TECHNISCHE UNIVERSITÄT DRESDEN

Comprehensive Lustre I/O Tracing with Vampir

Lustre User Group 2010

Zellescher Weg 12

WIL A 208

Tel. +49 351 - 463 – 34217

Michael Kluge (michael.kluge@tu-dresden.de)

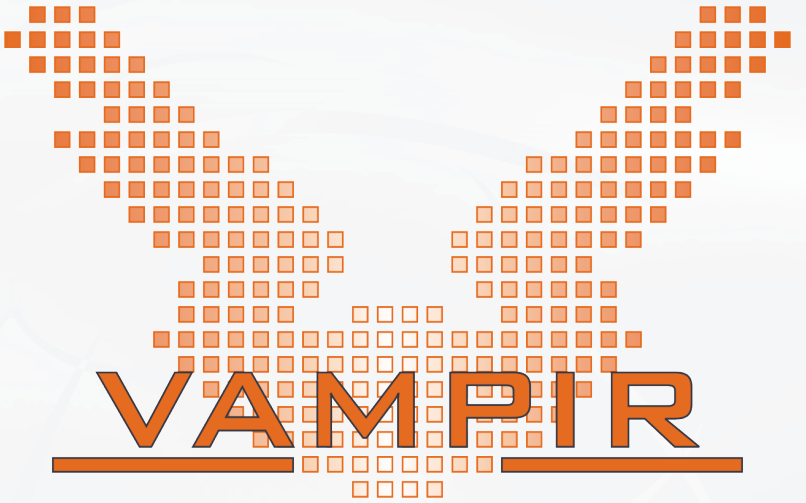
ZIH
Zentrum für Informationsdienste und Hochleistungsrechnen

Content

- Vampir Introduction
- VampirTrace introduction
- Enhancing I/O analysis
- Creation of a comprehensive system view
- Illustrating Examples

Vampir Introduction

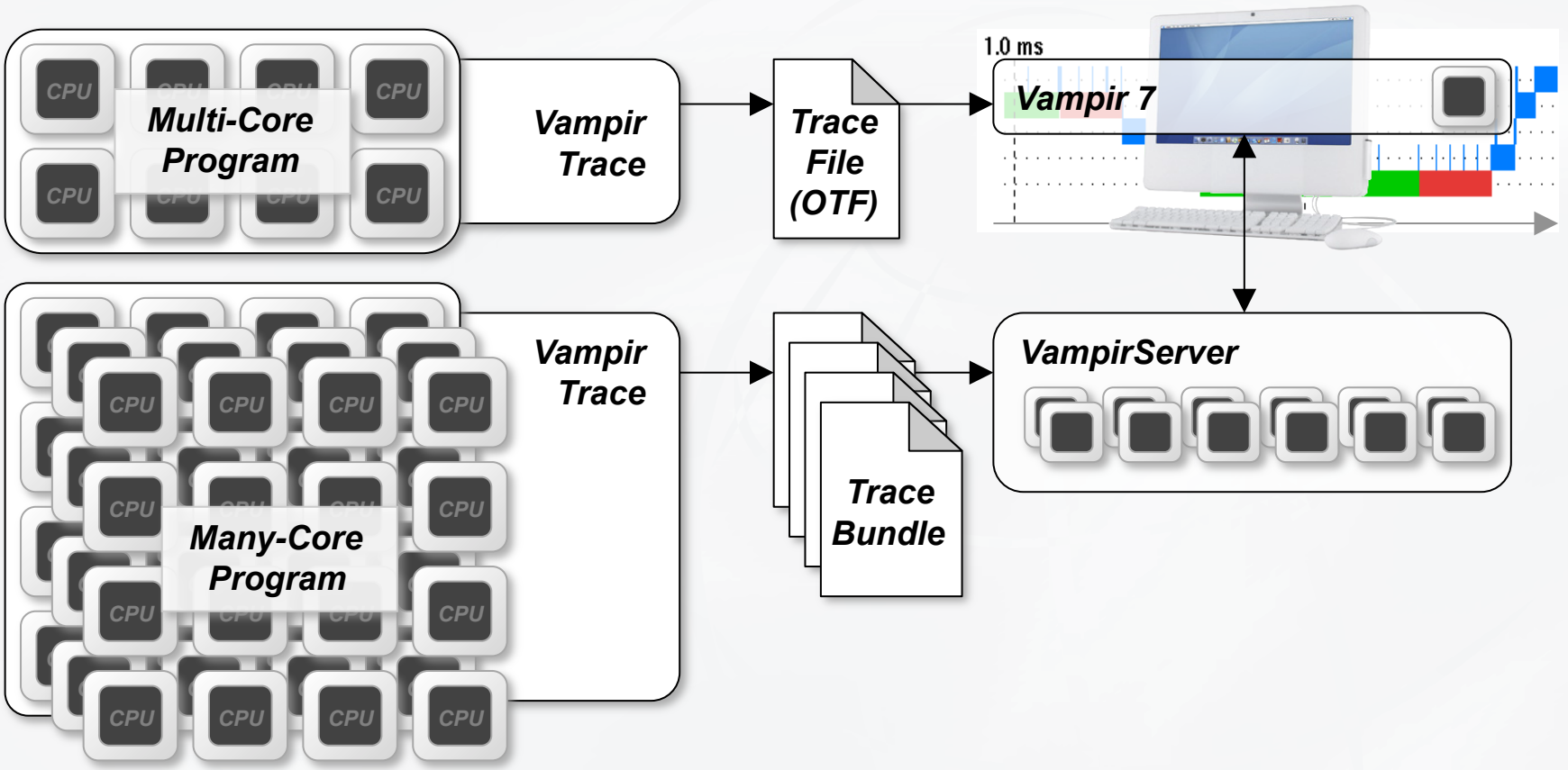
- Visualization of dynamics of complex parallel processes
- Requires two components
  - Monitor/Collector (VampirTrace)
  - Charts/Browser (Vampir)
- Available for major platforms
- Open Source (partially)



http://www.vampir.eu

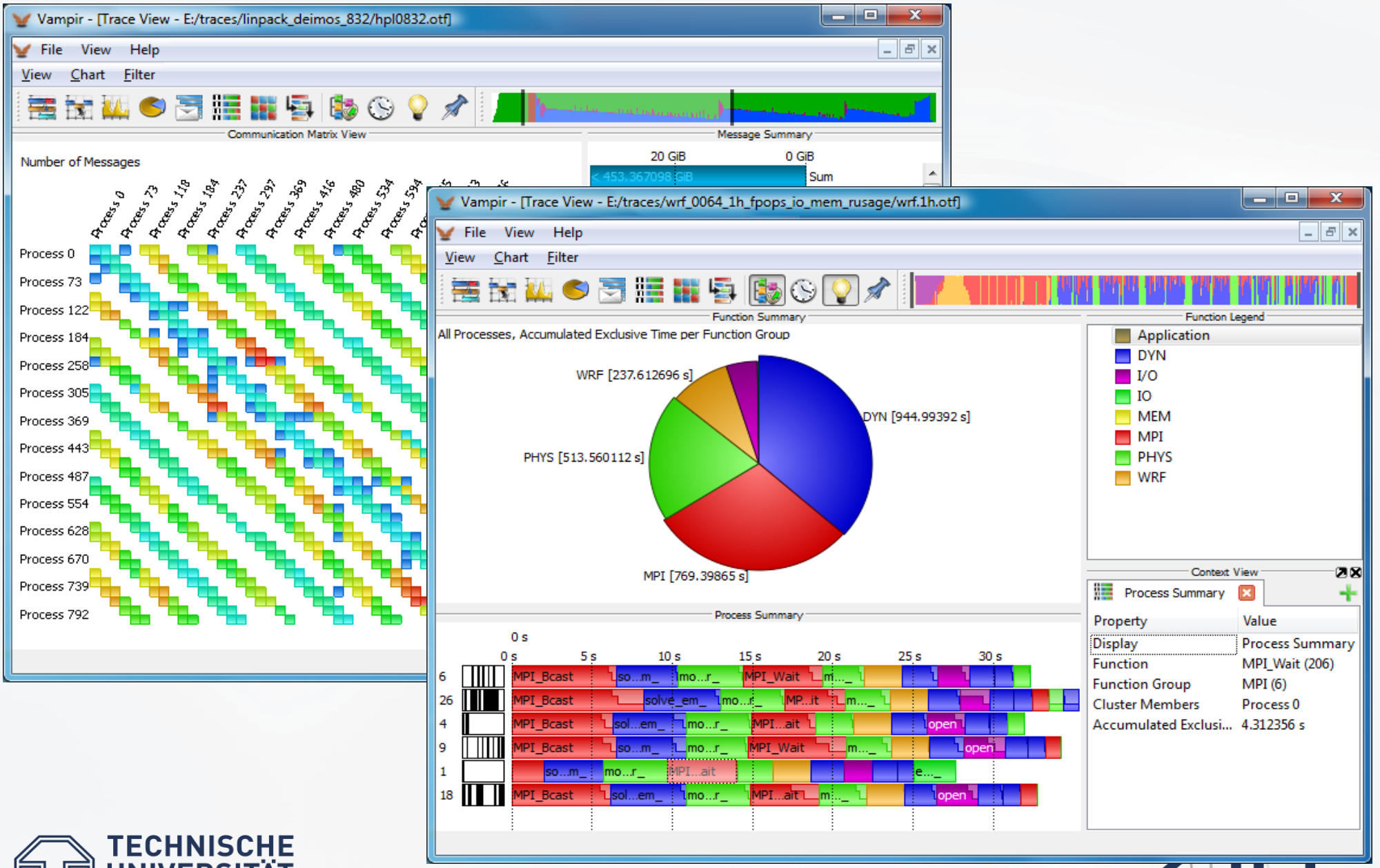
http://www.tu-dresden.de/zih/vampirtrace

