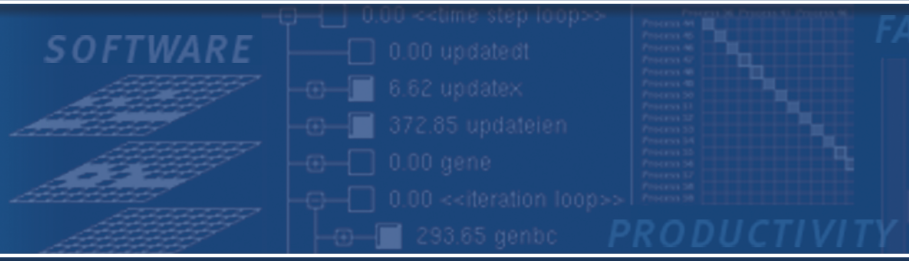
VI-HPS



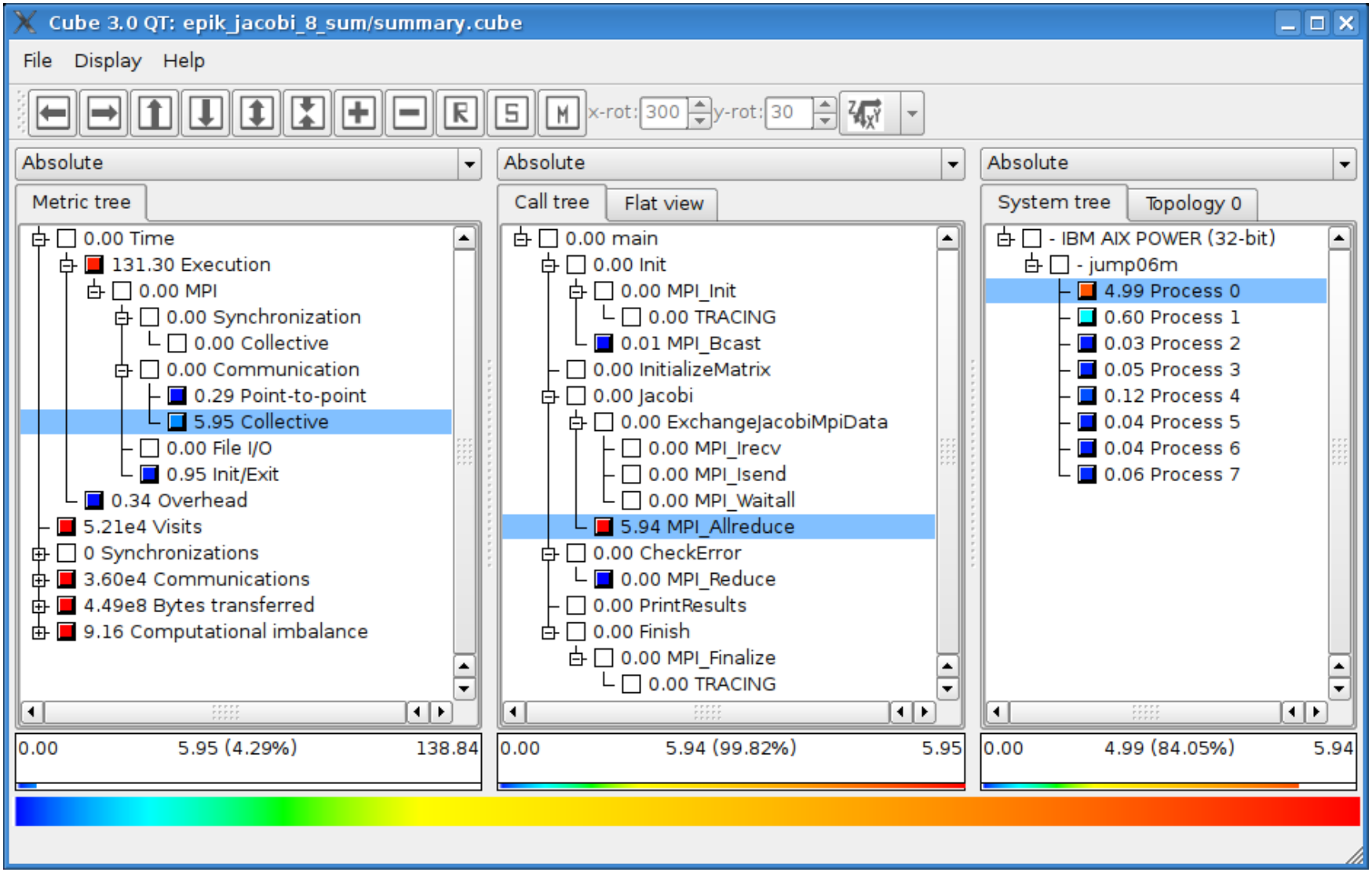
scalasca

Performance properties

"The metrics tour"

Markus Geimer & Brian Wylie
Jülich Supercomputing Centre
scalasca@fz-juelich.de
September 2011

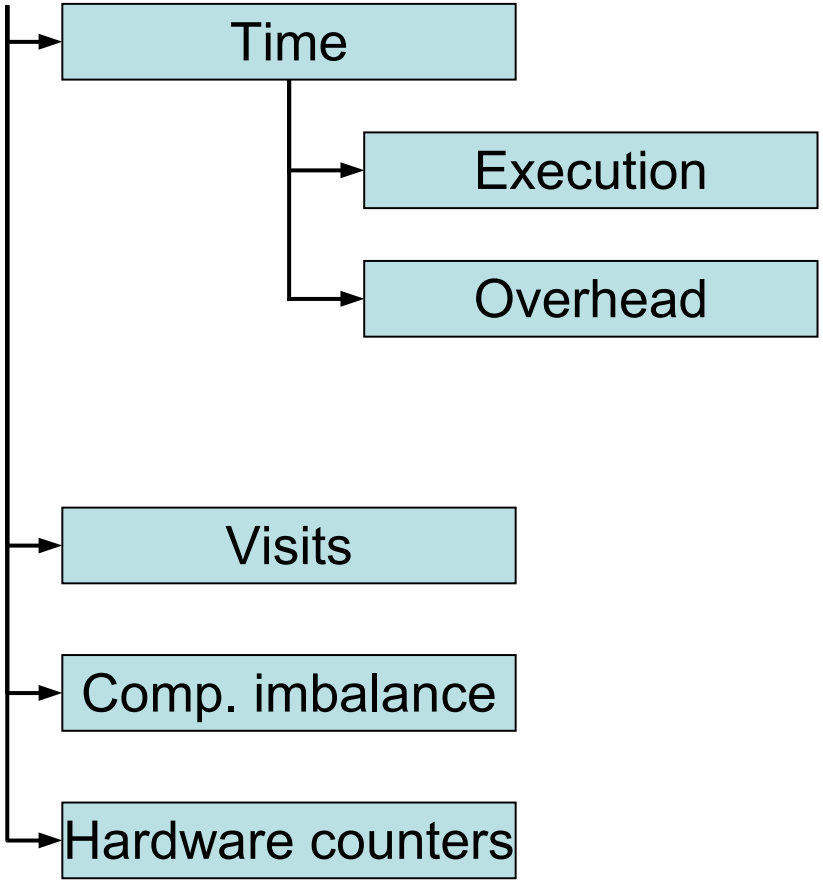
Scalasca analysis result



Confused?

Generic metrics

Generic metrics



Total CPU allocation time

Execution time without overhead

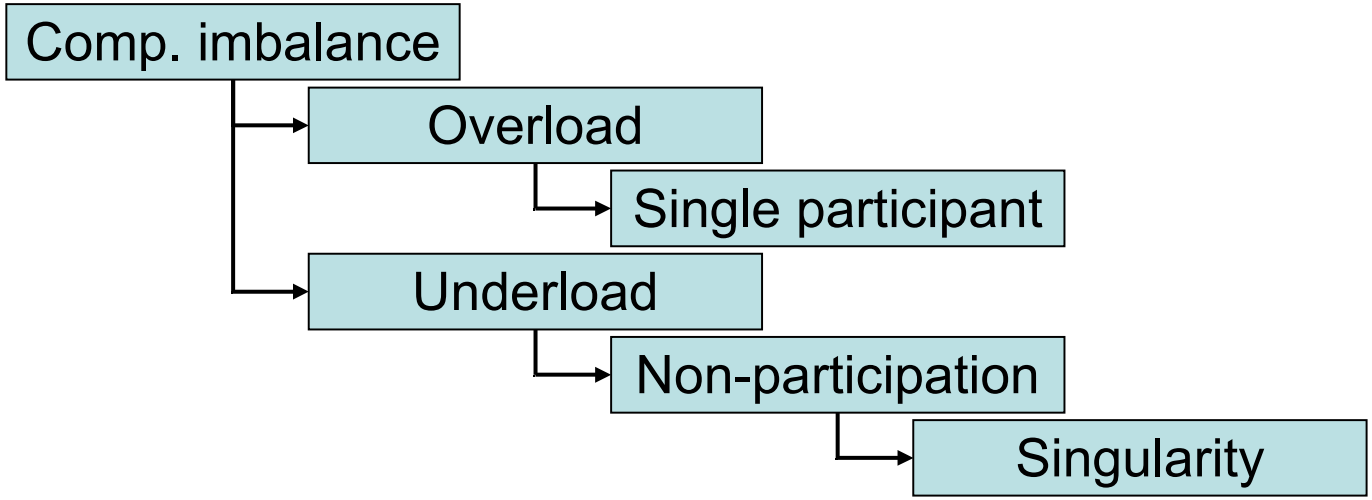
Time spent in tasks related to measurement (Does not include per-function perturbation!)

Number of times a function/region was executed

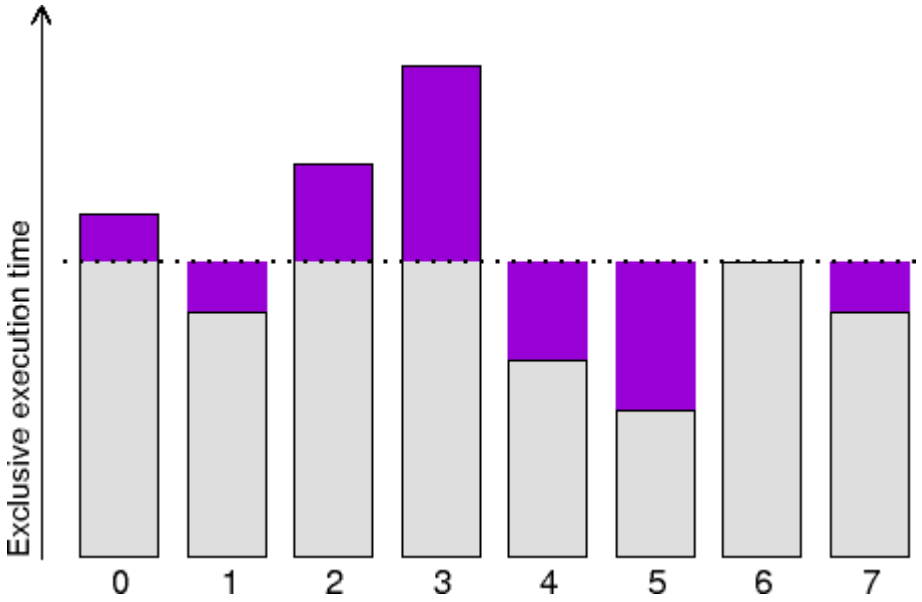
Simple load imbalance heuristics

Aggregated counter values for each function/region

Computational imbalance hierarchy

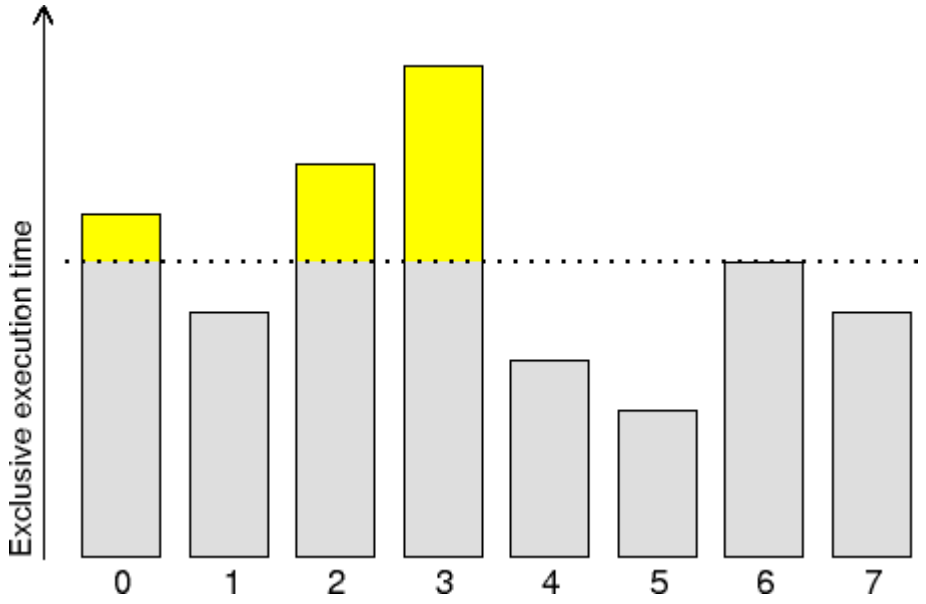


- Absolute difference to average exclusive execution time
  - Focusses only on computational parts
- Captures global imbalances
  - Based on entire measurement
  - Does not compare individual instances of function calls
- High value = Imbalance in the sub-calltree underneath
  - Expand the subtree to find the real location of the imbalance



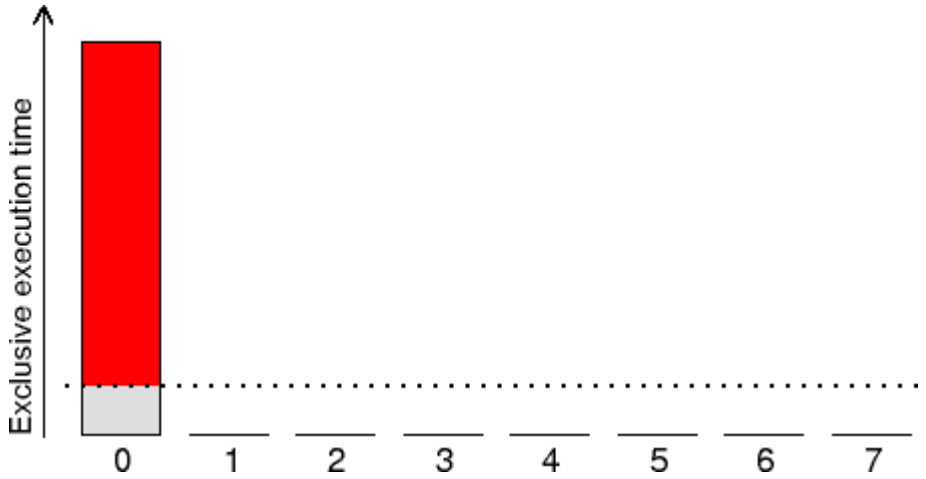
Overload

- Identifies processes/threads were exclusive execution time for the call-path was above average



