From Network to Application: Understanding Your Distributed System with Trace Compass

Tracing Summit,
Dusseldorf, Germany,
October 13, 2014

TRACE
COMPASS

Geneviève Bastien
Research Associate
Dorsal Laboraty
École Polytechnique de Montréal

Trace Compass

- Formerly known as TMF, the Linuxtools LTTng Eclipse plugin.
- Trace visualization tool
  - Standalone Rich Client Platform (RCP) application.
  - Also available as Eclipse plugins.
- Extendable framework
  - Add support for new trace types
  - Build trace analysis
  - With data-driven analysis, it's now easier than ever

Trace Compass

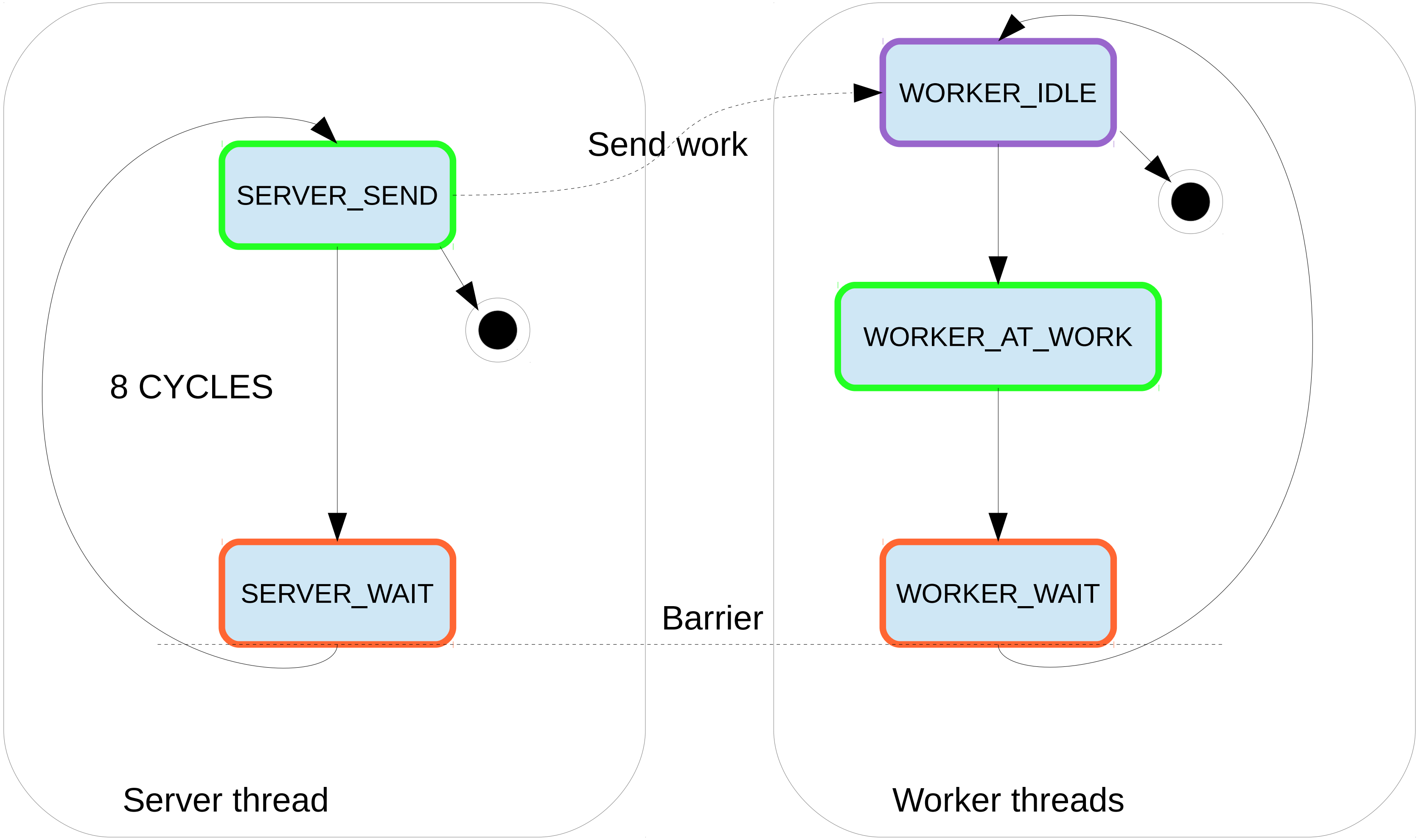
- Now goes beyond Linux-only
  - Trace types:
    - LTTng / CTF
    - BTF
    - Custom text and XML
    - GDB
    - PCAP
    - Windows! (prototype with CTF converter)
  - Analysis:
    - LTTng Kernel: Control Flow View, Resources View, CPU usage
    - LTTng UST: Memory Usage (liblttng-libc-wrapper), CallStack View (-g -finstrument-functions)
    - PCAP: **Network Stream lists**
    - META: **Data-driven analysis, Network trace synchronization, Virtual Machine analysis,** Critical path analysis

Demo

- 1 distributed application : 3 use cases
  - Local only (show data-driven analysis)
  - On 2 machines on the network (show network analysis)
  - On 2 virtual machines on the same host (show virtual machine analysis)

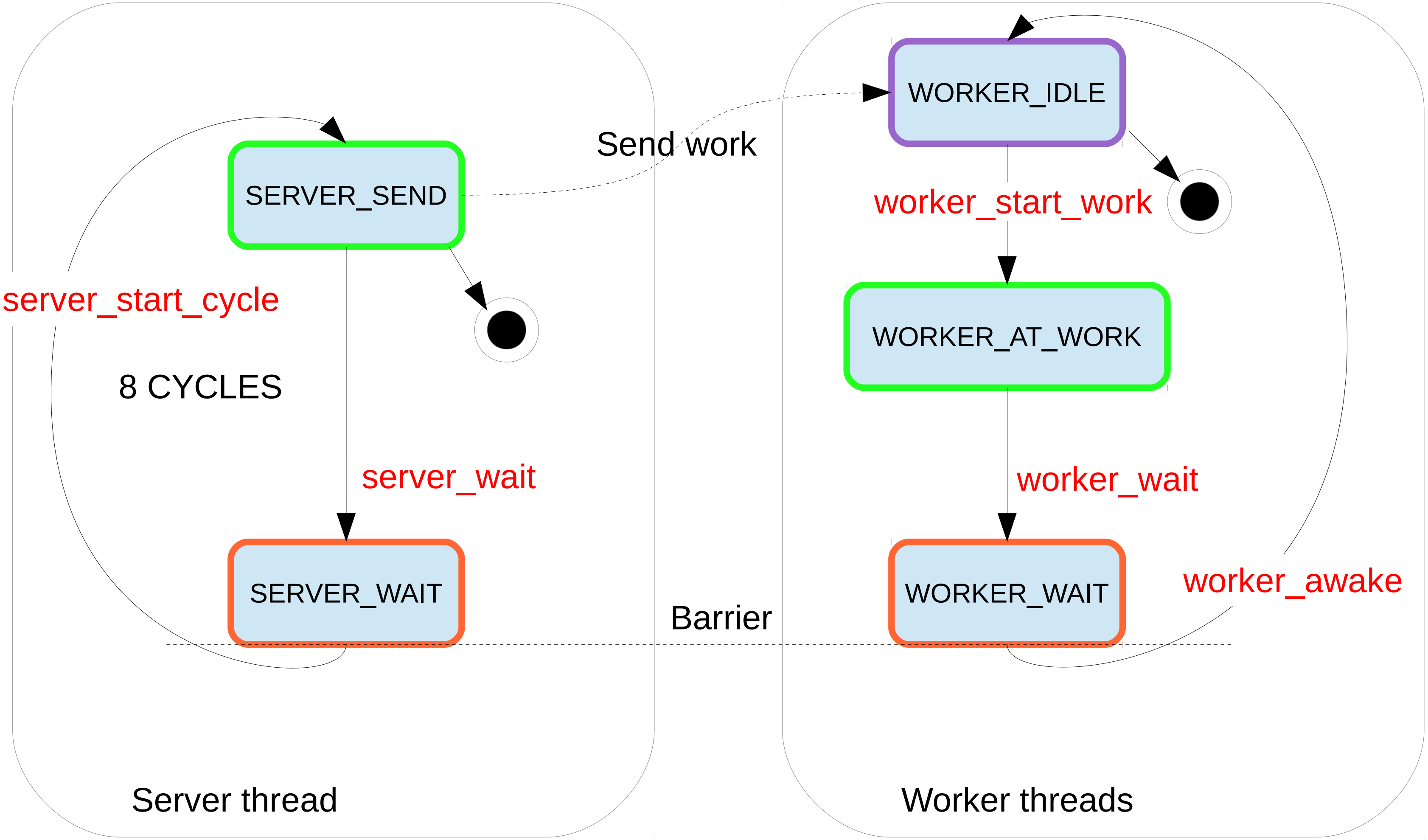
Demo Application

- MPI application: 5 worker threads + 1 server sending imbalanced workload to workers.

