TECHNISCHE UNIVERSITÄT DRESDEN

Center for Information Services and High Performance Computing (ZIH)

# MPI-focused Tracing with OTFX

## An MPI-aware In-memory Event Tracing Extension to the Open Trace Format 2

EuroMPI 2015, Bordeaux, France

Michael Wagner, Jens Doleschal, and Andreas Knüpfer

michael.wagner@zih.tu-dresden.de

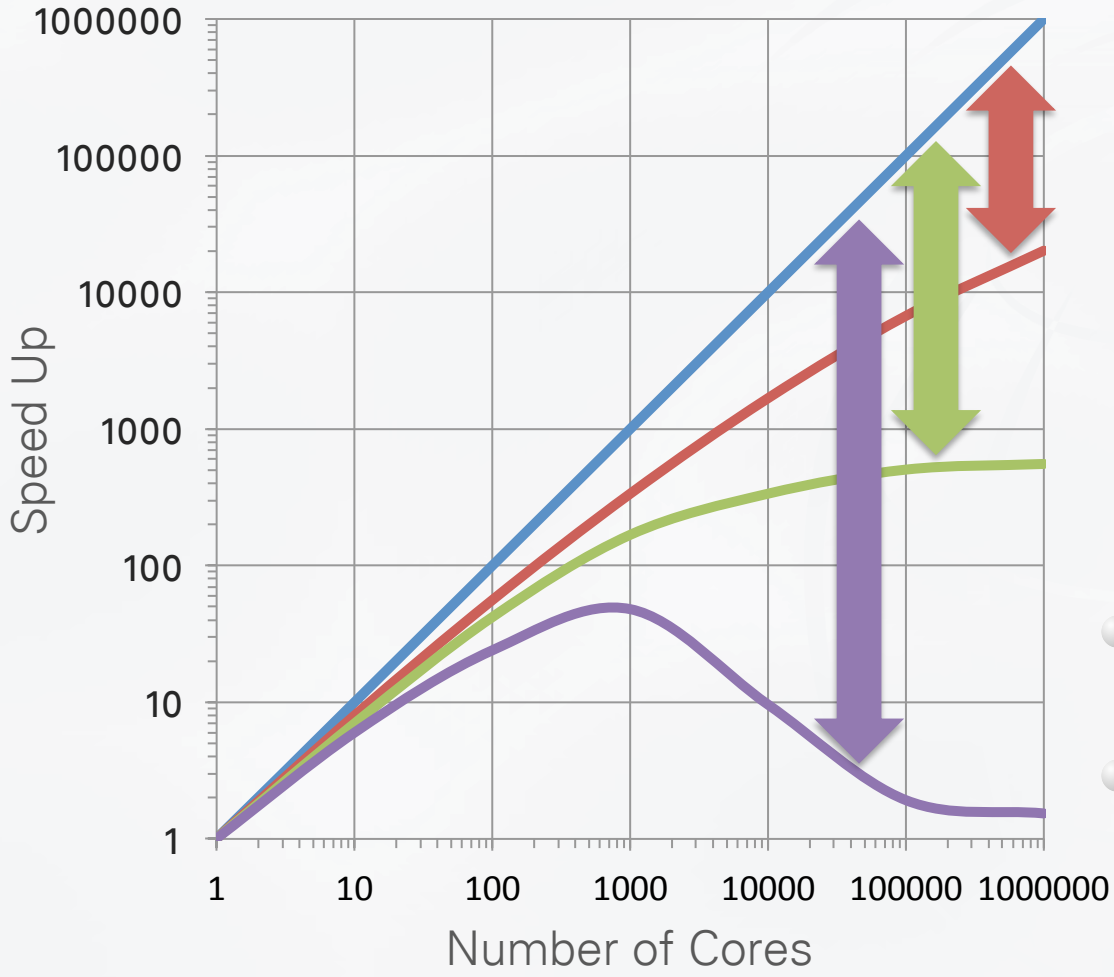
ZIH

Center for Information Services & High Performance Computing

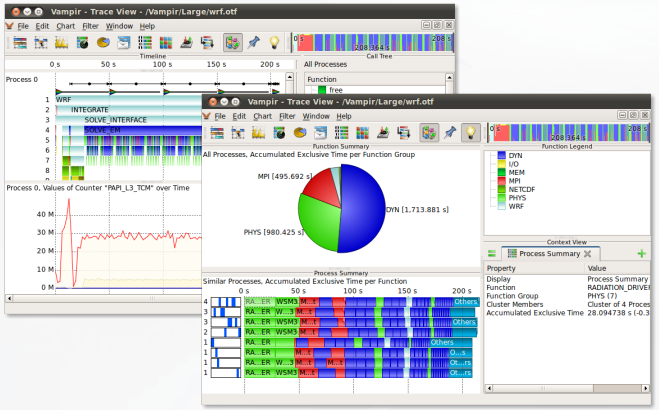
# Outline

- Introduction
- Motivation: the impact of uncoordinated intermediate memory buffer flushes
- MPI-aware in-memory event tracing
- Evaluation
- Conclusion