Vampir Toolset Architecture



Vampir 7: Timelines

Vampir 7: Summaries



I/O in HPC Environments

- Thousands of clients connected to a file server farm
- Dedicated infrastructure, large number of supporting components
- File systems often accessed through high level libraries
- File systems typically tuned for large I/O requests
- No backup

- Beside the 'large&fast' file system HPC systems have
  - slow home directory
  - even more slow archive space

- Multi cluster file systems

I/O Subsystem Sizes

- TU Dresden HRSK System (installed 2006):
  - 2*68 TB: 12 GB/s, >1300 disks, 12 file servers, 48 FC cables
  - 700+ nodes

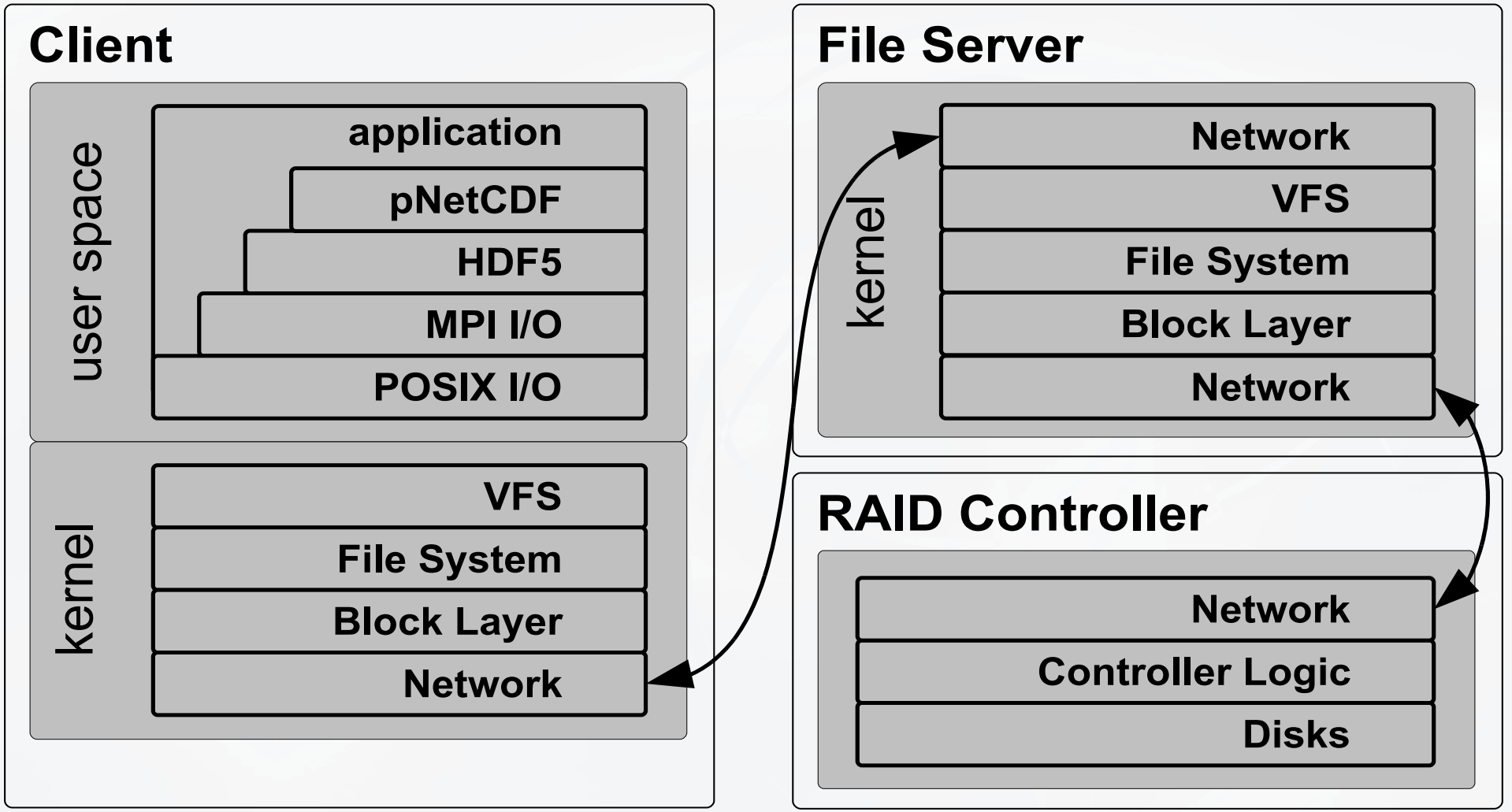
- DARPA HPCS Project (targeted 2010):
  - High Productivity Computing with millions of components
  - 50.000 spinning disks, 30.000 nodes, 100 PB file systems
  - millions of cores, 10 billion files per directory , 1 PB single file size
  - 40,000 file creates/sec from a single client node
  - 30 GB/sec single client, 1.5 TB/sec total bandwidth
  - 1 PB single file size
  - end-to-end data integrity

How to optimize I/O?

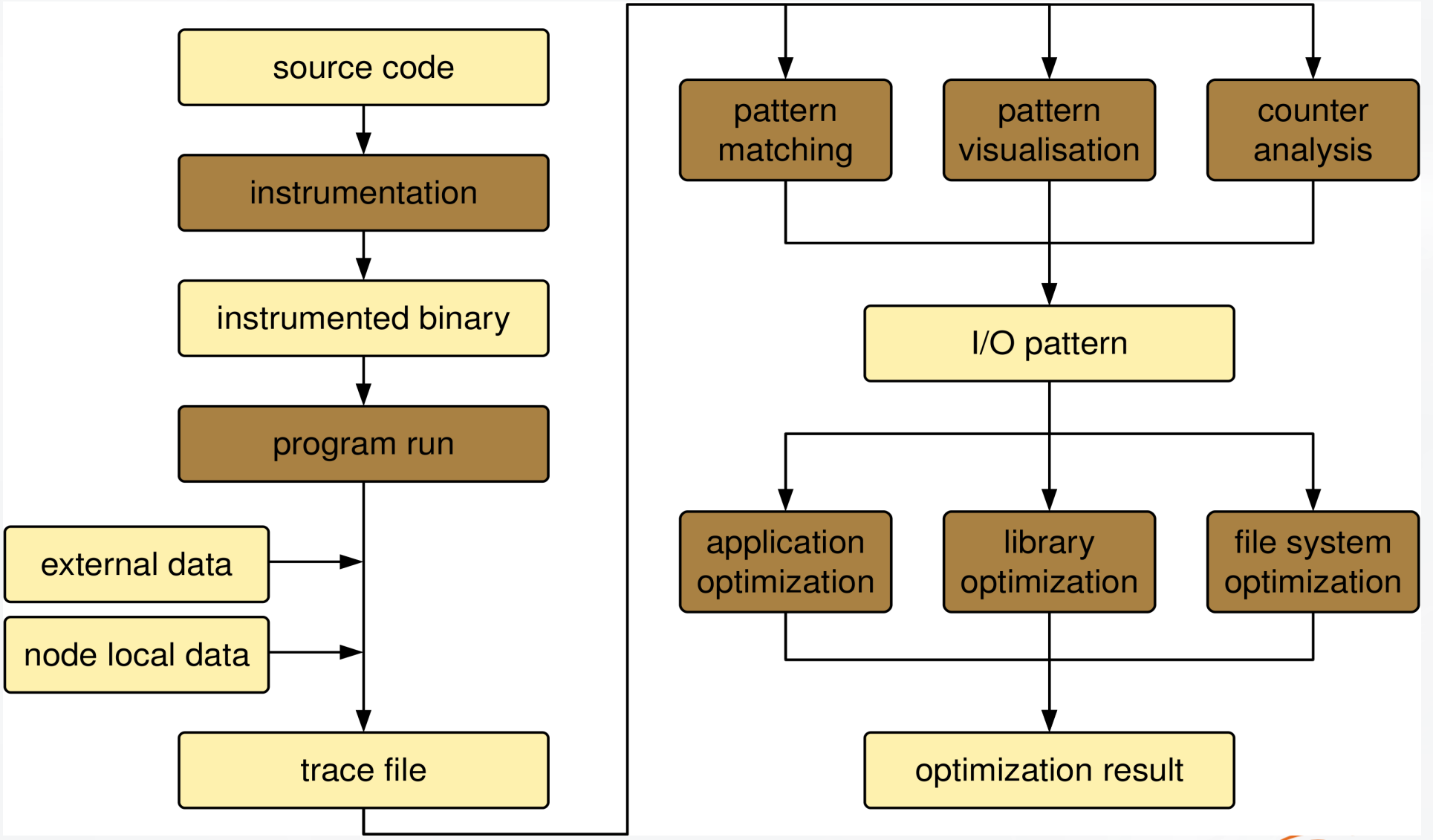
- Instrument all available data sources
- Make tools usable for production mode
- Create a global view onto the system
- Instrument the users application
- Run the user code
- Correlate and analyze the collected data
- Find an optimization strategy
- Decide about the layer to implement it
- Implement and test again