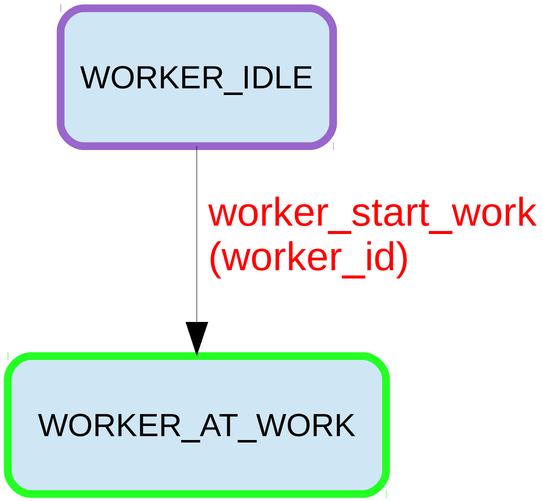


Data-Driven Analysis

- MPI application: 5 worker threads + 1 server sending imbalanced workload to workers.



Data-Driven Analysis



State change:

Worker/<worker_id> = WORKER_AT_WORK

```
...
<stateProvider id="mpi.imbalance.sp">
    ...
    <definedValue name="WORKER_AT_WORK" value="2" />
    ...
    <eventHandler eventName="mpi_imbalance:worker_start_work">
        <stateChange>
            <stateAttribute type="constant" value="Worker" />
            <stateAttribute type="eventField" value="worker_id" />
            <stateValue type="int" value="$WORKER_AT_WORK" />
        </stateChange>
    </eventHandler>
</stateProvider>
...
```

Data-Driven Analysis

- Visualization of the thread's states: time graph views or XY views

```
<timeGraphView id="mpi.imbalance.view.timegraph">

    <definedValue name="WORKER_AT_WORK" value="2" color="#66FF33" />
    <definedValue name="WORKER_WAIT" value="3" color="#FF3300" />
    <definedValue name="WORKER_IDLE" value="4" color="#CC66FF" />

    <entry path="Worker/*">
        <display type="self" />
    </entry>
</timeGraphView>
```

Future work

- Data-driven analysis:
  - Define visually, with state diagrams
  - Smart filters and user-defined actions on those filters
  - And much much more!
- GPU traces and analysis
- Compare traces from different executions, for CPU/Memory usage, etc.
- Live tracing
- Improve performances with large experiments

Questions

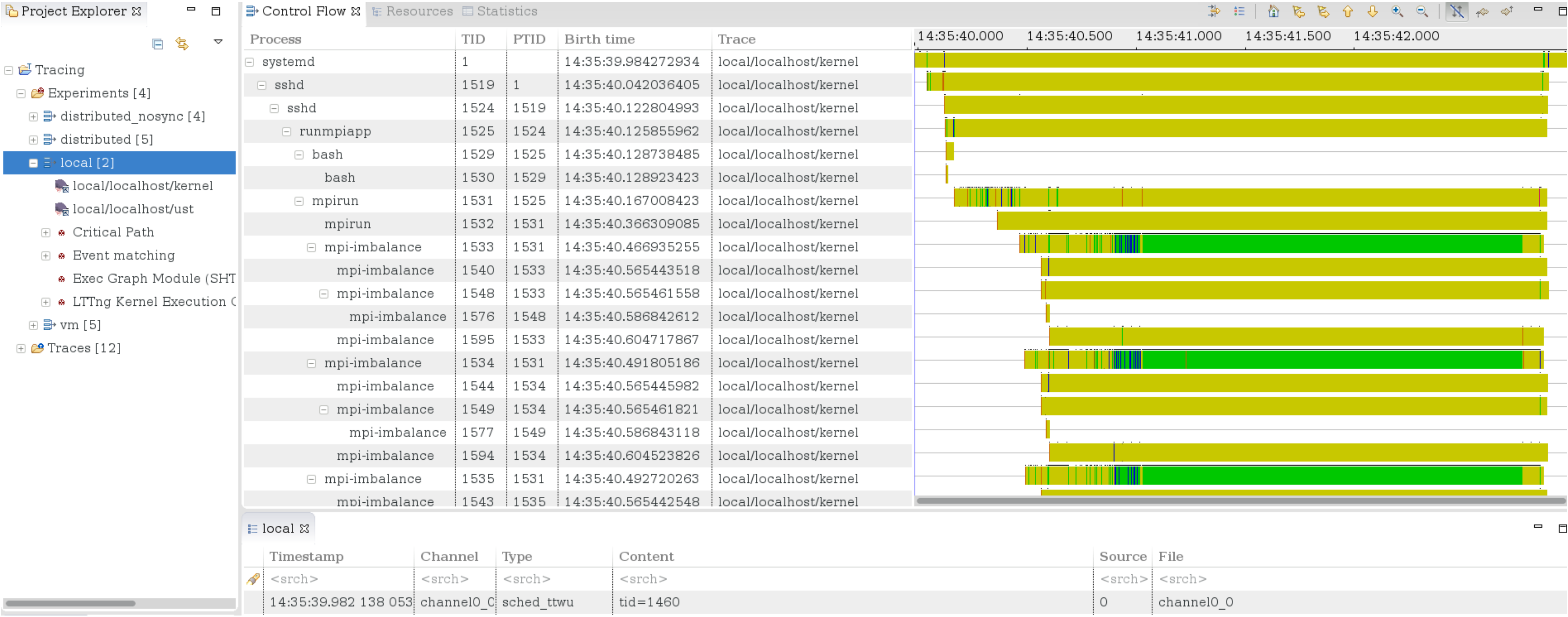
Resources

- Home Page: http://www.eclipse.org/tracecompass
- Mailing List: https://dev.eclipse.org/mailman/listinfo/tracecompass-dev
- Trace Compass standalone application used in this presentation: http://secretaire.dorsal.polymtl.ca/~gbastien/TracingRCP/DorsalExperimental/
- Sources:
  - Master (coming soon): http://git.eclipse.org/c/tracecompass/org.eclipse.tracecompass.git
  - TMF in Linuxtools: (under the lttng folder) git://git.eclipse.org/gitroot/linuxtools/org.eclipse.linuxtools.git
  - Experimental: branch dorsal_experimental http://git.dorsal.polymtl.ca/~gbastien?p=linuxtools-tmf.git;a=summary
- Used in this demo:
  - Sample MPI traces and XML analysis: http://secretaire.dorsal.polymtl.ca/~gbastien/tracingSummit2014/
  - MPI-imbalance source code: branch cluster (folder cluster/mpi-imbalance) http://git.dorsal.polymtl.ca/~gbastien?p=workload-kit.git;a=summary
- IRC: #lttng on oftc
- More doc and links: http://lttng.org/eclipse

Annexes

(Screenshots in case Eclipse refuses to cooperate)

