TECHNISCHE UNIVERSITÄT DRESDEN

Center for Information Services and High Performance Computing (ZIH)

Vampir and Lustre

Understanding Boundaries in I/O Intensive Applications

Zellescher Weg 14

Treffz-Bau (HRSK-Anbau) - HRSK 151

Tel. +49 351 - 463 - 39871

Guido Juckeland (guido.juckeland@tu-dresden.de)

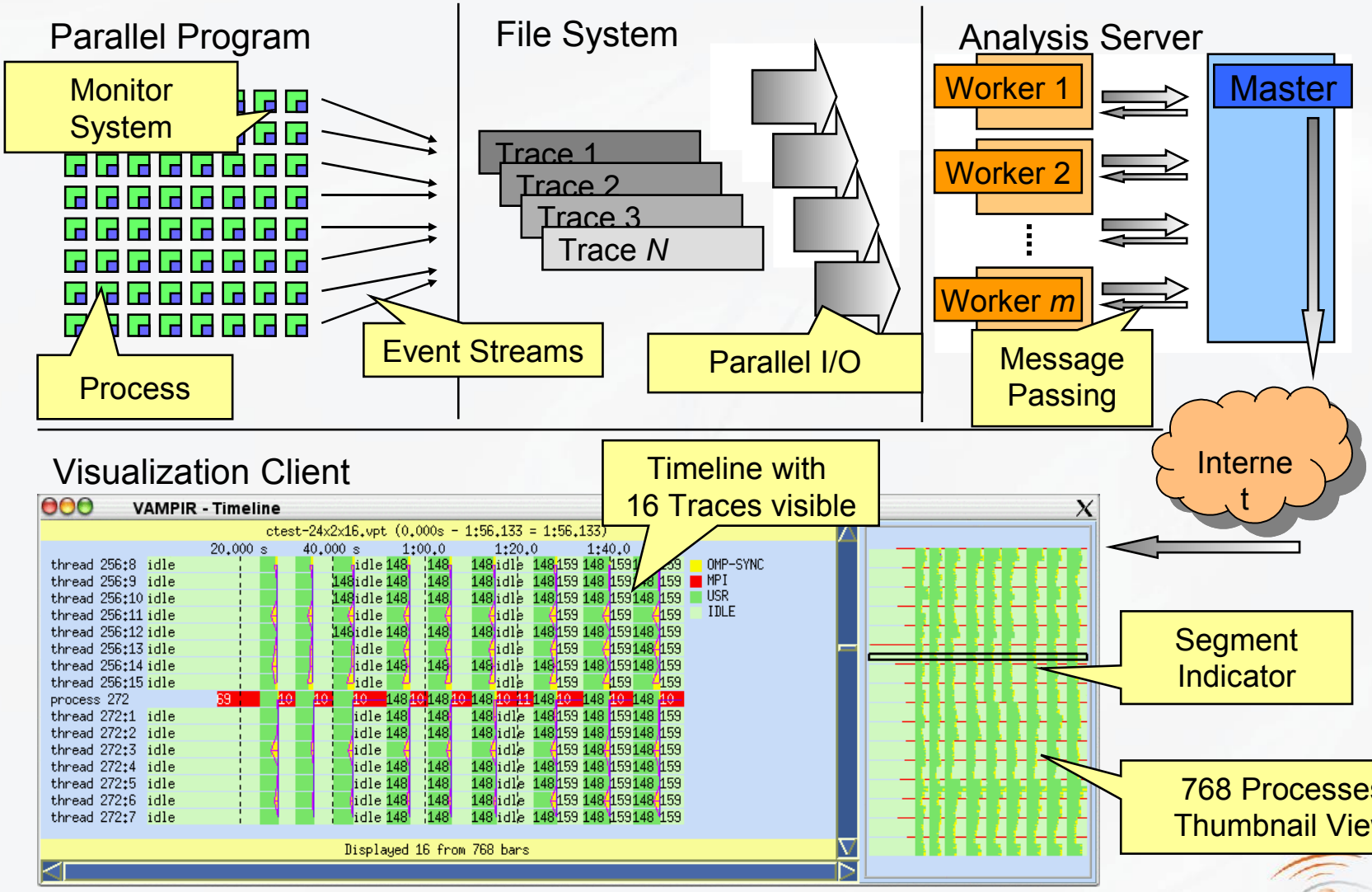
ZIH
Center for Information Services & High Performance Computing

Agenda

- Vampir – The Tool for Understanding Application Behavior
- ZIH – Focusing Resources around the Users' Data
- Tracing I/O – First Results and Further Steps

Vampir – The Tool for Understanding Application Behavior

Vampir Server Architecture



Requires: Source Code Instrumentation

```
int foo(void* arg){

   if (cond){

      return 1;
   }

   return 0;
}
```

```
int foo(void* arg){
   enter(7);
   if (cond){
      leave(7);
      return 1;
   }
   leave(7);
   return 0;
}
```

manually or automatically

What You Get: Parallel Trace (Open Trace Format (OTF))

```
10010 P 1 ENTER 5
10090 P 1 ENTER 6
10110 P 1 ENTER 12
10110 P 1 SEND TO 3 LEN 1024 ...
10330 P 1 LEAVE 12
10400 P 1 LEAVE 6
10520 P 1 ENTER 9
10550 P 1 LEAVE 9
...
```

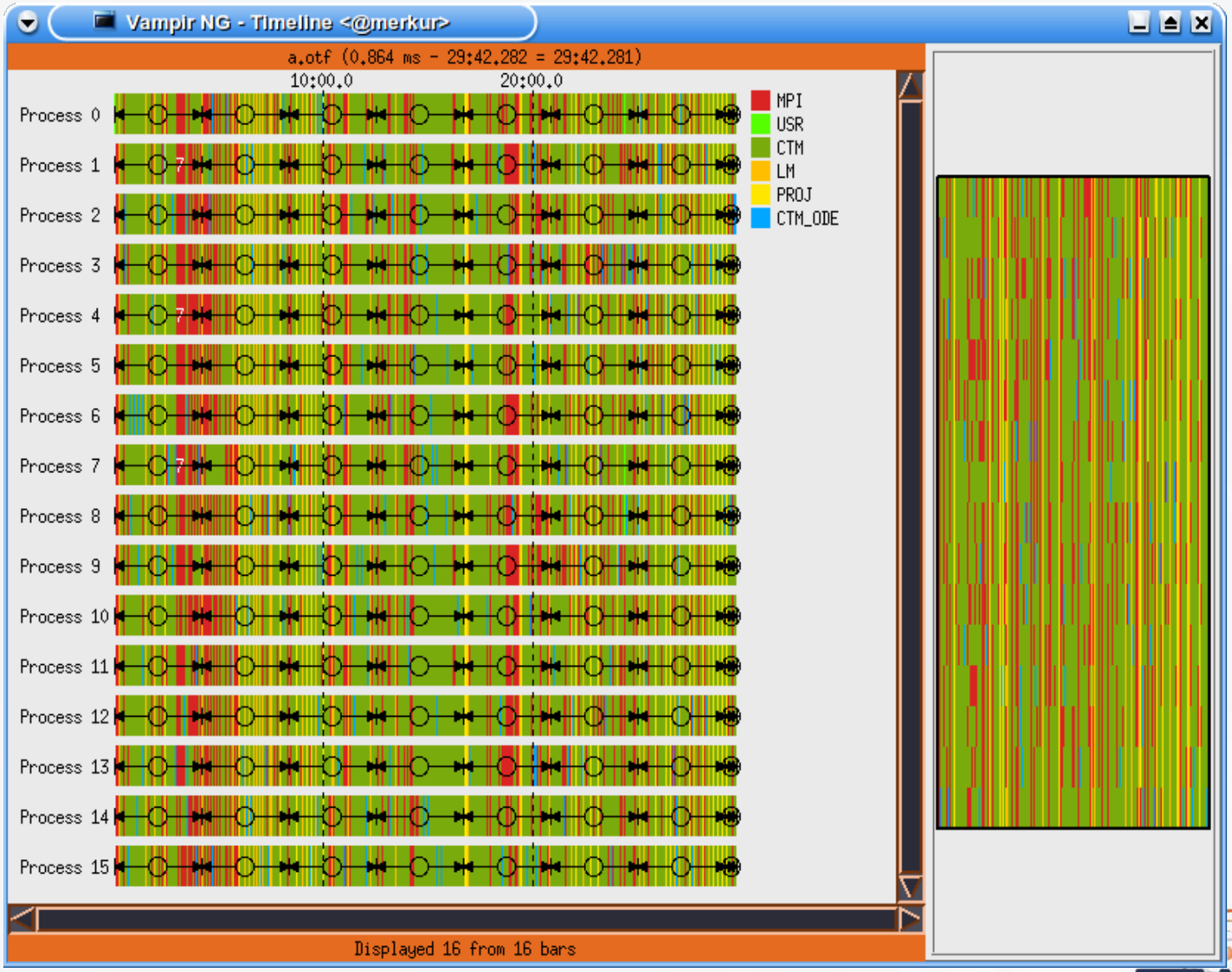
```
10020 P 2 ENTER 5
10095 P 2 ENTER 6
10120 P 2 ENTER 13
10300 P 2 RECV FROM 3 LEN 1024
10350 P 2 LEAVE 13
10450 P 2 LEAVE 6
10620 P 2 ENTER 9
10650 P 2 LEAVE 9
...
```

```
DEF TIMERRES 1000000000
DEF PROCESS 1 `Master`
DEF PROCESS 1 `Slave`
DEF FUNCTION 5 `main`
DEF FUNCTION 6 `foo`
DEF FUNCTION 9 `bar`
DEF FUNCTION 12 `MPI_Send`
DEF FUNCTION 13 `MPI_Recv`
```