- Check if you killed someone else‘s performance

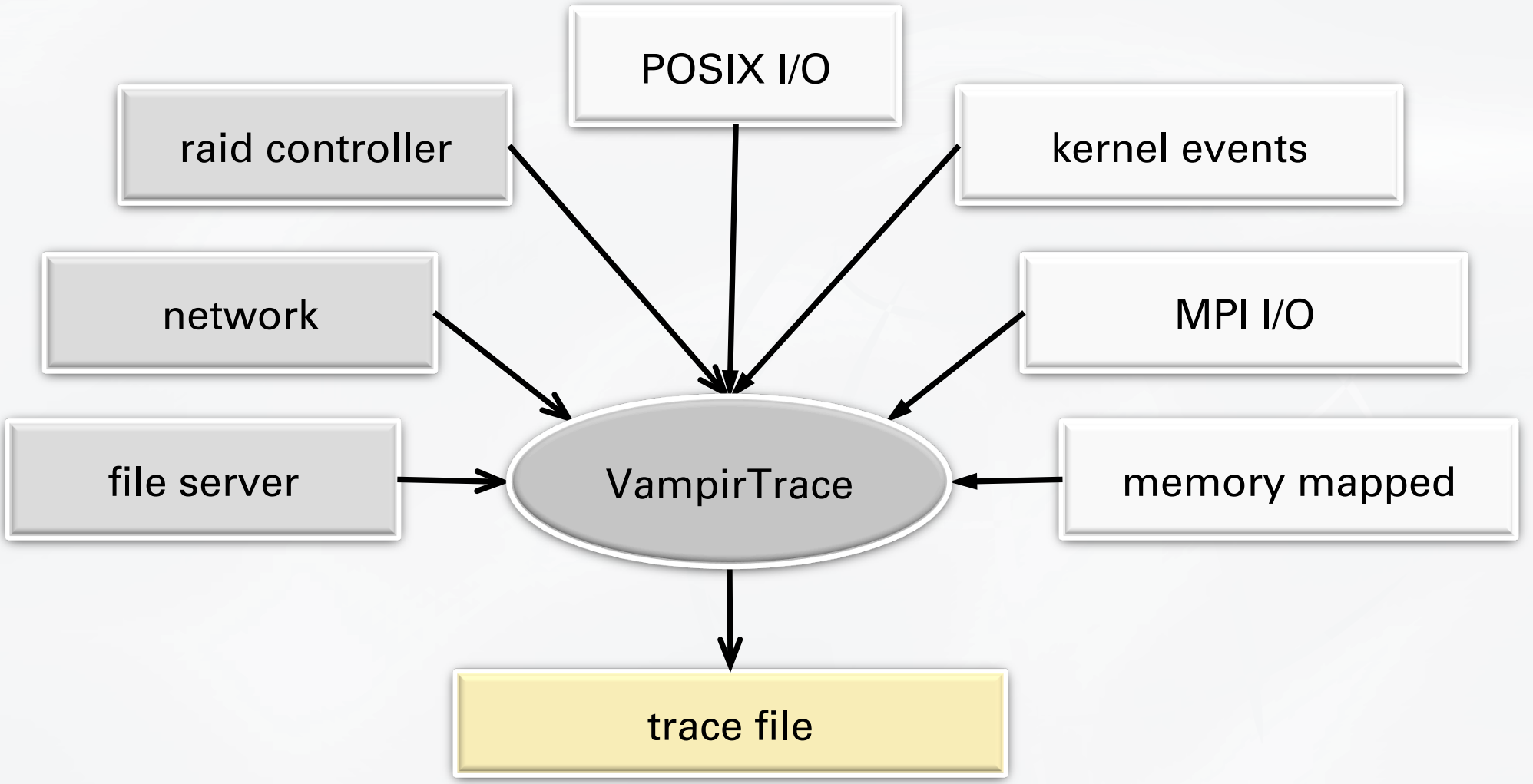
I/O Stack in HPC Environments



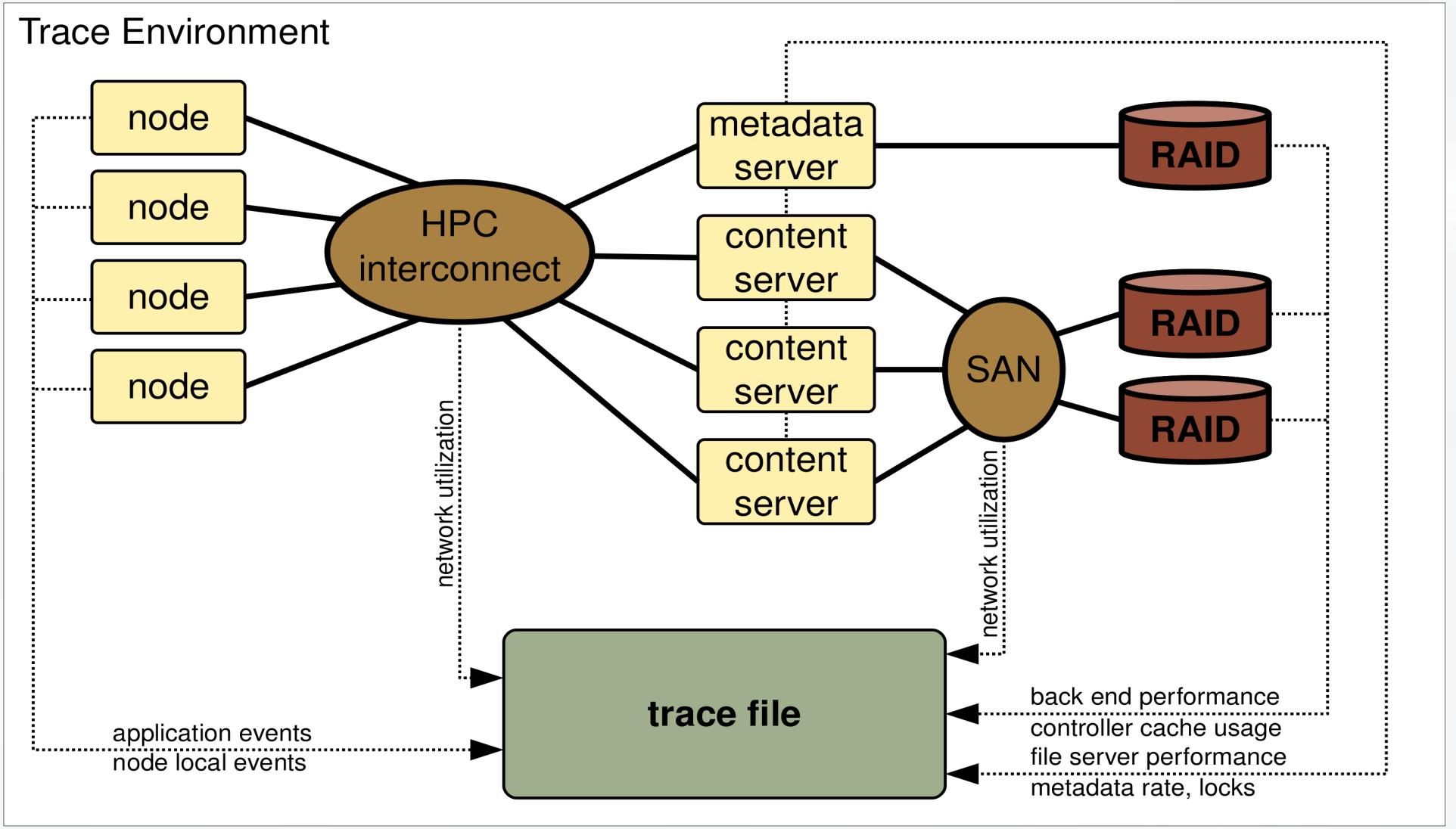