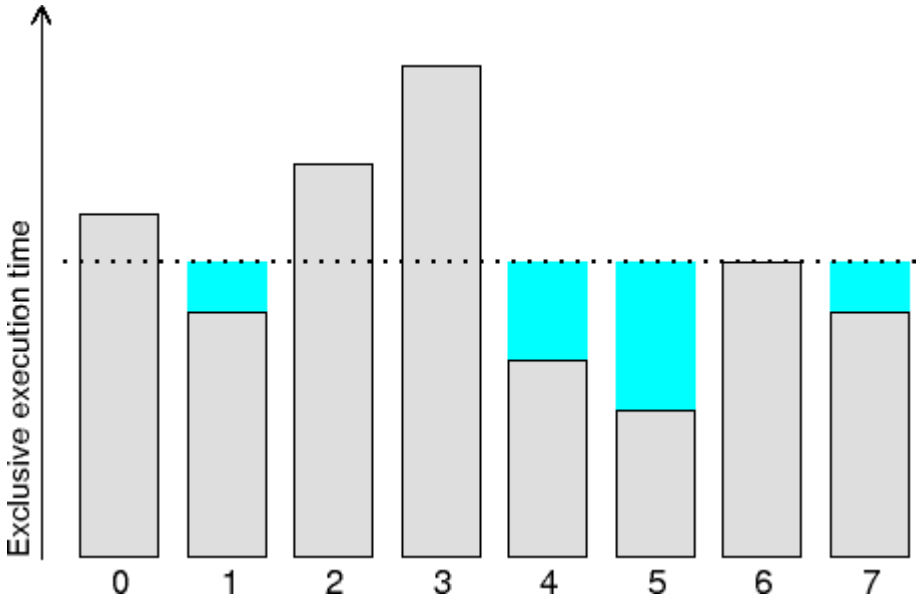
Overload, Single participant

- Identifies call-paths executed by single process/thread



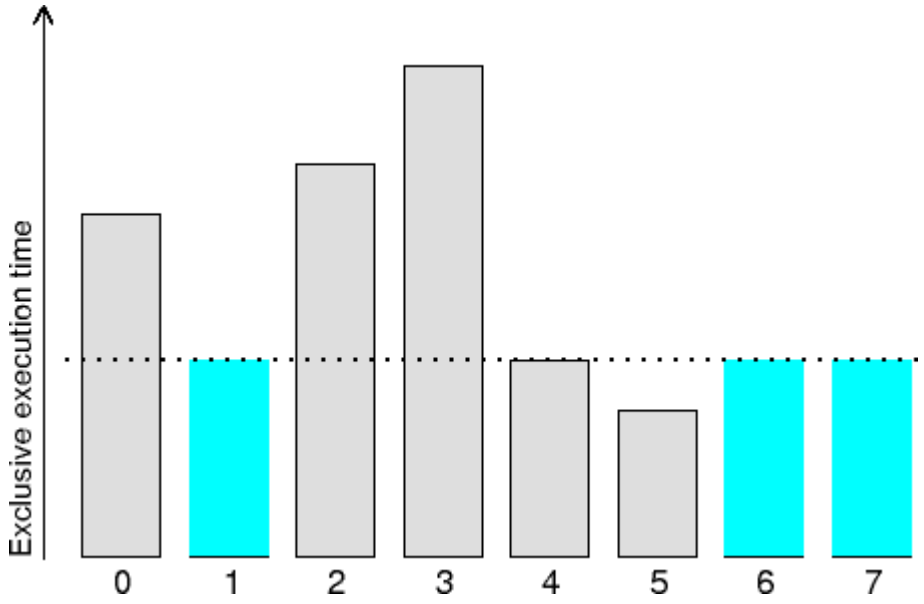
Underload

- Identifies processes/threads were exclusive execution time for the call-path was below average



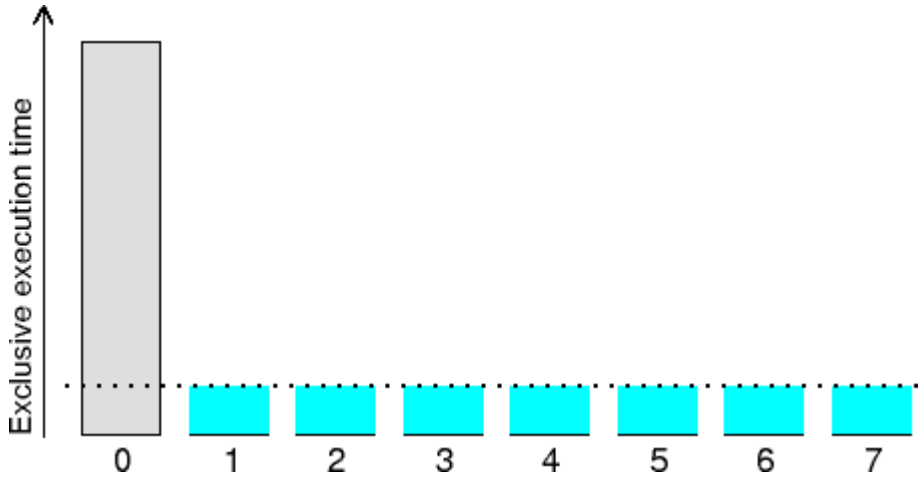
Underload, Non-participation

- Identifies call-paths not executed by a subset of processes/threads



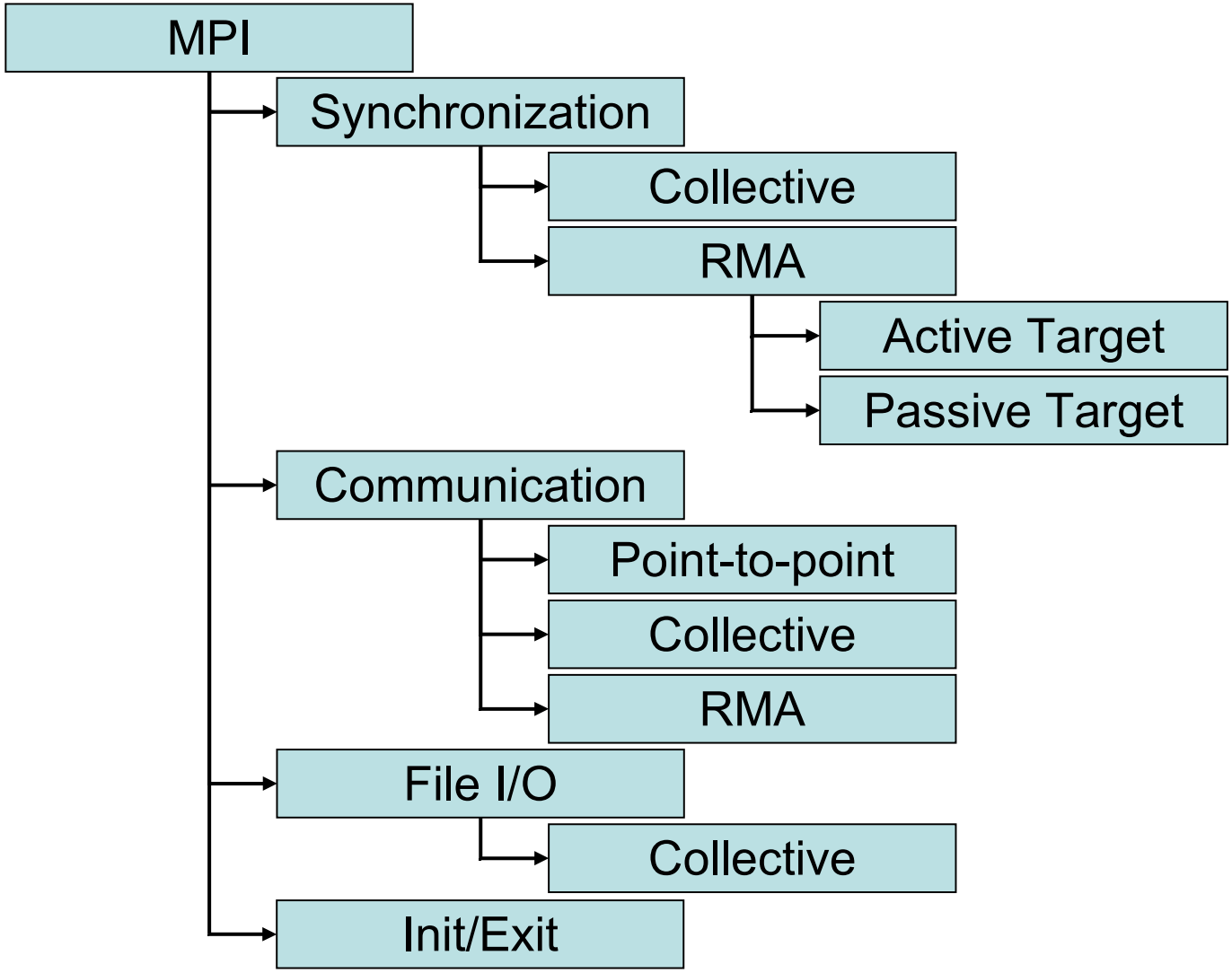
Underload, Singularity

- Identifies call-paths not executed by all but a single process/thread



MPI-related metrics

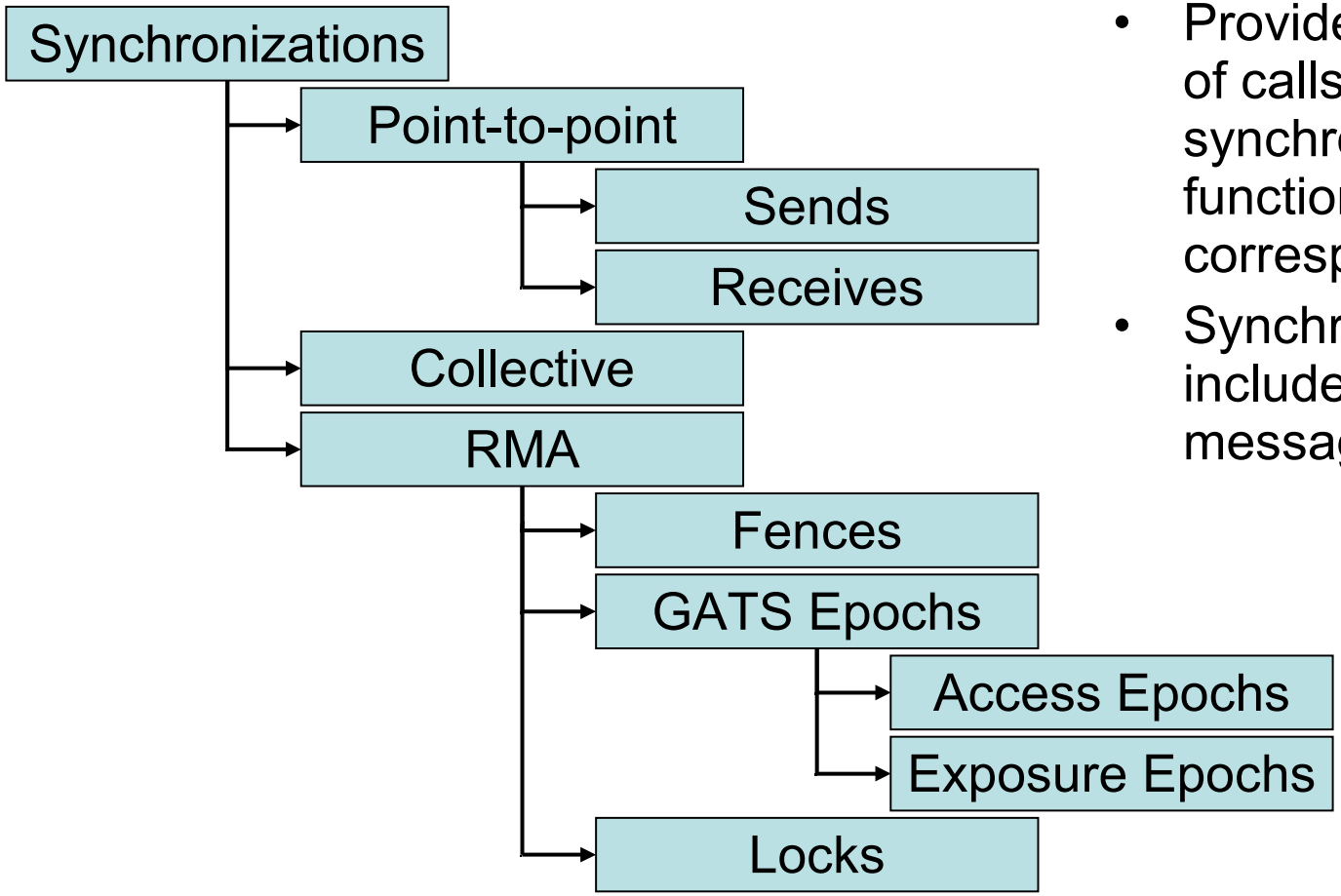
MPI Time hierarchy



MPI Time hierarchy details

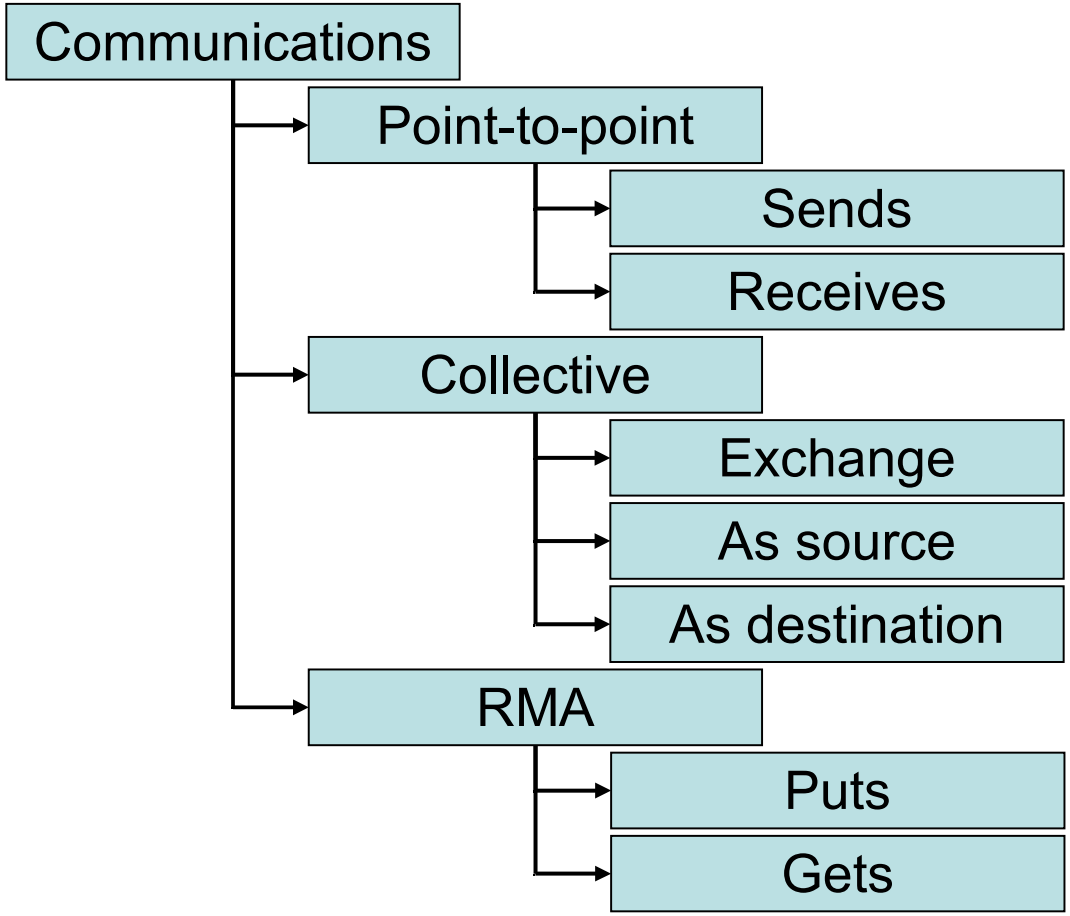
MPI	Time spent in pre-instrumented MPI functions
Synchronization	Time spent in calls to `MPI_Barrier` or Remote Memory Access synchronization calls
Communication	Time spent in MPI communication calls, subdivided into collective, point-to-point and RMA
File I/O	Time spent in MPI file I/O functions, with specialization for collective I/O calls
Init/Exit	Time spent in `MPI_Init` and `MPI_Finalize`

MPI Synchronizations hierarchy



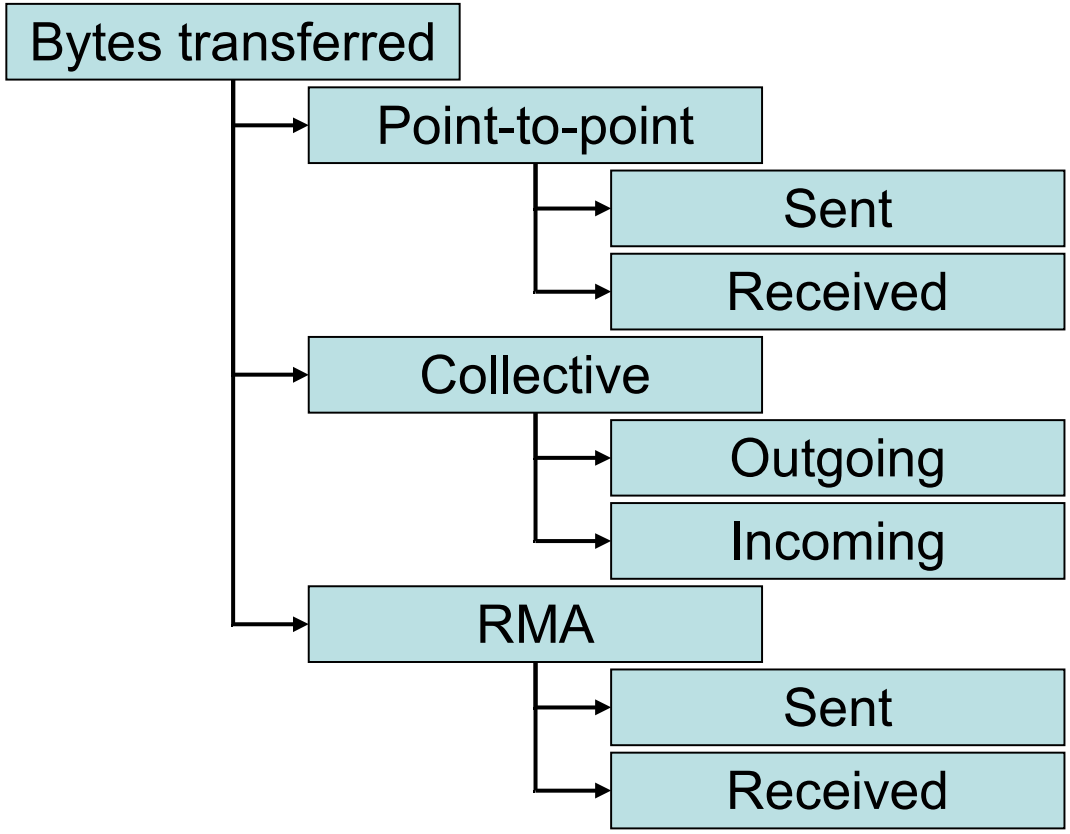
- Provides the number of calls to an MPI synchronization function of the corresponding class
- Synchronizations include zero-sized message transfers!