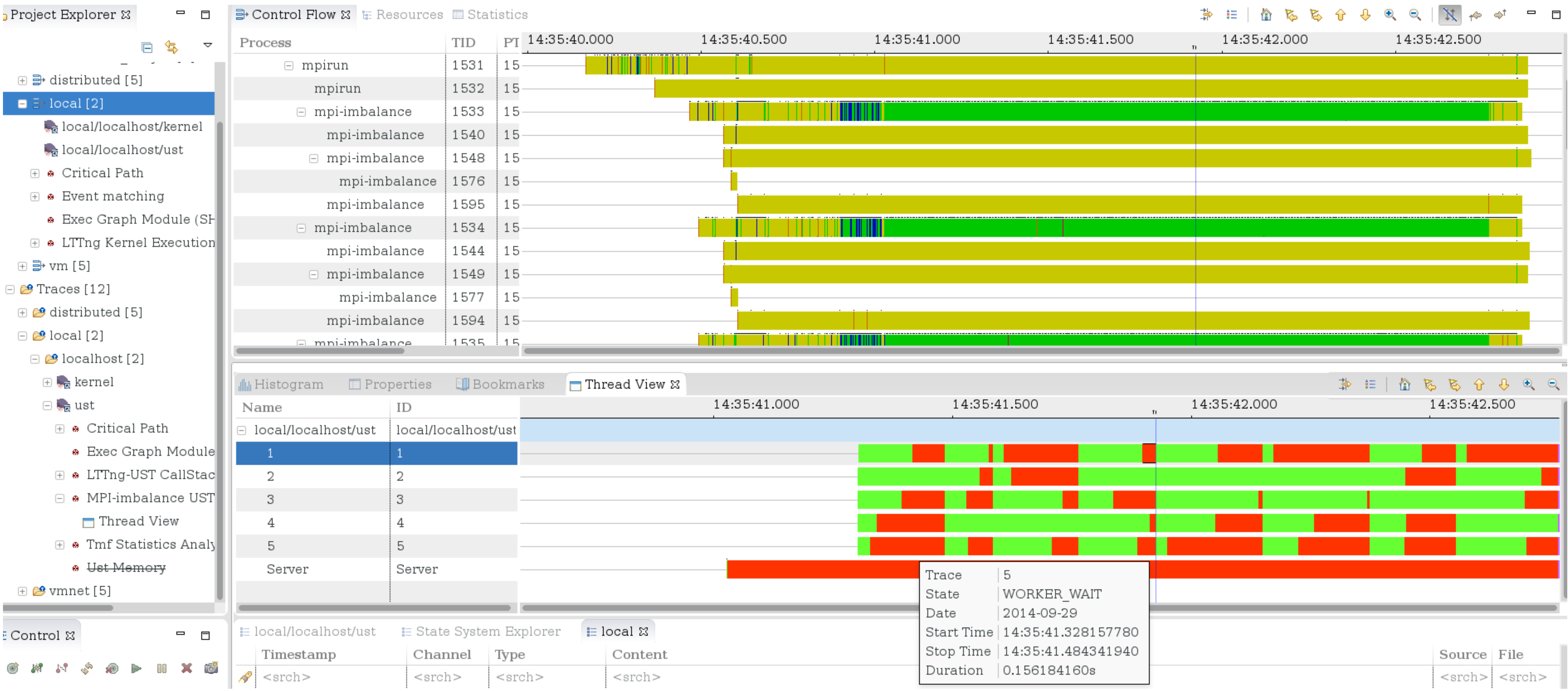
Experiment 1: Local: Control Flow View



Process	TID	PTID	Birth time	Trace
systemd	1		14:35:39.984272934	local/localhost/kernel
sshd	1519	1	14:35:40.042036405	local/localhost/kernel
sshd	1524	1519	14:35:40.122804993	local/localhost/kernel
runmpiapp	1525	1524	14:35:40.125855962	local/localhost/kernel
bash	1529	1525	14:35:40.128738485	local/localhost/kernel
bash	1530	1529	14:35:40.128923423	local/localhost/kernel
mpirun	1531	1525	14:35:40.167008423	local/localhost/kernel
mpirun	1532	1531	14:35:40.366309085	local/localhost/kernel
mpi-imbalance	1533	1531	14:35:40.466935255	local/localhost/kernel
mpi-imbalance	1540	1533	14:35:40.565443518	local/localhost/kernel
mpi-imbalance	1548	1533	14:35:40.565461558	local/localhost/kernel
mpi-imbalance	1576	1548	14:35:40.586842612	local/localhost/kernel
mpi-imbalance	1595	1533	14:35:40.604717867	local/localhost/kernel
mpi-imbalance	1534	1531	14:35:40.491805186	local/localhost/kernel
mpi-imbalance	1544	1534	14:35:40.565445982	local/localhost/kernel
mpi-imbalance	1549	1534	14:35:40.565461821	local/localhost/kernel
mpi-imbalance	1577	1549	14:35:40.586843118	local/localhost/kernel
mpi-imbalance	1594	1534	14:35:40.604523826	local/localhost/kernel
mpi-imbalance	1535	1531	14:35:40.492720263	local/localhost/kernel
mpi-imbalance	1543	1535	14:35:40.565442548	local/localhost/kernel

Timestamp	Channel	Type	Content	Source	File
<srch>	<srch>	<srch>	<srch>	<srch>	<srch>
14:35:39.982 138 053	channel0_C	sched_ttwu	tid=1460	0	channel0_0

Experiment 1: Local: Thread View



- distributed [5]
- local [2]
  - local/localhost/kernel
  - local/localhost/ust
  - Critical Path
  - Event matching
  - Exec Graph Module (SH
  - LTTng Kernel Execution
- vm [5]
- Traces [12]
  - distributed [5]
  - local [2]
    - localhost [2]
      - kernel
      - ust
        - Critical Path
        - Exec Graph Module
        - LTTng-UST CallStac
        - MPI-imbalance UST
          - Thread View
        - Tmf Statistics Analy
        - ~~Ust Memory~~
  - vmnet [5]

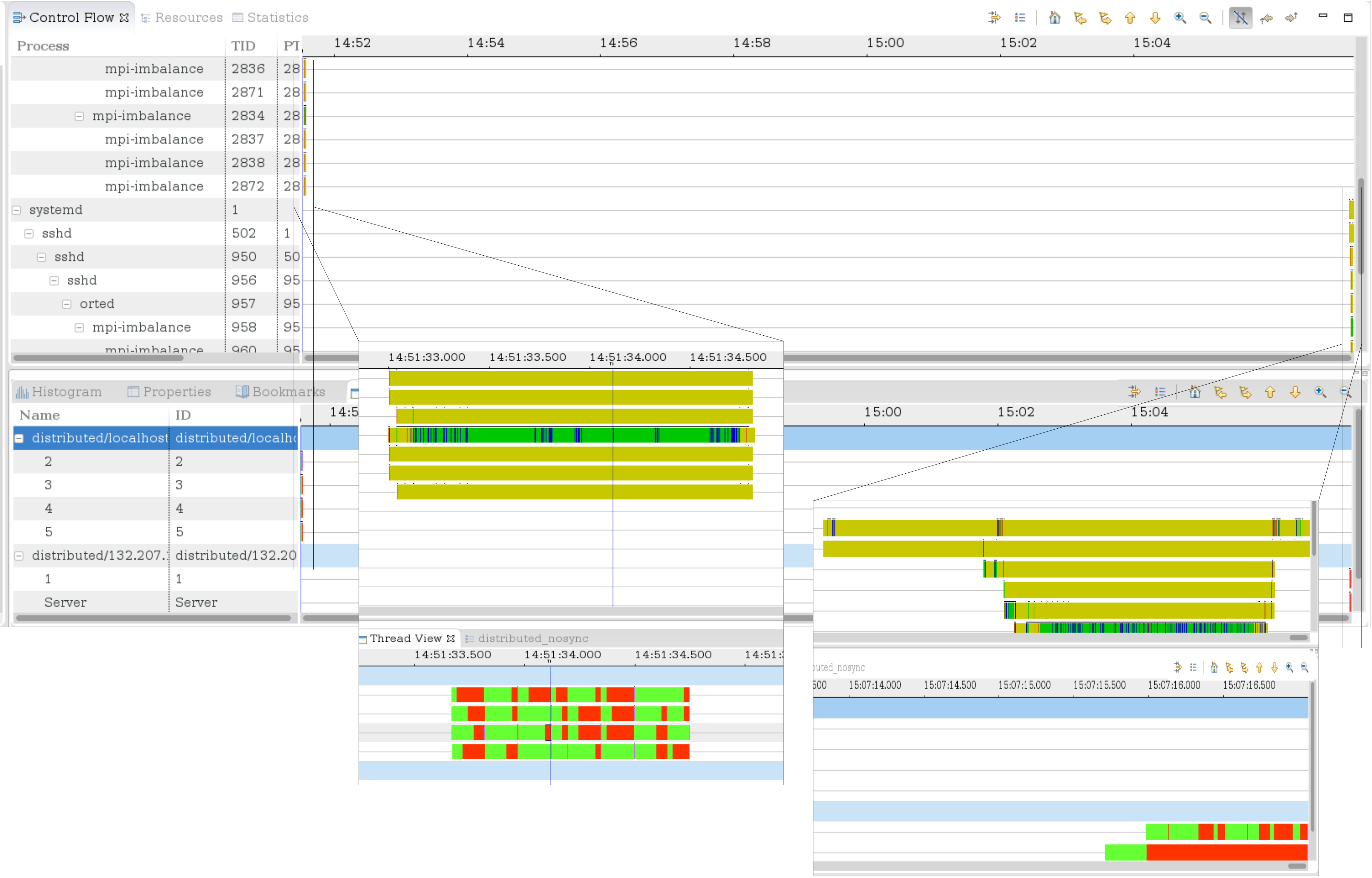
Process	TID	PT
mpirun	1531	15
mpirun	1532	15
mpi-imbalance	1533	15
mpi-imbalance	1540	15
mpi-imbalance	1548	15
mpi-imbalance	1576	15
mpi-imbalance	1595	15
mpi-imbalance	1534	15
mpi-imbalance	1544	15
mpi-imbalance	1549	15
mpi-imbalance	1577	15
mpi-imbalance	1594	15
mpi-imbalance	1535	15

Name	ID
local/localhost/ust	local/localhost/ust
1	1
2	2
3	3
4	4
5	5
Server	Server

Trace	5
State	WORKER_WAIT
Date	2014-09-29
Start Time	14:35:41.328157780
Stop Time	14:35:41.484341940
Duration	0.156184160s

Timestamp	Channel	Type	Content	Source	File
<srch>	<srch>	<srch>	<srch>	<srch>	<srch>

