

Visualization of Large Scale I/O Patterns

Chris Muelder

VIDi @ U. C. Davis

HEC FSIO 2010



Disclaimer

- I'm a 'vis' guy

Introduction

- Parallel file system optimization
- Analyze usage patterns
 - Bottlenecks
 - Inefficient codes
 - Trends
- Scalable
- Two steps process
 - Trace collection
 - Trace analysis

Trace Collection

- Various existing tracing libraries
 - MPE, Tau
 - Extend them for our purpose
- Focus on I/O benchmarks
 - IOR
- Run on several systems
 - Jazz, Intrepid/Eureka

Trace Analysis

- Existing visual tools
 - Jumpshot, PerfExplorer, etc...
 - Primarily Gantt chart or histogram based
 - Scalability?
- A new visualization approach
 - Our existing scalable visualization (InfoVis '09)
 - I/O specific extensions

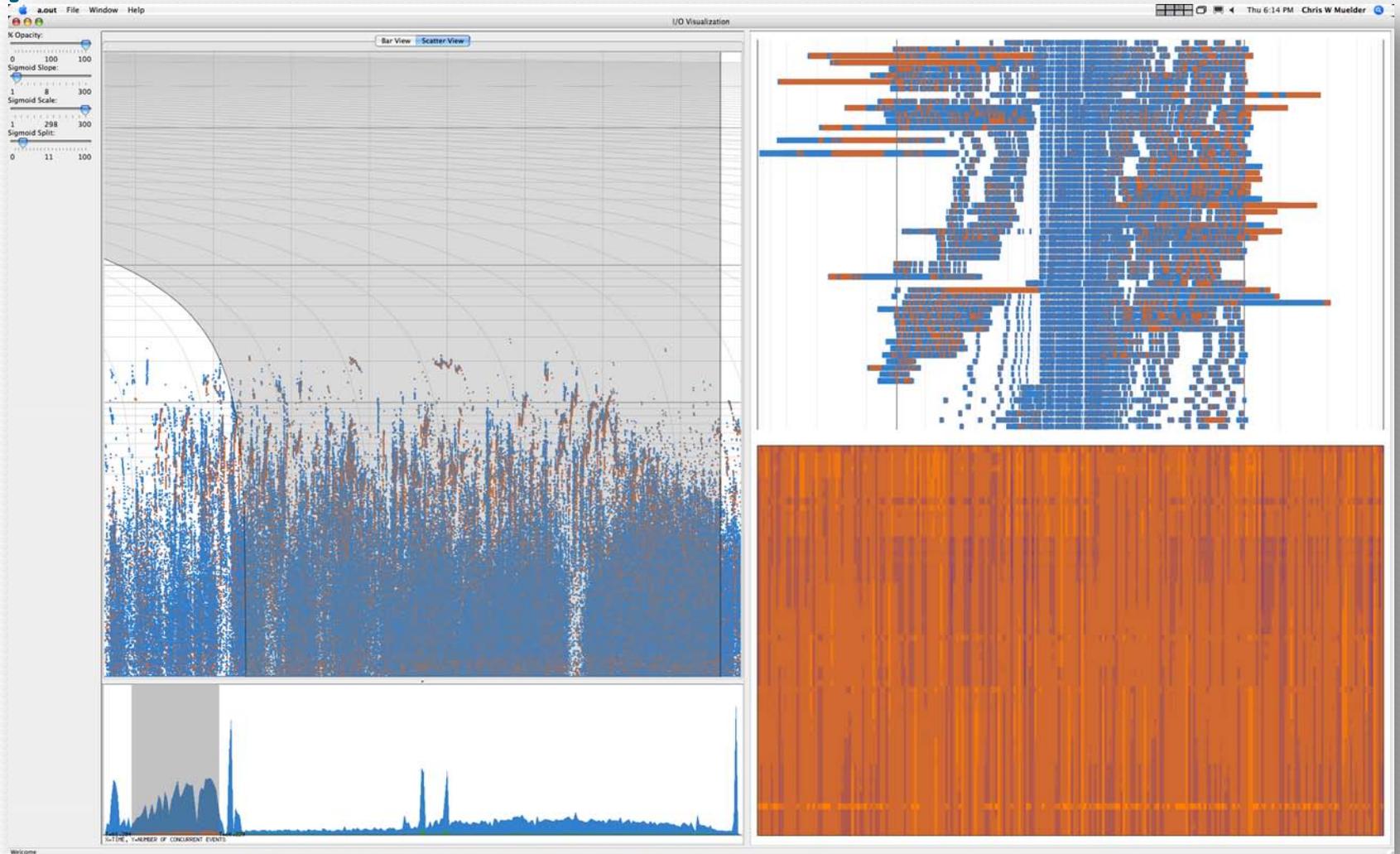
MPI Performance Visualization



MPI Performance Visualization

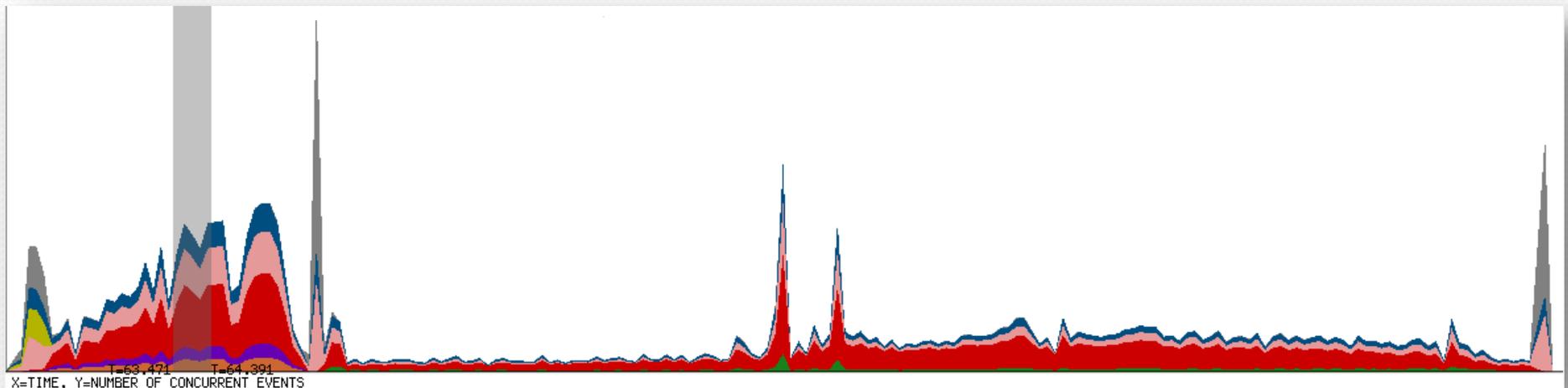
- Apply the method to I/O traces?
 - Good scalability
 - Poor handling of relational information
- Extend the approach
 - Additional representations