# Parallelization – Ideal vs. Reality

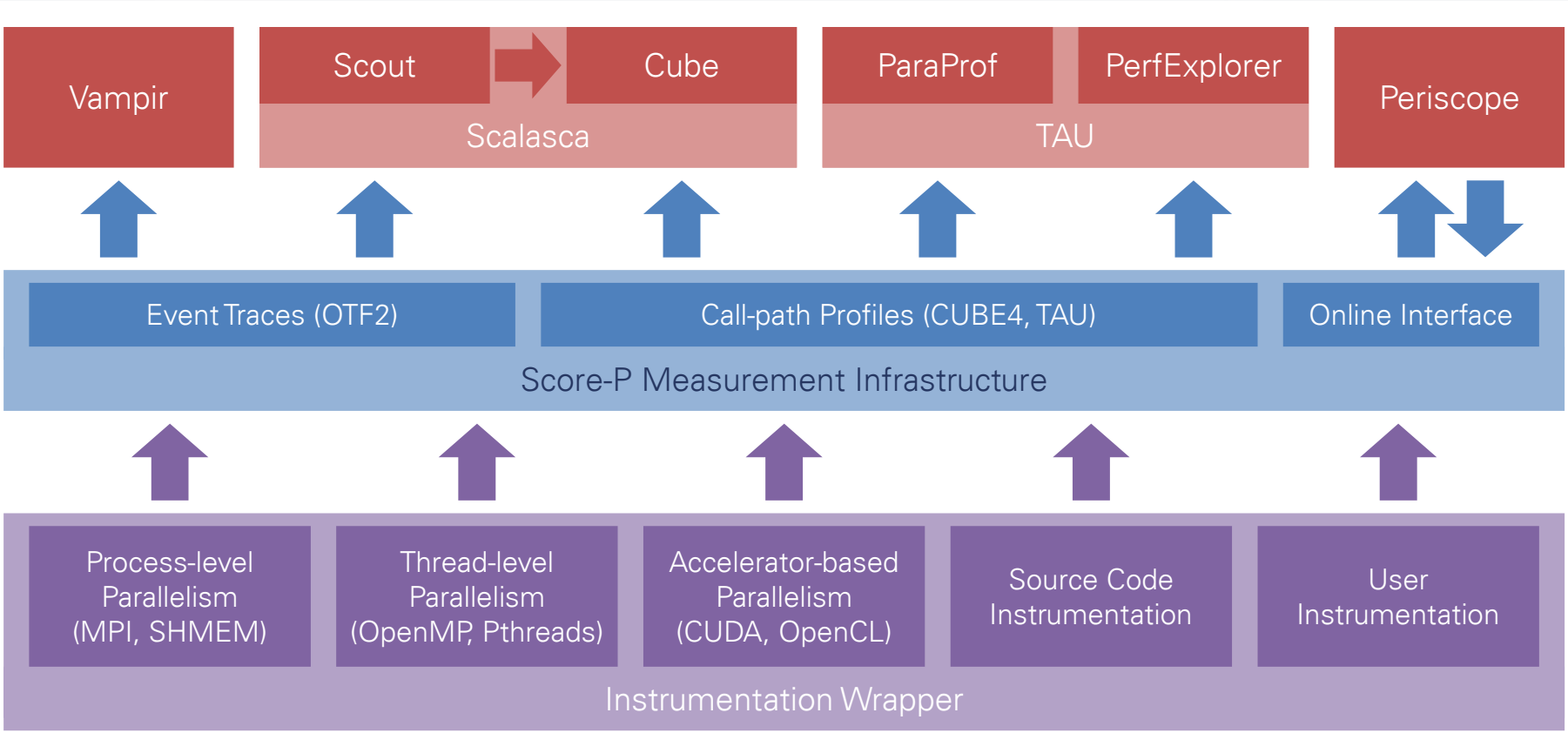



## Performance Analysis Tools



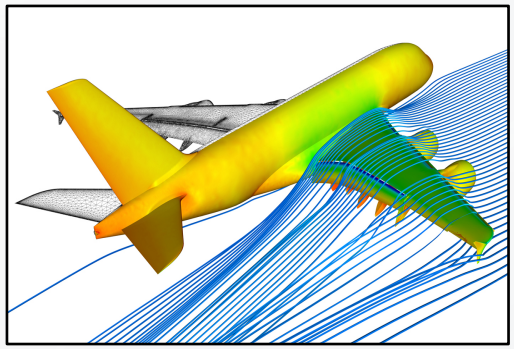
- Help to better understand the application behavior
- Identify performance critical code sections

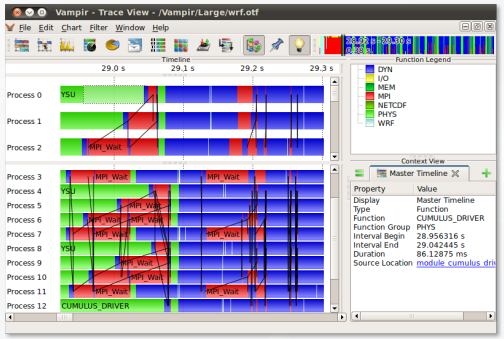
# Score-P and OTF2




- State-of-the-art open source event monitor
- Captures all levels of parallelism simultaneously
- Provides event traces and profiles for Vampir, Scalasca, and Tau