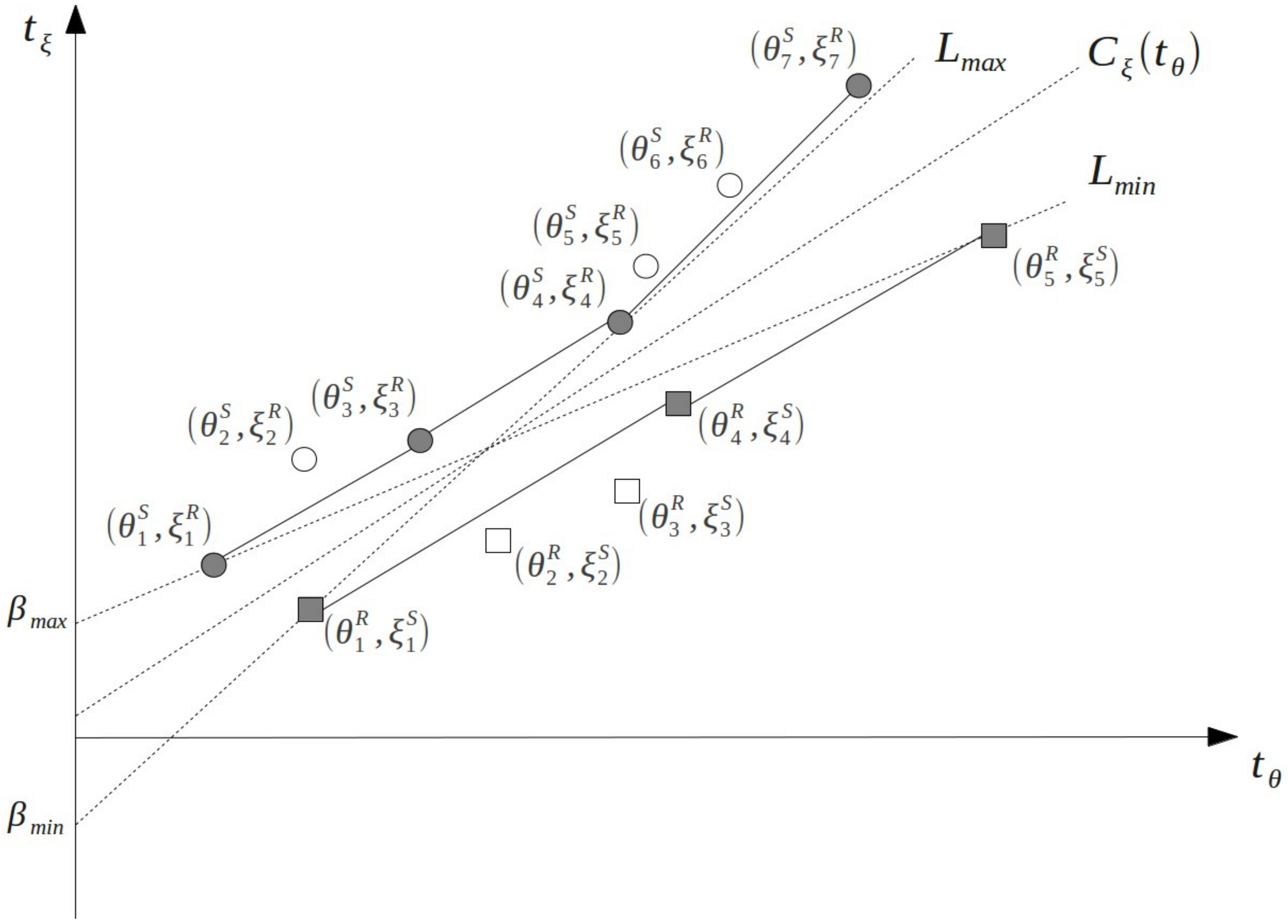
Experiment 2: Distributed Network: Control Flow View and Worker View



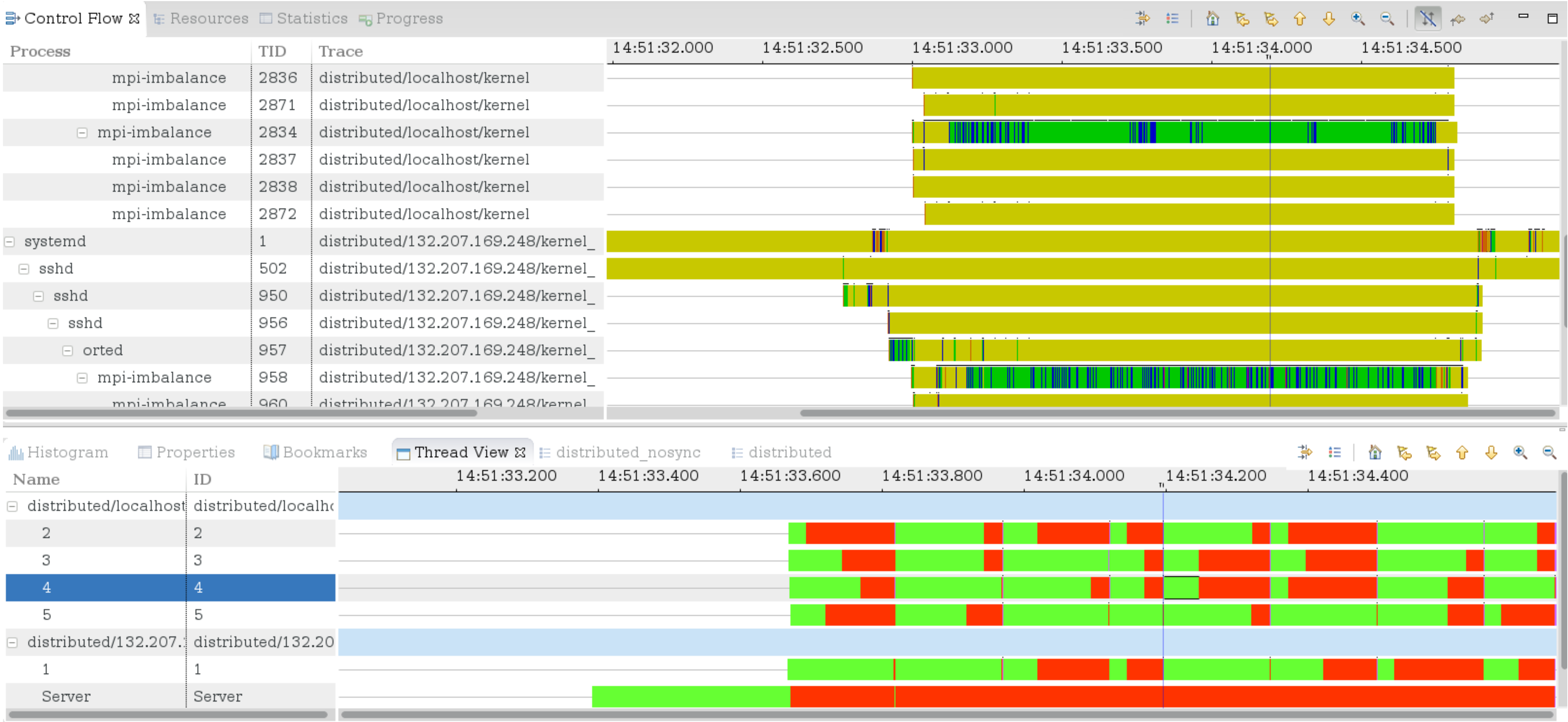
Process	TID	PI
mpi-imbalance	2836	28
mpi-imbalance	2871	28
mpi-imbalance	2834	28
mpi-imbalance	2837	28
mpi-imbalance	2838	28
mpi-imbalance	2872	28
systemd	1	
sshd	502	1
sshd	950	50
sshd	956	95
orted	957	95
mpi-imbalance	958	95
mpi-imbalance	960	95

Name	ID
distributed/localhost	distributed/localh
2	2
3	3
4	4
5	5
distributed/132.207.	distributed/132.20
1	1
Server	Server

Convex-Hull Synchronization Algorithm



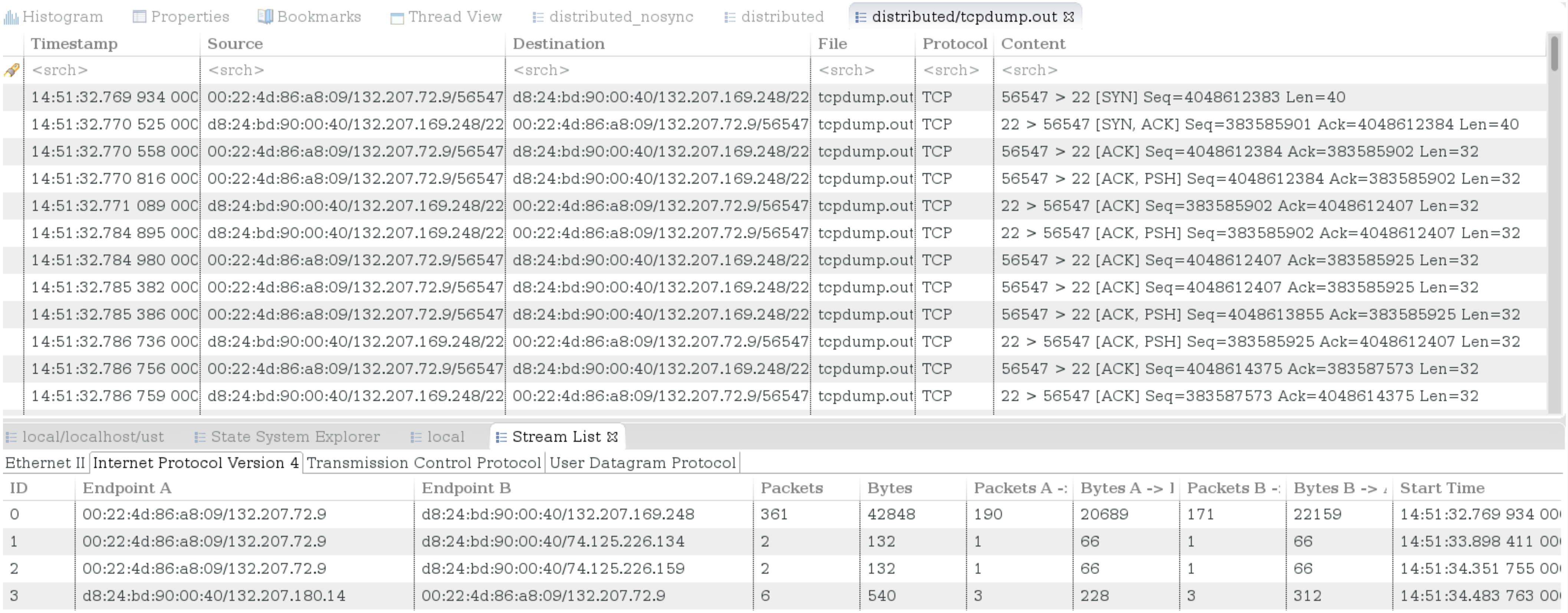
Experiment 2: Distributed Network: Synchronized View