I/O Performance Visualization

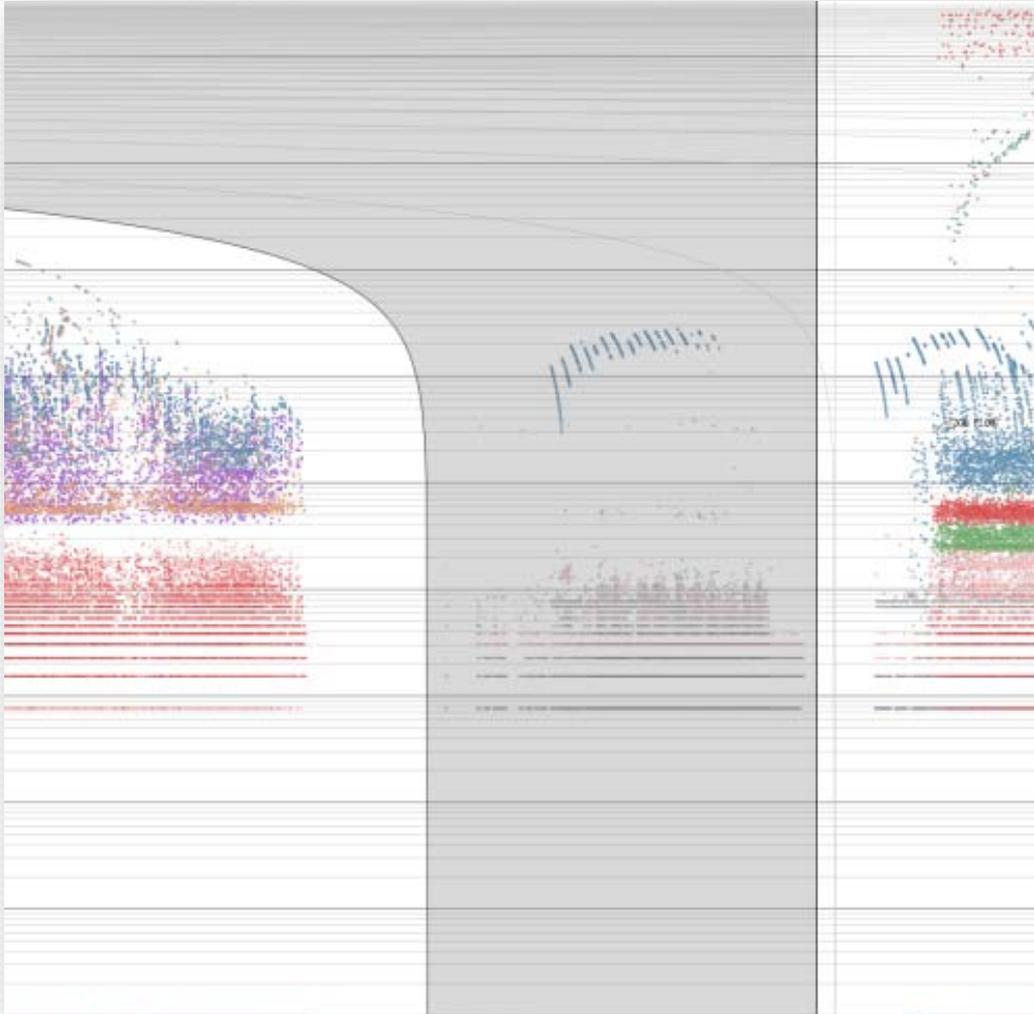


Timeline View

- Overall I/O server activity level over time
 - Each area is a particular function or type of function
 - Height is proportional to time spent in function and number of processors
 - Select region to analyze in more detail