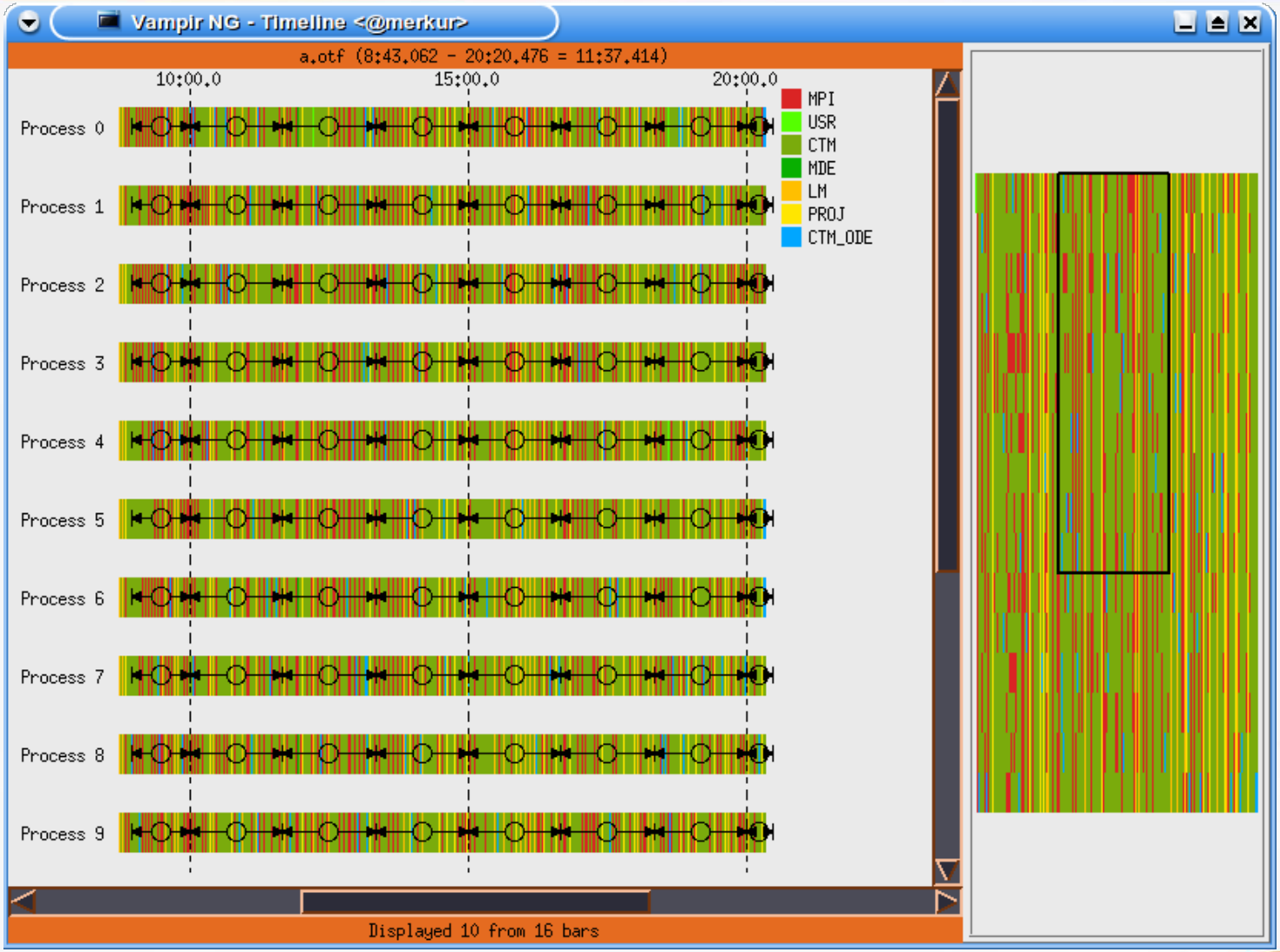
Common Event Types

- enter/leave of function/routine/region
  - time stamp, process/thread, function ID
- send/receive of P2P message (MPI)
  - time stamp, sender, receiver, length, tag, communicator
- collective communication (MPI)
  - time stamp, process, root, communicator, # bytes
- hardware performance counter values
  - time stamp, process, counter ID, value
- etc.

