	12	13	14	0	1	2	3	4	5	6	7
6	7	8	9	10	11	12	13	14	0	1	2	3	4	5	6
5	6	7	8	9	10	11	12	13	14	0	1	2	3	4	5
4	5	6	7	8	9	10	11	12	13	14	0	1	2	3	4
3	4	5	6	7	8	9	10	11	12	13	14	0	1	2	3
2	3	4	5	6	7	8	9	10	11	12	13	14	0	1	2
1	2	3	4	5	6	7	8	9	10	11	12	13	14	0	1



Allreduce - Dissemination

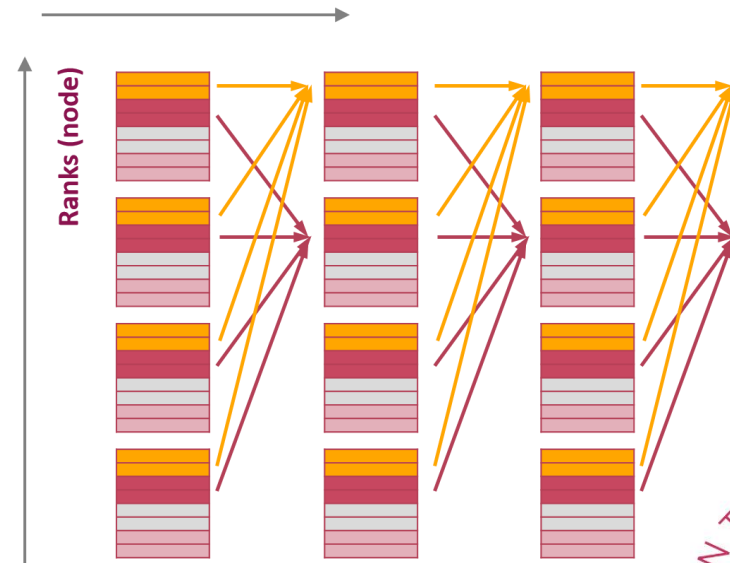
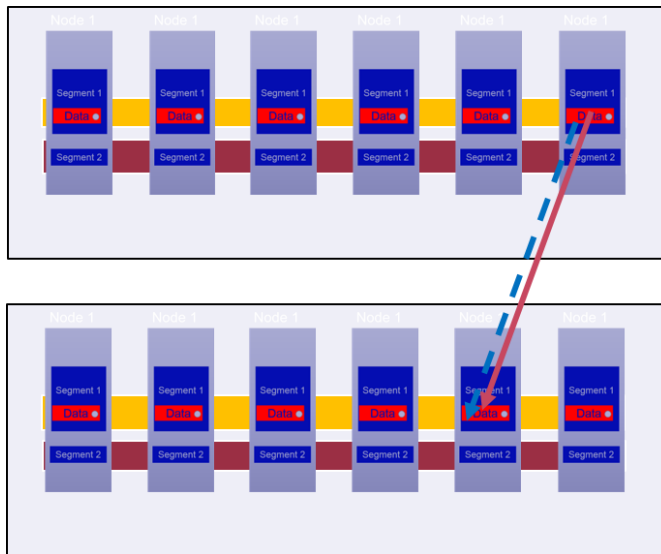


Allgather(V)

Testbed for notified communication - Collectives.

Allgather(V) with GASPI shared segments

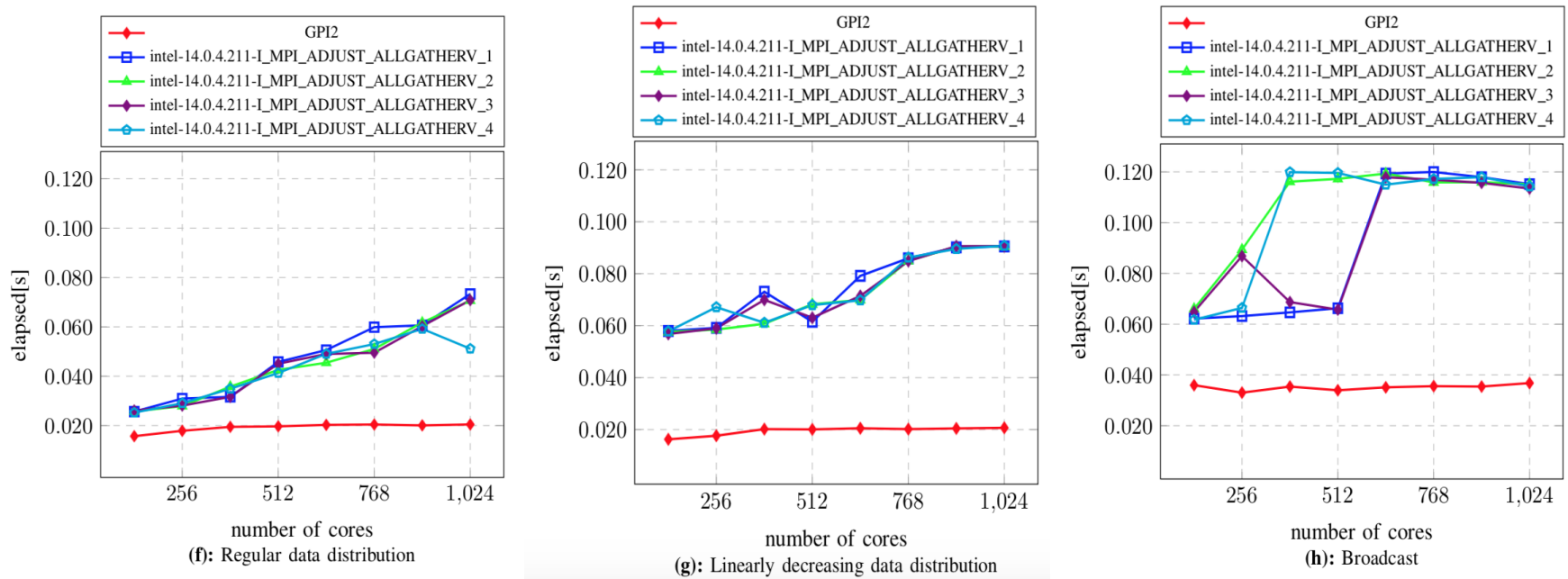
- Irregular data distribution
- Pipelined ring with dynamic load distribution



Allgather(V) : Performance results

Data distribution

- Regular, linear decreasing, broadcast





Questions?

Thank you for your attention



This project is funded from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 671602.