# Event-based Performance Analysis Workflow






Application

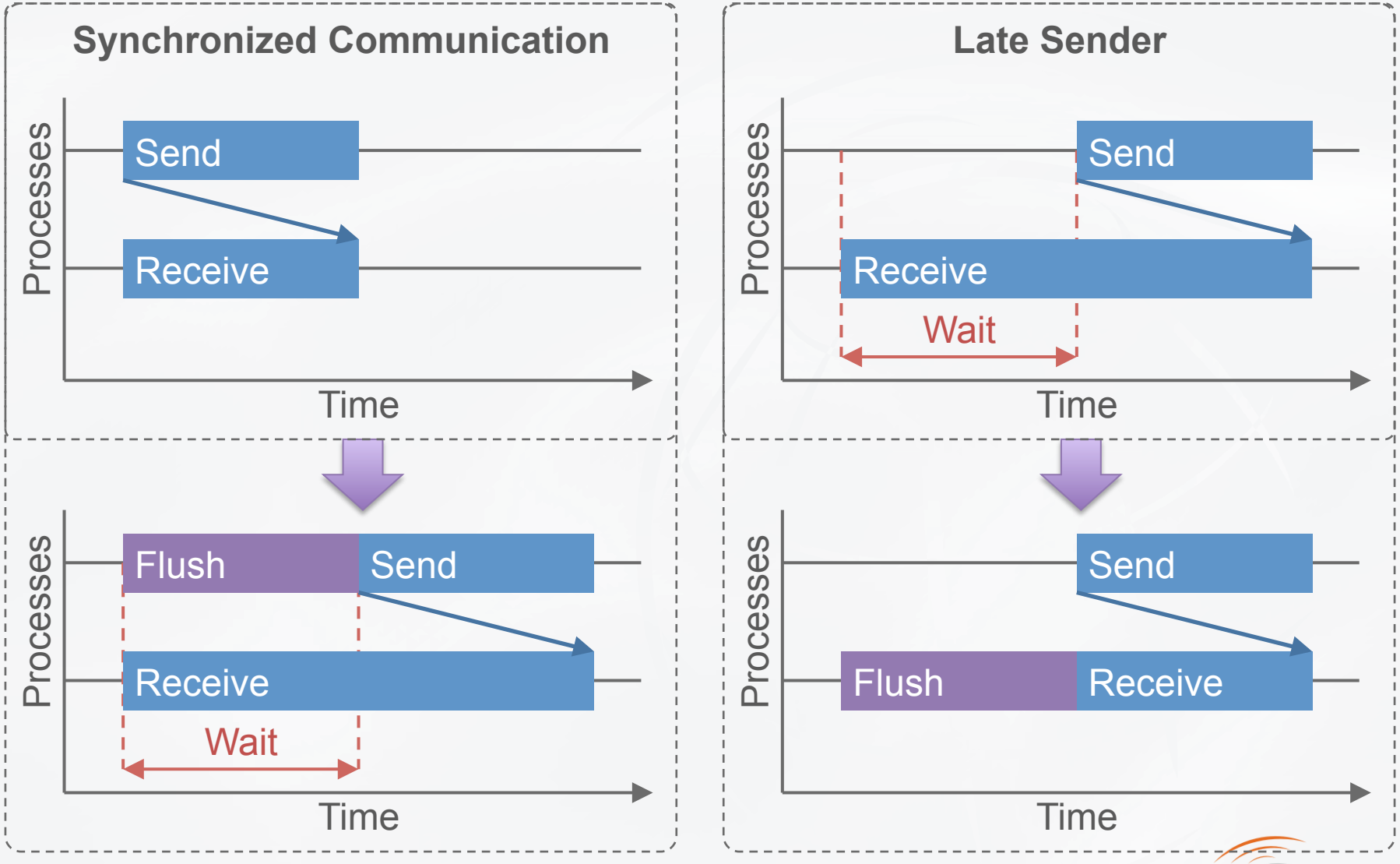
Measurement Tool

Analysis

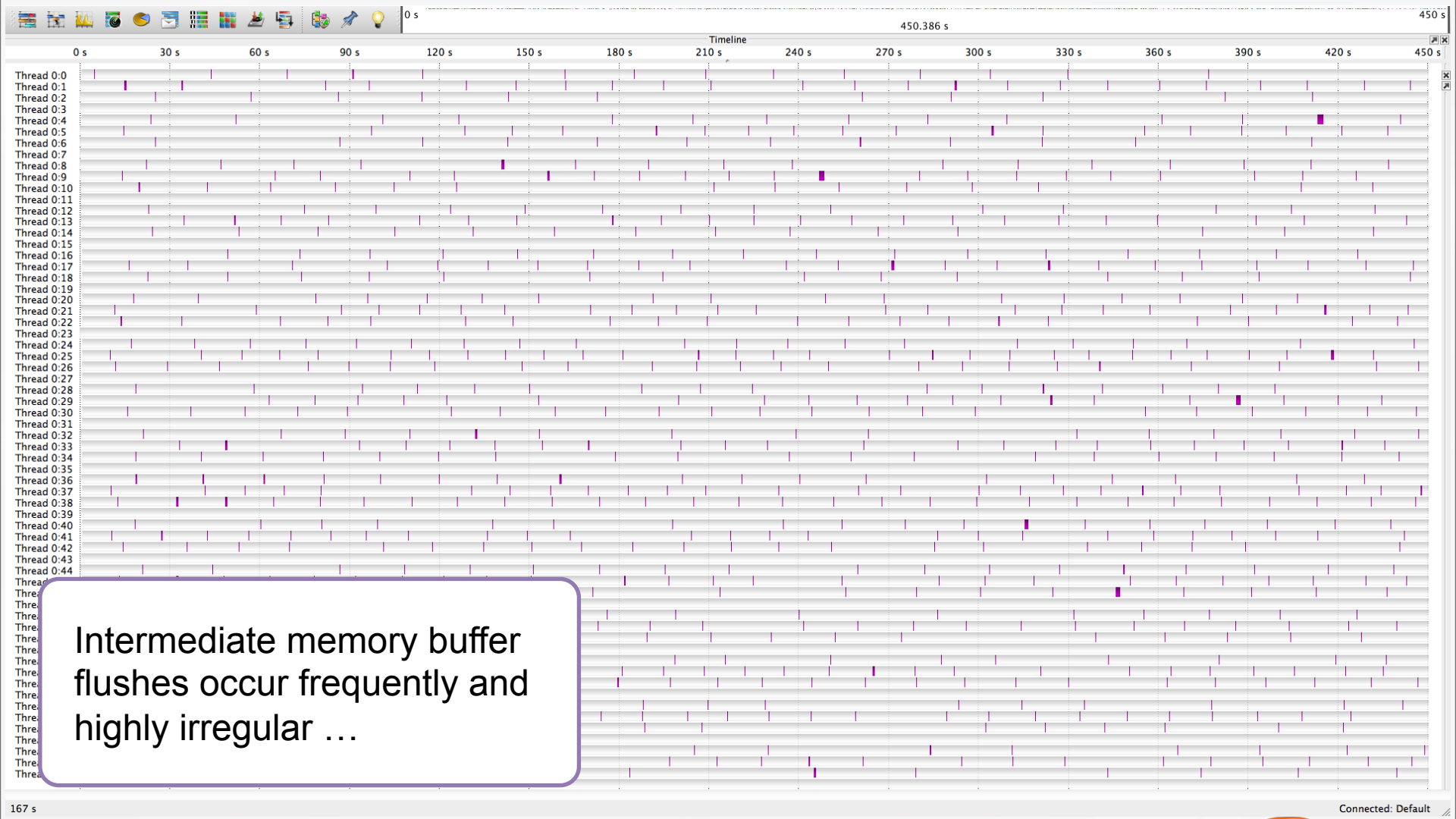
Analysis Tool

File System

# Intermediate Memory Buffer Flushes



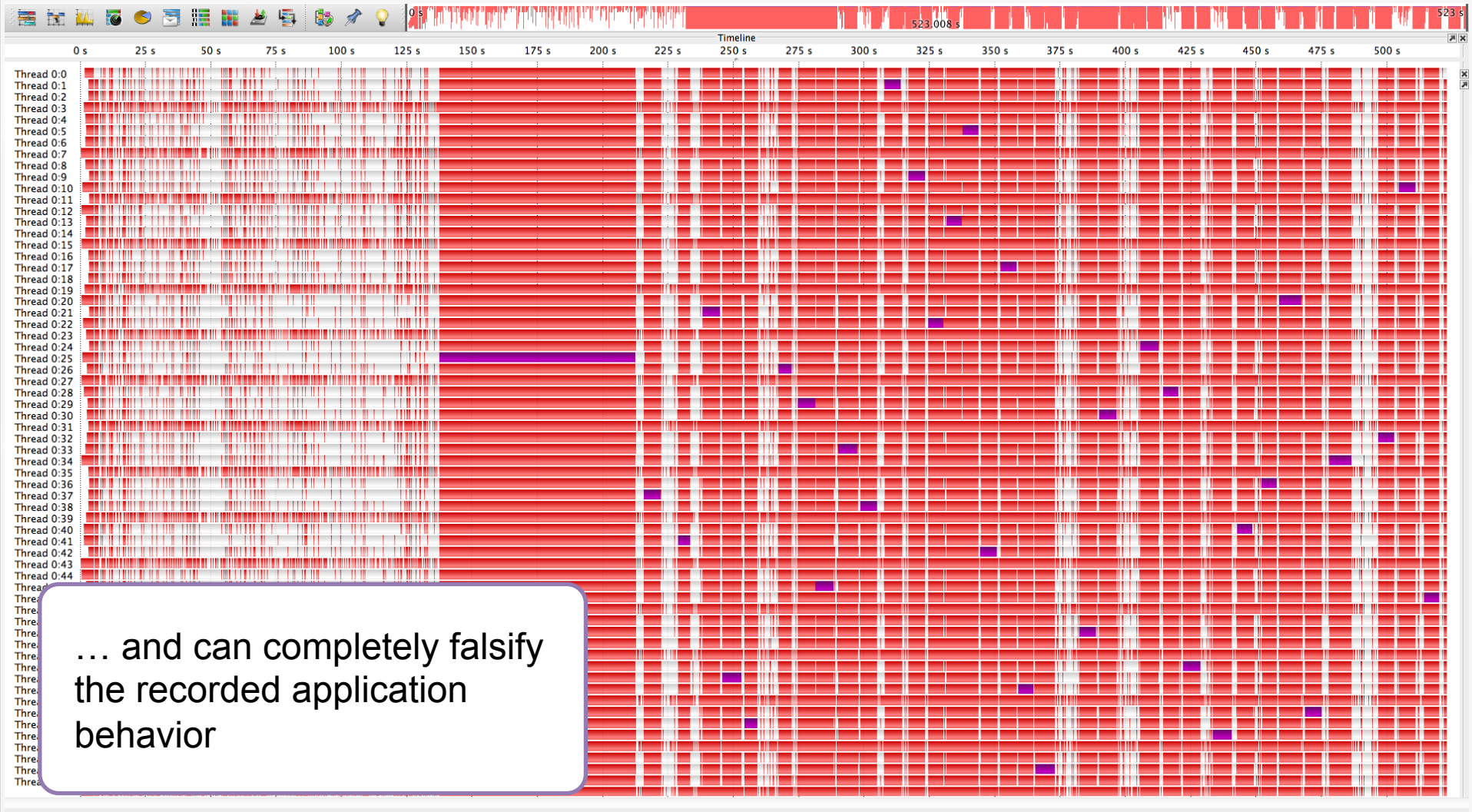
# Intermediate Memory Buffer Flushes – Distribution




Intermediate memory buffer flushes occur frequently and highly irregular …

- Buffer flush, buffer size: 50 MiB

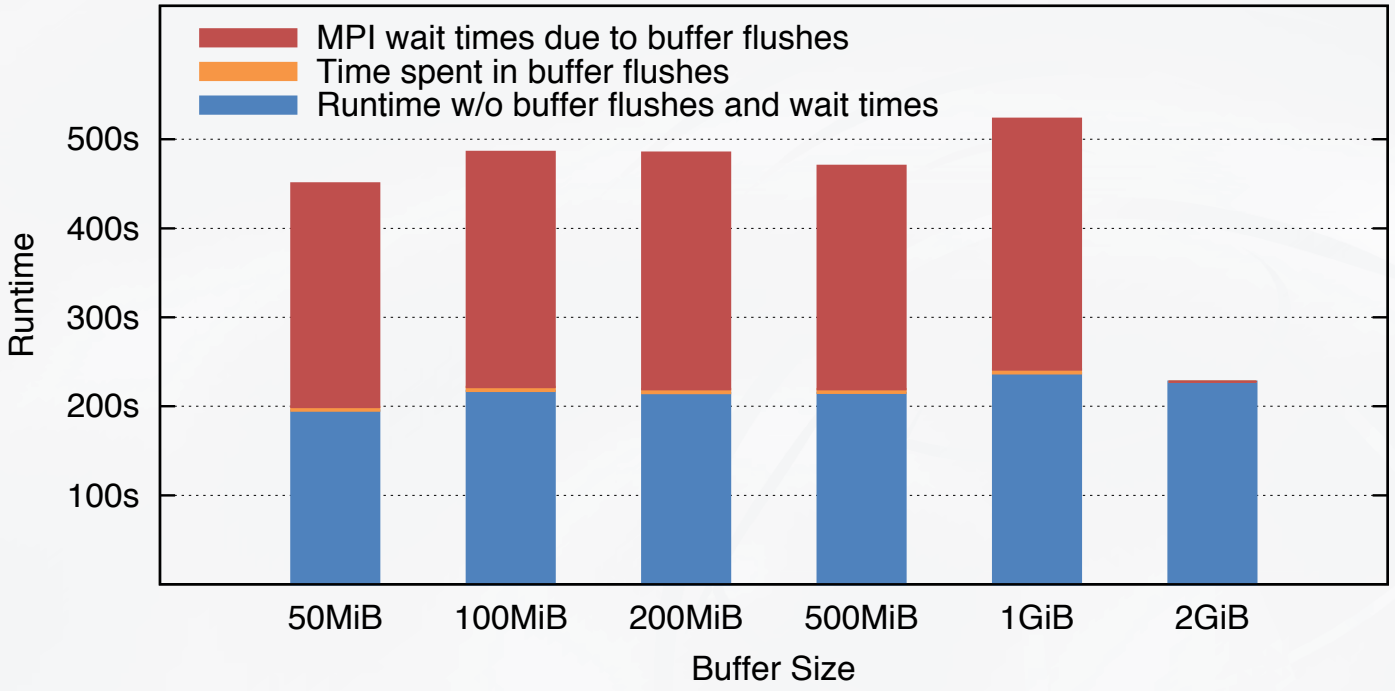
# Intermediate Memory Buffer Flushes – Distribution




… and can completely falsify the recorded application behavior

- Buffer flush, buffer size: 1 GiB
- MPI

# Intermediate Buffer Flushes




- MPI wait times due to buffer flushes account for 55% of runtime
- Mainly in **`MPI_SendRecv, MPI_Recv, MPI_Waitall`** – not in collectives

# MPI-only Tracing

| Application | Trace size (per process) | | |
|---|---|---|---|
| | OTF2 | | MPI-only |
| gromacs | 1.7 GB | | 9.8 MB |
| cosmo-specs | 1.5 GB | | 80 KB |
| 3dbox | 919 MB | ? | 8.8 MB |
| pipe | 817 MB | | 8.5 MB |
| colloid | 900 MB | | 12 MB |
| lennard-jones | 1.8 GB | | 690 kB |
| rigid | 709 MB | | 680 kB |

- MPI-only tracing drastically reduces trace size
- Communication events lose their context in the application behavior

# State-of-the-art

**Complete trace**

May contain falsified information

?