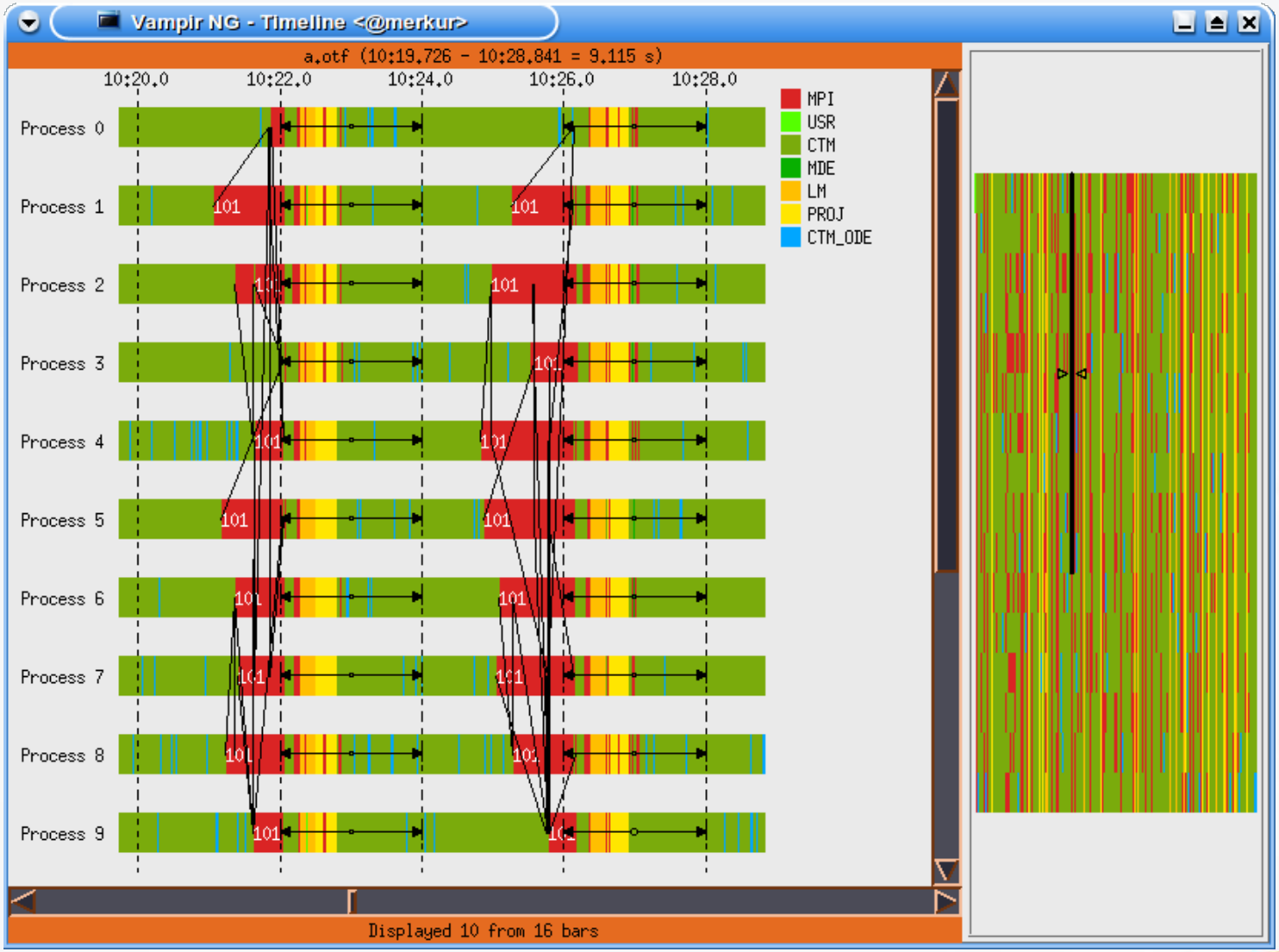
Vampir: Global Timeline



Vampir: Global Timeline

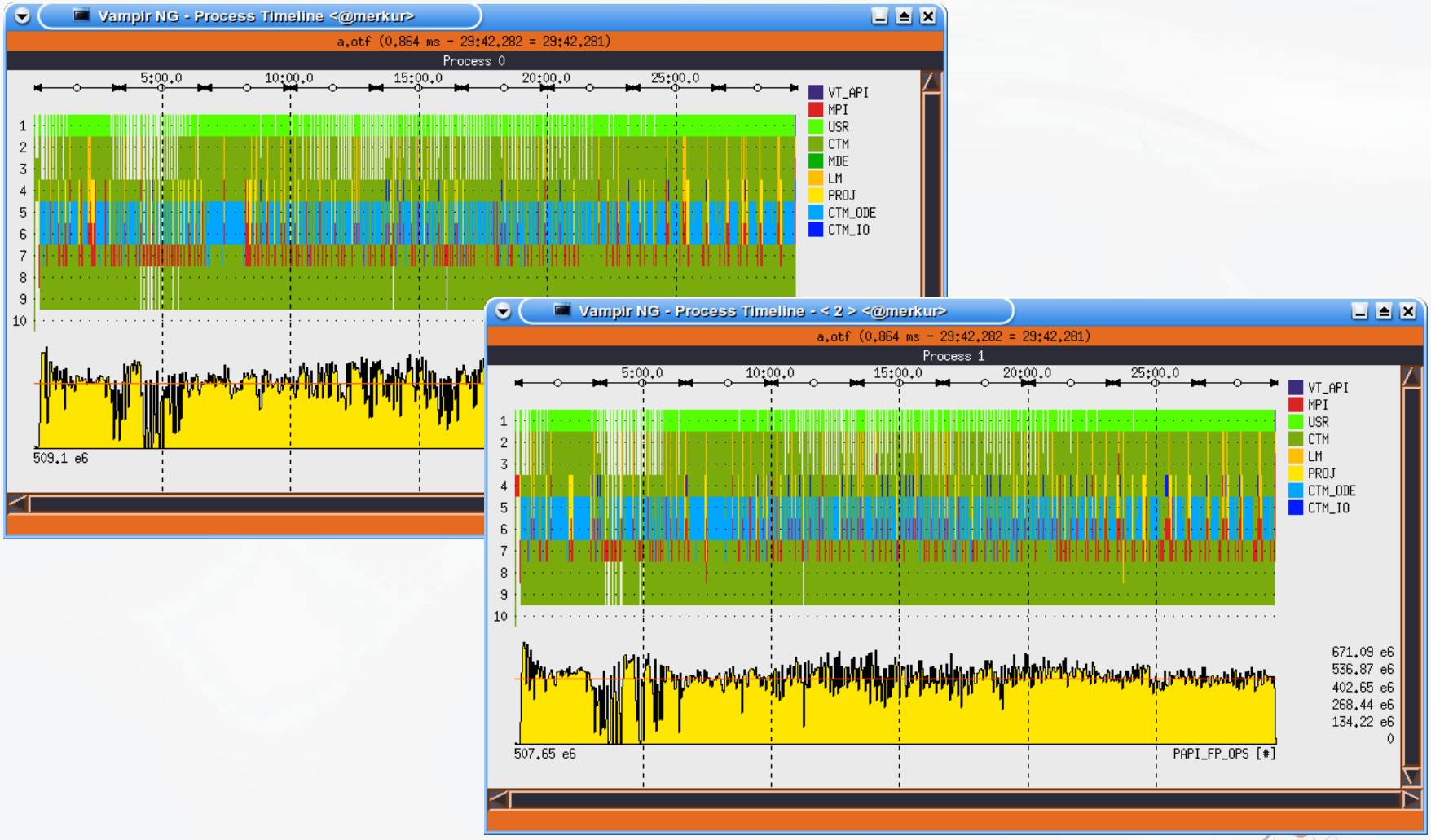


Vampir: Global Timeline

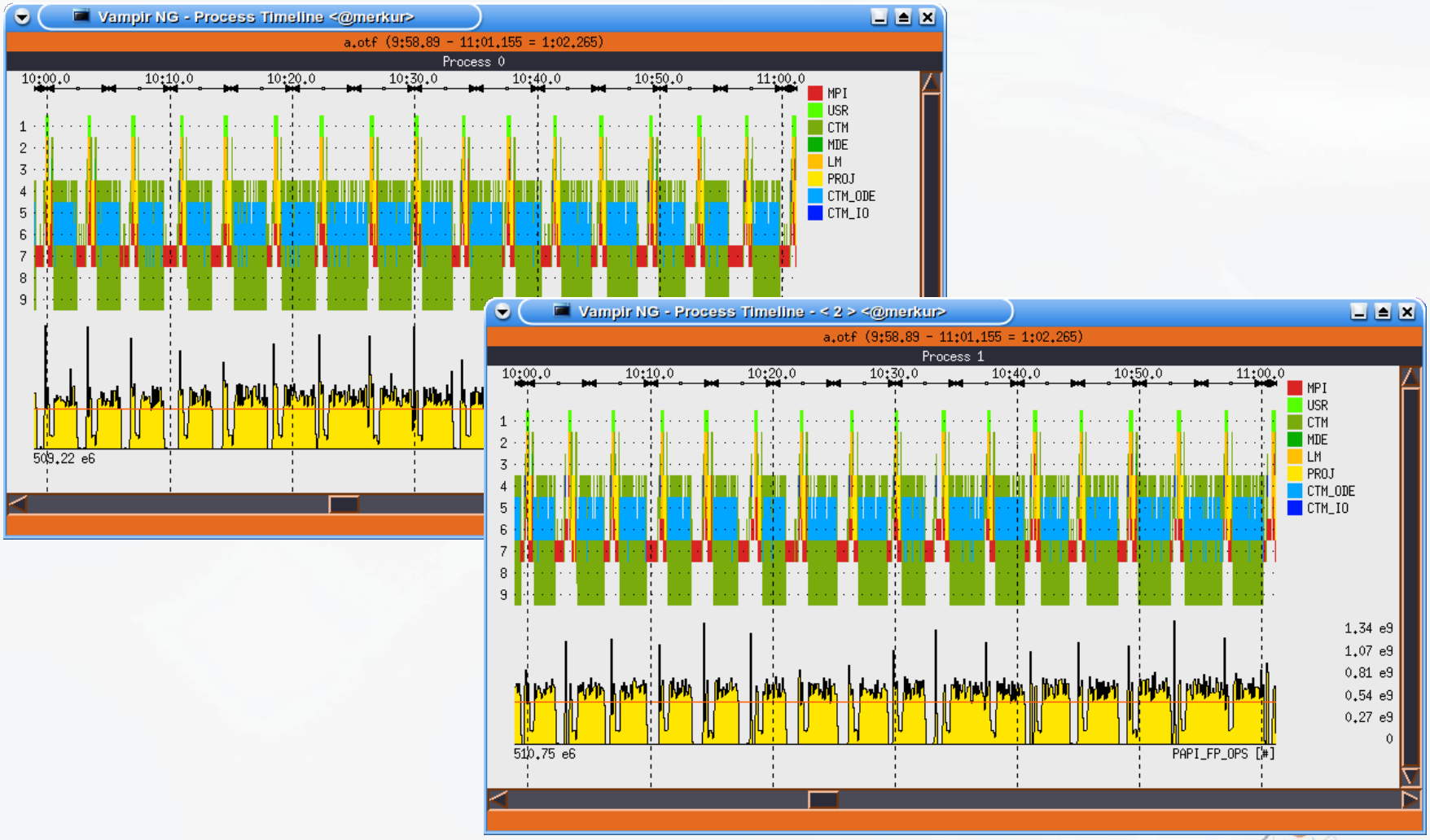


Vampir: Global Timeline

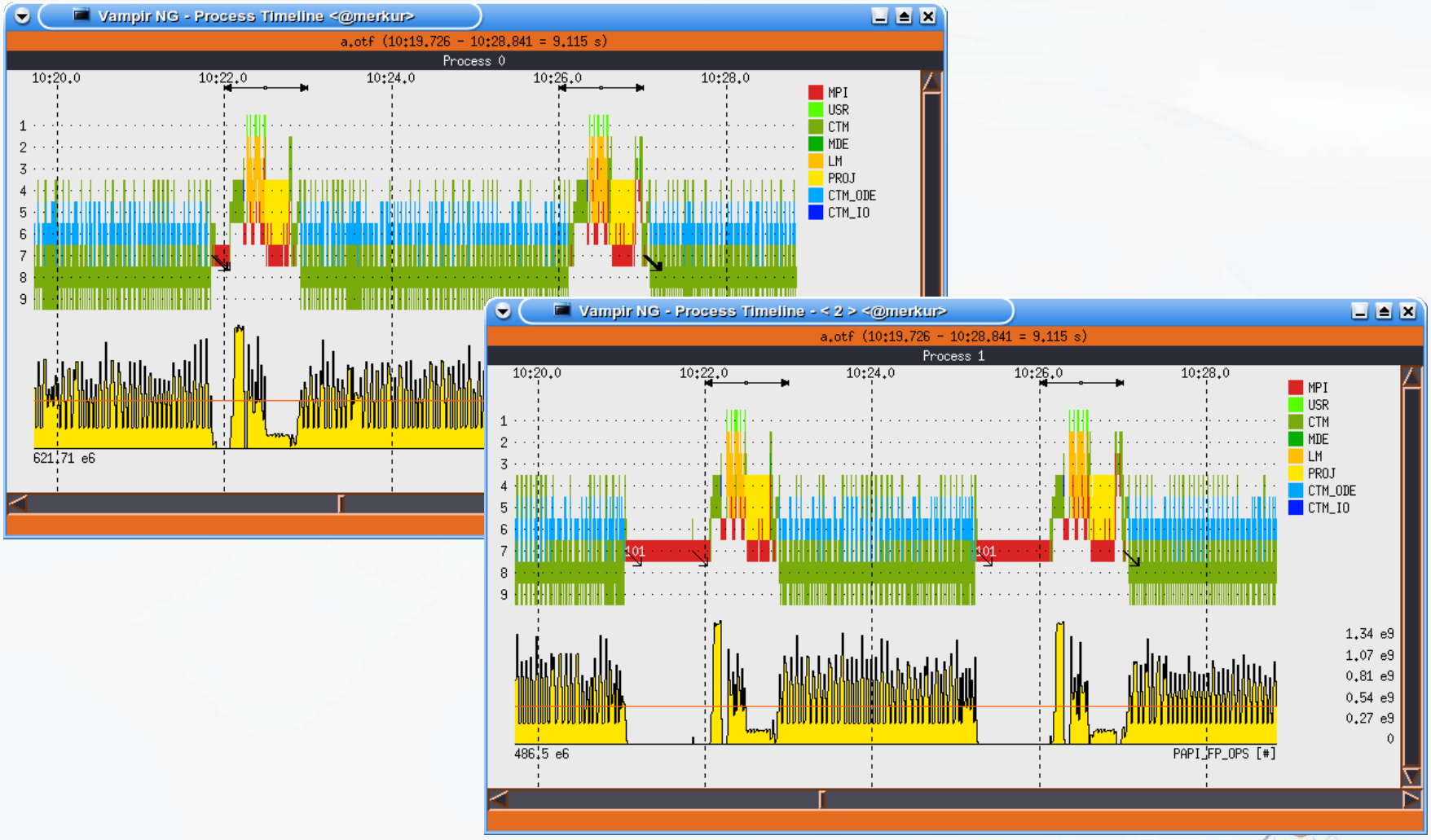
Vampir: Process Timeline



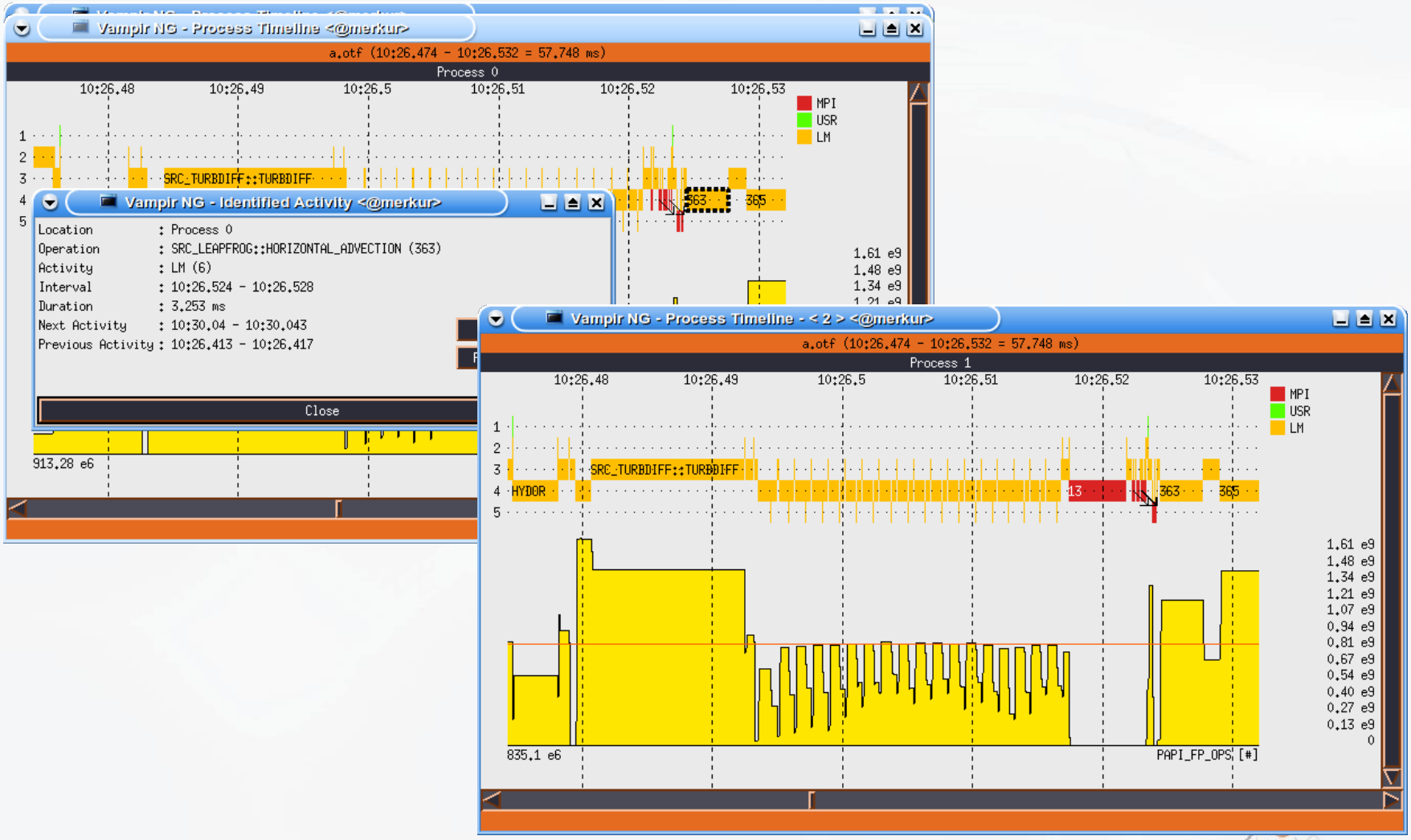