I/O Analysis Workflow



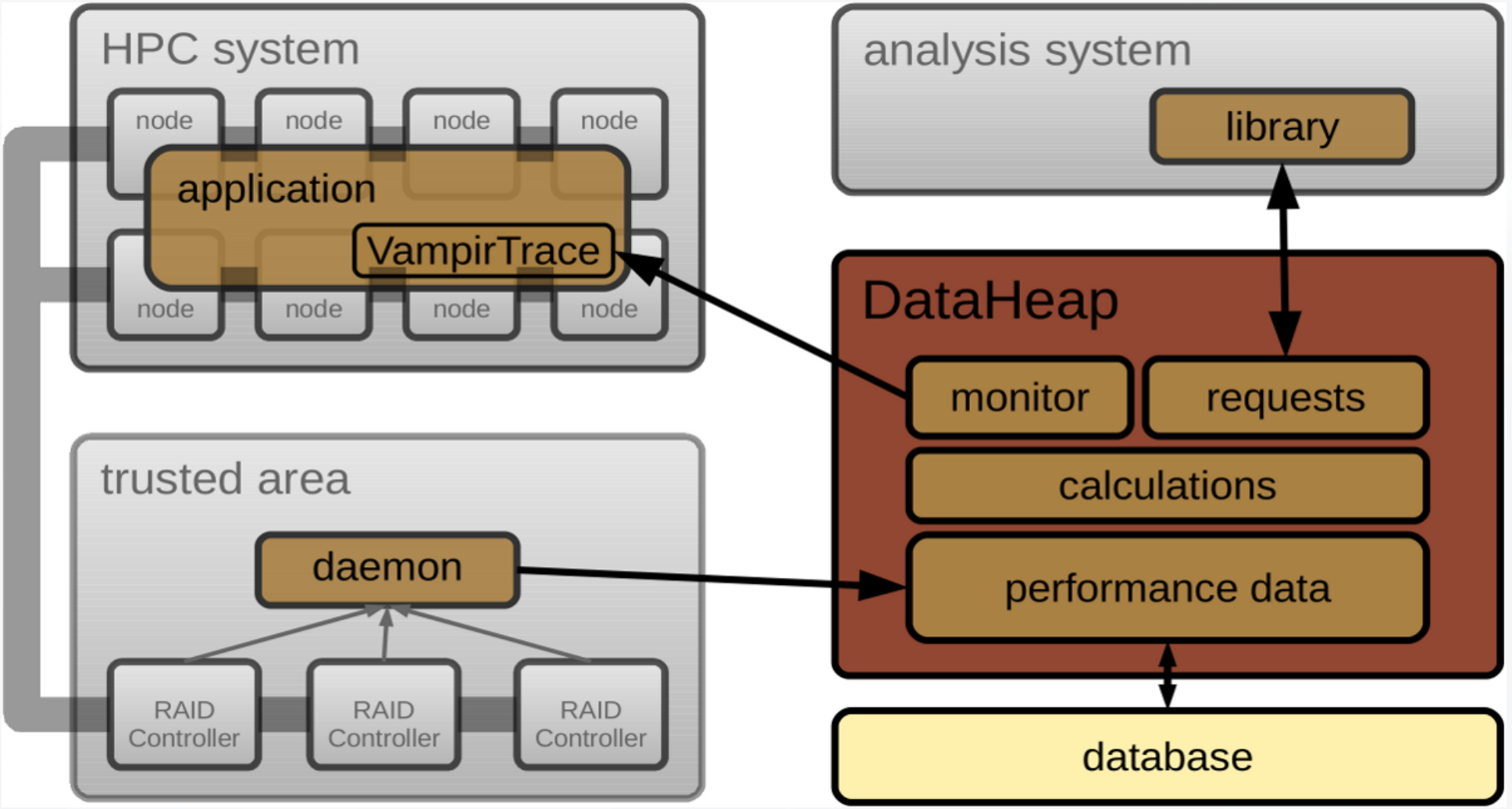
I/O Analysis Support in VampirTrace



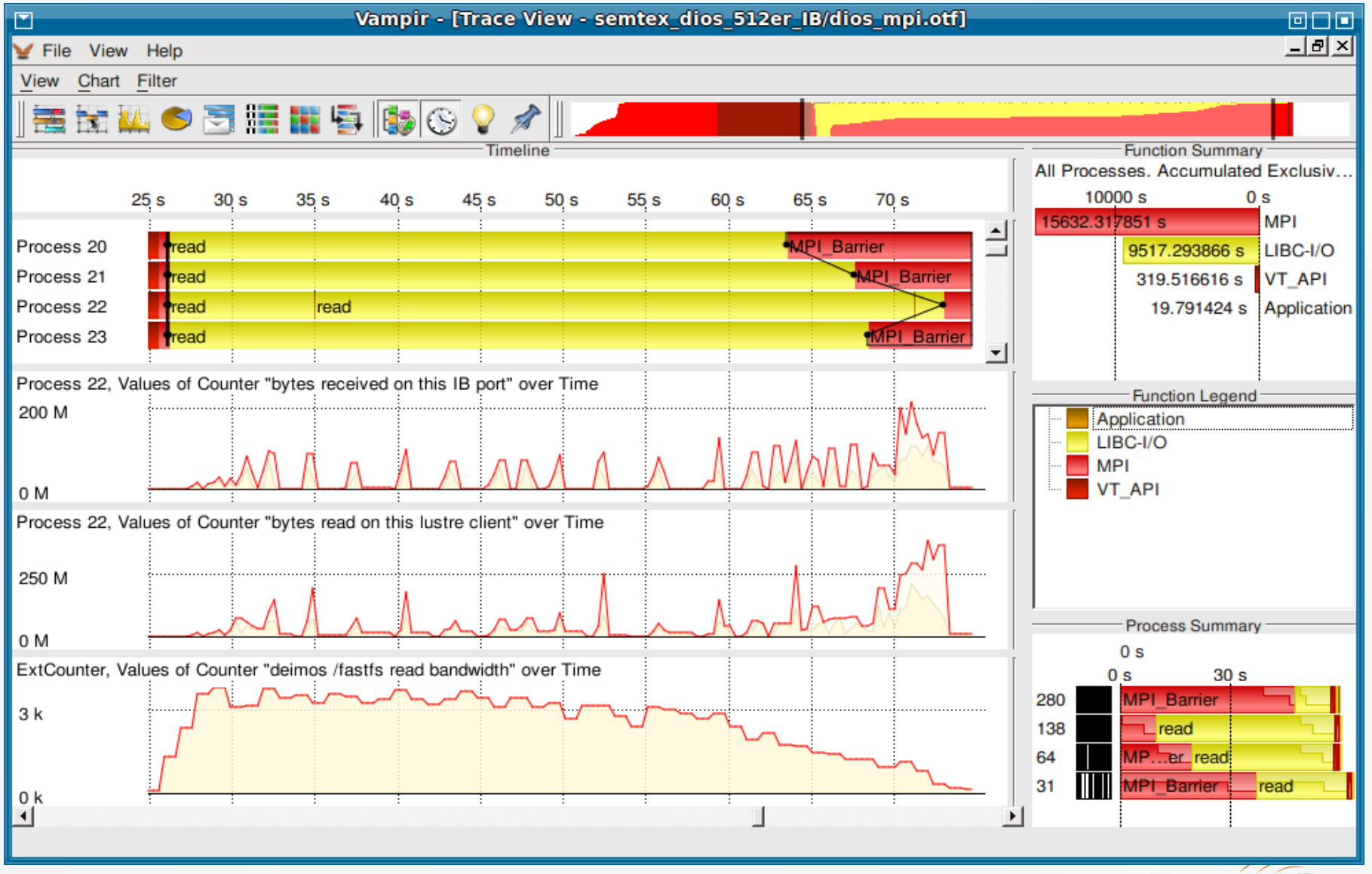
Instrumented System