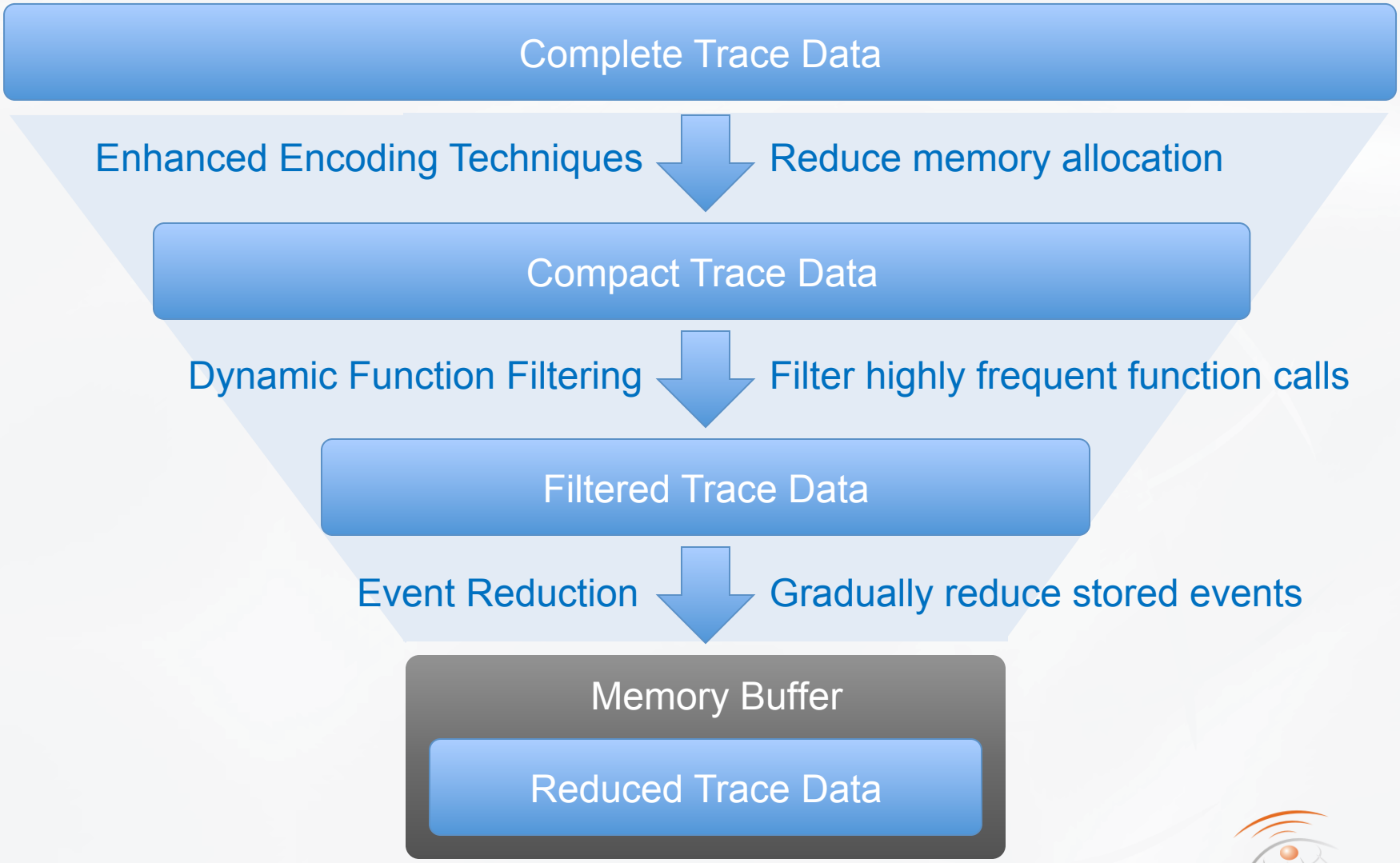
**MPI-only trace**

Allows correct but context-free communication analysis

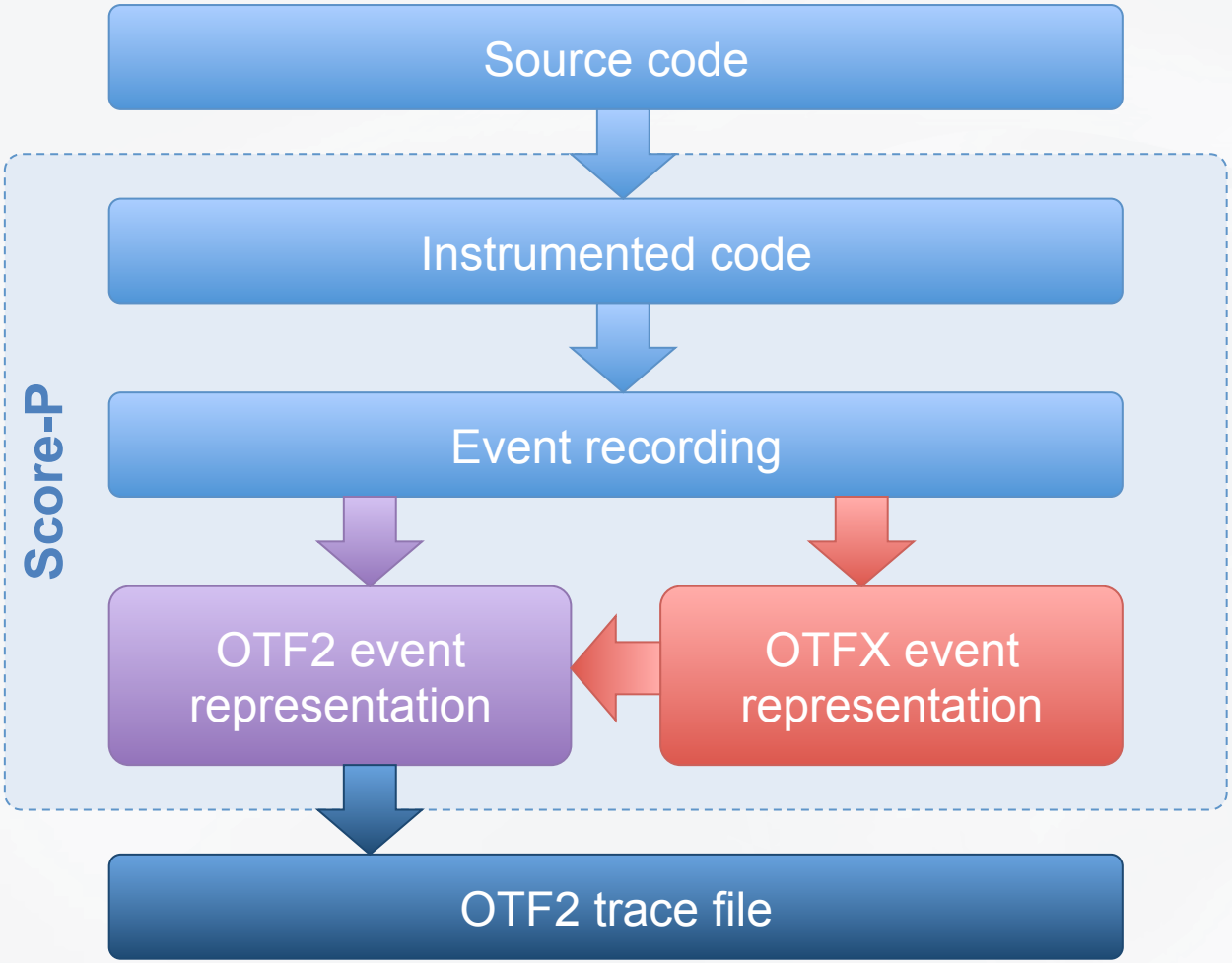
**MPI-focused Tracing – A compromise**