Process	TID	Trace
mpi-imbalance	2836	distributed/localhost/kernel
mpi-imbalance	2871	distributed/localhost/kernel
⊟ mpi-imbalance	2834	distributed/localhost/kernel
mpi-imbalance	2837	distributed/localhost/kernel
mpi-imbalance	2838	distributed/localhost/kernel
mpi-imbalance	2872	distributed/localhost/kernel
⊟ systemd	1	distributed/132.207.169.248/kernel_
⊟ sshd	502	distributed/132.207.169.248/kernel_
⊟ sshd	950	distributed/132.207.169.248/kernel_
⊟ sshd	956	distributed/132.207.169.248/kernel_
⊟ orted	957	distributed/132.207.169.248/kernel_
⊟ mpi-imbalance	958	distributed/132.207.169.248/kernel_
mpi-imbalance	960	distributed/132.207.169.248/kernel

Name	ID
⊟ distributed/localhost	distributed/localhc
2	2
3	3
4	4
5	5
⊟ distributed/132.207.:	distributed/132.20
1	1
Server	Server

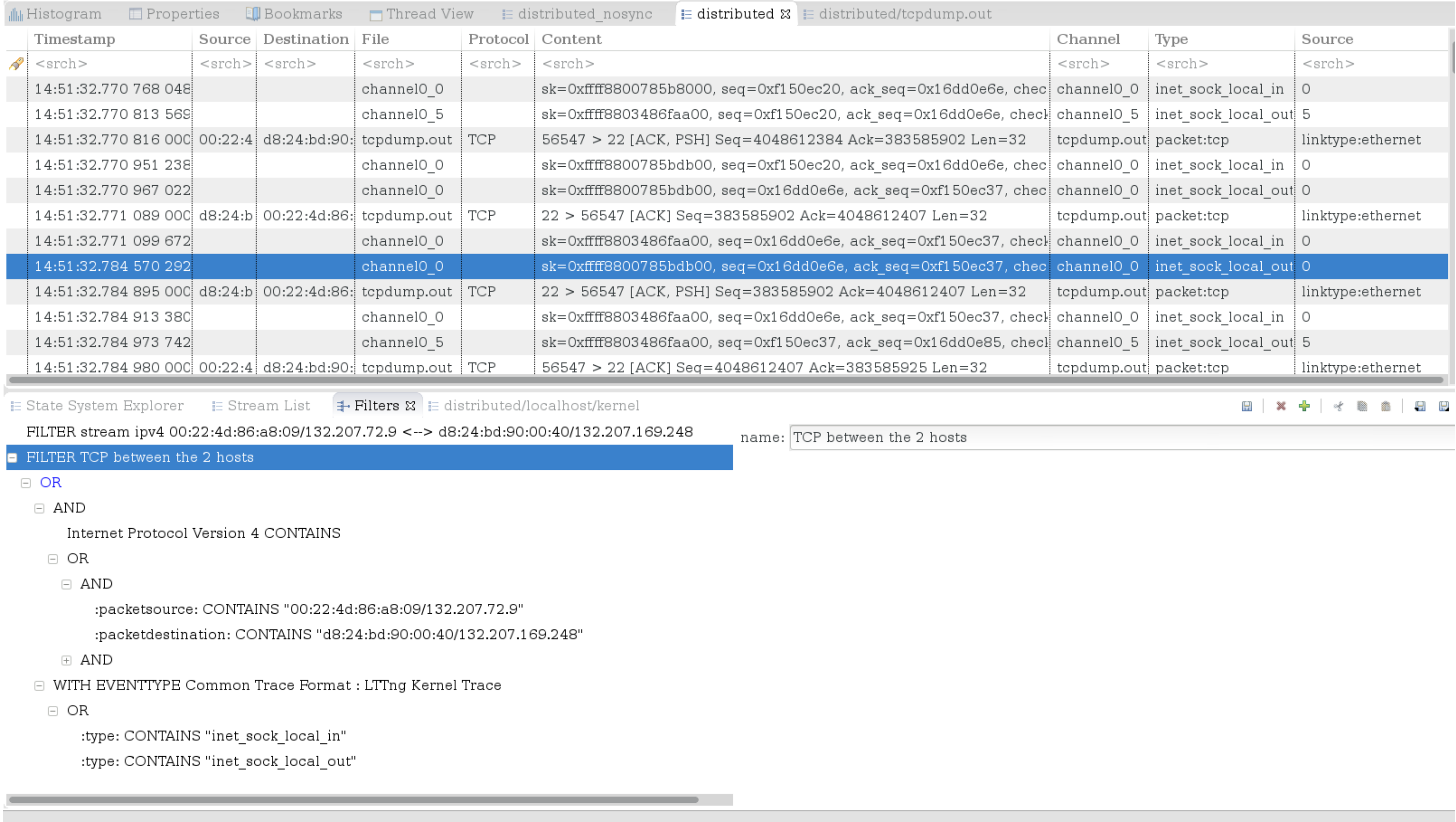
Experiment 2: Distributed Network: PCap traces



Timestamp	Source	Destination	File	Protocol	Content
<srch>	<srch>	<srch>	<srch>	<srch>	<srch>
14:51:32.769 934 00	00:22:4d:86:a8:09/132.207.72.9/56547	d8:24:bd:90:00:40/132.207.169.248/22	tcpdump.out	TCP	56547 > 22 [SYN] Seq=4048612383 Len=40
14:51:32.770 525 00	d8:24:bd:90:00:40/132.207.169.248/22	00:22:4d:86:a8:09/132.207.72.9/56547	tcpdump.out	TCP	22 > 56547 [SYN, ACK] Seq=383585901 Ack=4048612384 Len=40
14:51:32.770 558 00	00:22:4d:86:a8:09/132.207.72.9/56547	d8:24:bd:90:00:40/132.207.169.248/22	tcpdump.out	TCP	56547 > 22 [ACK] Seq=4048612384 Ack=383585902 Len=32
14:51:32.770 816 00	00:22:4d:86:a8:09/132.207.72.9/56547	d8:24:bd:90:00:40/132.207.169.248/22	tcpdump.out	TCP	56547 > 22 [ACK, PSH] Seq=4048612384 Ack=383585902 Len=32
14:51:32.771 089 00	d8:24:bd:90:00:40/132.207.169.248/22	00:22:4d:86:a8:09/132.207.72.9/56547	tcpdump.out	TCP	22 > 56547 [ACK] Seq=383585902 Ack=4048612407 Len=32
14:51:32.784 895 00	d8:24:bd:90:00:40/132.207.169.248/22	00:22:4d:86:a8:09/132.207.72.9/56547	tcpdump.out	TCP	22 > 56547 [ACK, PSH] Seq=383585902 Ack=4048612407 Len=32
14:51:32.784 980 00	00:22:4d:86:a8:09/132.207.72.9/56547	d8:24:bd:90:00:40/132.207.169.248/22	tcpdump.out	TCP	56547 > 22 [ACK] Seq=4048612407 Ack=383585925 Len=32
14:51:32.785 382 00	00:22:4d:86:a8:09/132.207.72.9/56547	d8:24:bd:90:00:40/132.207.169.248/22	tcpdump.out	TCP	56547 > 22 [ACK] Seq=4048612407 Ack=383585925 Len=32
14:51:32.785 386 00	00:22:4d:86:a8:09/132.207.72.9/56547	d8:24:bd:90:00:40/132.207.169.248/22	tcpdump.out	TCP	56547 > 22 [ACK, PSH] Seq=4048613855 Ack=383585925 Len=32
14:51:32.786 736 00	d8:24:bd:90:00:40/132.207.169.248/22	00:22:4d:86:a8:09/132.207.72.9/56547	tcpdump.out	TCP	22 > 56547 [ACK, PSH] Seq=383585925 Ack=4048612407 Len=32
14:51:32.786 756 00	00:22:4d:86:a8:09/132.207.72.9/56547	d8:24:bd:90:00:40/132.207.169.248/22	tcpdump.out	TCP	56547 > 22 [ACK] Seq=4048614375 Ack=383587573 Len=32
14:51:32.786 759 00	d8:24:bd:90:00:40/132.207.169.248/22	00:22:4d:86:a8:09/132.207.72.9/56547	tcpdump.out	TCP	22 > 56547 [ACK] Seq=383587573 Ack=4048614375 Len=32

ID	Endpoint A	Endpoint B	Packets	Bytes	Packets A -	Bytes A ->	Packets B -	Bytes B ->	Start Time
0	00:22:4d:86:a8:09/132.207.72.9	d8:24:bd:90:00:40/132.207.169.248	361	42848	190	20689	171	22159	14:51:32.769 934 00
1	00:22:4d:86:a8:09/132.207.72.9	d8:24:bd:90:00:40/74.125.226.134	2	132	1	66	1	66	14:51:33.898 411 00
2	00:22:4d:86:a8:09/132.207.72.9	d8:24:bd:90:00:40/74.125.226.159	2	132	1	66	1	66	14:51:34.351 755 00
3	d8:24:bd:90:00:40/132.207.180.14	00:22:4d:86:a8:09/132.207.72.9	6	540	3	228	3	312	14:51:34.483 763 00