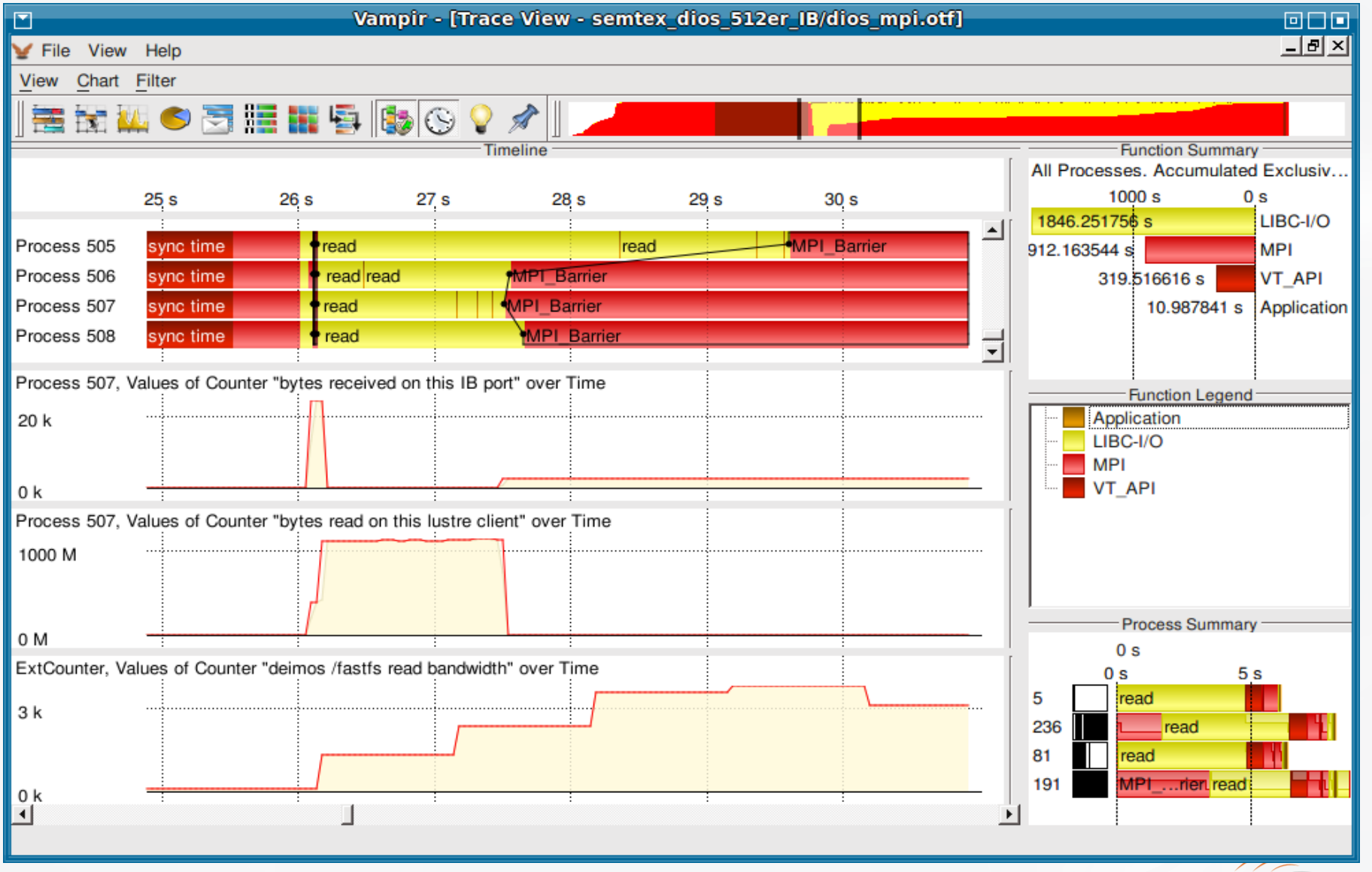
External Tool Architecture



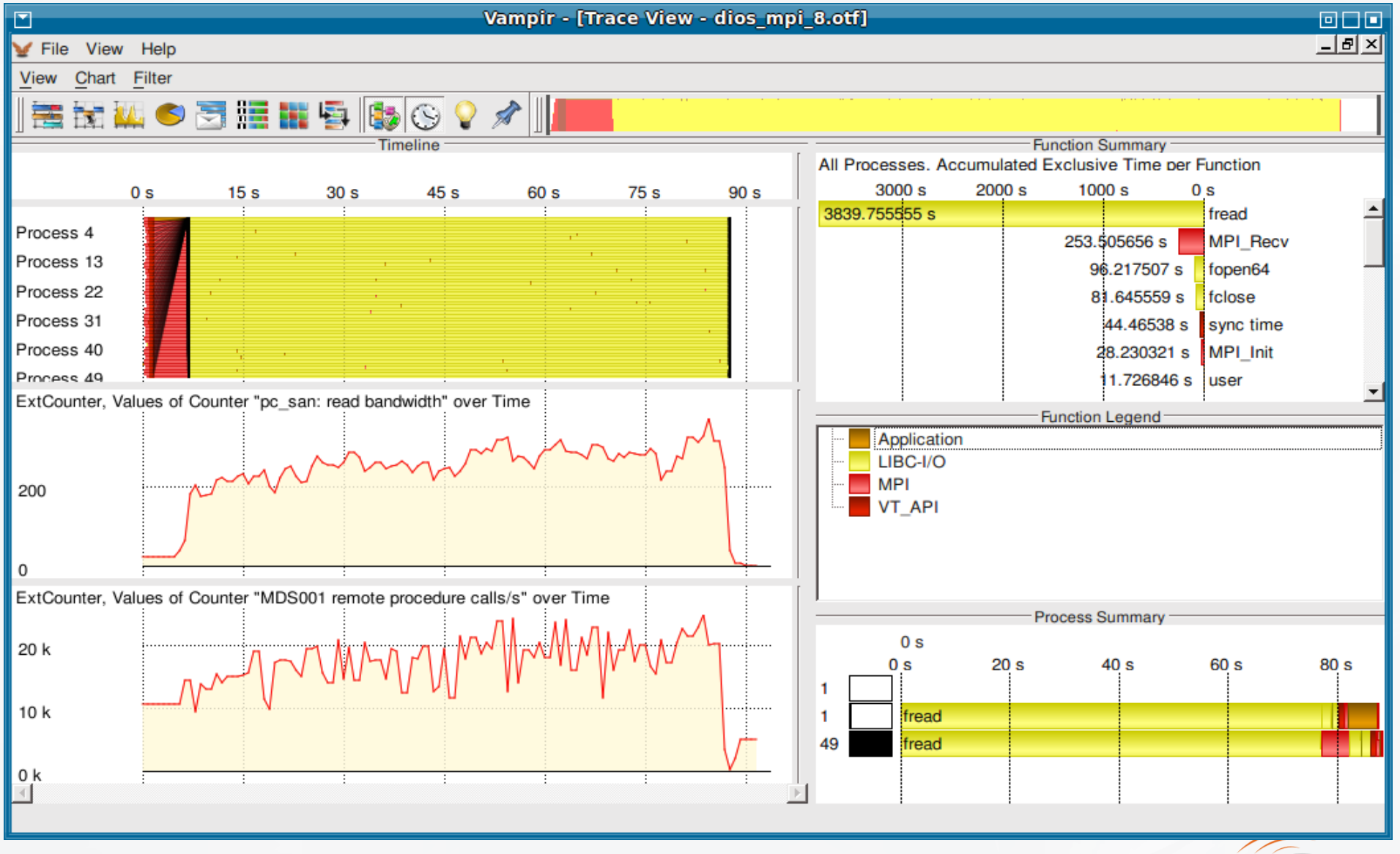
Example – Local and External Counter Values