Vampir: Process Timeline



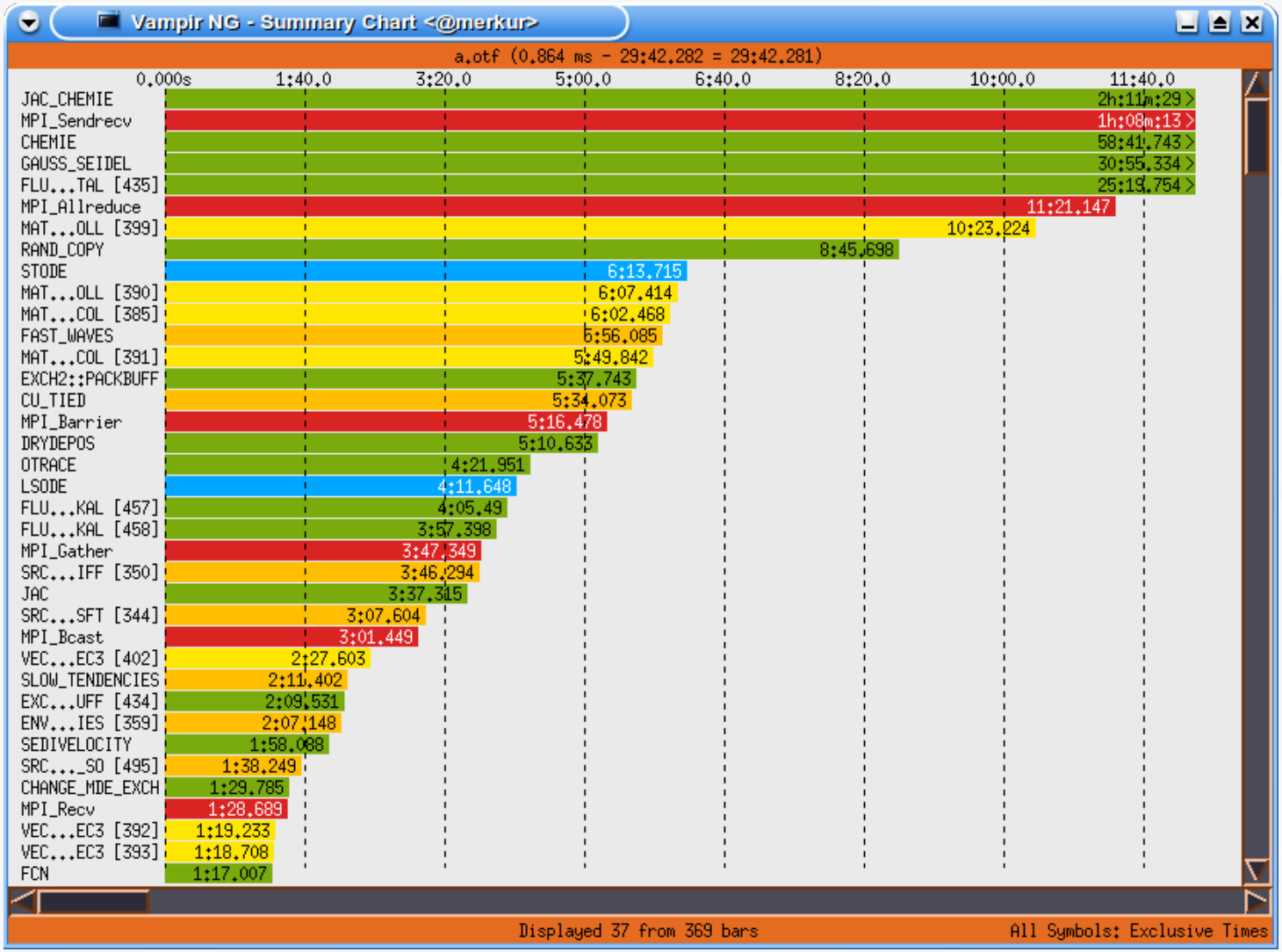
Vampir: Process Timeline



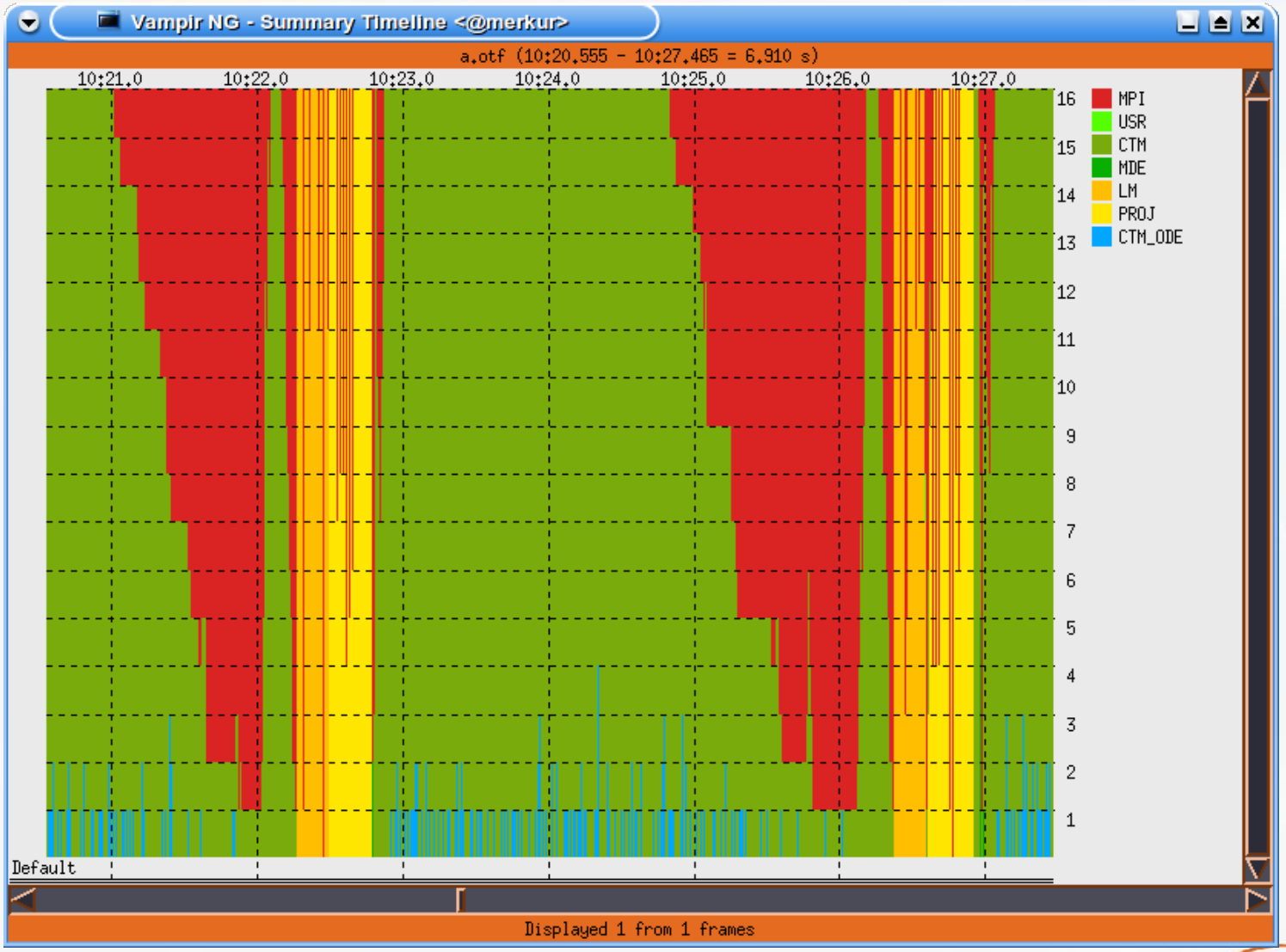
Vampir: Process Timeline



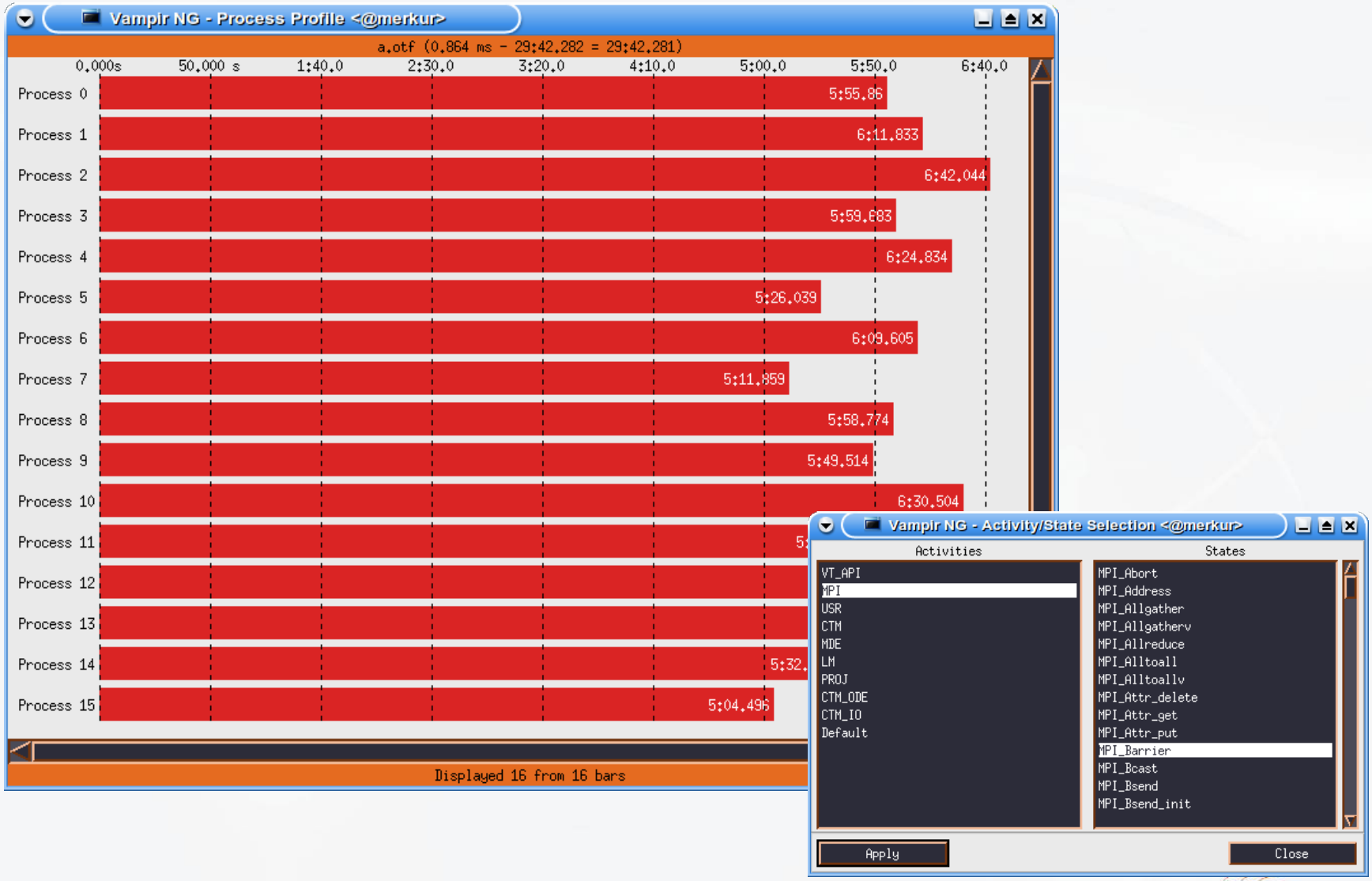
Vampir: Summary Chart



Vampir: Summary Timeline

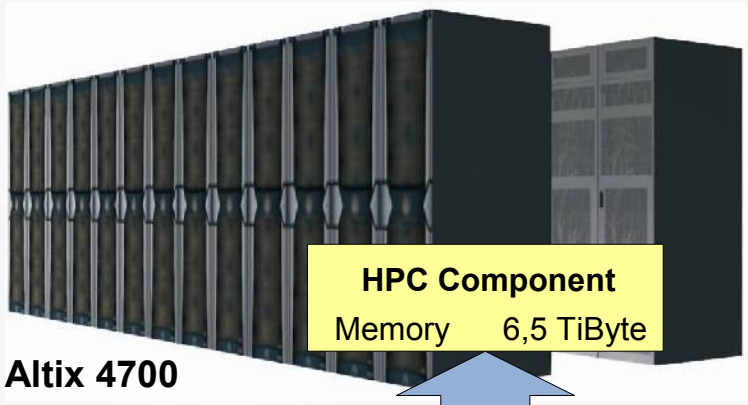


Vampir: Process Profile



ZIH – Focusing Resources around the Users' Data

Realization of the Concept for Data Intensive Computing



HPC Component
Memory 6,5 TiByte

SGI Altix 4700

1024 x Sockets mit Itanium2
Montecito Dual-Core CPUs
(1.6 GHz/9MB L3 Cache)

PC Farm

LNXI PC-Farm

AMD Opteron X85 Dual Core
Chip mit 2,6 GHz
384x Single CPU Nodes
232x Dual CPU Nodes
112x Quad CPU Nodes