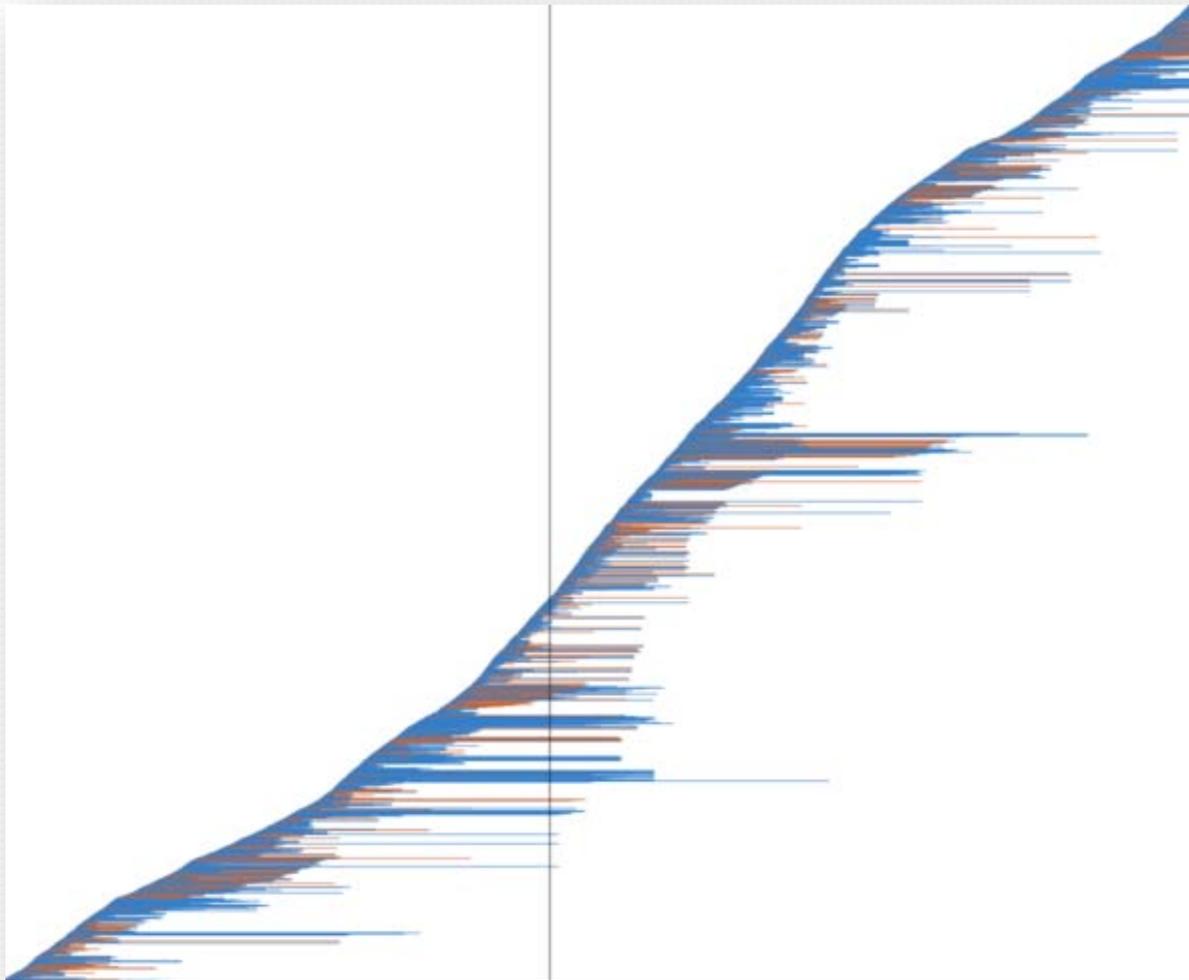


Mid-Level Views



- Plot of traced calls
 - Log(duration) vs time
 - Color is function type
 - Logarithmic opacity map
 - Lines or points
- Scalable
- No relational information