Provide complete MPI communication and reduce the application events to fit into a single buffer

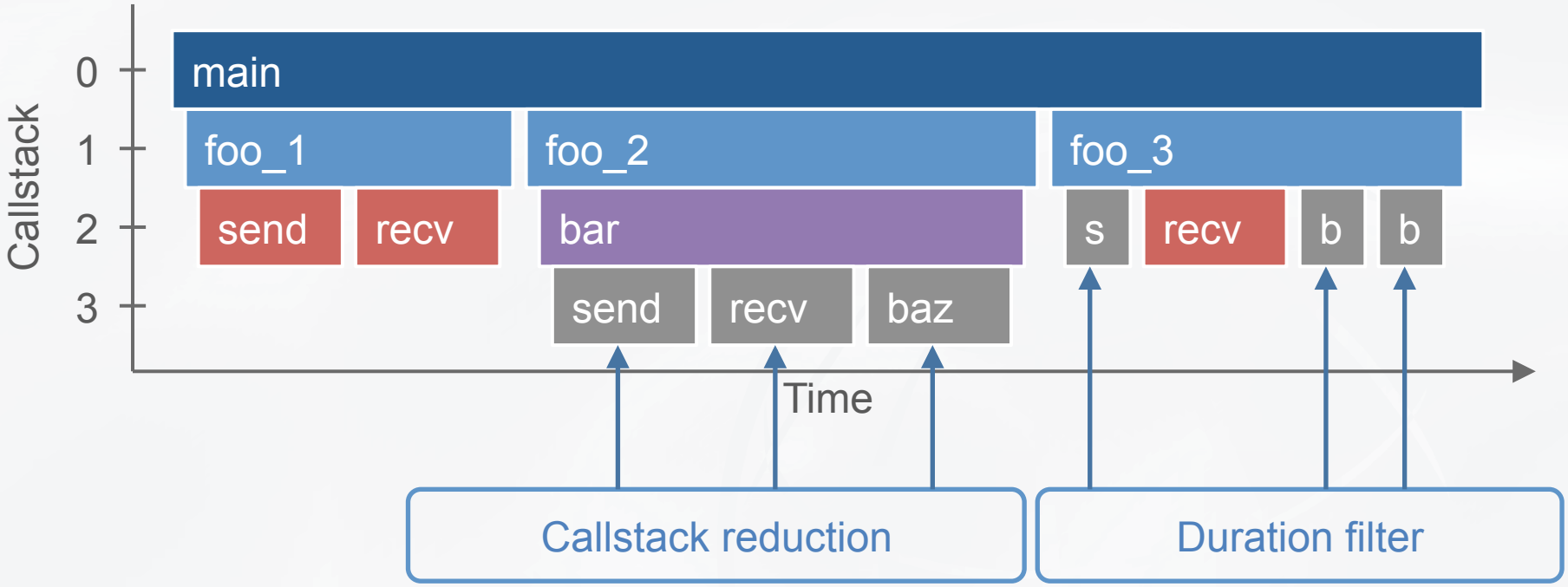
# OTFX – In-memory Event Tracing Extension to OTF2