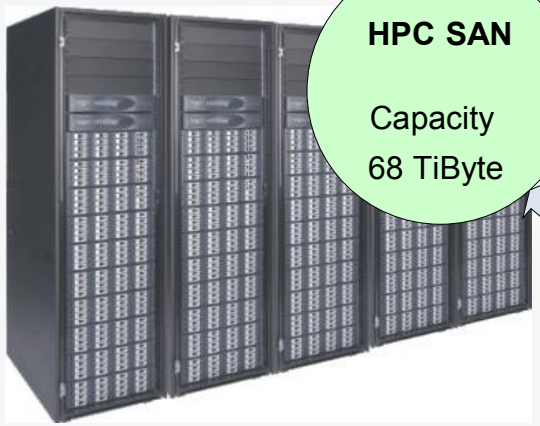
8 GiB/s

4 GiB/s

4 GiB/s

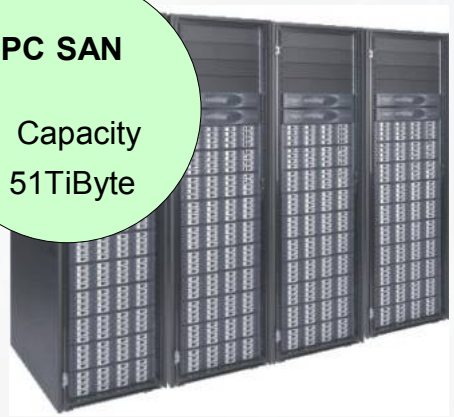
HPC SAN

Capacity
68 TiByte

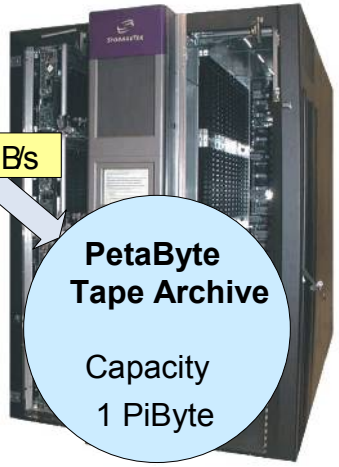


PC SAN

Capacity
51TiByte



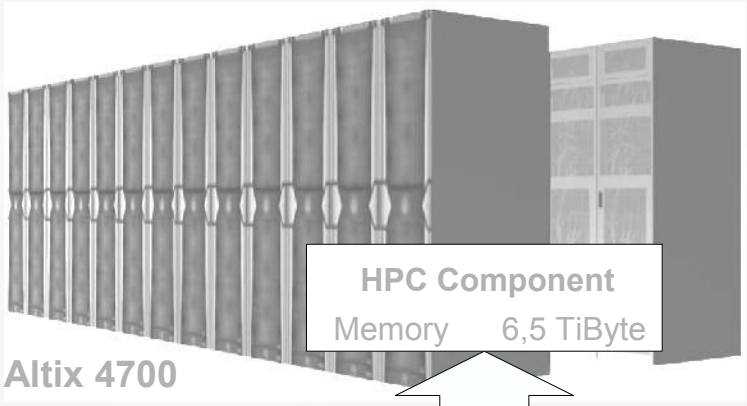
1,8 GiB/s



PetaByte Tape Archive

Capacity
1 PiByte

Lustre Relevant Components



HPC Component
Memory 6,5 TiByte

SGI Altix 4700

1024 x Sockets mit Itanium2
Montecito Dual-Core CPUs
(1.6 GHz/9MB L3 Cache)

PC Farm

LNXI PC-Farm

AMD Opteron X85 Dual Core
Chip mit 2,6 GHz
384x Single CPU Nodes
232x Dual CPU Nodes
112x Quad CPU Nodes

8 GiB/s

4 GiB/s

4 GiB/s

HPC SAN

Capacity
68 TiByte

PC SAN

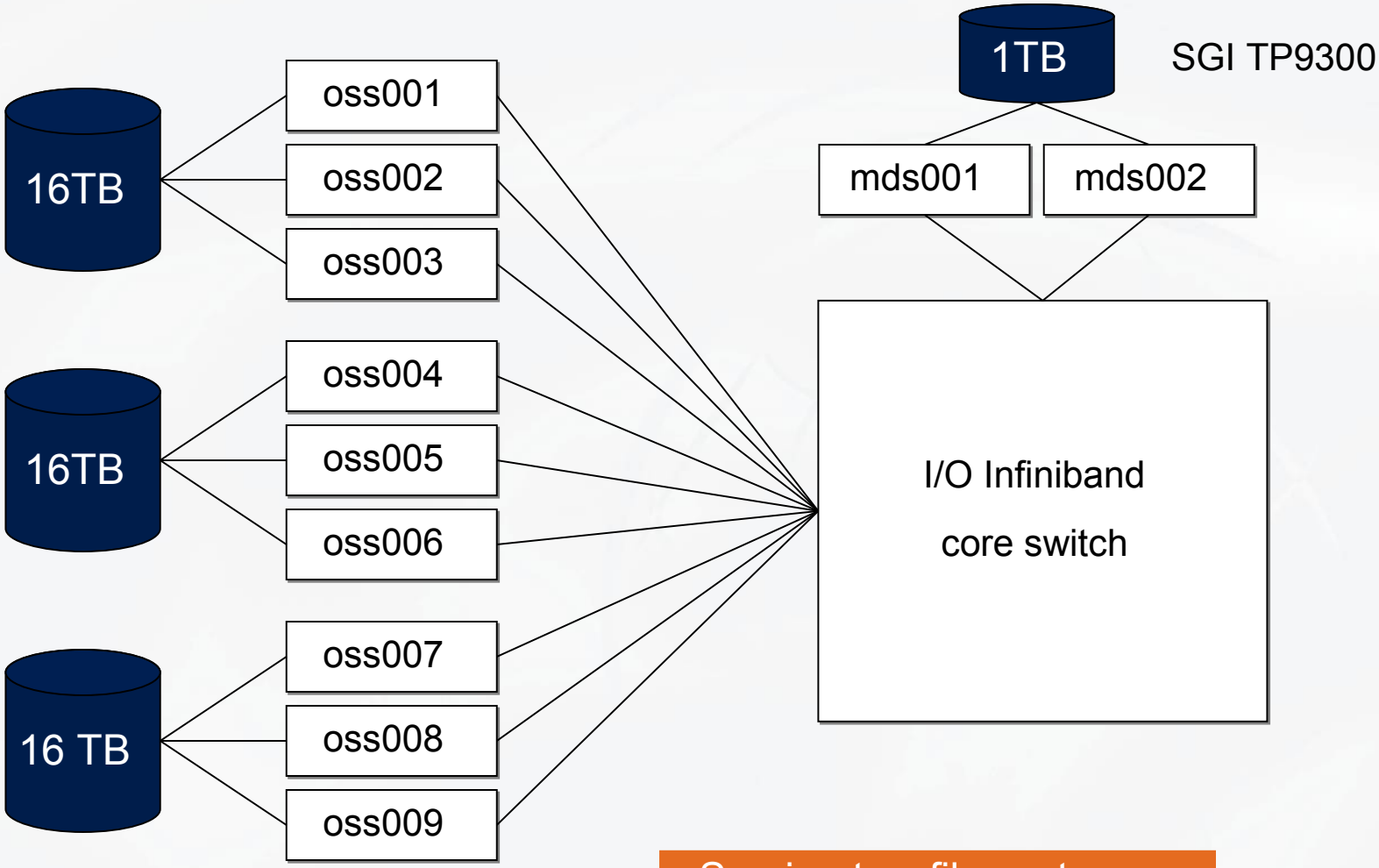
Capacity
51TiByte

1,8 GiB/s

PetaByte
Tape Archive

Capacity
1 PiByte

Lustre Setup



SGI InfiniteStorage 6700

Serving two file systems: /work and /fastfs

User profile: Image Analysis in Biology (MPI-CBG)

- Quantitative image analysis
  - Identify cells and cell components in images
  - Derive statistical information about the cell components
- Functional genomics assays
  - High resolution images of large number of cells
- Live cell imaging
  - Low resolution images in time series