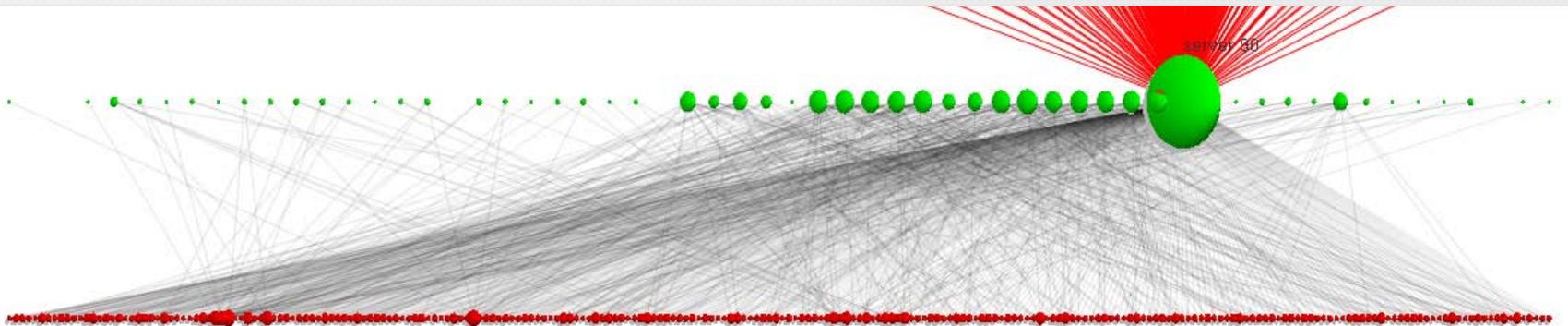
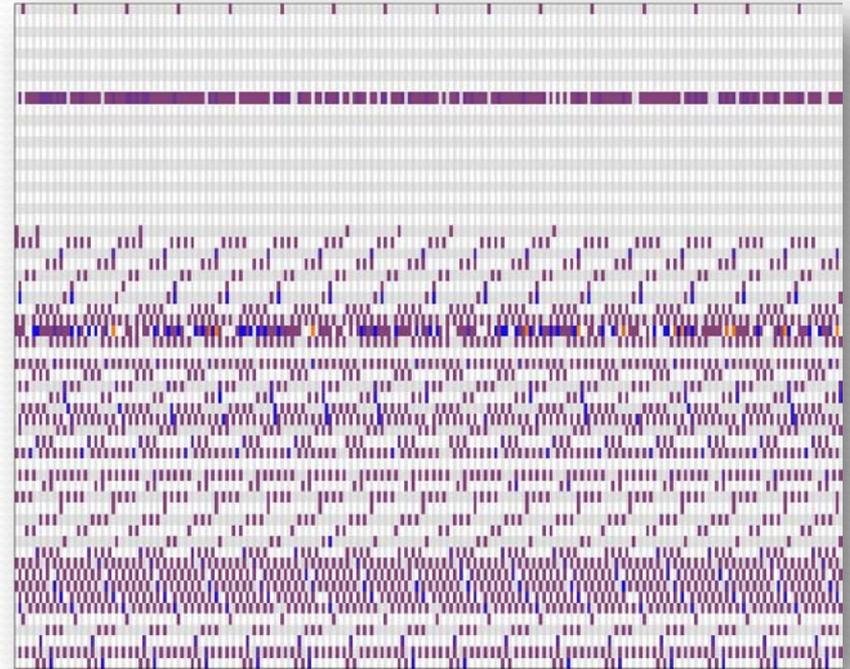
Mid-Level Views



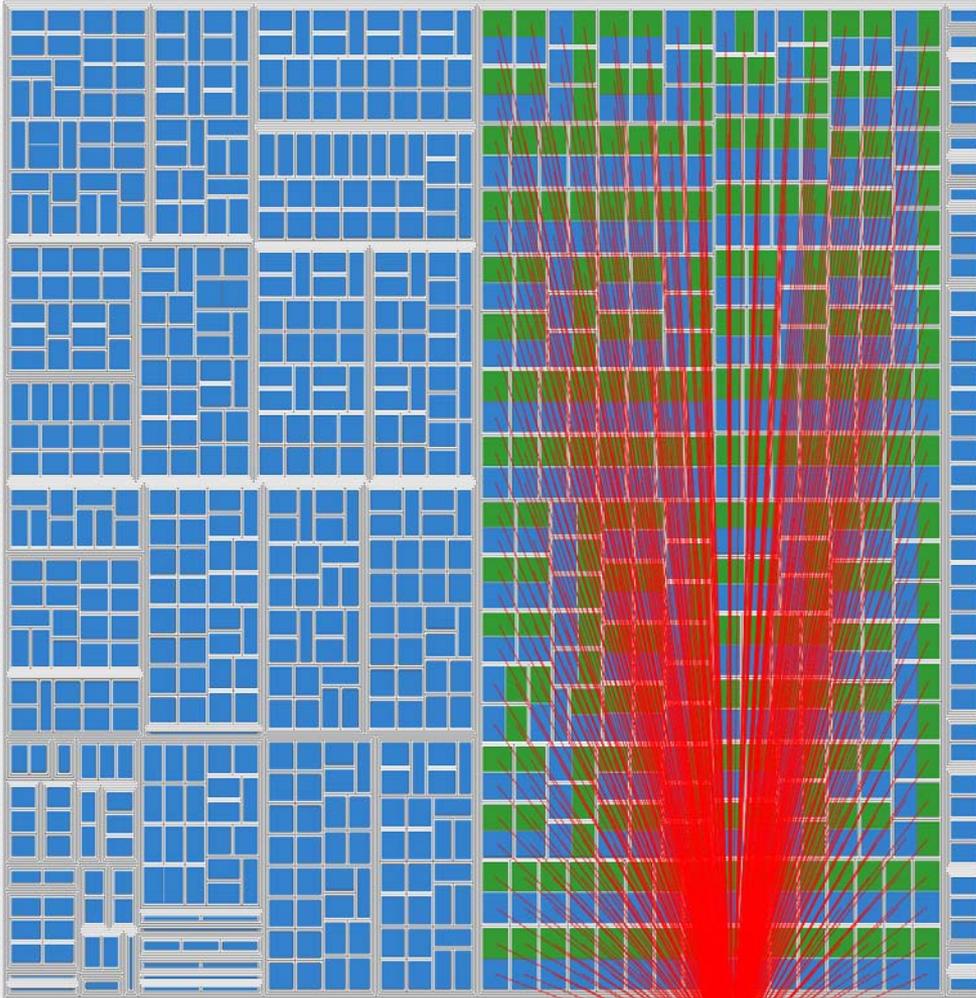
- Sorted line plot
 - Sort by start/end
 - Slope is I/O op handling rate
- Somewhat scalable
- Still no relations