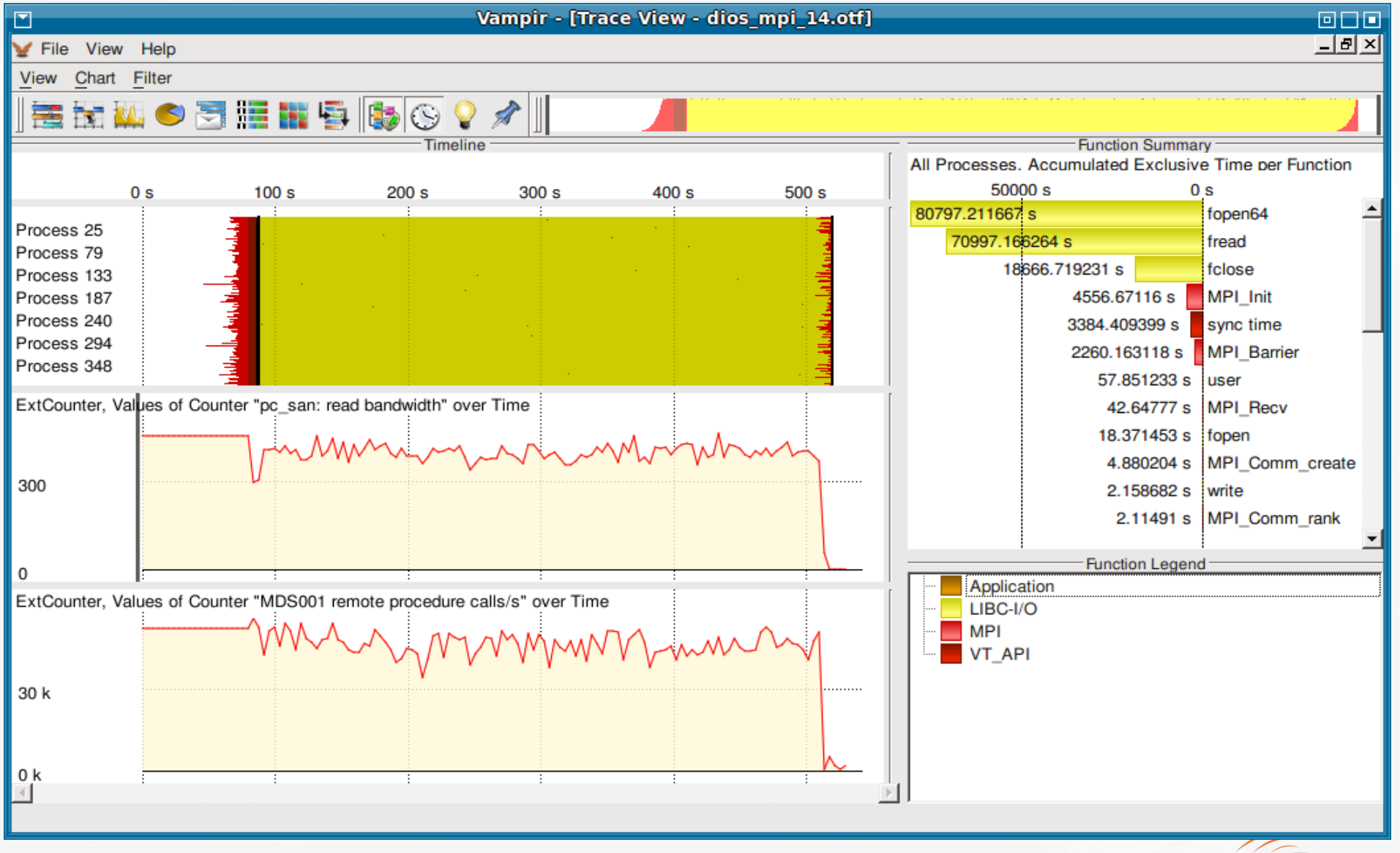
Example – Local and External Counter Values



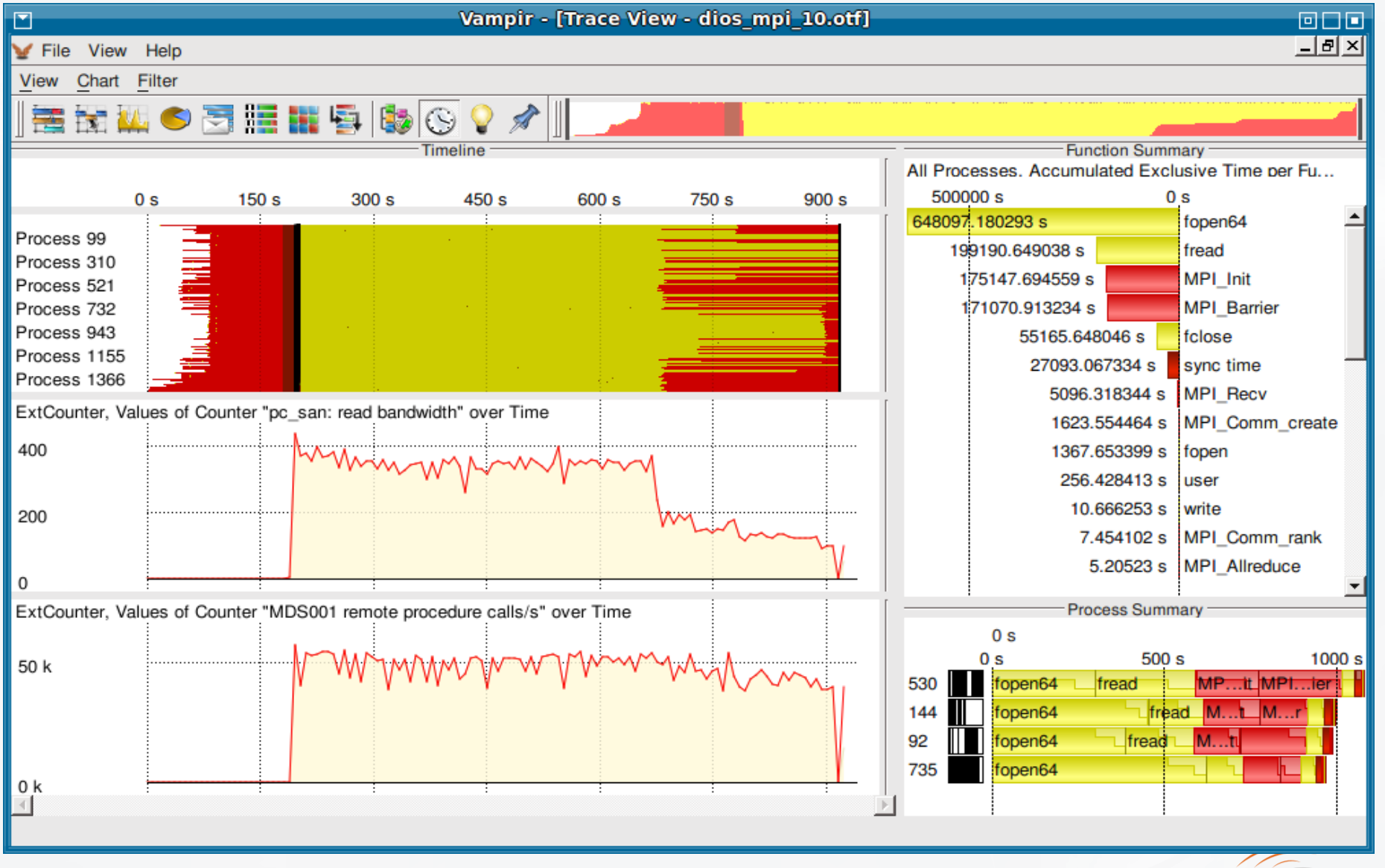
Example – threader simulation with 50 processes