MPI Communications hierarchy



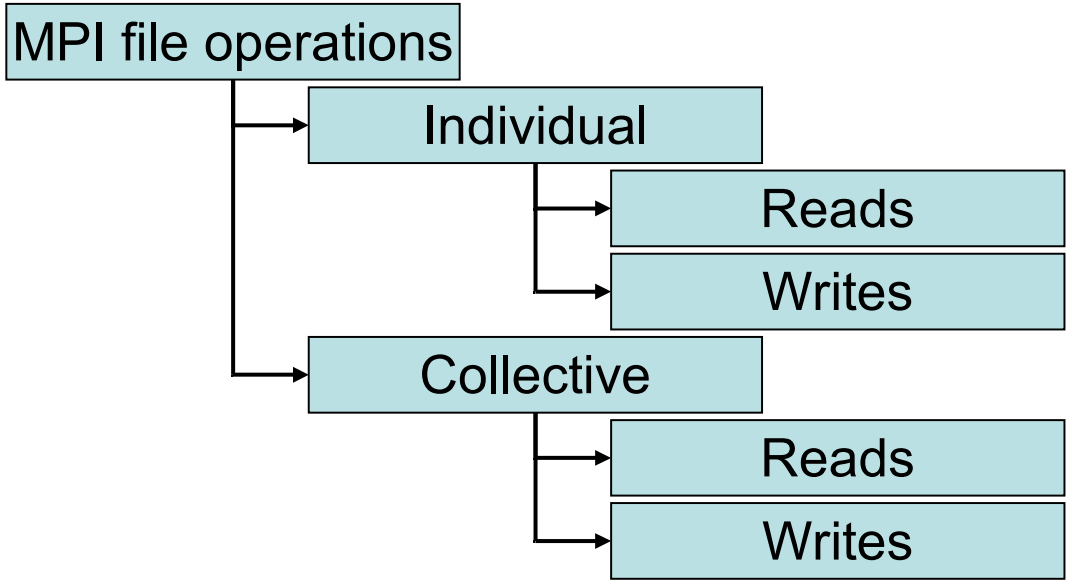
- Provides the number of calls to an MPI communication function of the corresponding class
- Zero-sized message transfers are considered synchronization!

MPI Transfer hierarchy



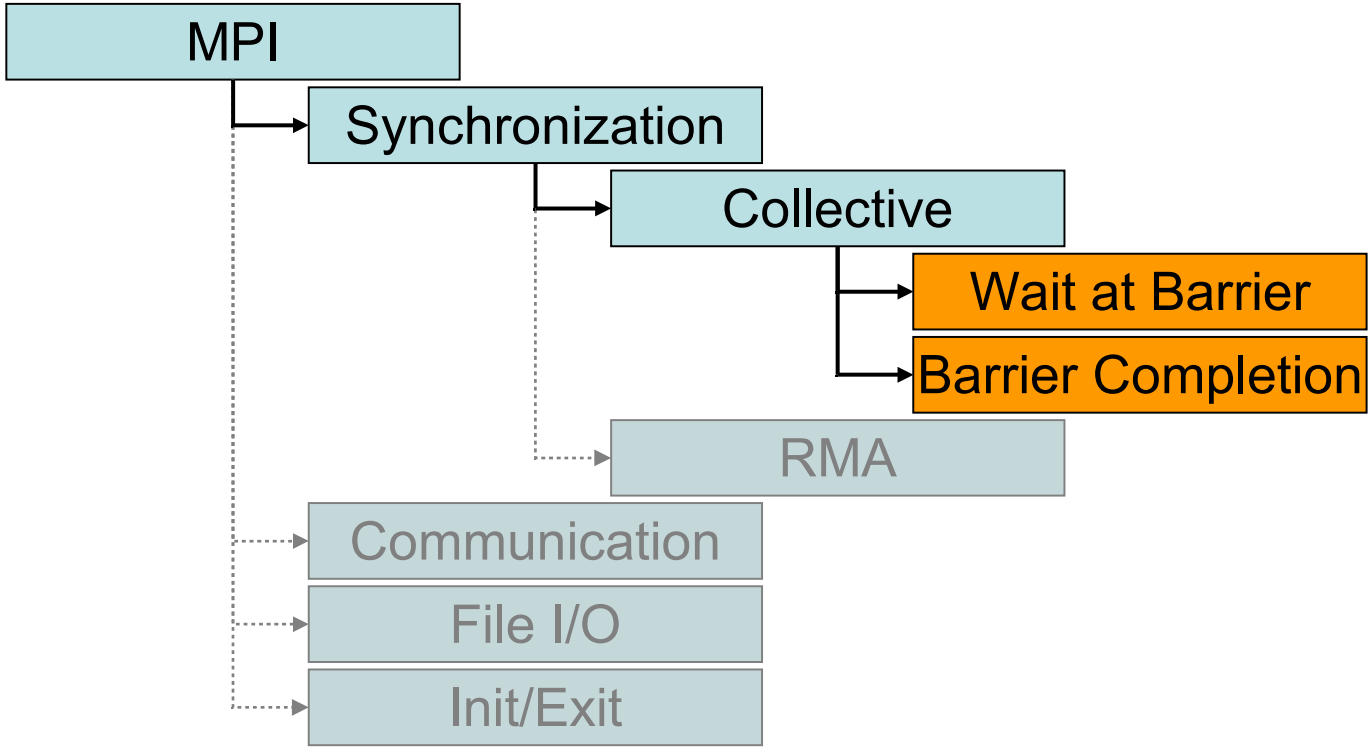
- Provides the number of bytes transferred by an MPI communication function of the corresponding class

MPI File operations hierarchy

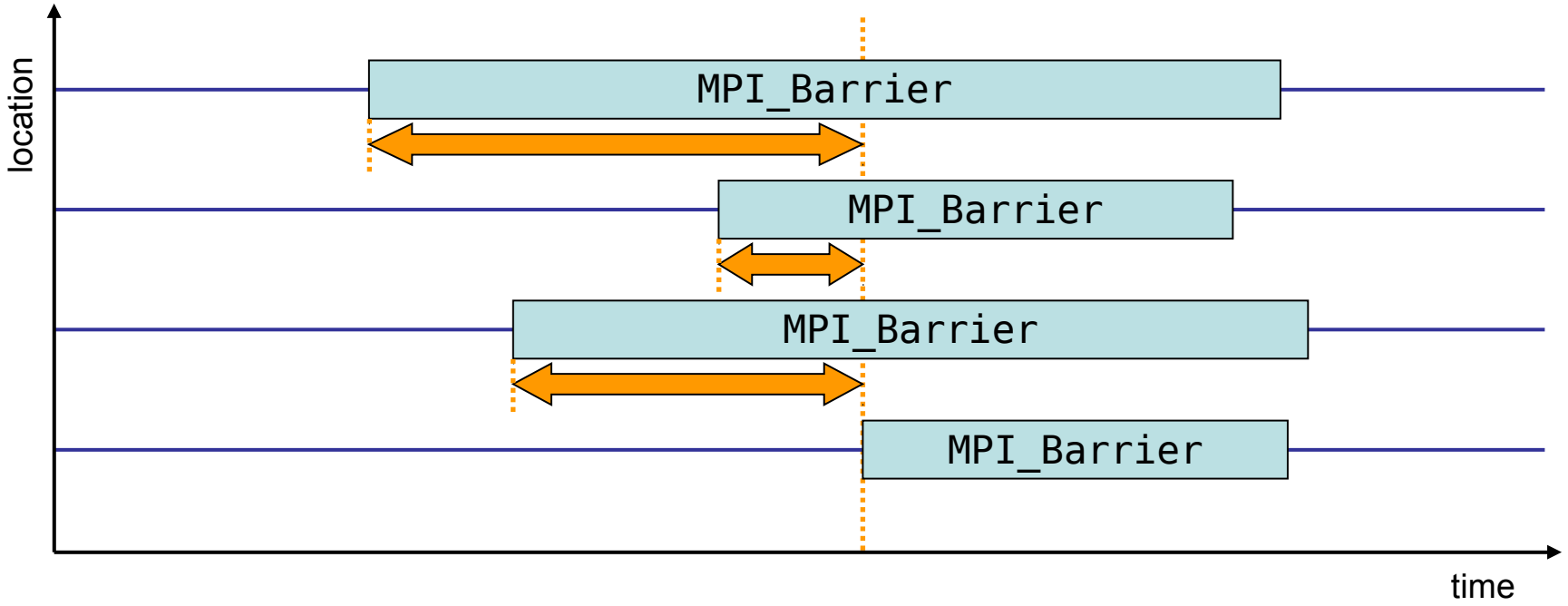


- Provides the number of calls to MPI file I/O functions of the corresponding class

MPI collective synchronization time

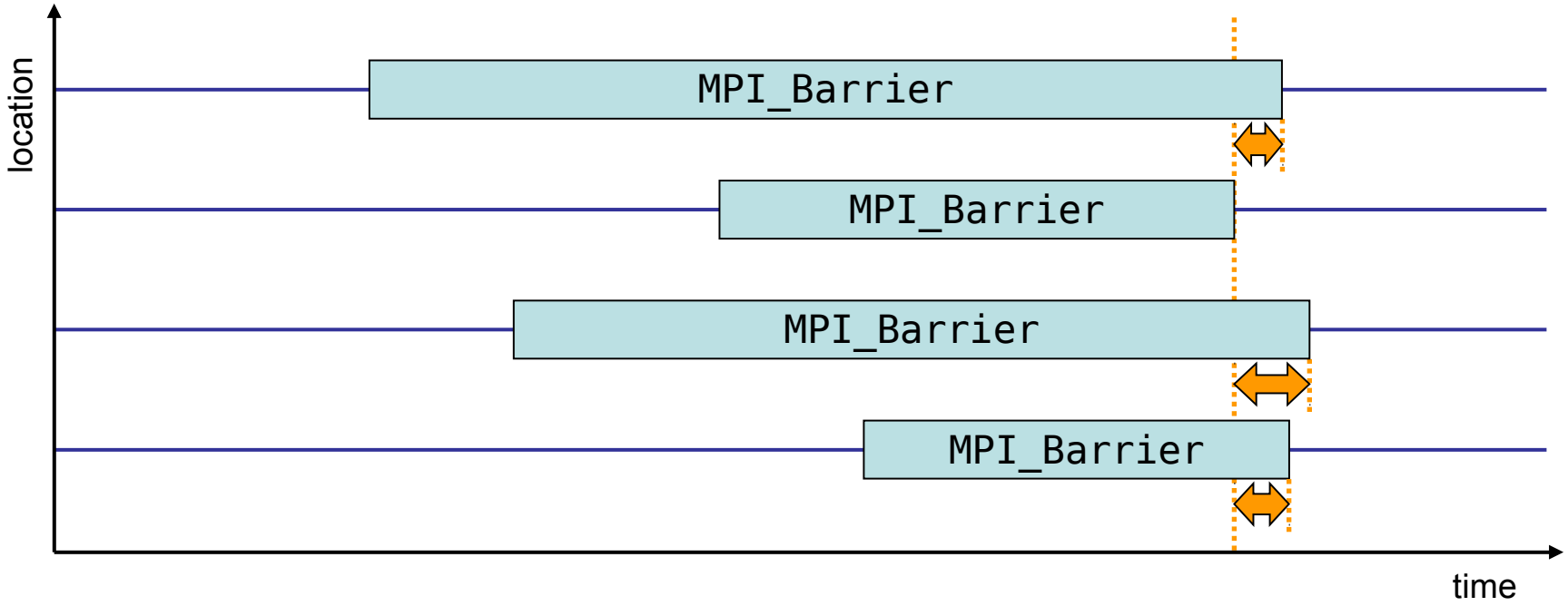


Wait at Barrier



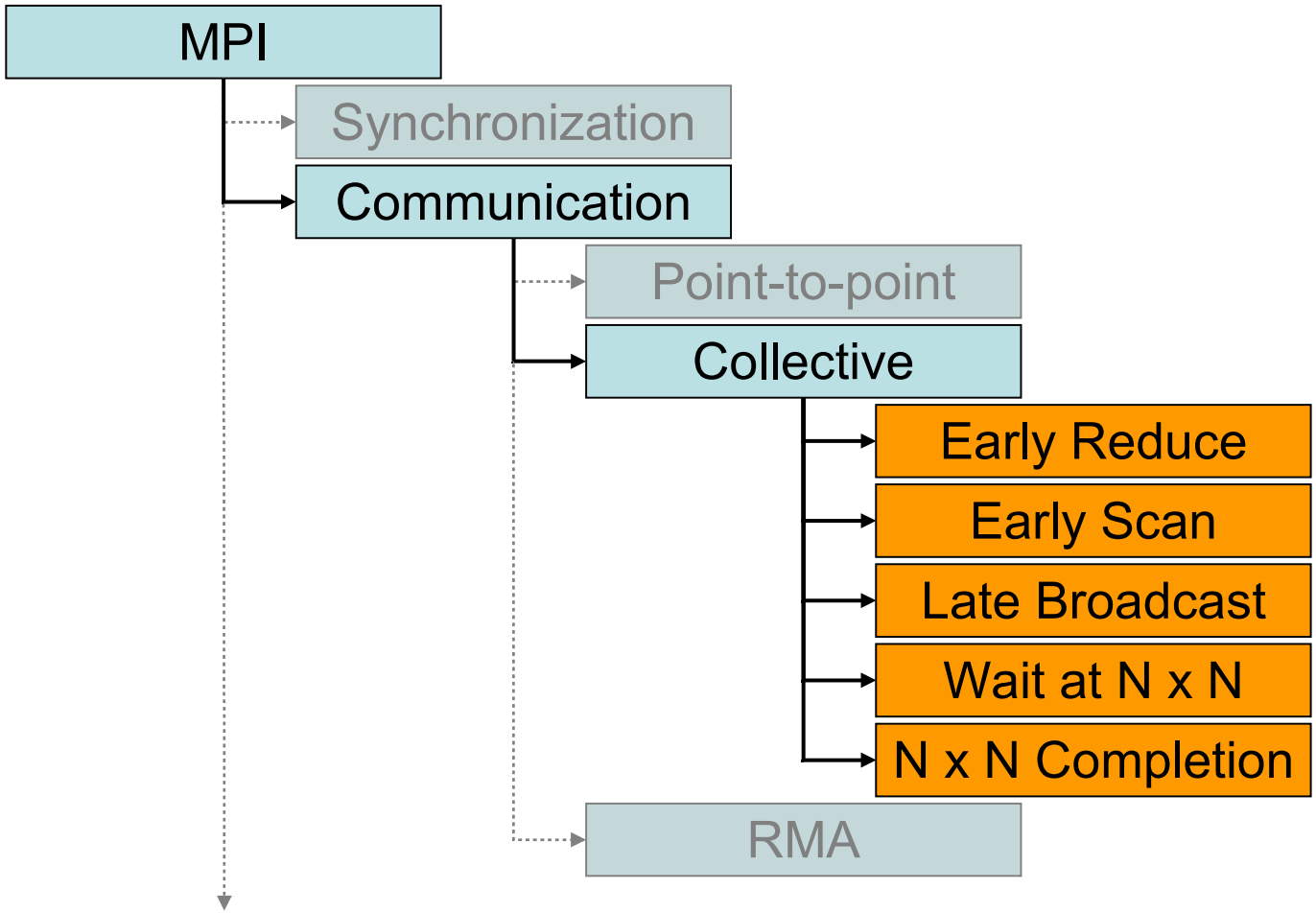
- Time spent waiting in front of a barrier call until the last process reaches the barrier operation
- Applies to: `MPI_Barrier`