# Workflow OTFX

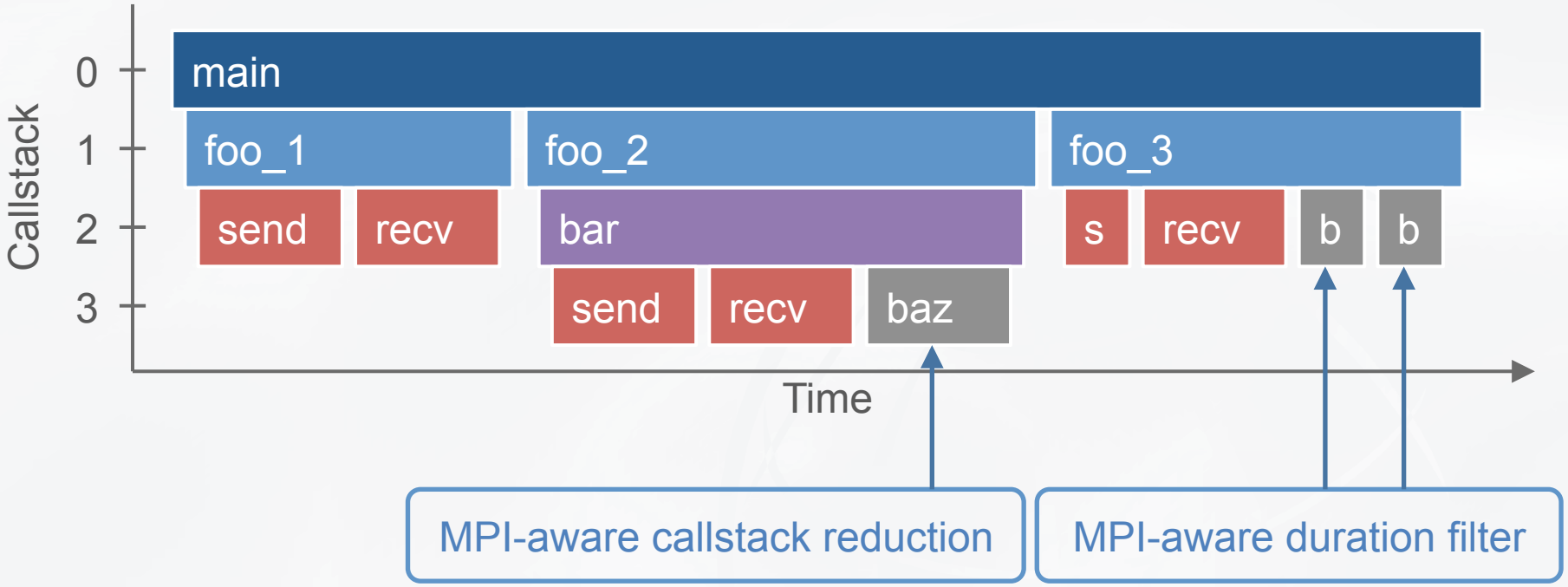


# Non-MPI-aware OTFX



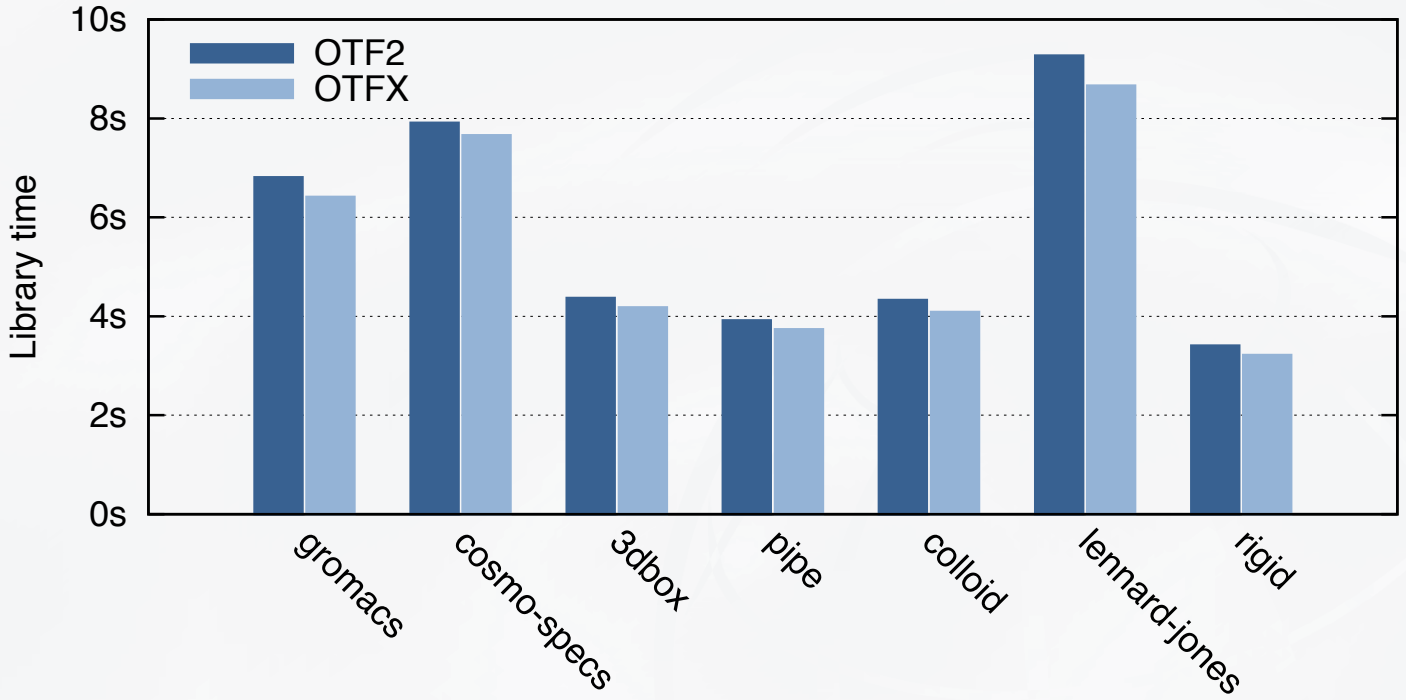

- Standard reduction and filtering eliminates also matching MPI events

# MPI-aware OTFX




- MPI-aware reduction and filtering omits MPI events

# Evaluation: Runtime overhead



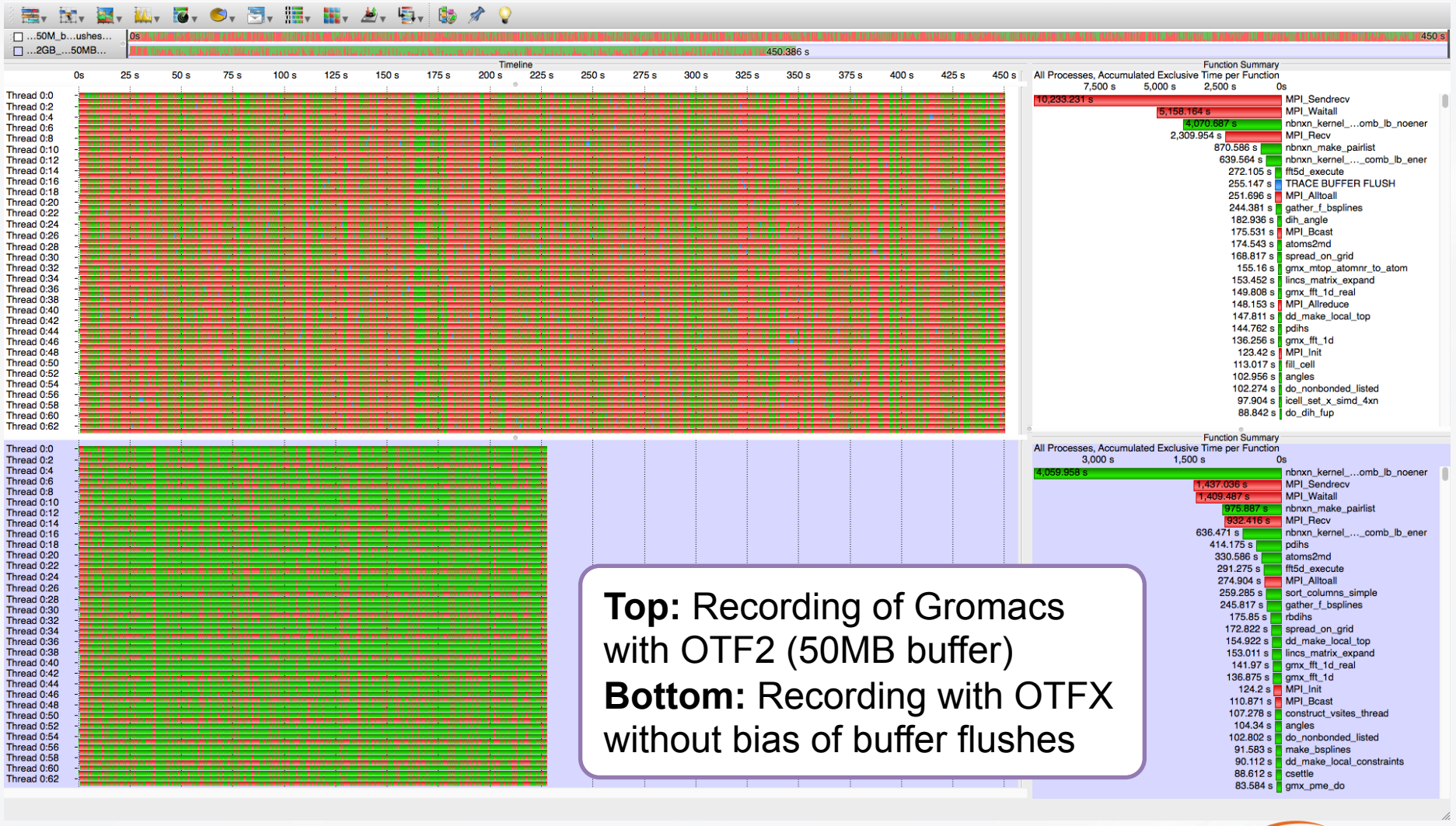

- Trace replay to ensure equal input data for both libraries
- In average 5.1% faster than OTF2
- Library time of OTFX accounts for 7.8% of overall runtime

# Evaluation: Trace Sizes

| Application | Trace size (per process) | | | |
|---|---|---|---|---|
| | OTF2 | OTFX | +Filter (1µs) | MPI-only |
| gromacs | 1.7 GB | 603 MB | 127 MB | 9.8 MB |
| cosmo-specs | 1.5 GB | 514 MB | 21 MB | 80 KB |
| 3dbox | 919 MB | 297 MB | 116 MB | 8.8 MB |
| pipe | 817 MB | 267 MB | 88 MB | 8.5 MB |
| colloid | 900 MB | 266 MB | 40 MB | 12 MB |
| lennard-jones | 1.8 GB | 546 MB | 4.1 MB | 690 kB |
| rigid | 709 MB | 203 MB | 23 MB | 680 kB |