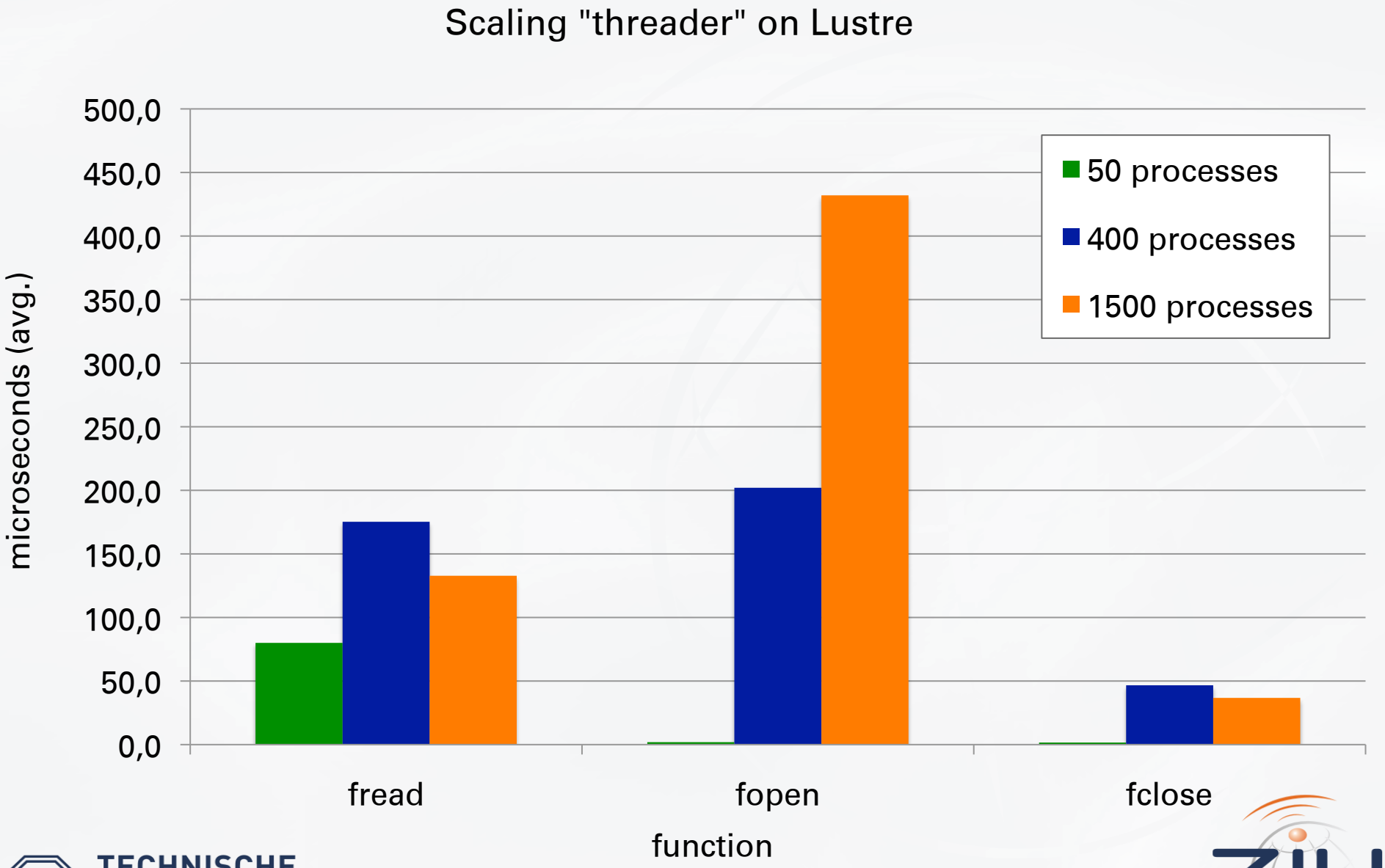
Example – threader simulation with 400 processes



Example – threader simulation with 1500 processes

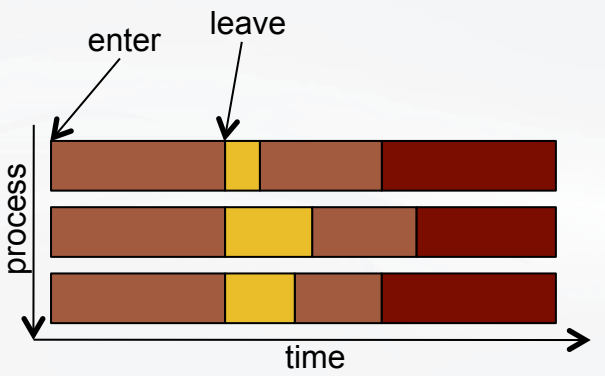


Comparing the Numbers

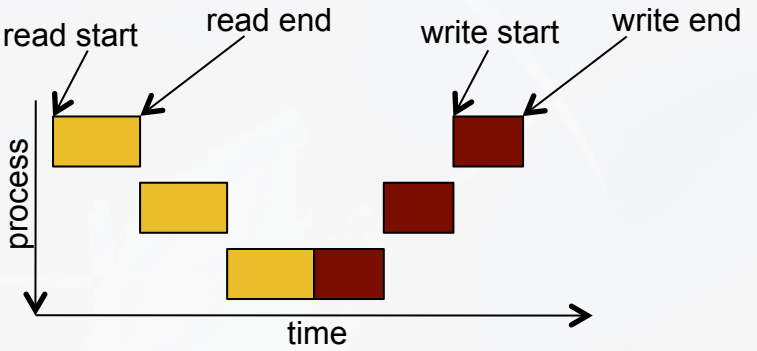


File Access Pattern Visualization

- what does the Vampir Master Timeline show:
  - X axis: time
  - Y axis: processes



- what is needed for file access patterns:
  - X axis: byte range within a single file
  - Y axis: processes



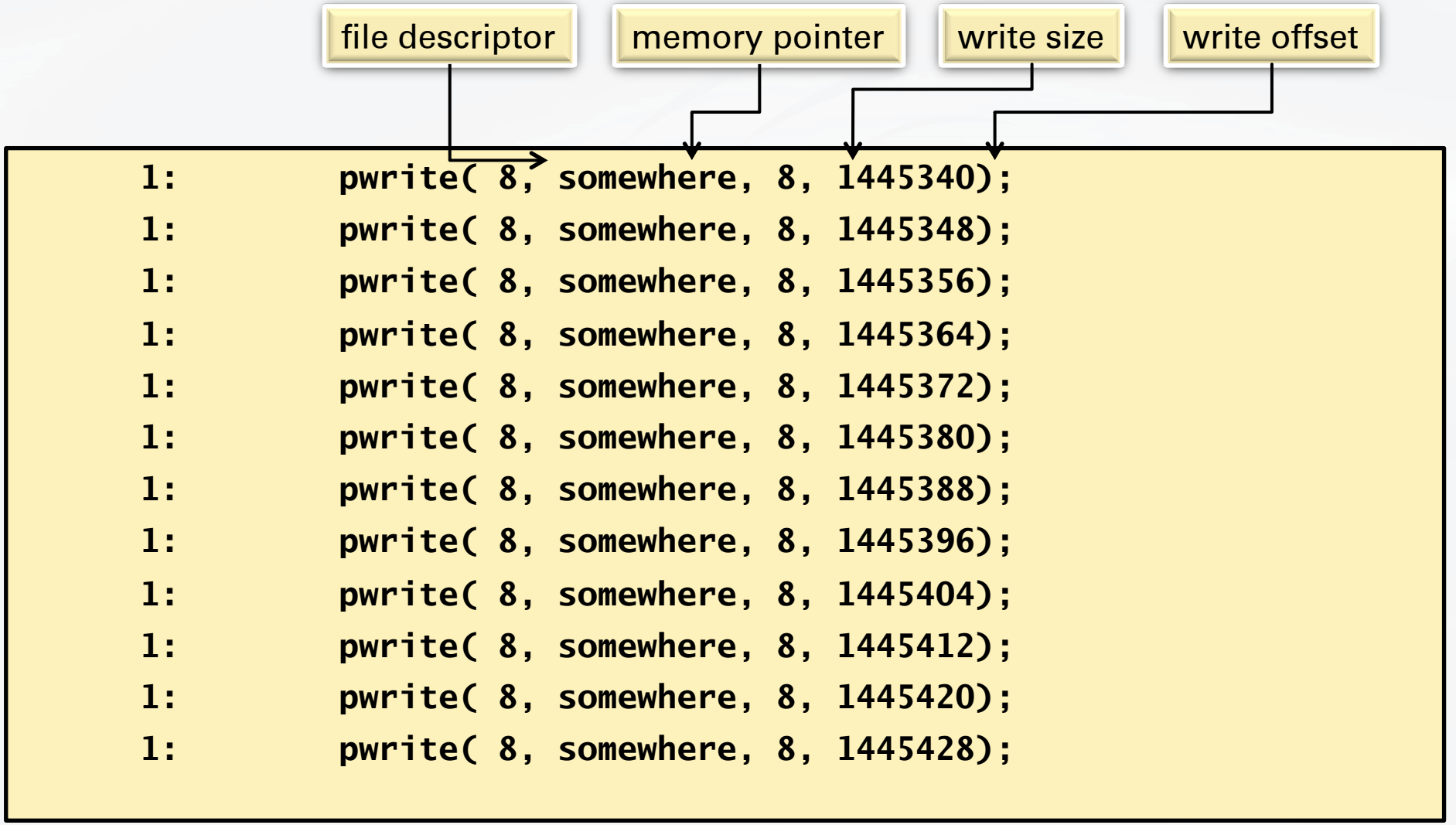
- record each write/read access for each byte range in a vector
- merge consecutive accesses
- write an appropriate enter/leave pair based on the process number and the access offset/access range into a new trace file

Flash I/O and HDF5

- Multiphysics astronomy code
- solves a broad range of astrophysical problems
- uses HDF5 or pNetCDF for parallel storage access
- I/O kernel available since 2001
- revised implementation available since 2006