Barrier Completion

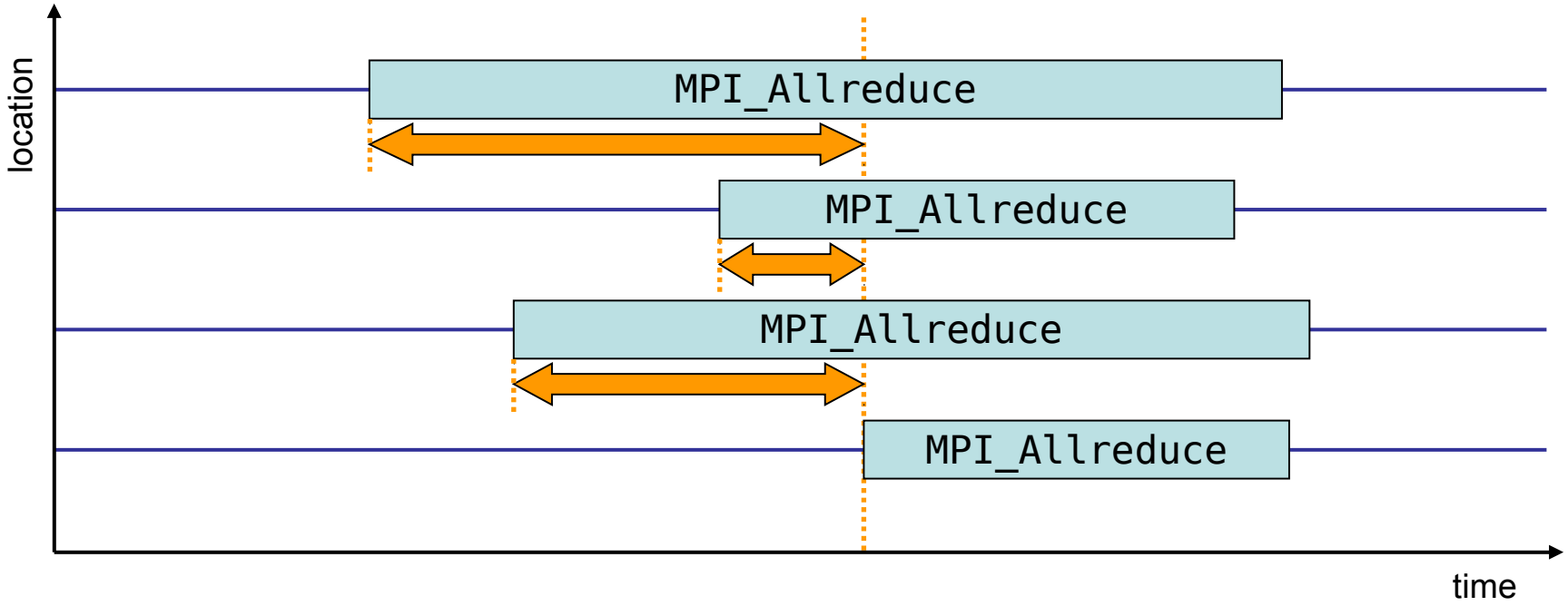


- Time spent in barrier after the first process has left the operation
- Applies to: `MPI_Barrier`

MPI collective communication time

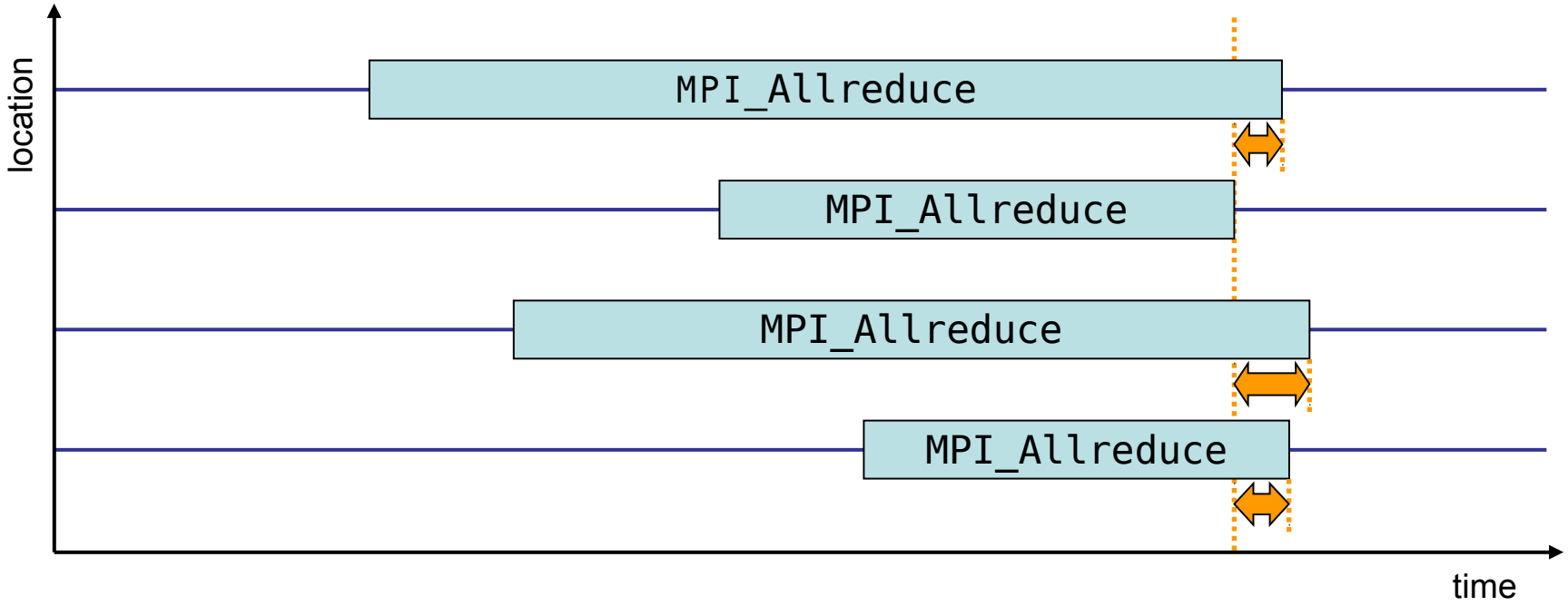


Wait at N x N



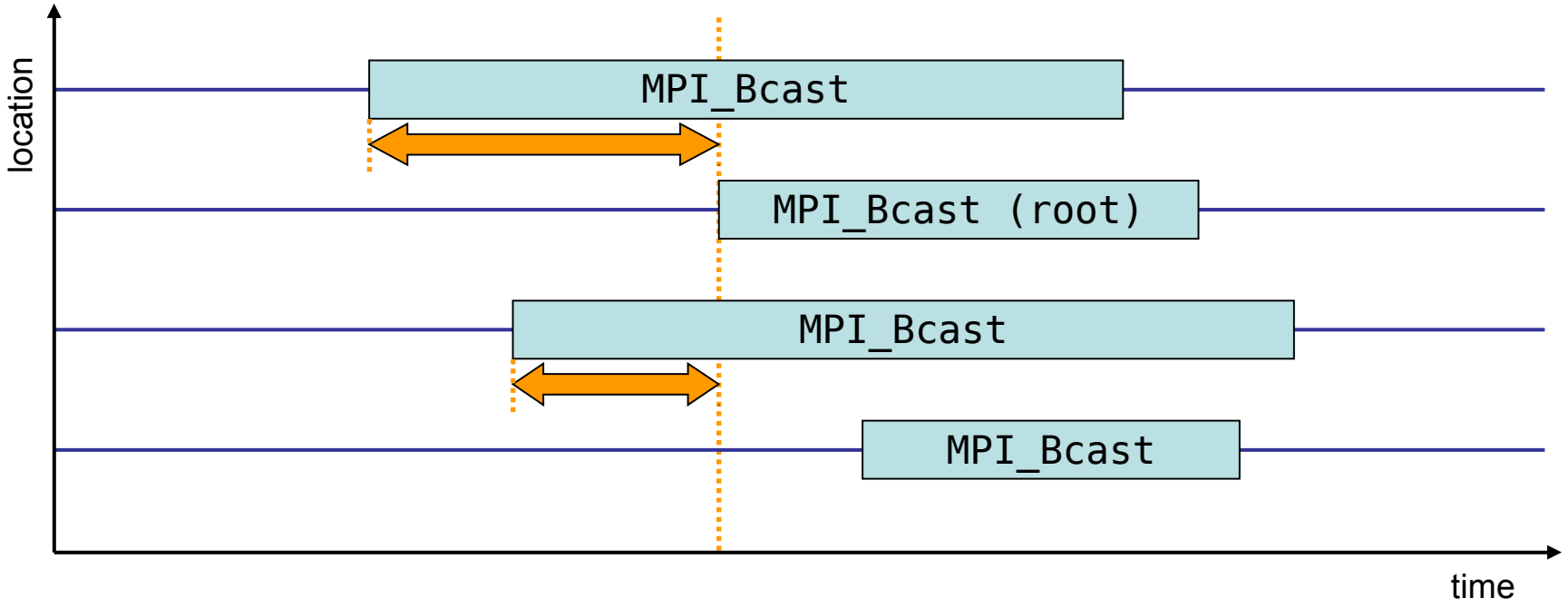
- Time spent waiting in front of a synchronizing collective operation call until the last process reaches the operation
- Applies to:
  `MPI_Allgather`, `MPI_Allgatherv`,
  `MPI_Allreduce`, `MPI_Alltoall`,
  `MPI_Reduce_scatter`, `MPI_Reduce_scatter_block`

N x N Completion



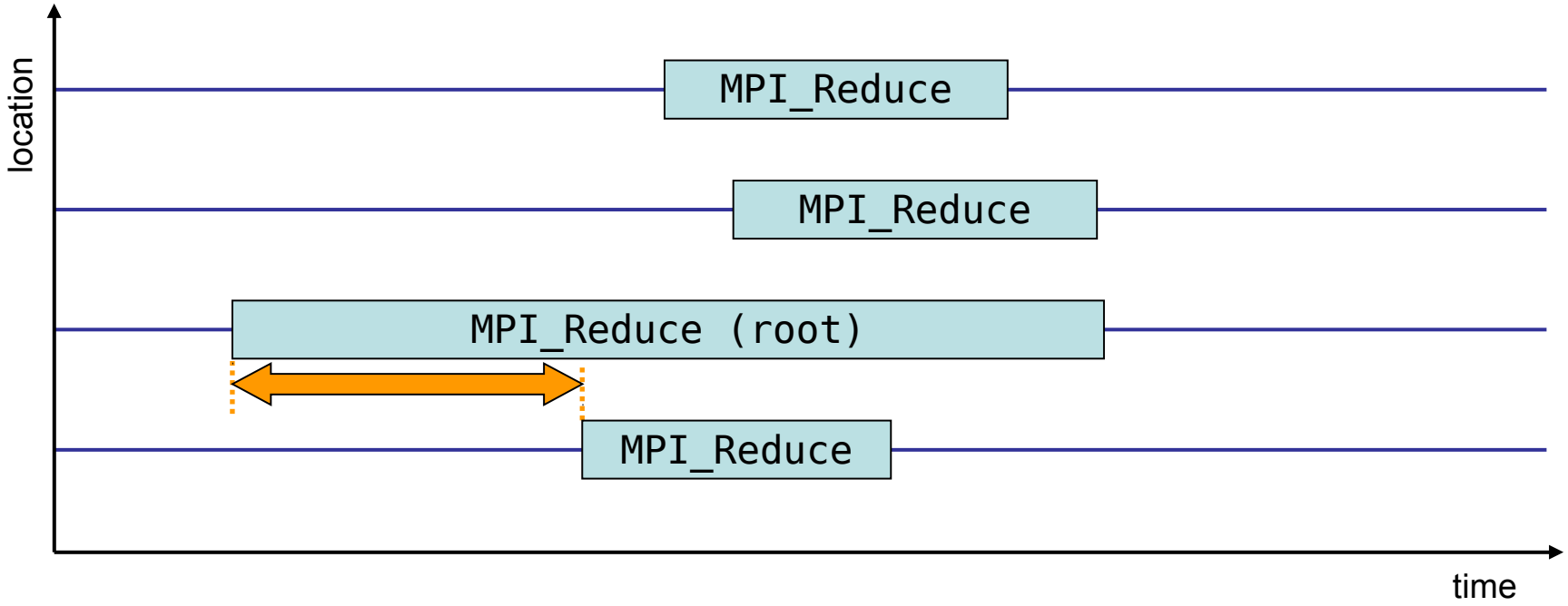
- Time spent in synchronizing collective operations after the first process has left the operation
- Applies to:
  `MPI_Allgather`, `MPI_Allgatherv`,
  `MPI_Allreduce`, `MPI_Alltoall`,
  `MPI_Reduce_scatter`, `MPI_Reduce_scatter_block`

Late Broadcast



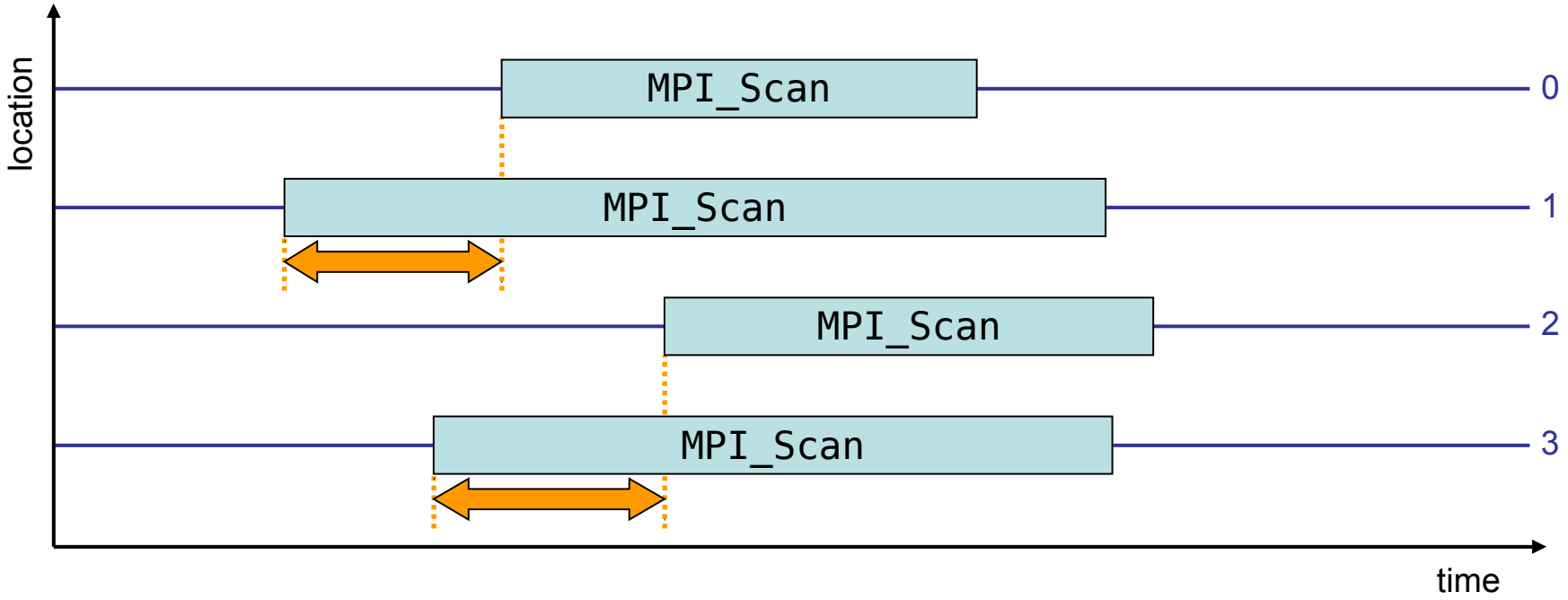
- Waiting times if the destination processes of a collective 1-to-N communication operation enter the operation earlier than the source process (root)
- Applies to: `MPI_Bcast`, `MPI_Scatter`, `MPI_Scatterv`