Experiment 2: Distributed Network: PCap stream filter

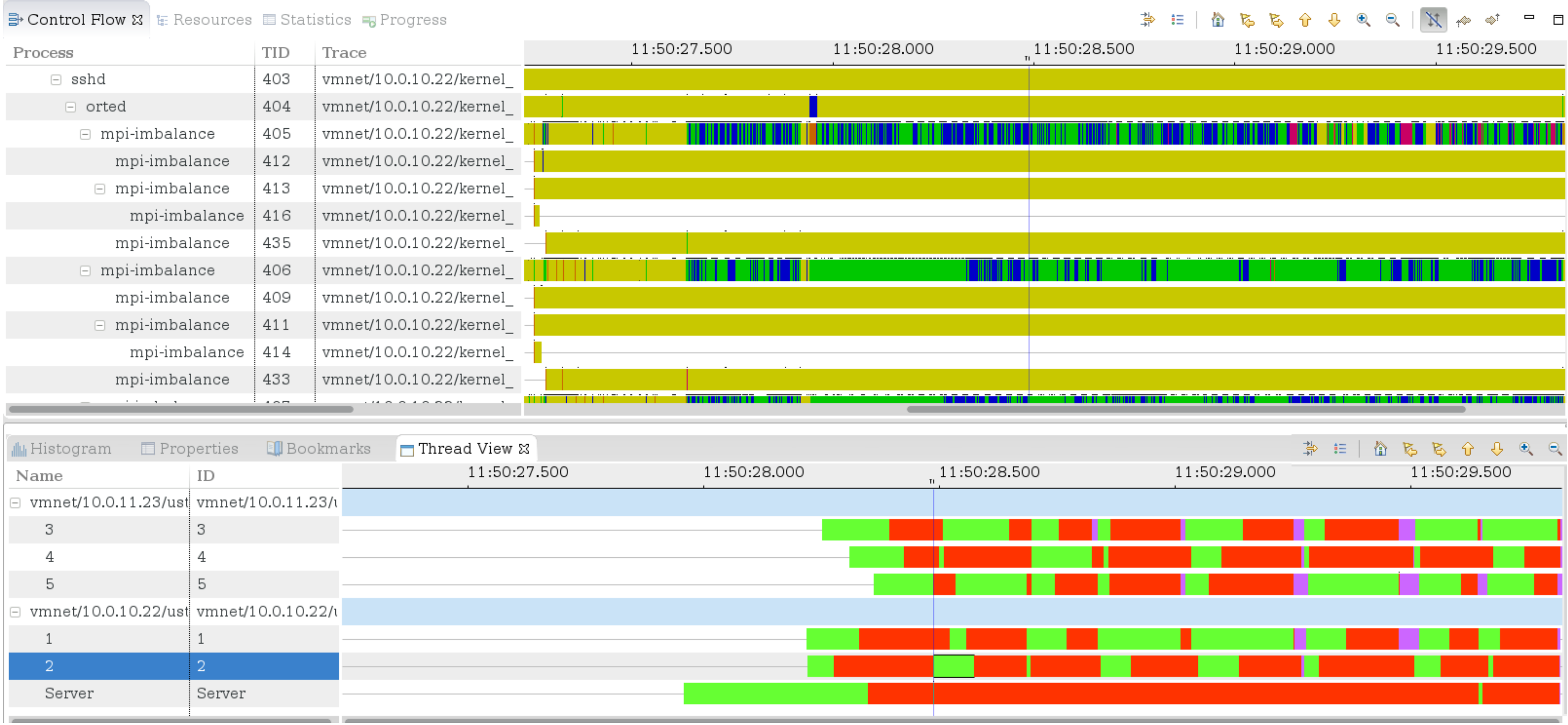


Timestamp	Source	Destination	File	Protocol	Content	Channel	Type	Source
<srch>	<srch>	<srch>	<srch>	<srch>	<srch>	<srch>	<srch>	<srch>
14:51:32.770 768 048			channel0_0		sk=0xffff8800785b8000, seq=0xf150ec20, ack_seq=0x16dd0e6e, chec	channel0_0	inet_sock_local_in	0
14:51:32.770 813 569			channel0_5		sk=0xffff8803486faa00, seq=0xf150ec20, ack_seq=0x16dd0e6e, checl	channel0_5	inet_sock_local_out	5
14:51:32.770 816 000	00:22:4	d8:24:bd:90:	tcpdump.out	TCP	56547 > 22 [ACK, PSH] Seq=4048612384 Ack=383585902 Len=32	tcpdump.out	packet:tcp	linktype:ethernet
14:51:32.770 951 238			channel0_0		sk=0xffff8800785bdb00, seq=0xf150ec20, ack_seq=0x16dd0e6e, chec	channel0_0	inet_sock_local_in	0
14:51:32.770 967 022			channel0_0		sk=0xffff8800785bdb00, seq=0x16dd0e6e, ack_seq=0xf150ec37, chec	channel0_0	inet_sock_local_out	0
14:51:32.771 089 000	d8:24:b	00:22:4d:86:	tcpdump.out	TCP	22 > 56547 [ACK] Seq=383585902 Ack=4048612407 Len=32	tcpdump.out	packet:tcp	linktype:ethernet
14:51:32.771 099 672			channel0_0		sk=0xffff8803486faa00, seq=0x16dd0e6e, ack_seq=0xf150ec37, checl	channel0_0	inet_sock_local_in	0
14:51:32.784 570 292			channel0_0		sk=0xffff8800785bdb00, seq=0x16dd0e6e, ack_seq=0xf150ec37, chec	channel0_0	inet_sock_local_out	0
14:51:32.784 895 000	d8:24:b	00:22:4d:86:	tcpdump.out	TCP	22 > 56547 [ACK, PSH] Seq=383585902 Ack=4048612407 Len=32	tcpdump.out	packet:tcp	linktype:ethernet
14:51:32.784 913 380			channel0_0		sk=0xffff8803486faa00, seq=0x16dd0e6e, ack_seq=0xf150ec37, checl	channel0_0	inet_sock_local_in	0
14:51:32.784 973 742			channel0_5		sk=0xffff8803486faa00, seq=0xf150ec37, ack_seq=0x16dd0e85, checl	channel0_5	inet_sock_local_out	5
14:51:32.784 980 000	00:22:4	d8:24:bd:90:	tcpdump.out	TCP	56547 > 22 [ACK] Seq=4048612407 Ack=383585925 Len=32	tcpdump.out	packet:tcp	linktype:ethernet

```
FILTER stream ipv4 00:22:4d:86:a8:09/132.207.72.9 <--> d8:24:bd:90:00:40/132.207.169.248
⊟ FILTER TCP between the 2 hosts
  ⊟ OR
    ⊟ AND
        Internet Protocol Version 4 CONTAINS
      ⊟ OR
        ⊟ AND
            :packetsource: CONTAINS "00:22:4d:86:a8:09/132.207.72.9"
            :packetdestination: CONTAINS "d8:24:bd:90:00:40/132.207.169.248"
        ⊞ AND
    ⊟ WITH EVENTTYPE Common Trace Format : LTTng Kernel Trace
      ⊟ OR
          :type: CONTAINS "inet_sock_local_in"
          :type: CONTAINS "inet_sock_local_out"
```

name: TCP between the 2 hosts

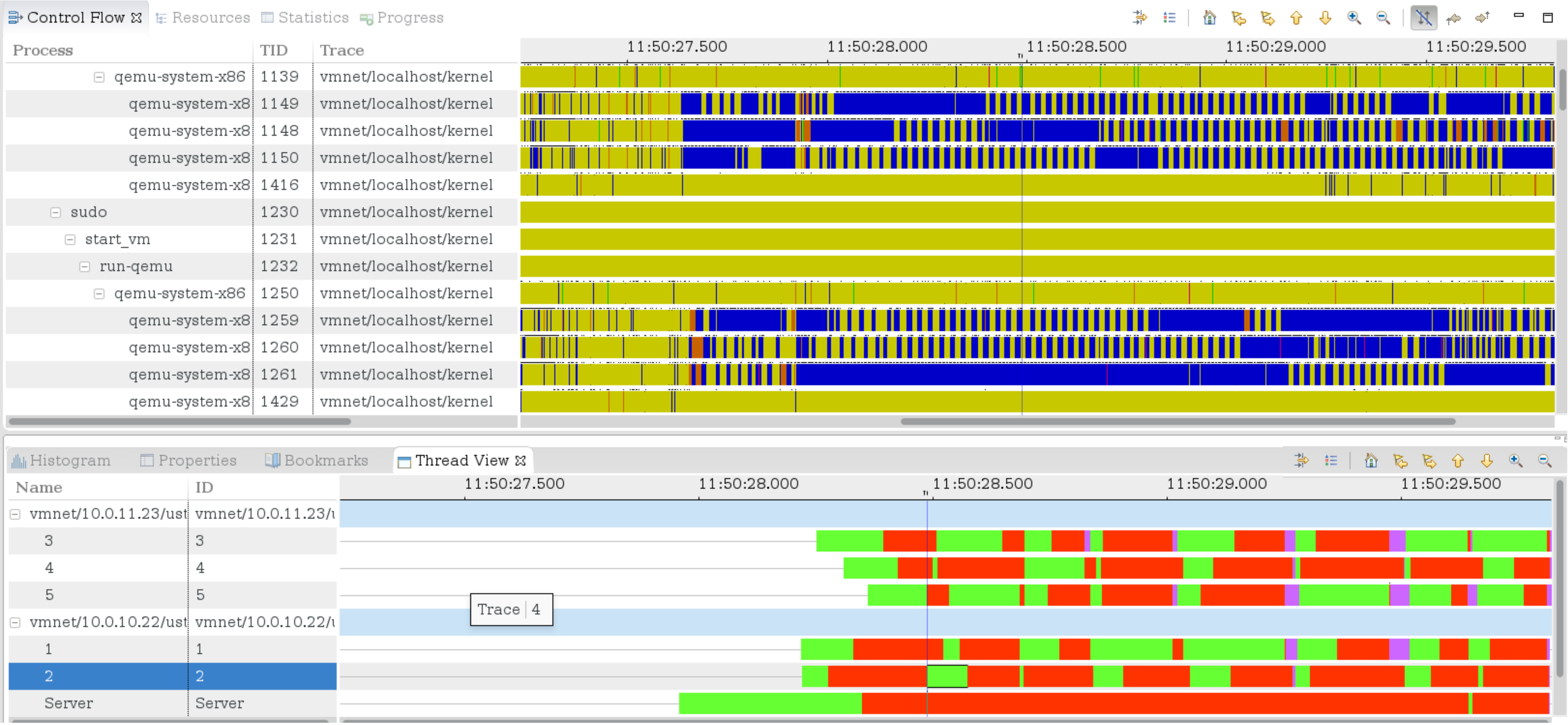
Experiment 3: Virtual Machines: Control Flow View and Thread View