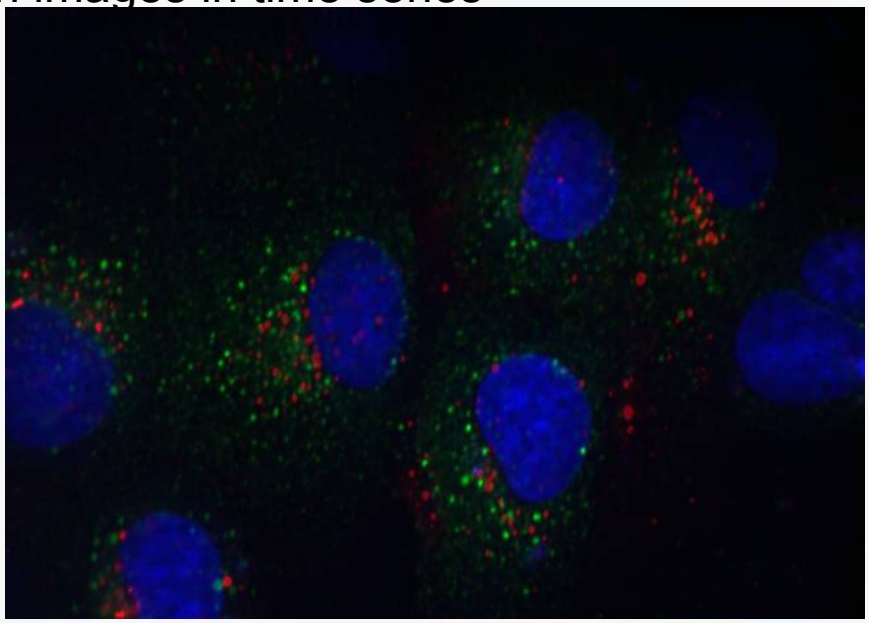


Image Analysis Workflow

1 Preparing Plates

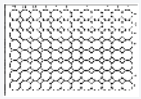
2 Acquiring Images

3 Prepro-cessing Images

4 Adjusting Parameters

5 Structure Identification

6 Calculating Statistics



Plates

Image Packages

Images

Parameters

Images

Structure Information

Results

384 well plate

384 image packages

~ 3000 image files

~18 GiB

1 parameter file

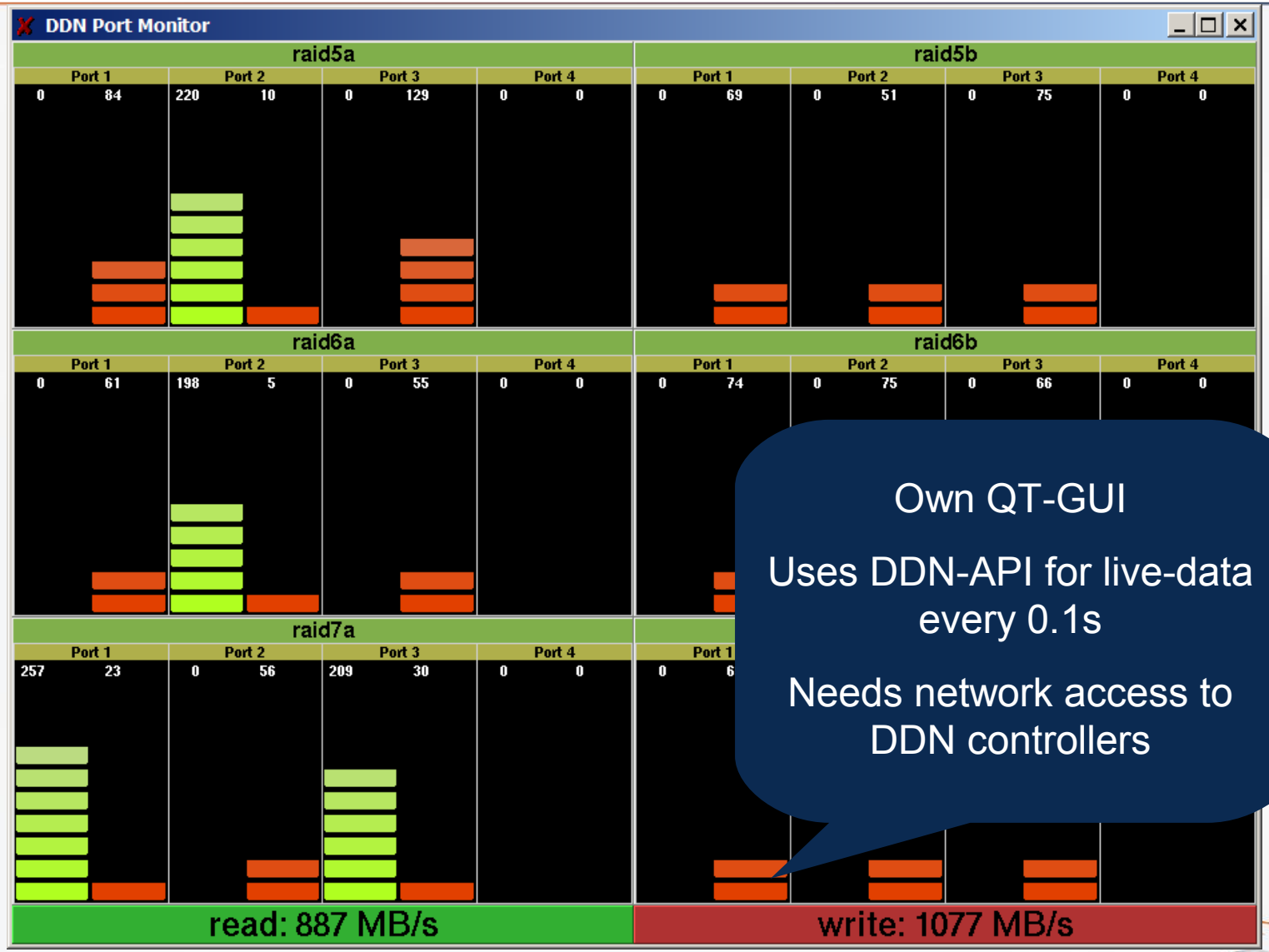
~ 3000 image files

~3 GiB

~ 50 global statistics

Tracing I/O – First results and further steps

Getting Data from the DDN-RAIDs

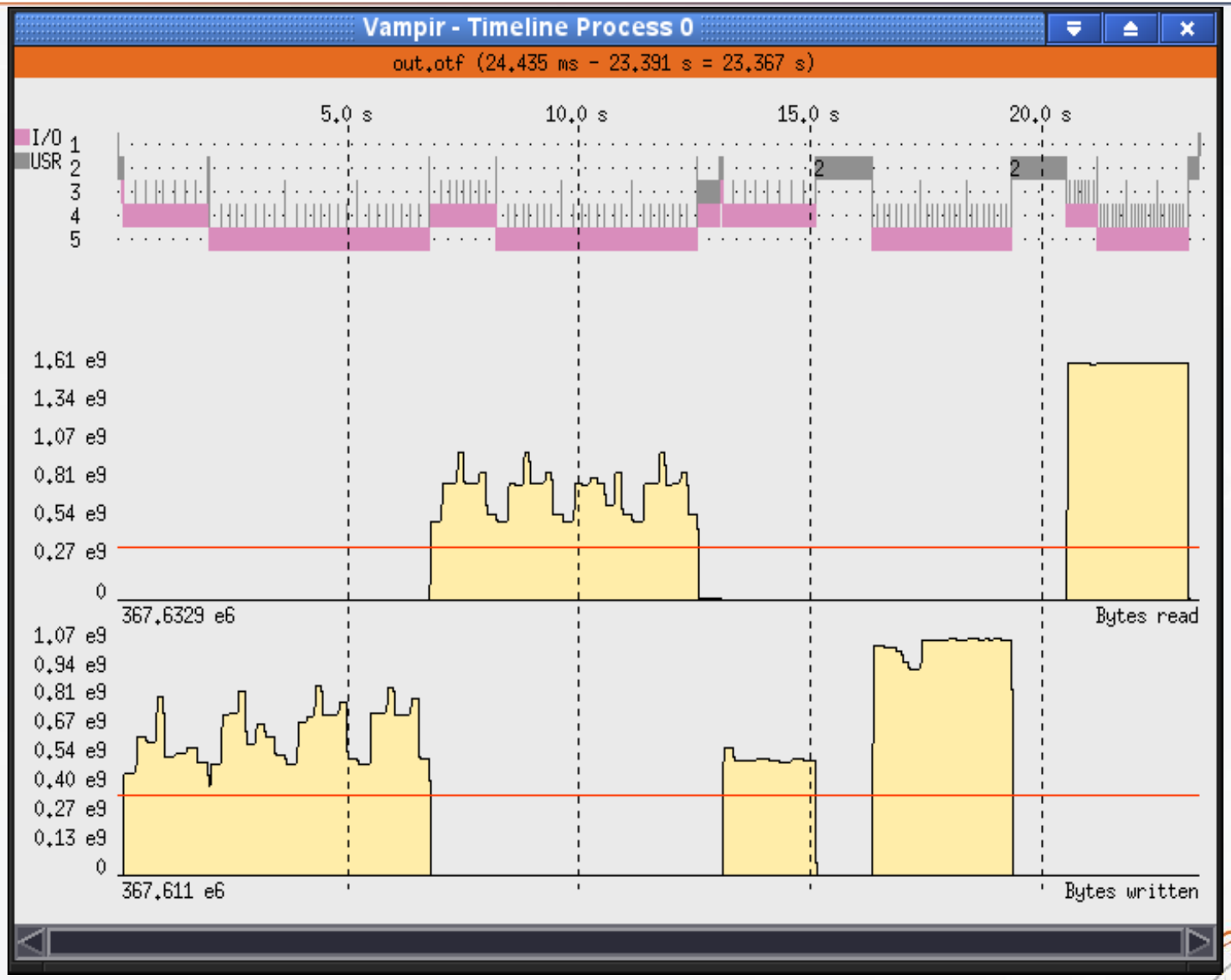


Getting Data from the Application

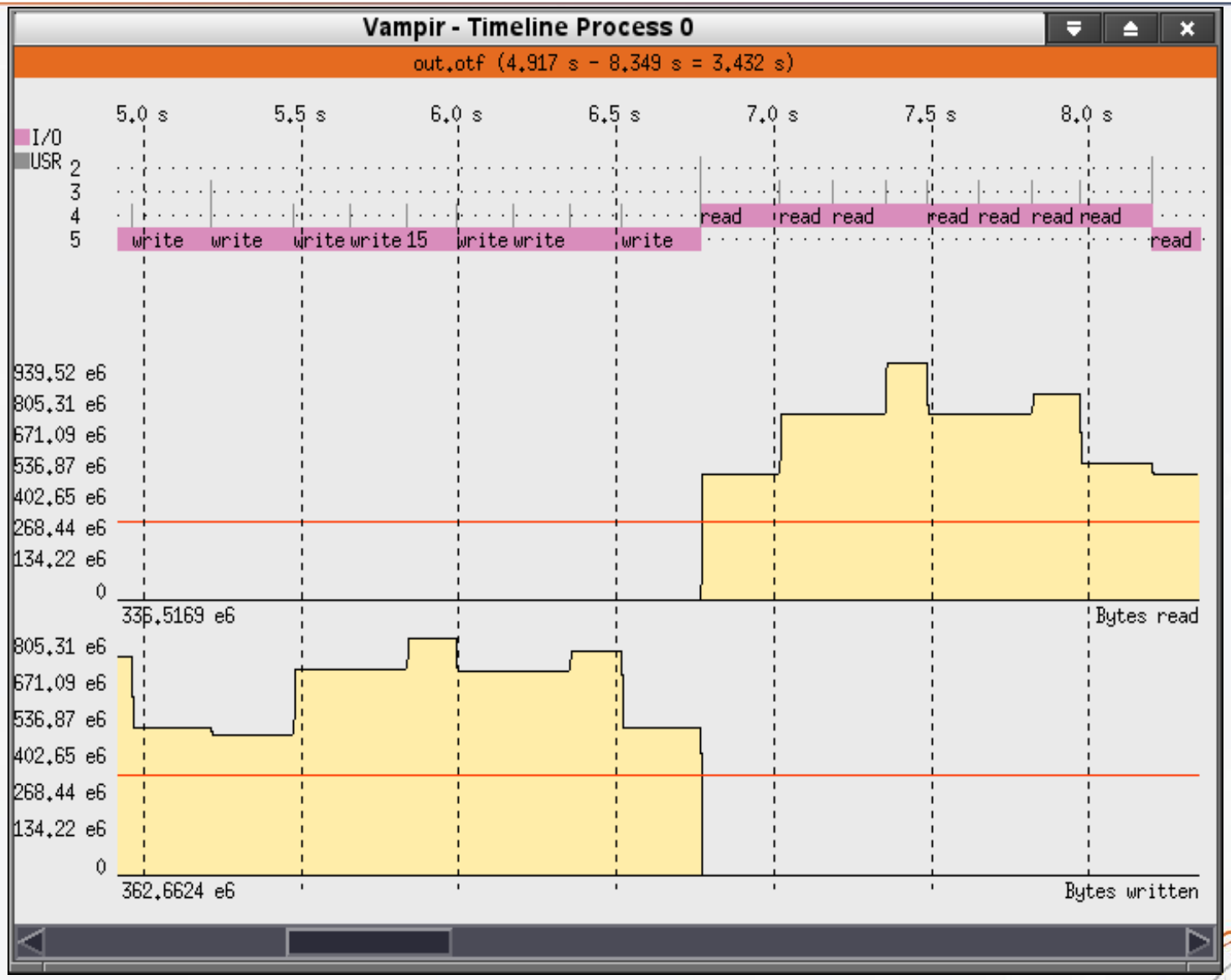
- Catching all I/O calls from the application
  - Adding them to the trace as performance counter data
  - Include filenames, offsets, and request sizes as OTF comments
  - Currently done with LD_PRELOAD library
  - Data needs to be merged with application OTF trace
  - Captured data: open (filename, filedescriptor), read/write (filedescriptor, size), close/dup (filedescriptor), seek (filedescriptor, position)

- Tracing I/O requests within the kernel
  - To follow the path of the request to the devices

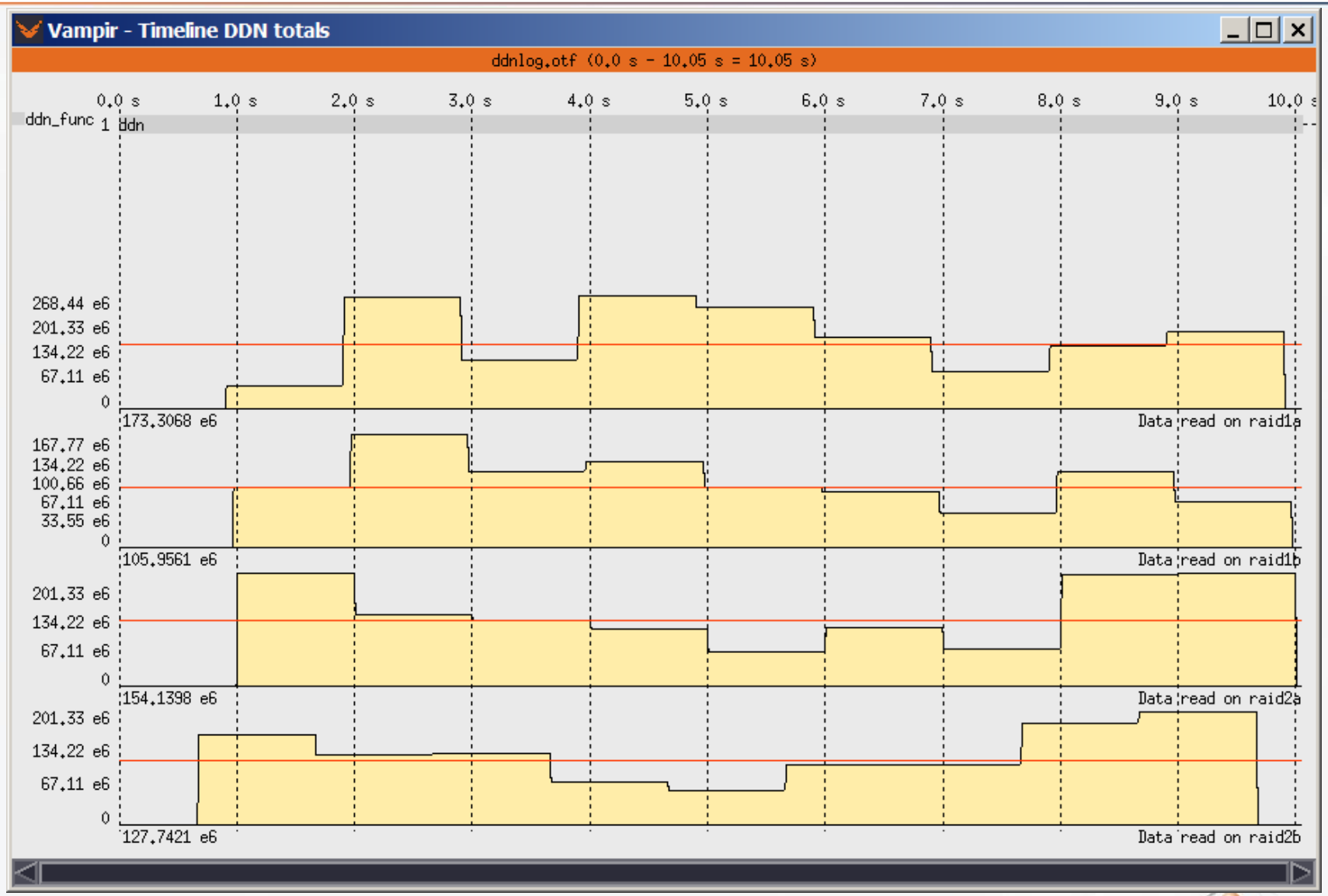
Vampir I/O Stats per Process (work in progress)