Early Reduce



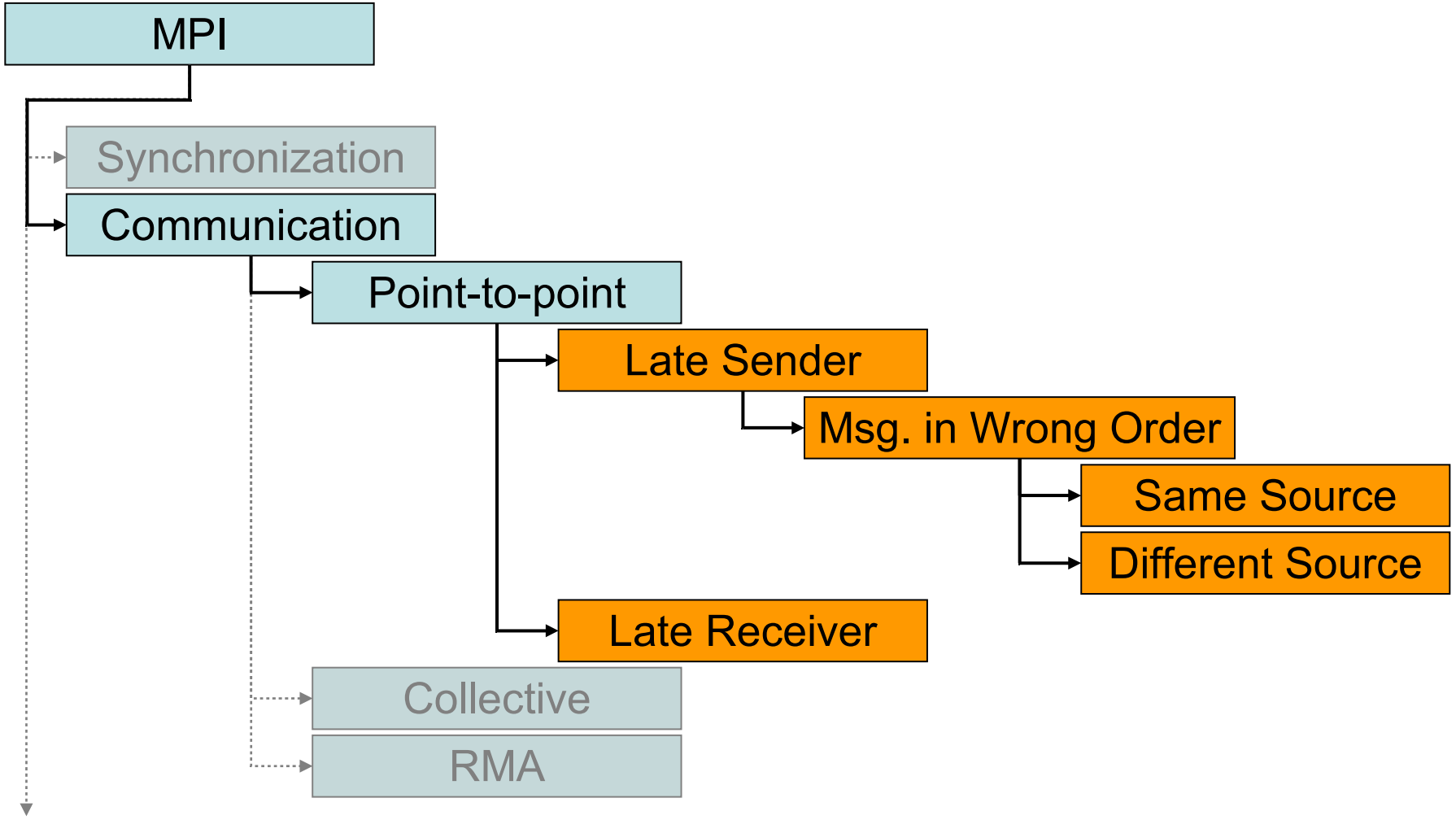
- Waiting time if the destination process (root) of a collective N-to-1 communication operation enters the operation earlier than its sending counterparts
- Applies to: `MPI_Reduce`, `MPI_Gather`, `MPI_Gatherv`

Early Scan

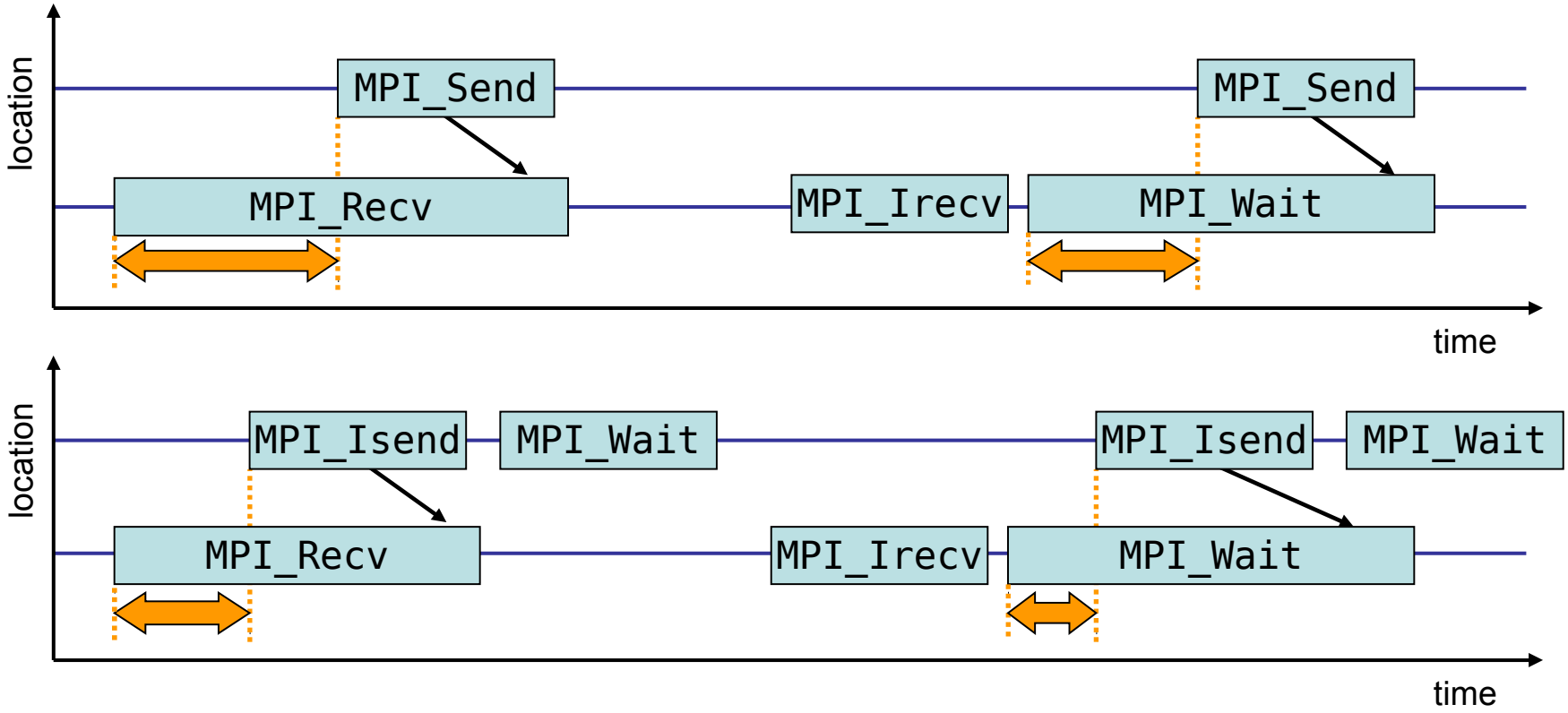


- Waiting time if process *n* enters a prefix reduction operation earlier than its sending counterparts (i.e., ranks 0..*n*-1)
- Applies to: `MPI_Scan, MPI_Exscan`

MPI point-to-point communication time

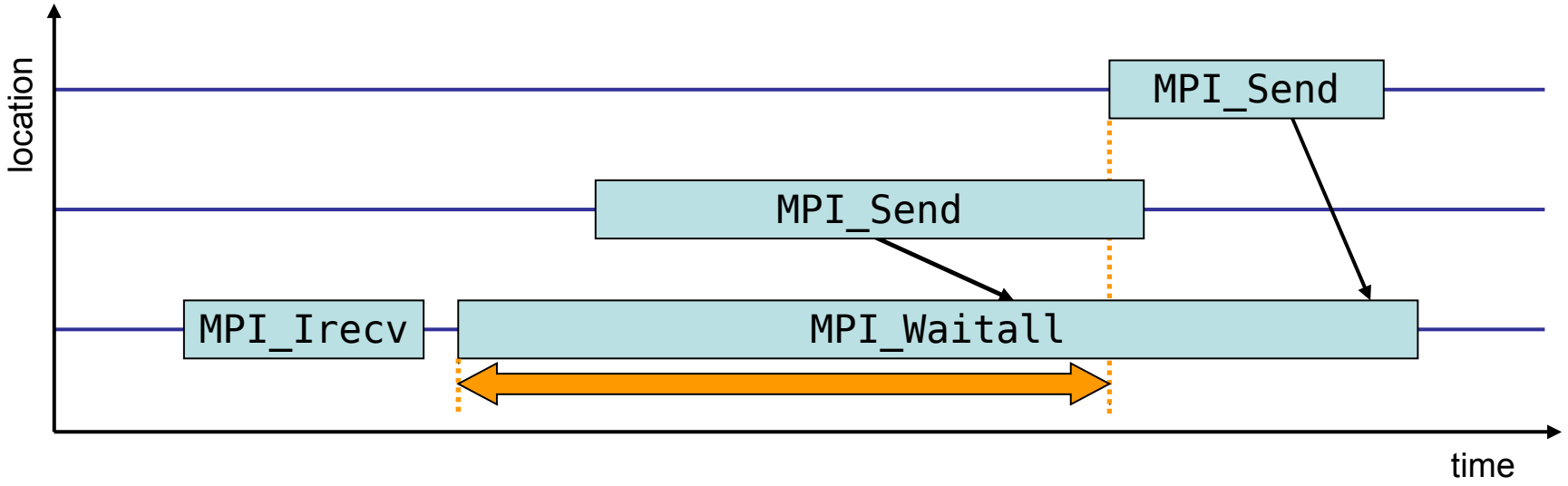


Late Sender



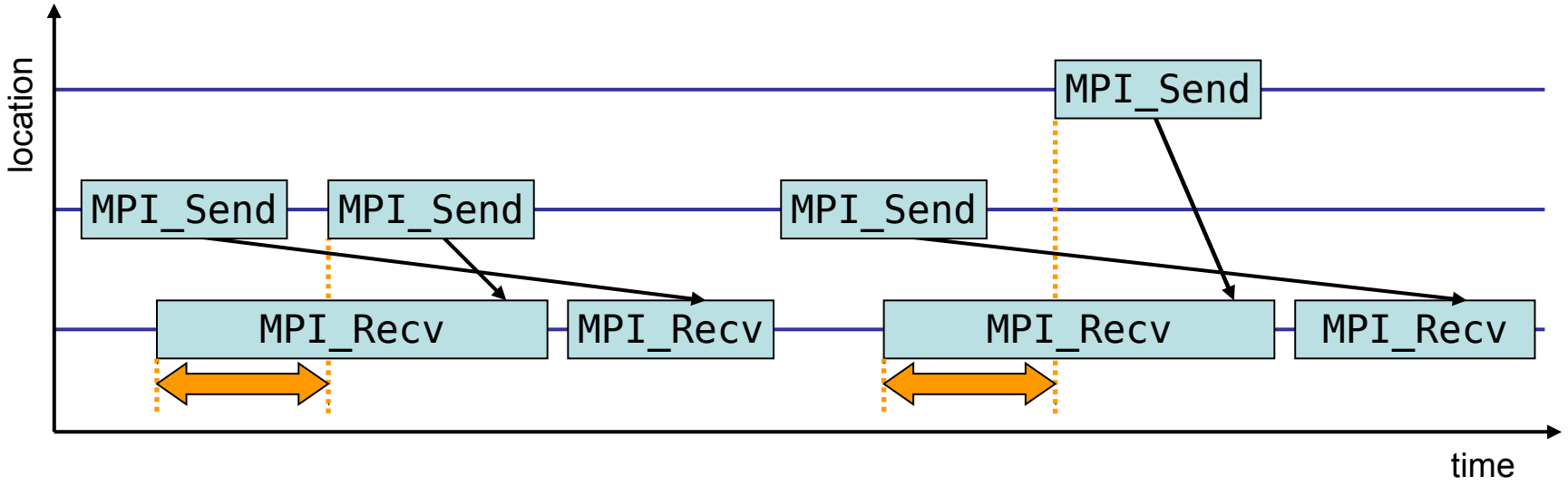
- Waiting time caused by a blocking receive operation posted earlier than the corresponding send operation
- Applies to blocking as well as non-blocking communication

Late Sender (II)



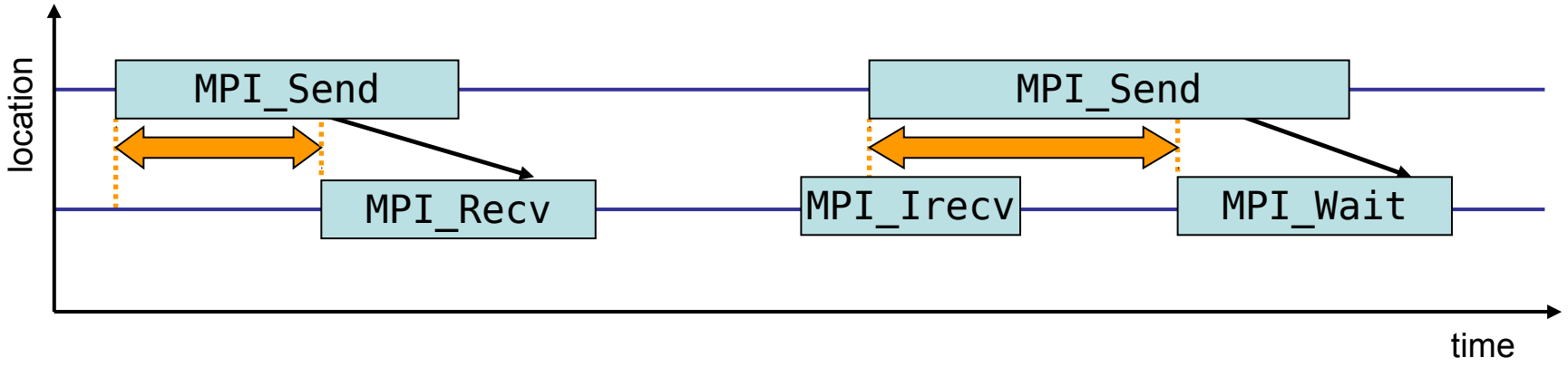
- While waiting for several messages, the maximum waiting time is accounted
- Applies to: `MPI_Waitall`, `MPI_Waitsome`

Late Sender, Messages in Wrong Order



- Refers to Late Sender situations which are caused by messages received in wrong order
- Comes in two flavours:
  - Messages sent from same source location
  - Messages sent from different source locations

Late Receiver

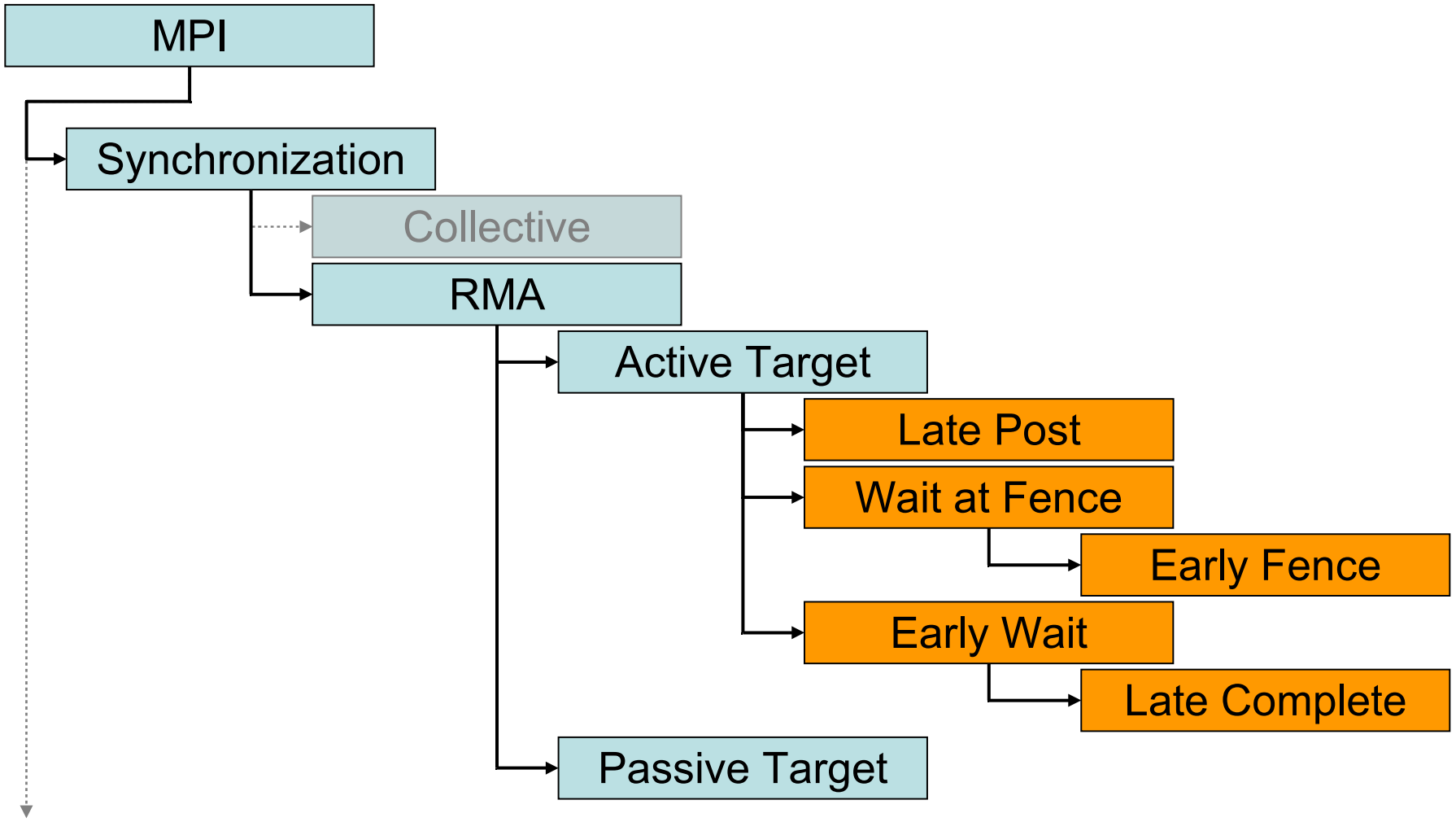


- Waiting time caused by a blocking send operation posted earlier than the corresponding receive operation
- Calculated by receiver but waiting time attributed to sender
- Does currently not apply to non-blocking sends