- What has been changed?
  - instrument both versions
  - check differences in the I/O patterns

Pattern - 2001 Version

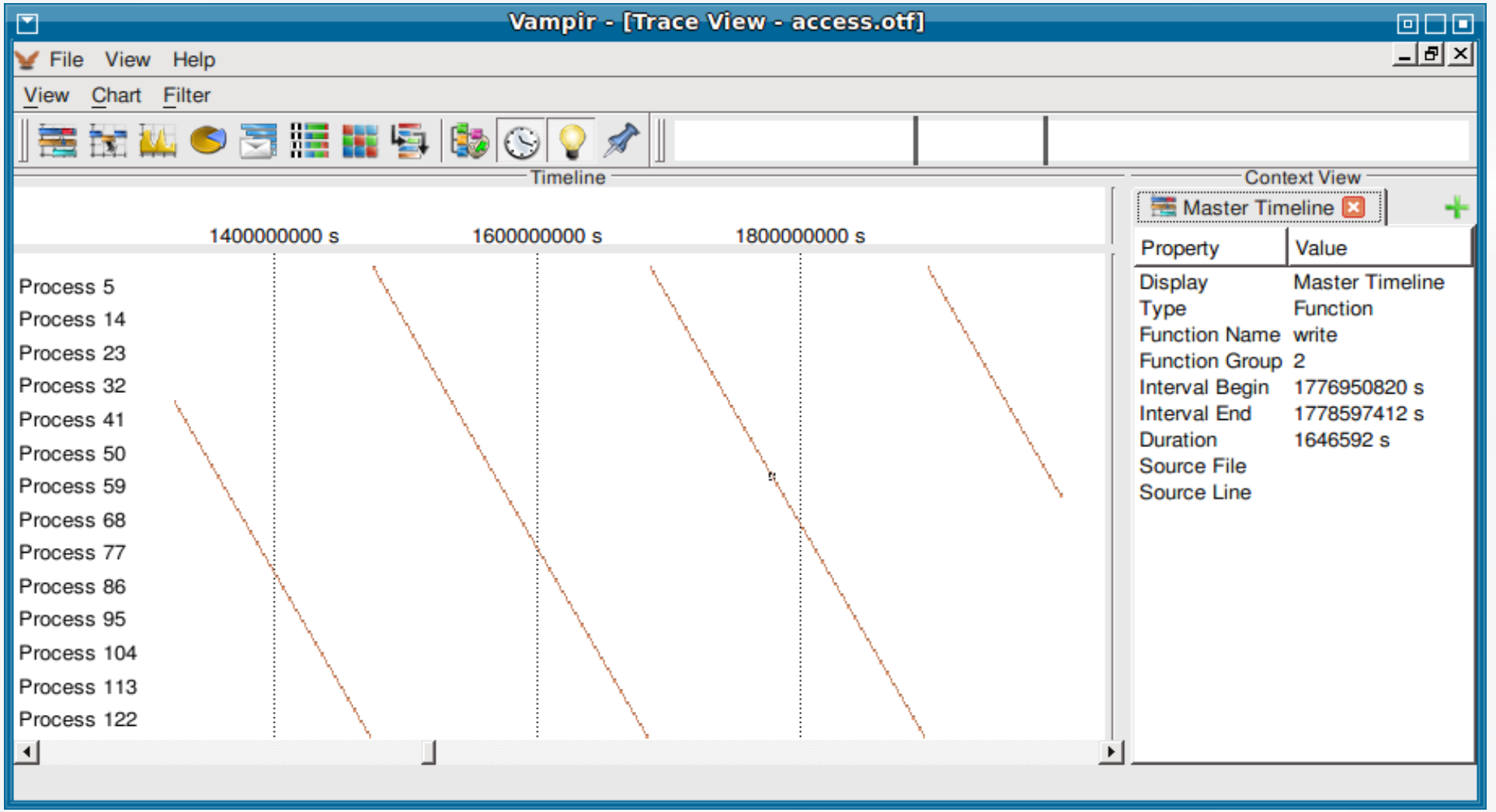


Pattern - 2006 Version

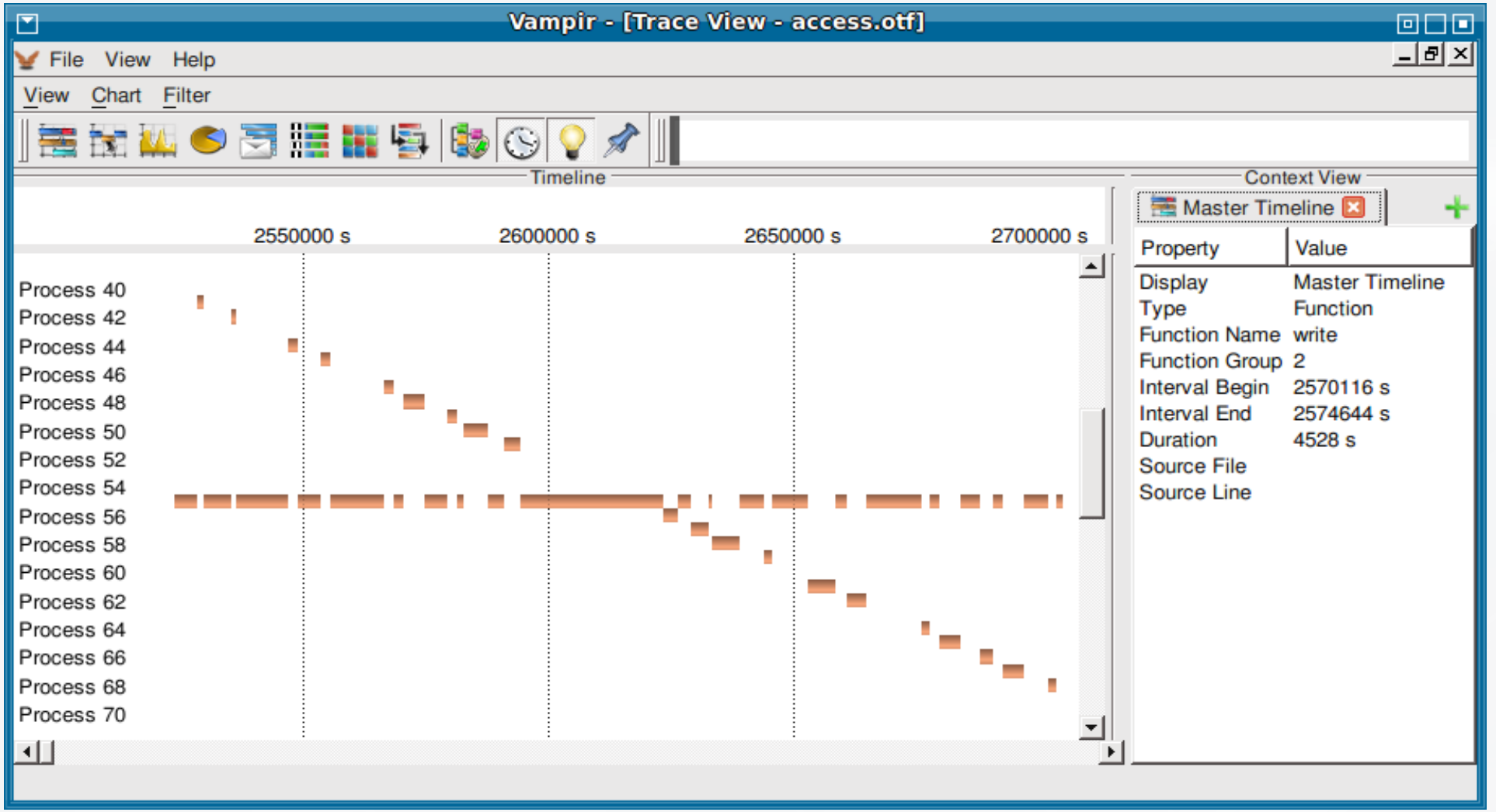
```
1:        pwrite( 8, somewhere, 1638400, 1054266916);
1:        pwrite( 8, somewhere, 1638400, 1264502308);
1:        pwrite( 8, somewhere, 1638400, 1474737700);
1:        pwrite( 8, somewhere, 1638400, 1684975140);
1:        pwrite( 8, somewhere, 1638400, 1895210532);
1:        pwrite( 8, somewhere, 1638400, 2105445924);
1:        pwrite( 8, somewhere, 1638400, 2315681316);
1:        pwrite( 8, somewhere, 1638400, 2525916708);

          ...
```

Access Pattern Visualization



Access Pattern Visualization



Time for Questions