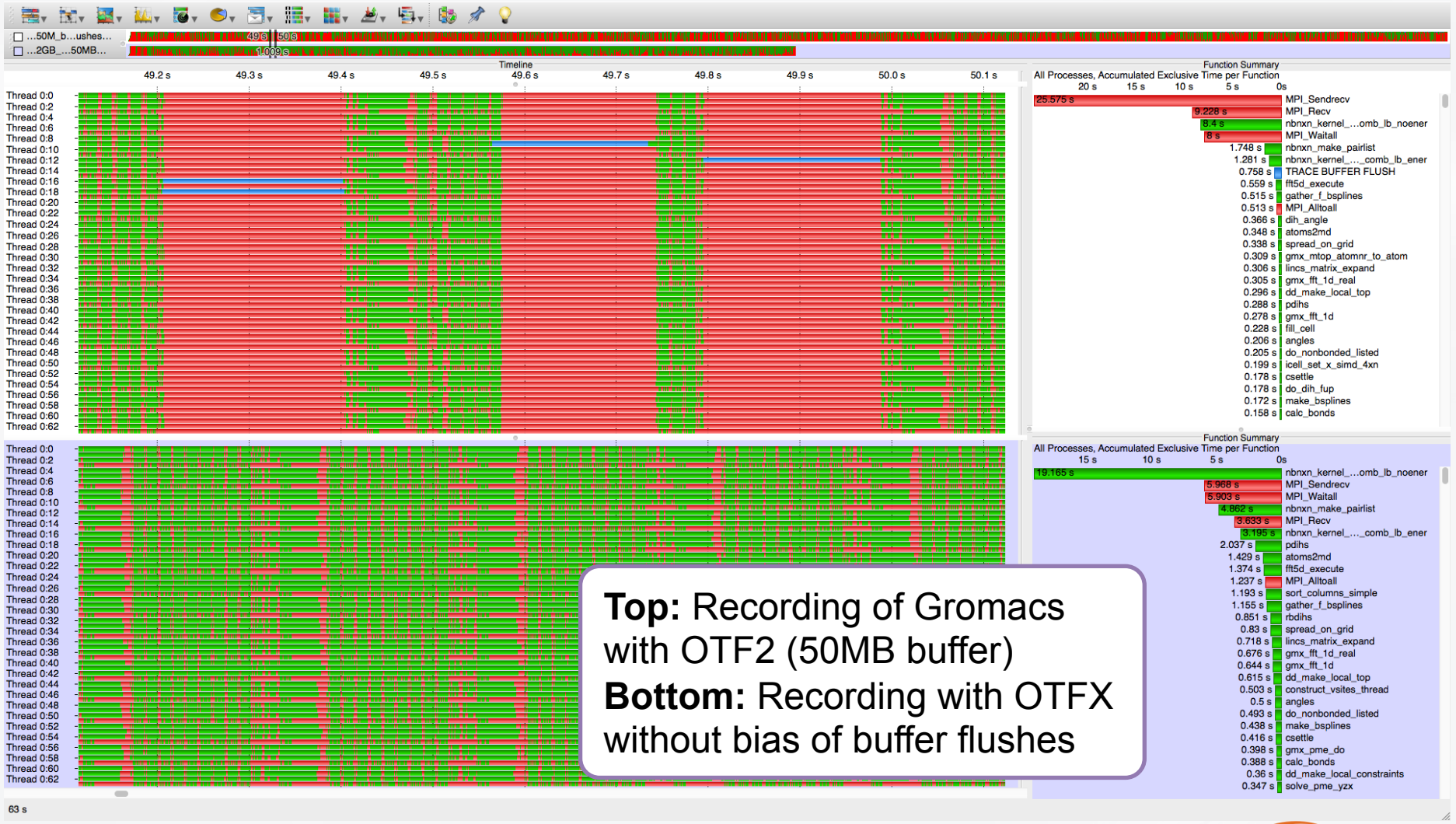
- OTFX compression results in 2.8x - 3.5x smaller traces
- Duration filter reduces trace to 0.2% - 12.6% of original size
- For gromacs and nek5000 (3dbox, pipe) event reduction is triggered

# Evaluation: Measurement Bias




**Top:** Recording of Gromacs with OTF2 (50MB buffer)
**Bottom:** Recording with OTFX without bias of buffer flushes

# Evaluation: Measurement Bias




**Top:** Recording of Gromacs with OTF2 (50MB buffer)
**Bottom:** Recording with OTFX without bias of buffer flushes

# Conclusions

- Tracing collects large amounts of data
  - Complete trace may prevent a correct analysis due to buffer flushes
  - MPI-only trace does not provide application context
- MPI-focused tracing provides complete MPI communication and reduces the application events to fit into a single buffer
- Reduces overhead by 5% and trace size up to 3 orders of magnitude
- Allows a meaningful, as well as correct, analysis