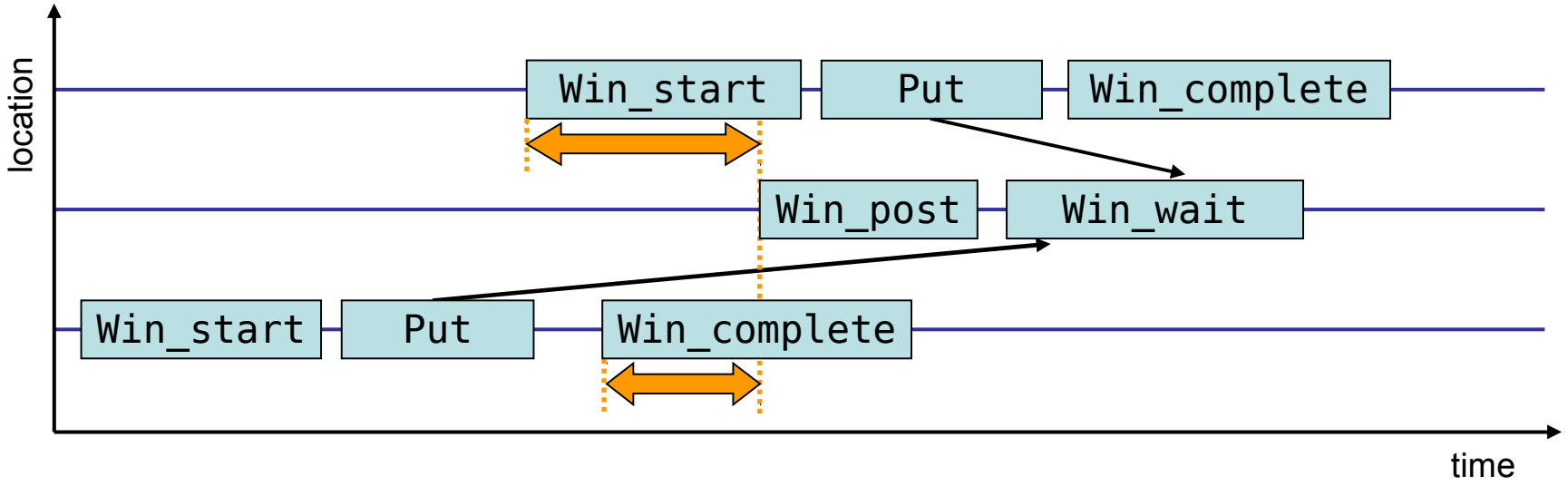
Late Sender/Receiver Counts

- The number of Late Sender / Late Receiver instances are also available
- They are divided into communications & synchronizations and shown in the corresponding hierarchies

MPI RMA synchronization time

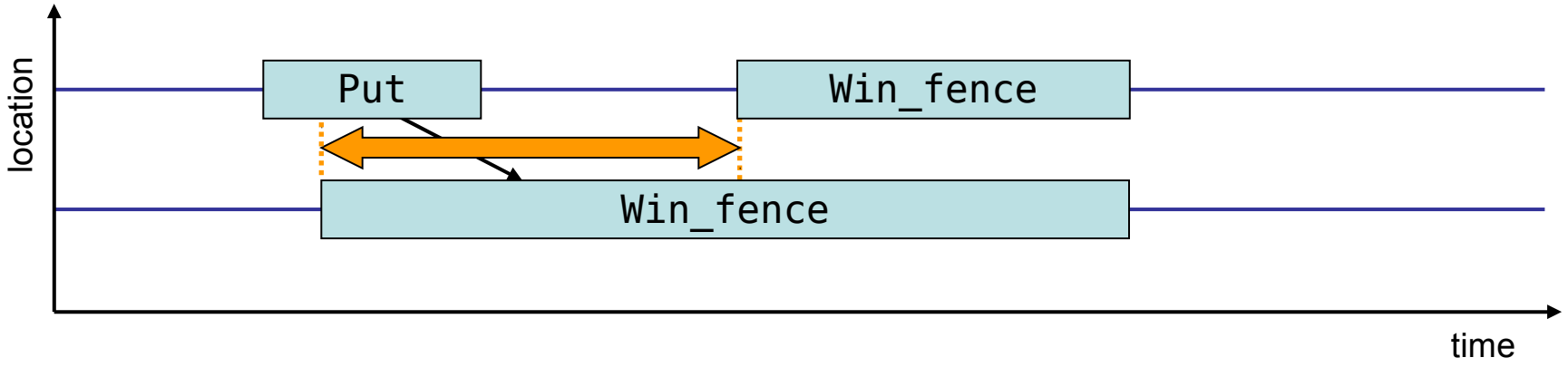


Late Post



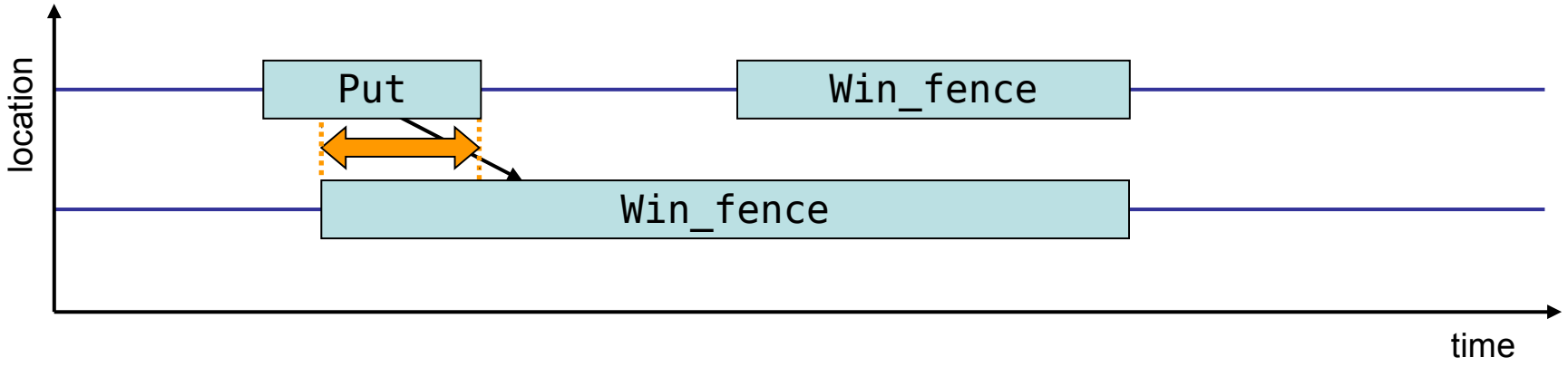
- `MPI_Win_start` (top) or `MPI_Win_complete` (bottom) wait until exposure epoch is opened by `MPI_Win_post`
- Which of the two calls blocks is implementation dependent

Wait at Fence



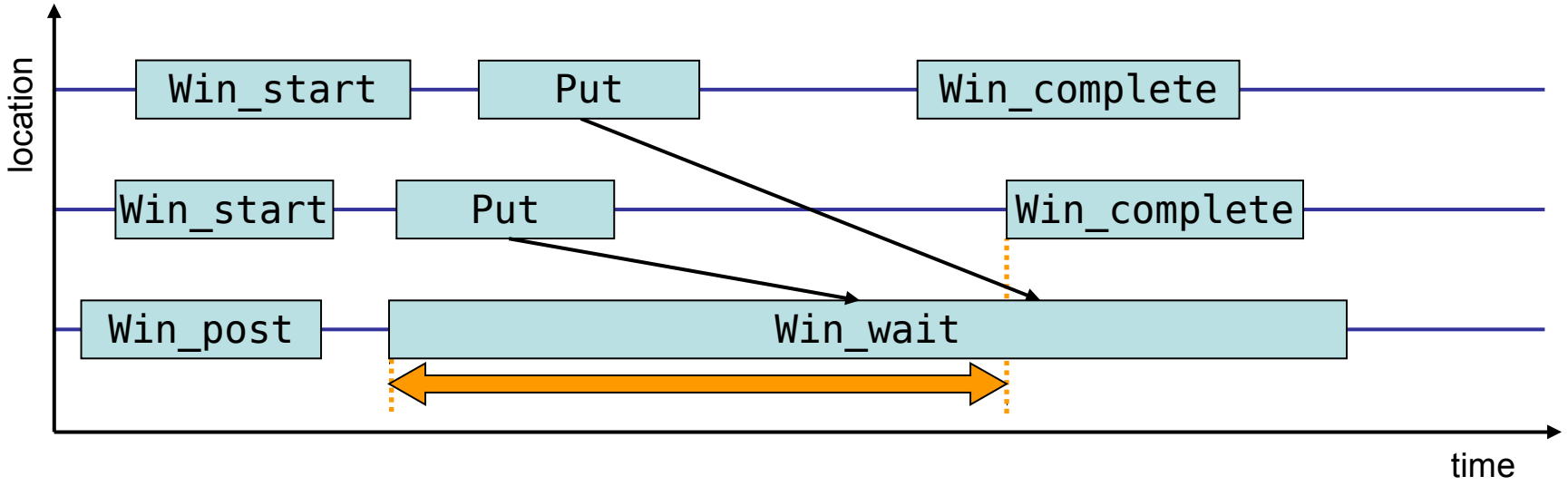
- Time spent waiting in front of a synchronizing `MPI_Win_fence` call until the last process reaches the fence operation
- Only triggered if at least one of the following conditions applies
  - Given assertion is 0
  - All fence calls overlap (heuristic)

Early Fence



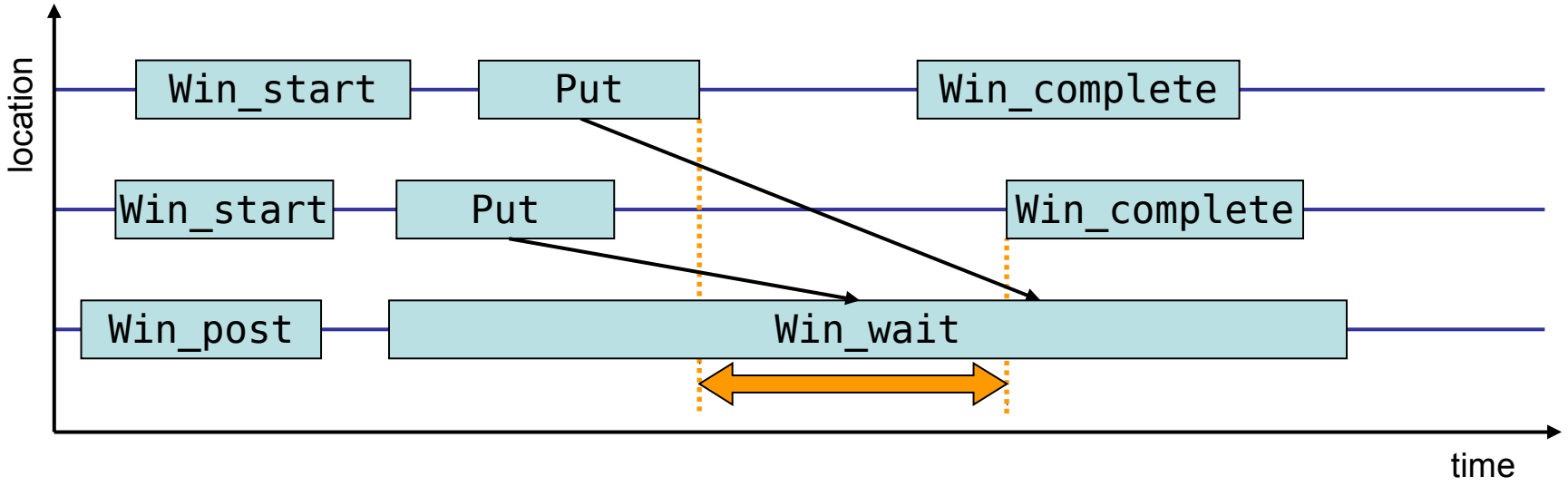
- Time spent waiting for exit of last RMA operation to target location
- Sub-pattern of Wait at Fence

Early Wait



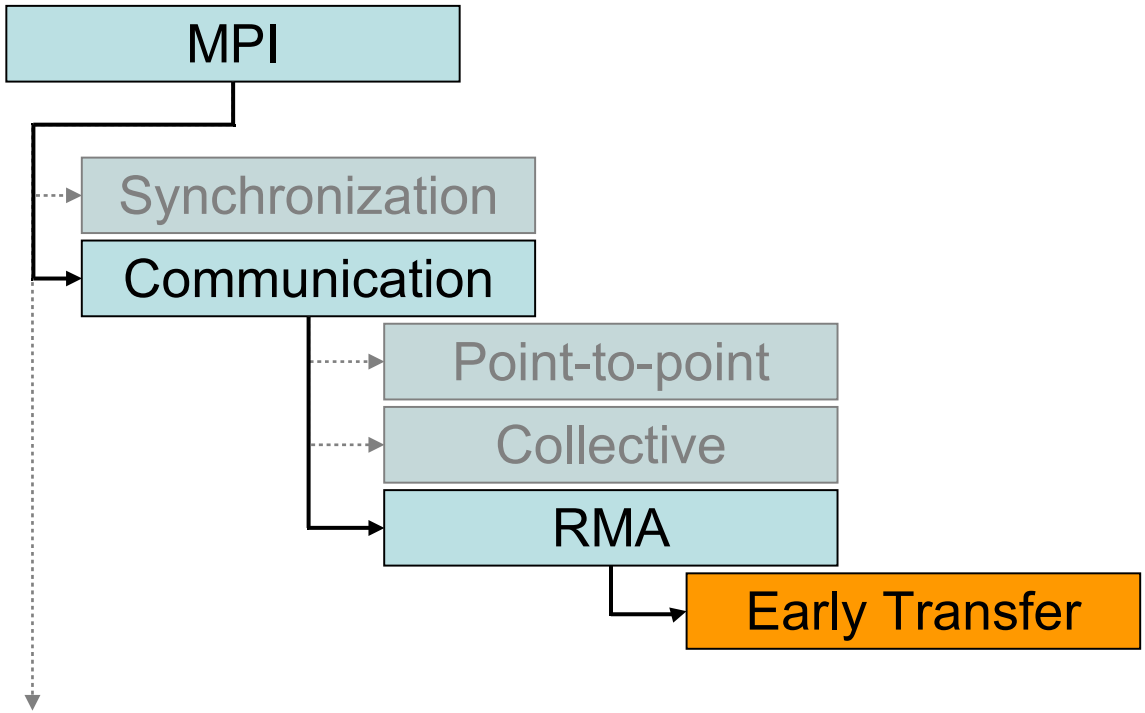
- Time spent in `MPI_Win_wait` until access epoch is closed by last `MPI_Win_complete`

Late Complete

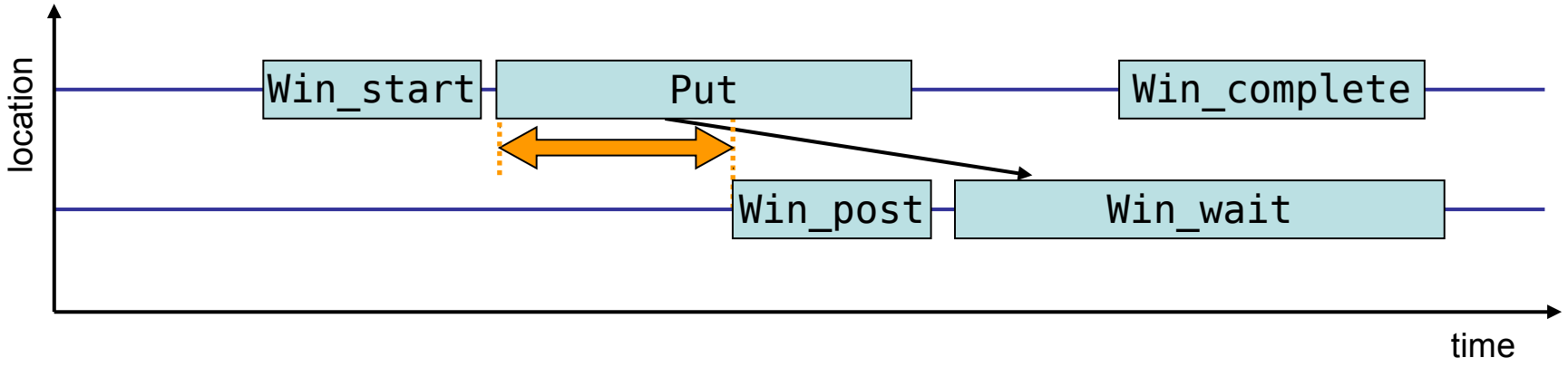


- Waiting time due to unnecessary pause between last RMA operation to target and closing the access epoch by last `MPI_Win_complete`
- Sub-pattern of Early Wait