Process	TID	Trace
sshd	403	vmnet/10.0.10.22/kernel_
orted	404	vmnet/10.0.10.22/kernel_
mpi-imbalance	405	vmnet/10.0.10.22/kernel_
mpi-imbalance	412	vmnet/10.0.10.22/kernel_
mpi-imbalance	413	vmnet/10.0.10.22/kernel_
mpi-imbalance	416	vmnet/10.0.10.22/kernel_
mpi-imbalance	435	vmnet/10.0.10.22/kernel_
mpi-imbalance	406	vmnet/10.0.10.22/kernel_
mpi-imbalance	409	vmnet/10.0.10.22/kernel_
mpi-imbalance	411	vmnet/10.0.10.22/kernel_
mpi-imbalance	414	vmnet/10.0.10.22/kernel_
mpi-imbalance	433	vmnet/10.0.10.22/kernel_

Name	ID
vmnet/10.0.11.23/ust	vmnet/10.0.11.23/ι
3	3
4	4
5	5
vmnet/10.0.10.22/ust	vmnet/10.0.10.22/ι
1	1
2	2
Server	Server

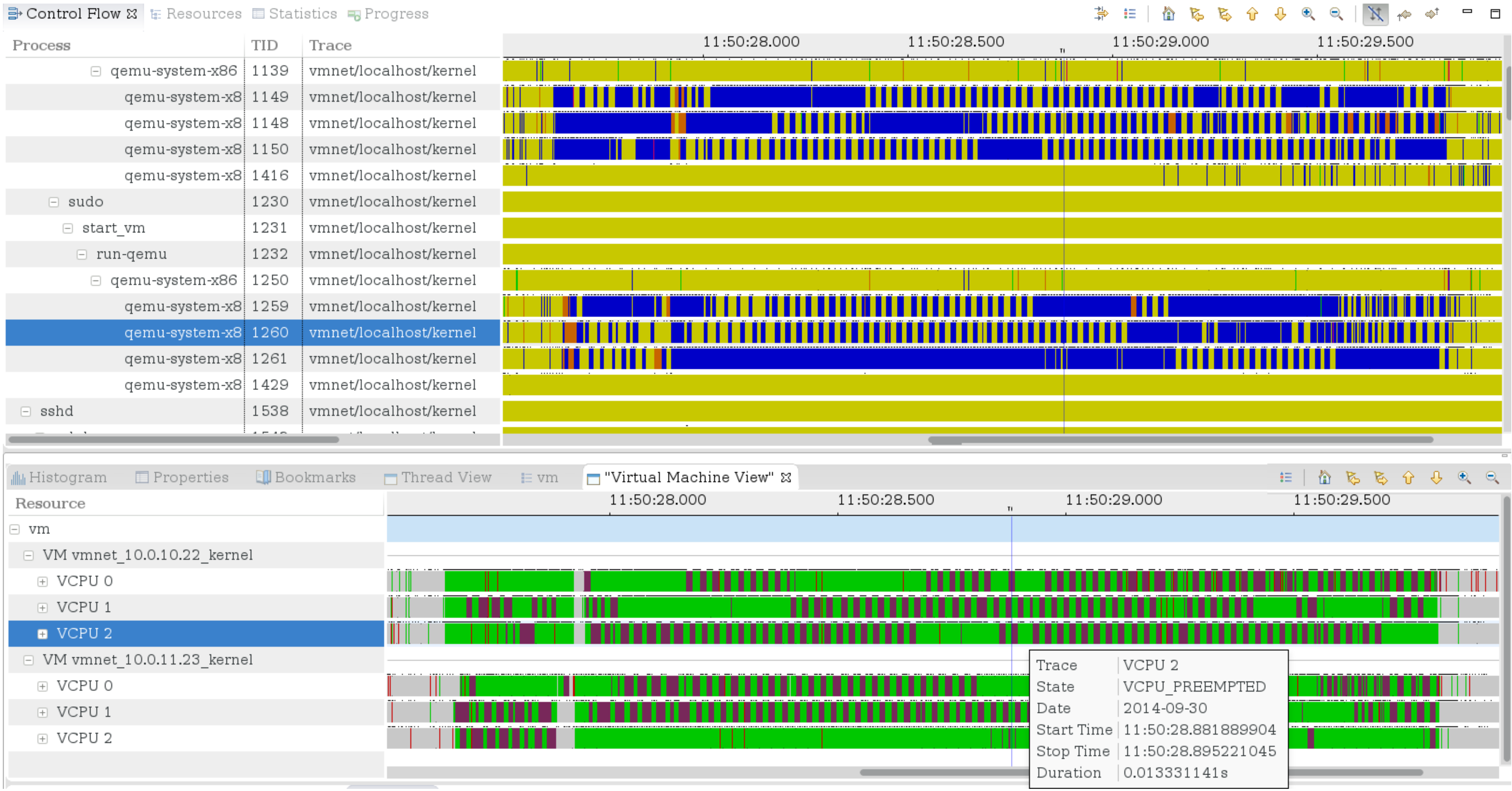
Experiment 3: Virtual Machines: qemu processes view



Process	TID	Trace
⊟ qemu-system-x86	1139	vmnet/localhost/kernel
qemu-system-x8	1149	vmnet/localhost/kernel
qemu-system-x8	1148	vmnet/localhost/kernel
qemu-system-x8	1150	vmnet/localhost/kernel
qemu-system-x8	1416	vmnet/localhost/kernel
⊟ sudo	1230	vmnet/localhost/kernel
⊟ start_vm	1231	vmnet/localhost/kernel
⊟ run-qemu	1232	vmnet/localhost/kernel
⊟ qemu-system-x86	1250	vmnet/localhost/kernel
qemu-system-x8	1259	vmnet/localhost/kernel
qemu-system-x8	1260	vmnet/localhost/kernel
qemu-system-x8	1261	vmnet/localhost/kernel
qemu-system-x8	1429	vmnet/localhost/kernel

Name	ID
⊟ vmnet/10.0.11.23/ust	vmnet/10.0.11.23/
3	3
4	4
5	5
⊟ vmnet/10.0.10.22/ust	vmnet/10.0.10.22/
1	1
2	2
Server	Server