Vampir I/O Stats per Process (work in progress)



Including DDN Statistics (work in progress)



Restraints and Requirements

- I/O Tracing cannot modify kernel (site requirements)
- DDN data is sampling data
- DDN data is global data
- Kernel data is global data

Therefore:

- Need existing FS (e.g. Lustre) to include on demand trace hooks

- Device mappings need to be known as well

Conclusions and Further Work

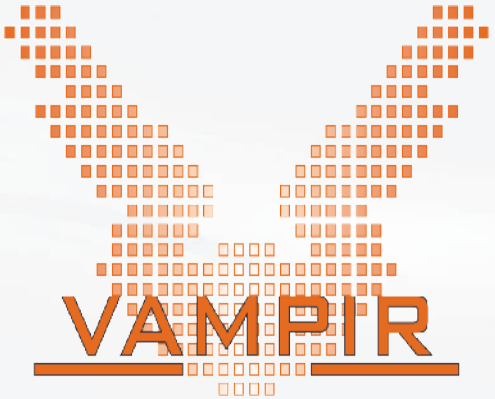
- Data intensive computing produces new bottlenecks
- New HPC user groups (e.g. biologists) produce new problems

- Vampir platform is/was extended to help indentify I/O bottlenecks

- DDN data needs to be correlated with application trace
- Should work with any device and any FS (Lustre very good starting point)
- Could include file object information
- Need to follow I/O request form application to device

Thanks and Contacts

- Vampir/OTF Introduction
  - Andreas Knüpfer (andreas.knuepfer@tu-dresden.de)
  - Holger Brunst (holger.brunst@tu-dresden.de)



- ZIH System Setup
  - Dr. Matthias Müller (matthias.mueller@tu-dresden.de)

- I/O Tracing Facilities
  - Michael Kluge (michael.kluge@tu-dresden.de)
  - Holger Mickler (holger.mickler@zih.tu-dresden.de)

Visit us at `www.tu-dresden.de/zih` and `www.vampir.eu`