MPI RMA communication time



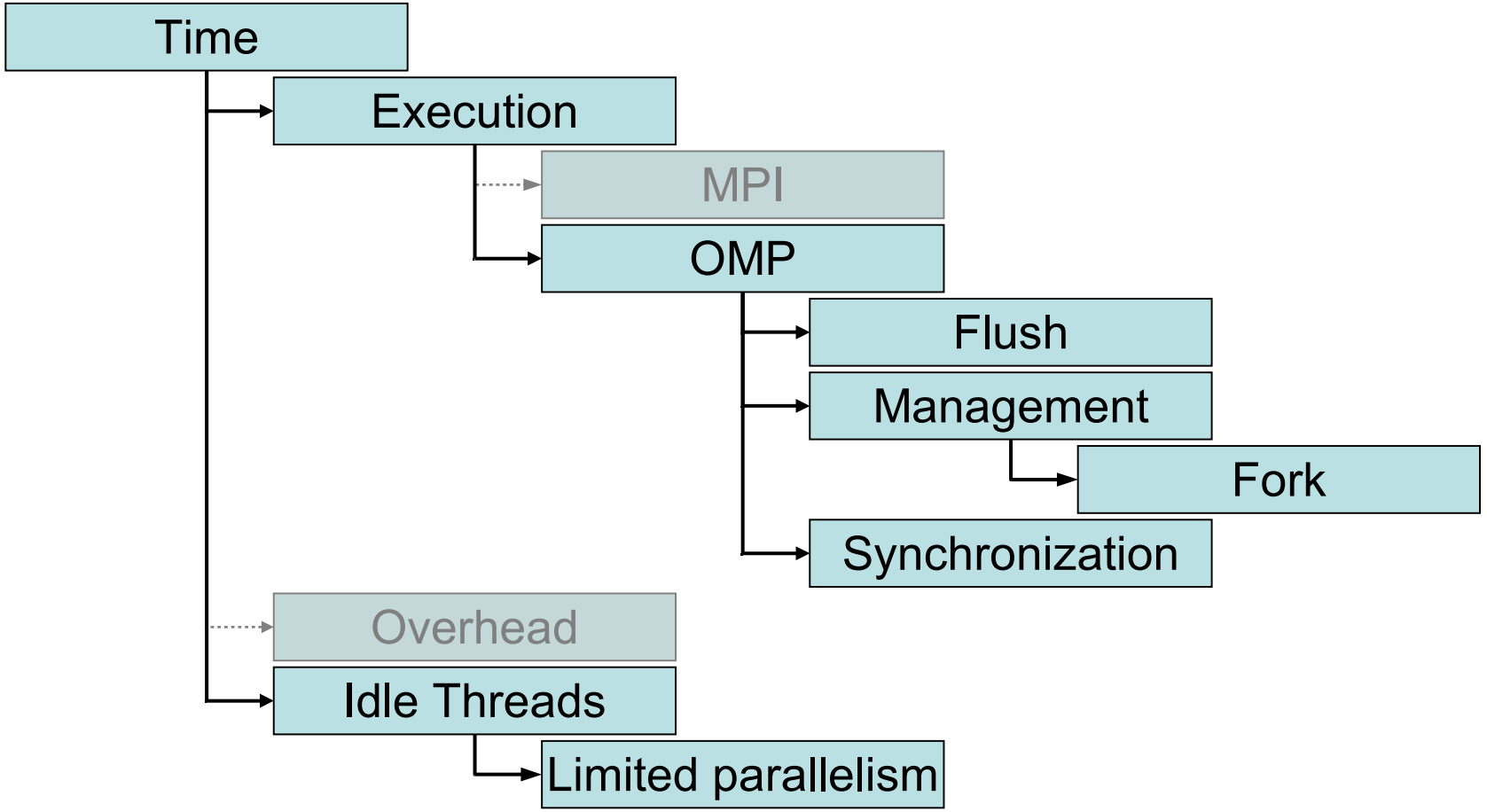
Early Transfer



- Time spent waiting in RMA operation on origin(s) started before exposure epoch was opened on target

OpenMP-related metrics

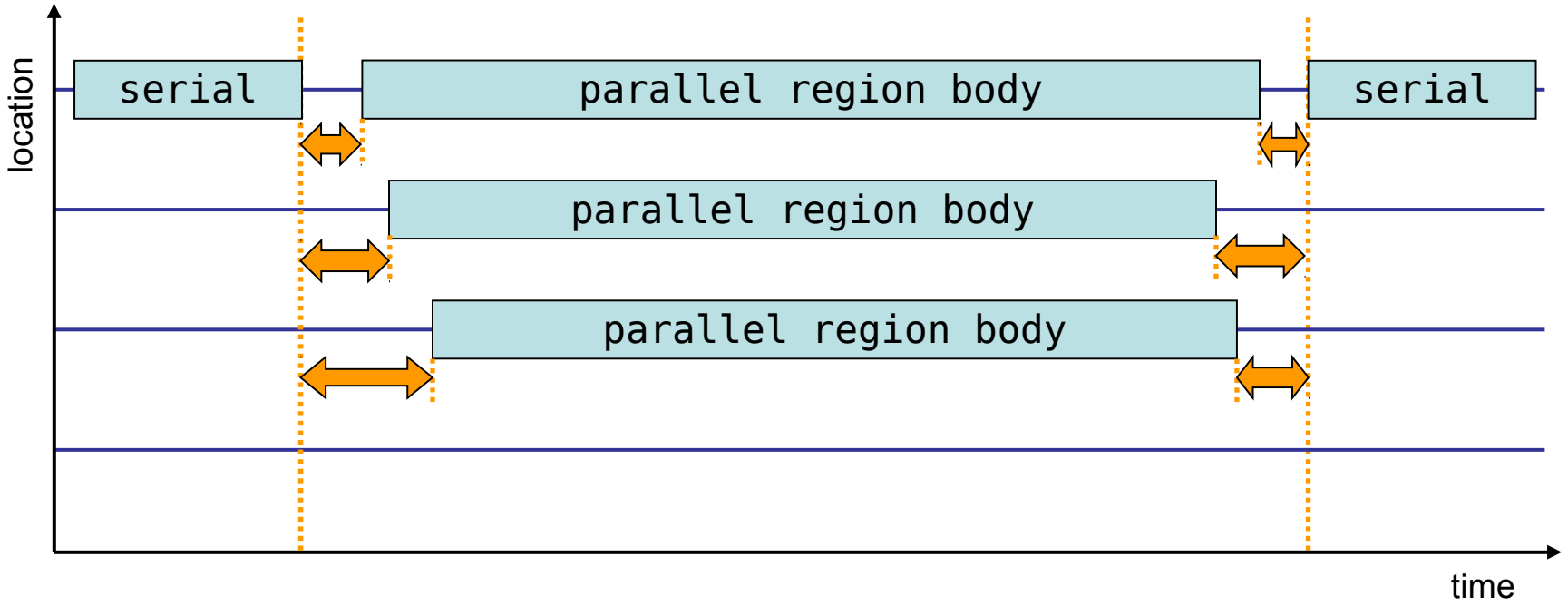
OpenMP Time hierarchy



OpenMP Time hierarchy details

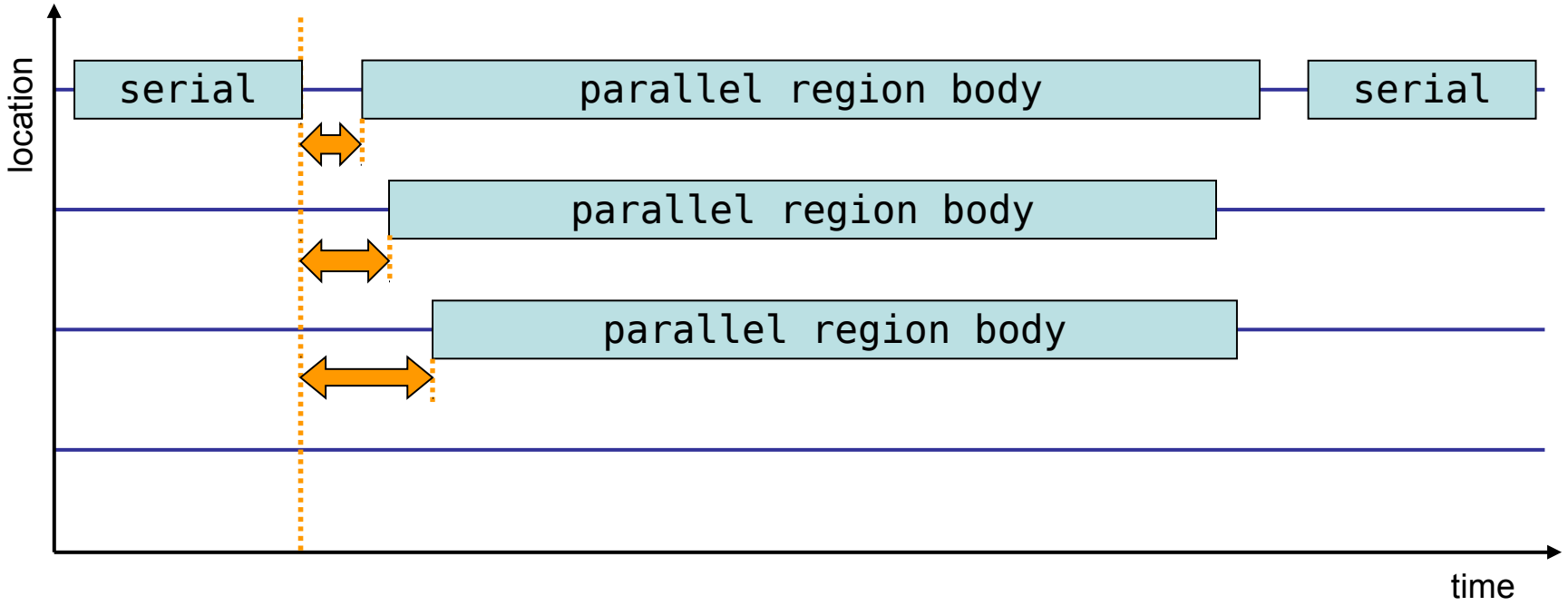
OMP	Time spent for OpenMP-related tasks
Flush	Time spent in OpenMP flush directives
Synchronization	Time spent to synchronize OpenMP threads

OpenMP Management Time



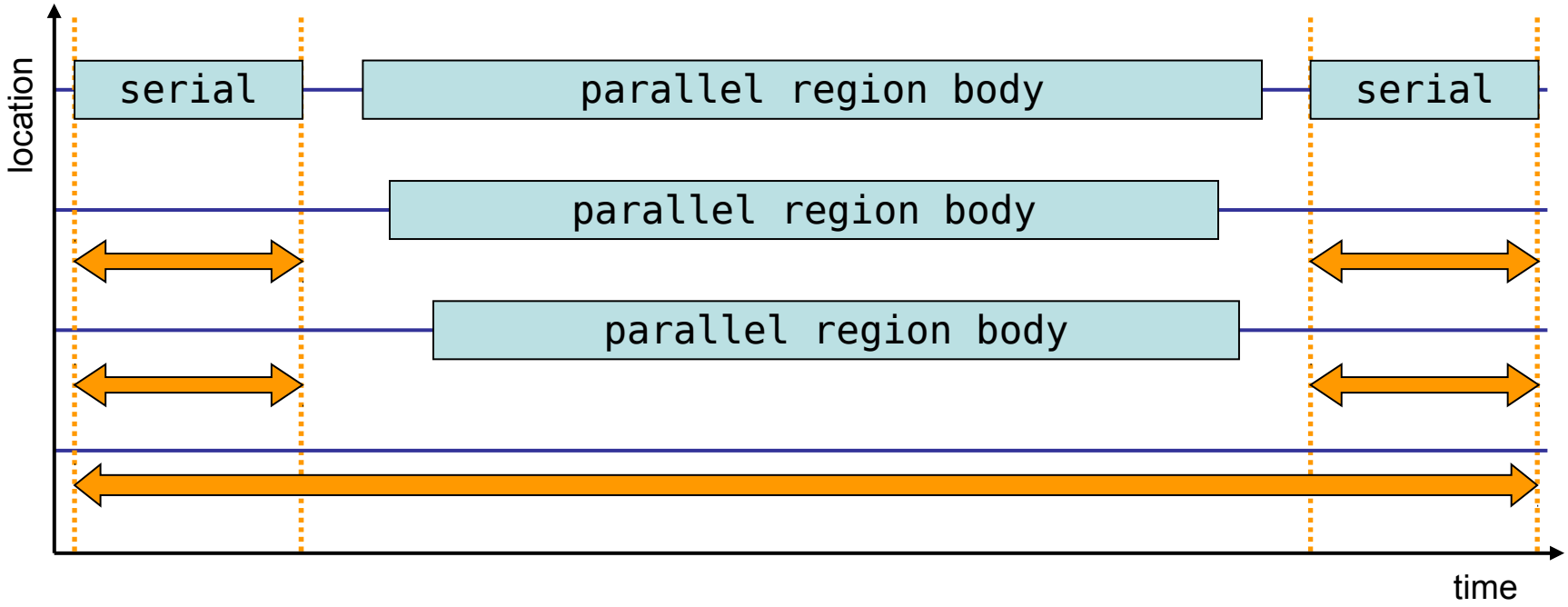
- Time spent on master thread for creating/destroying OpenMP thread teams

OpenMP Fork Time



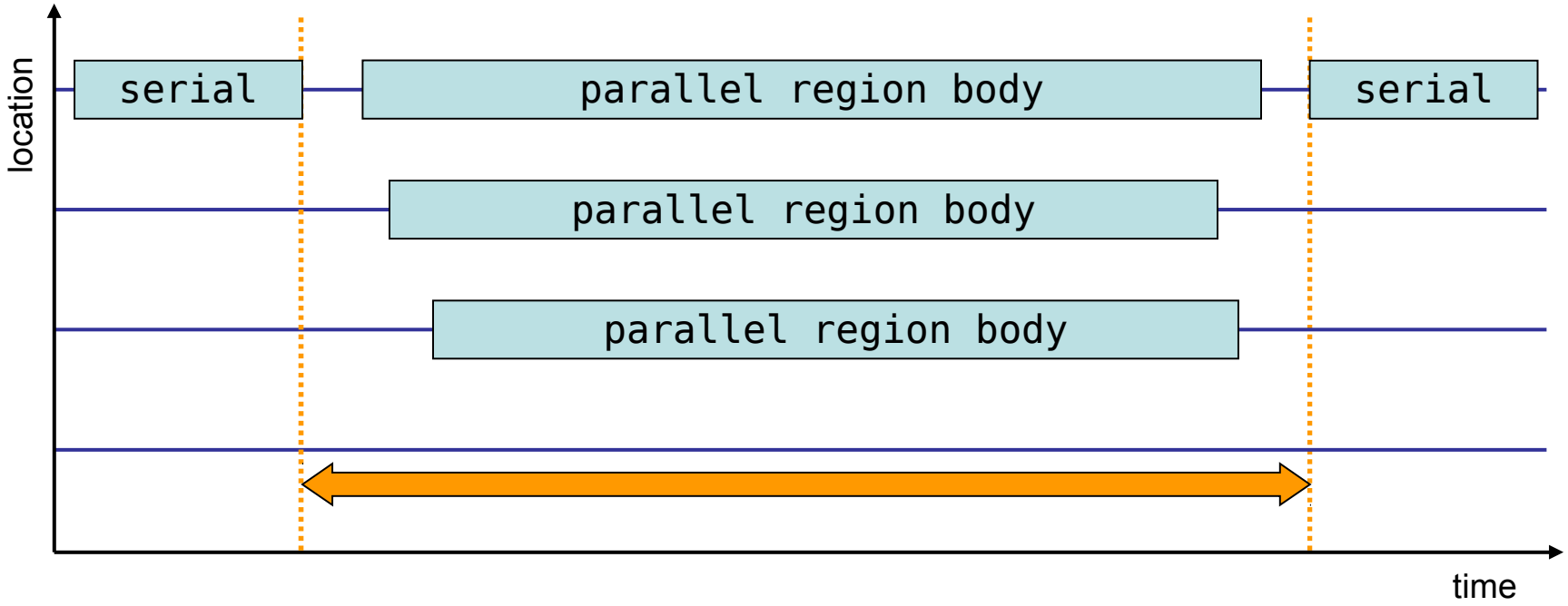
- Time spent on master threads for creating OpenMP thread teams