Experiment 3: Virtual Machines: VCPUs view



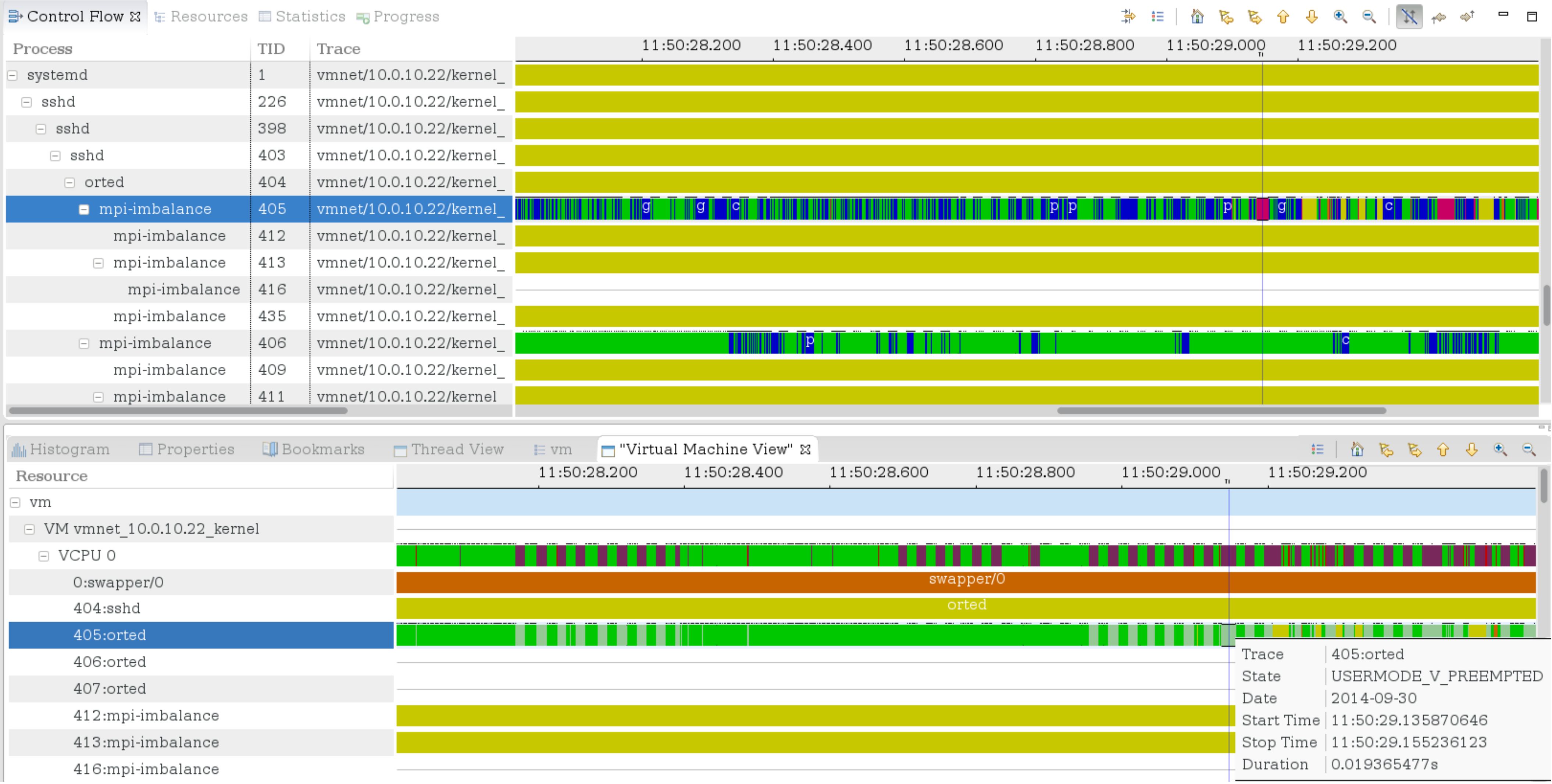
Process	TID	Trace
⊟ qemu-system-x86	1139	vmnet/localhost/kernel
qemu-system-x8	1149	vmnet/localhost/kernel
qemu-system-x8	1148	vmnet/localhost/kernel
qemu-system-x8	1150	vmnet/localhost/kernel
qemu-system-x8	1416	vmnet/localhost/kernel
⊟ sudo	1230	vmnet/localhost/kernel
⊟ start_vm	1231	vmnet/localhost/kernel
⊟ run-qemu	1232	vmnet/localhost/kernel
⊟ qemu-system-x86	1250	vmnet/localhost/kernel
qemu-system-x8	1259	vmnet/localhost/kernel
qemu-system-x8	1260	vmnet/localhost/kernel
qemu-system-x8	1261	vmnet/localhost/kernel
qemu-system-x8	1429	vmnet/localhost/kernel
⊟ sshd	1538	vmnet/localhost/kernel

Resource

- ⊟ vm
  - ⊟ VM vmnet_10.0.10.22_kernel
    - ⊞ VCPU 0
    - ⊞ VCPU 1
    - ⊞ VCPU 2
  - ⊟ VM vmnet_10.0.11.23_kernel
    - ⊞ VCPU 0
    - ⊞ VCPU 1
    - ⊞ VCPU 2

Trace	VCPU 2
State	VCPU_PREEMPTED
Date	2014-09-30
Start Time	11:50:28.881889904
Stop Time	11:50:28.895221045
Duration	0.013331141s

Experiment 3: Virtual Machines: 1 VCPU view

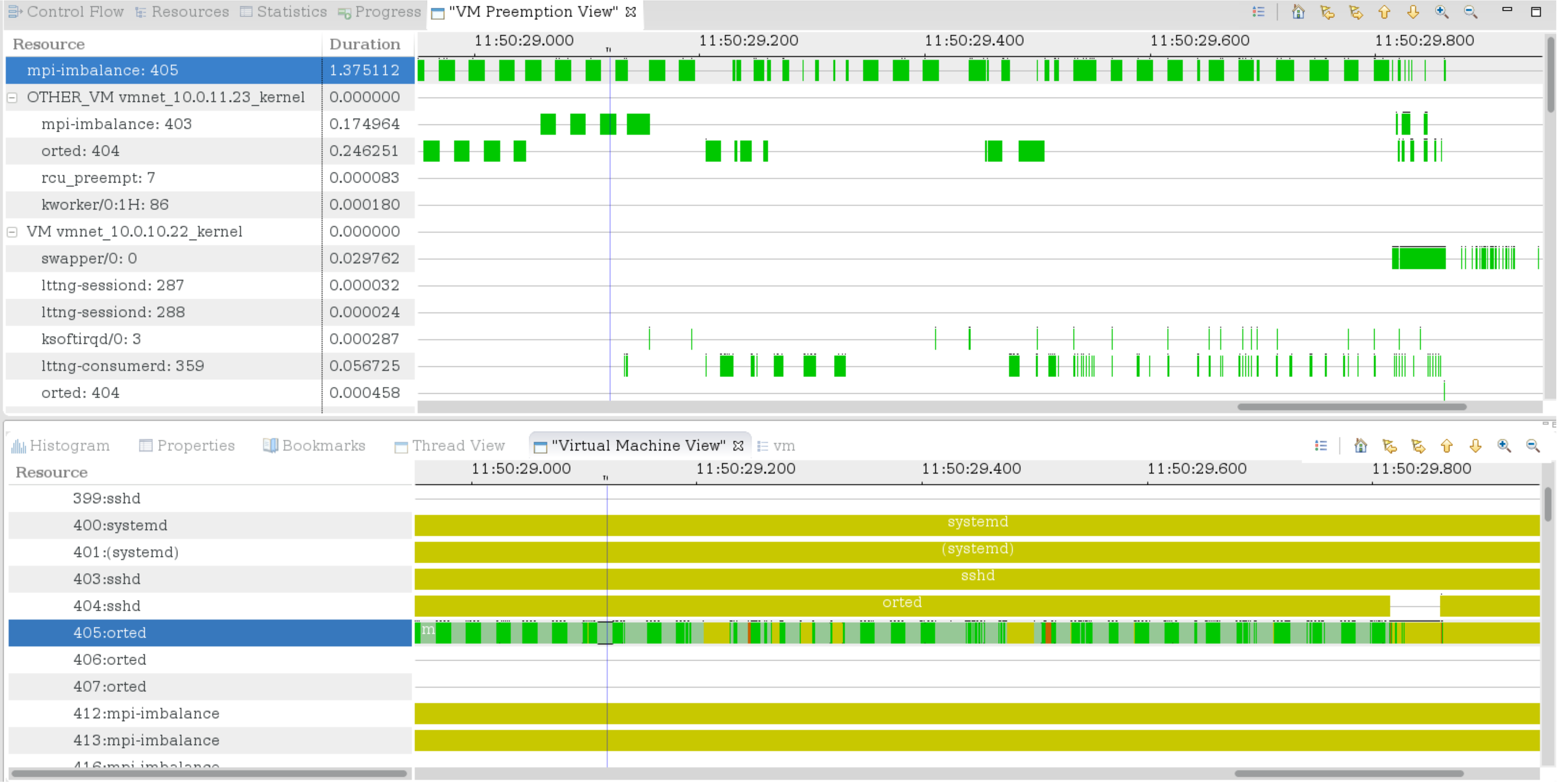


Process	TID	Trace
systemd	1	vmnet/10.0.10.22/kernel_
sshd	226	vmnet/10.0.10.22/kernel_
sshd	398	vmnet/10.0.10.22/kernel_
sshd	403	vmnet/10.0.10.22/kernel_
orted	404	vmnet/10.0.10.22/kernel_
mpi-imbalance	405	vmnet/10.0.10.22/kernel_
mpi-imbalance	412	vmnet/10.0.10.22/kernel_
mpi-imbalance	413	vmnet/10.0.10.22/kernel_
mpi-imbalance	416	vmnet/10.0.10.22/kernel_
mpi-imbalance	435	vmnet/10.0.10.22/kernel_
mpi-imbalance	406	vmnet/10.0.10.22/kernel_
mpi-imbalance	409	vmnet/10.0.10.22/kernel_
mpi-imbalance	411	vmnet/10.0.10.22/kernel

Resource
- vm
  - VM vmnet_10.0.10.22_kernel
    - VCPU 0
      - 0:swapper/0
      - 404:sshd
      - 405:orted
      - 406:orted
      - 407:orted
      - 412:mpi-imbalance
      - 413:mpi-imbalance
      - 416:mpi-imbalance

Trace	405:orted
State	USERMODE_V_PREEMPTED
Date	2014-09-30
Start Time	11:50:29.135870646
Stop Time	11:50:29.155236123
Duration	0.019365477s

Experiment 3: Virtual Machines: VM Preempt View



Resource	Duration
mpi-imbalance: 405	1.375112
⊟ OTHER_VM vmnet_10.0.11.23_kernel	0.000000
mpi-imbalance: 403	0.174964
orted: 404	0.246251
rcu_preempt: 7	0.000083
kworker/0:1H: 86	0.000180
⊟ VM vmnet_10.0.10.22_kernel	0.000000
swapper/0: 0	0.029762
lttng-sessiond: 287	0.000032
lttng-sessiond: 288	0.000024
ksoftirqd/0: 3	0.000287
lttng-consumerd: 359	0.056725
orted: 404	0.000458

Resource
399:sshd
400:systemd
401:(systemd)
403:sshd
404:sshd
405:orted
406:orted
407:orted
412:mpi-imbalance
413:mpi-imbalance
416:mpi-imbalance