OpenMP Idle Threads



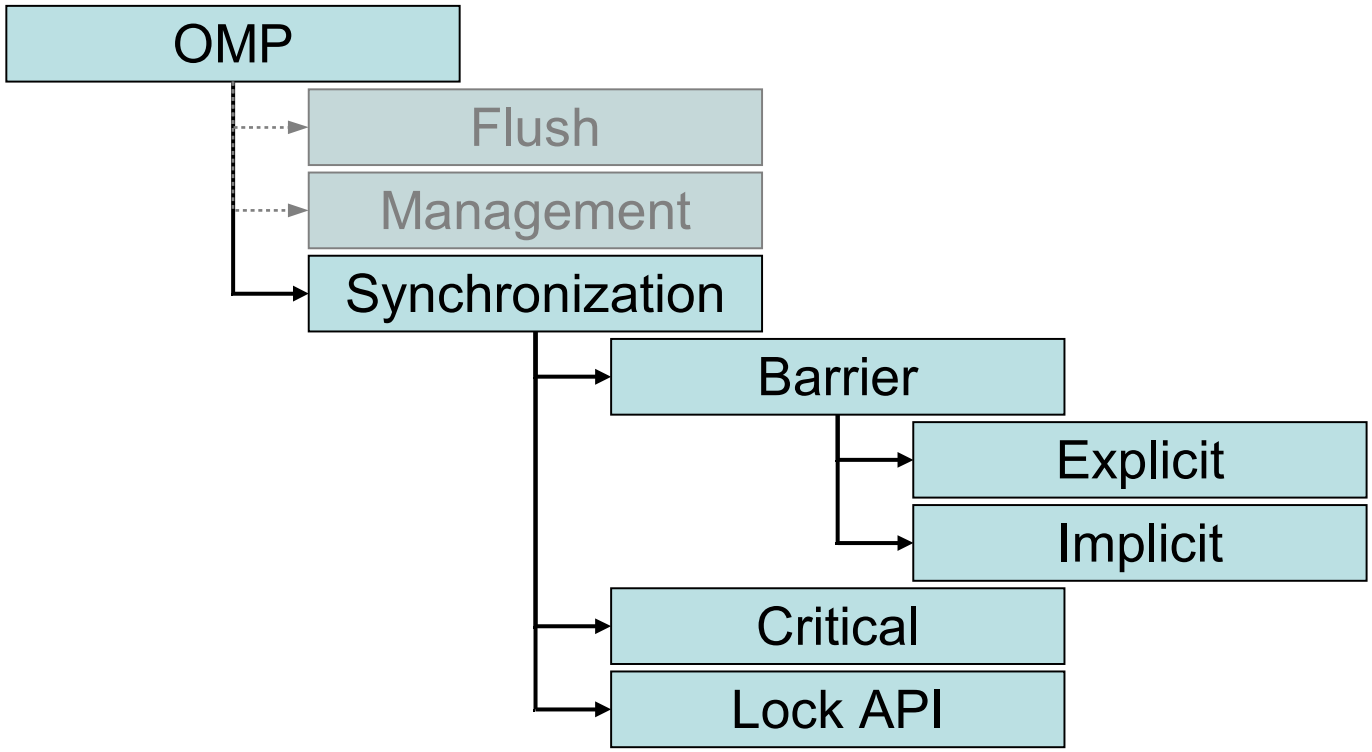
- Time spent idle on CPUs reserved for worker threads

OpenMP Limited Parallelism



- Time spent idle on worker threads within parallel regions

OpenMP Synchronization Time hierarchy

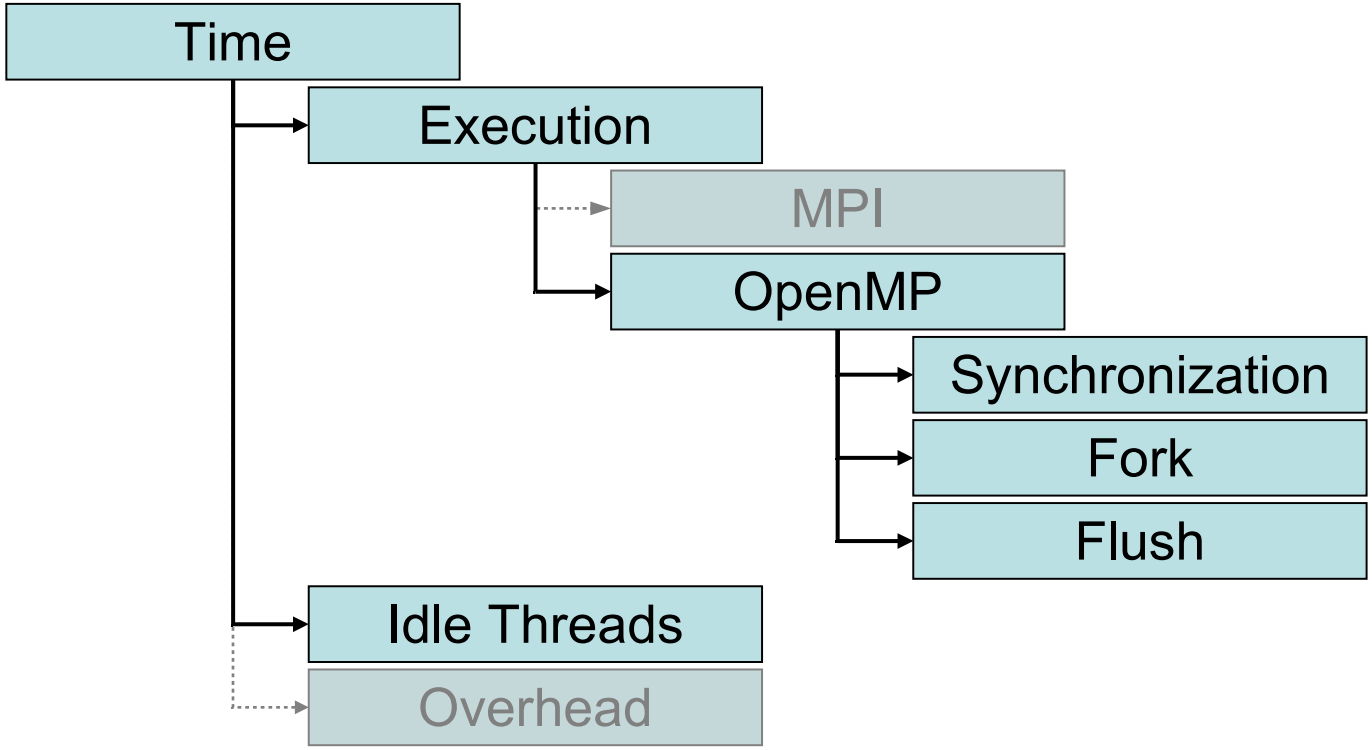


- Time spent in OpenMP atomic constructs is attributed to the “Critical” metric

OpenMP-related metrics

(as produced by Scalasca's sequential trace analyzer for OpenMP and hybrid MPI/OpenMP applications)

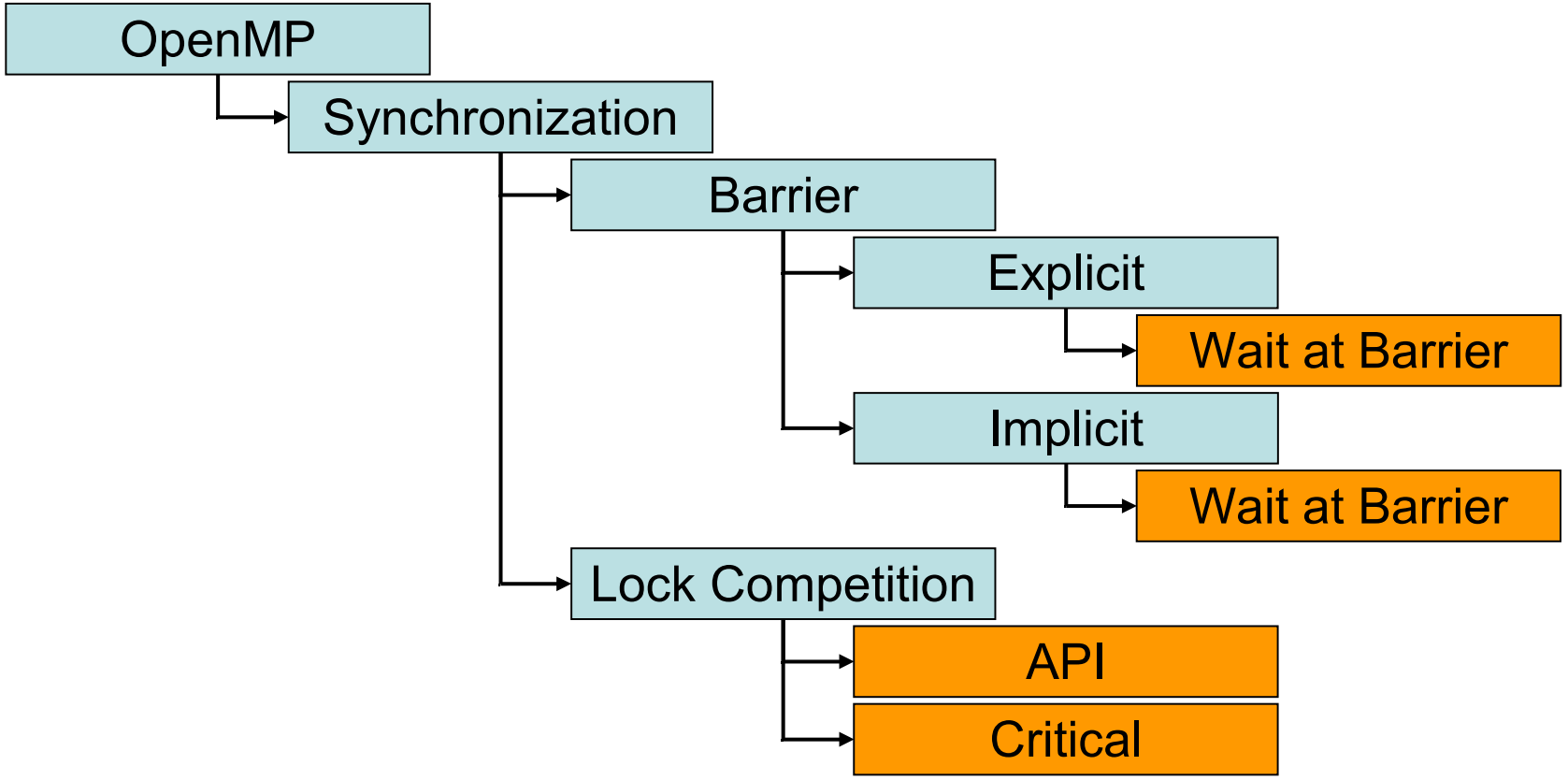
OpenMP Time hierarchy



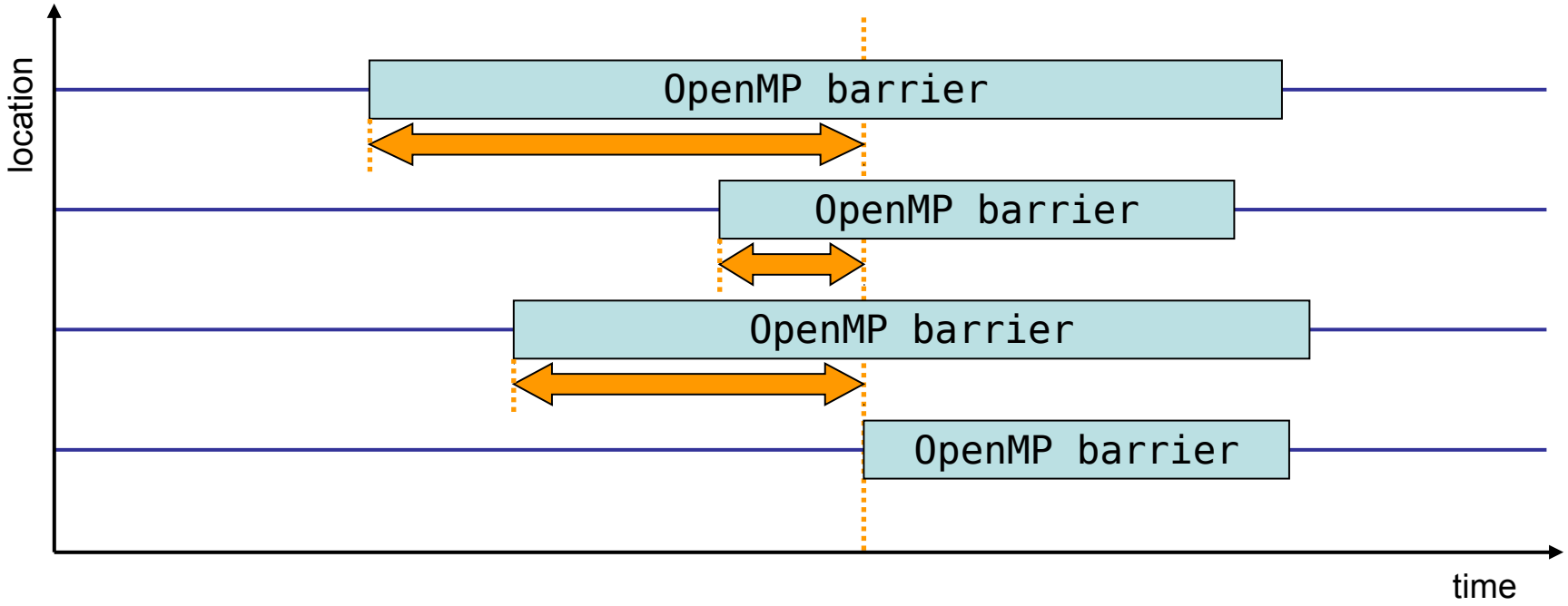
OpenMP Time hierarchy details

OpenMP	Time spent for OpenMP-related tasks
Synchronization	Time spent for synchronizing OpenMP threads
Fork	Time spent by master thread to create thread teams
Flush	Time spent in OpenMP flush directives
Idle Threads	Time spent idle on CPUs reserved for worker threads

OpenMP synchronization time

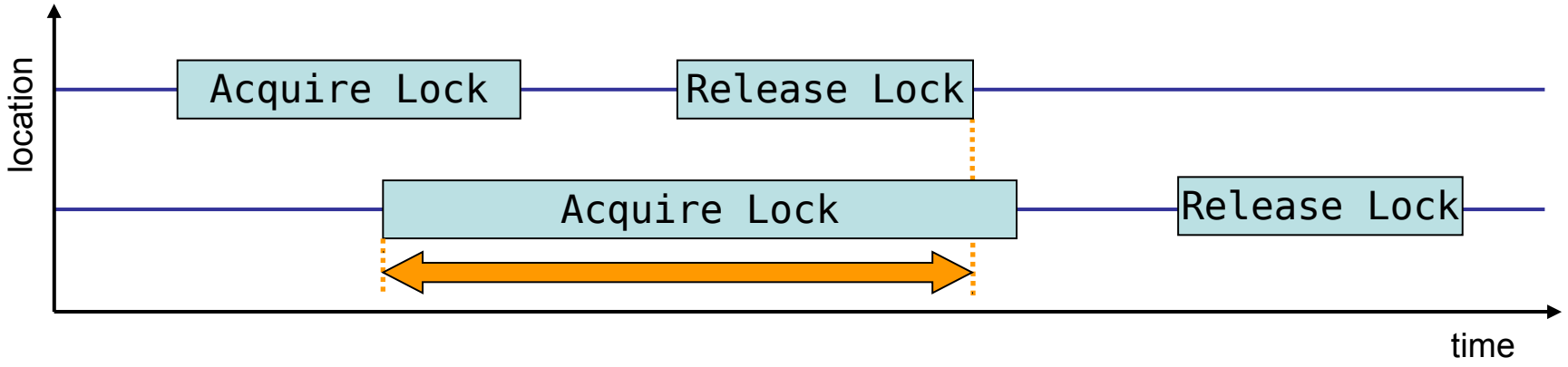


Wait at Barrier



- Time spent waiting in front of a barrier call until the last process reaches the barrier operation
- Applies to: Implicit/explicit barriers

Lock Competition



- Time spent waiting for a lock that has been previously acquired by another thread
- Applies to: critical sections, OpenMP lock API

Happy end...