Detail Views

- Connectivity graph/matrix
- I/O servers vs compute nodes

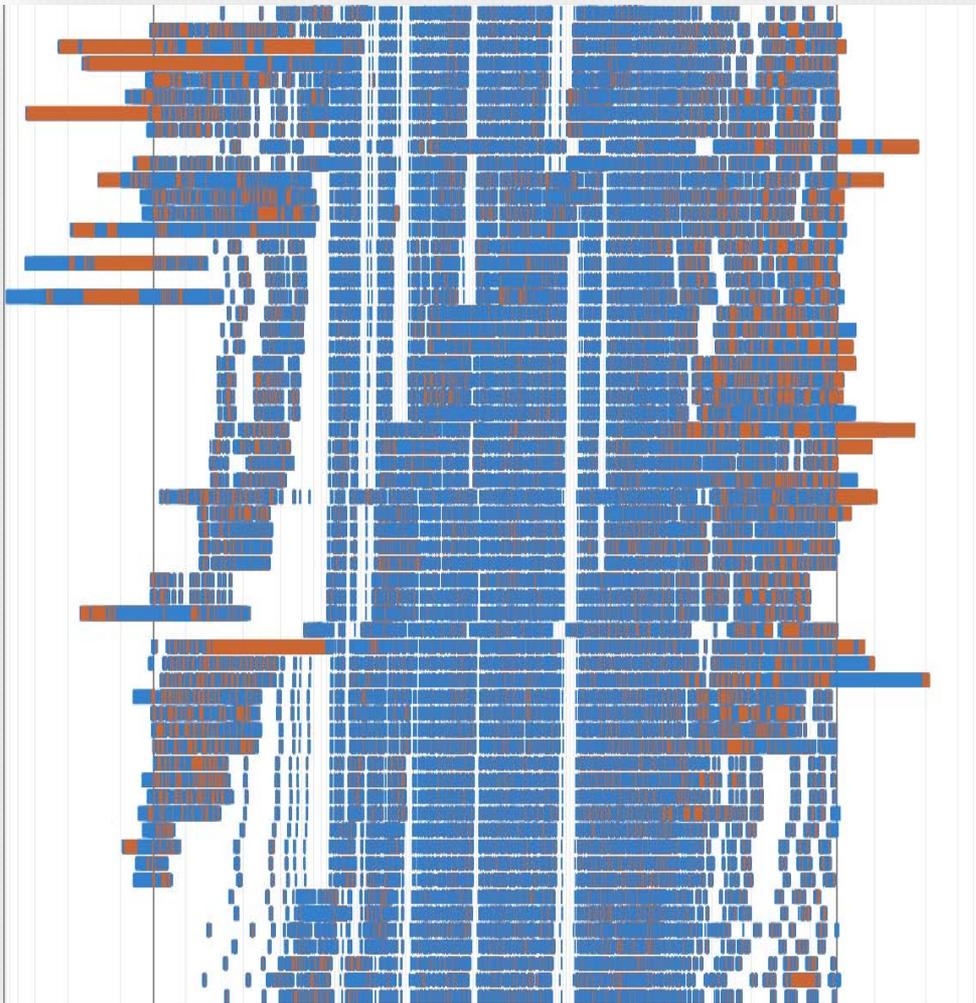


Detail Views



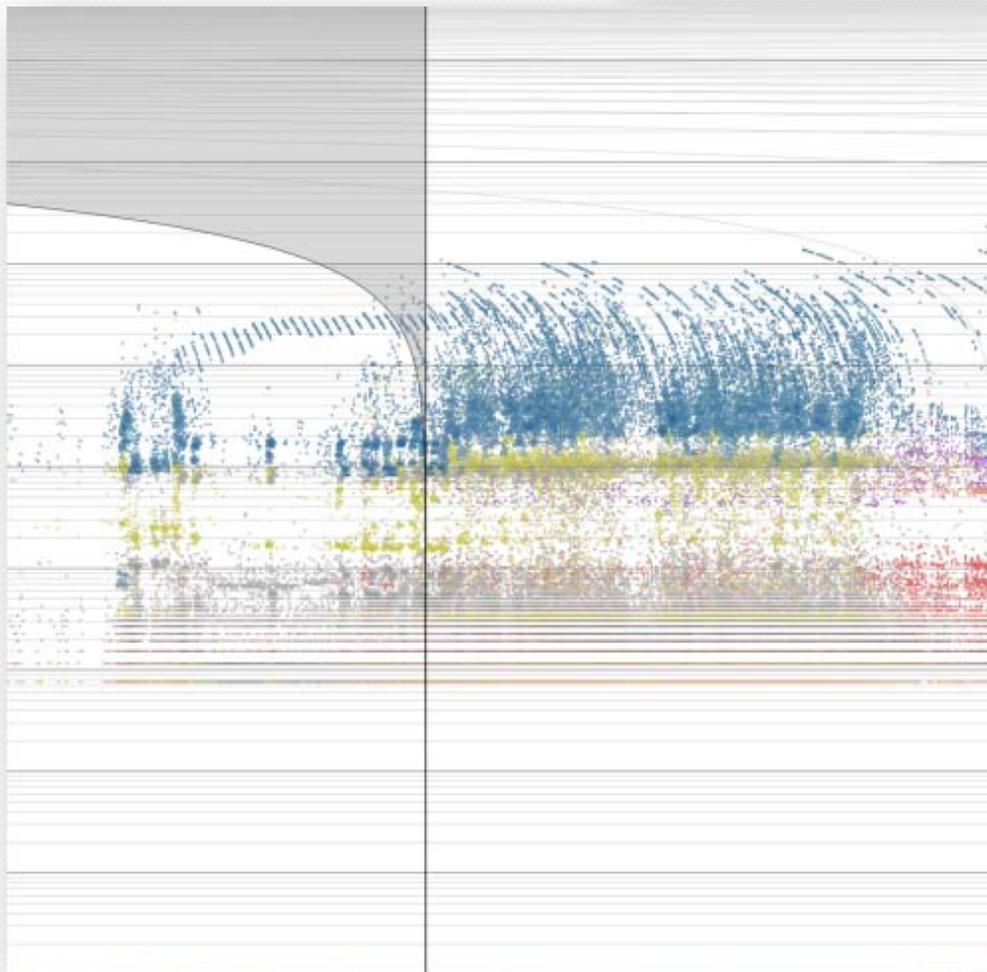
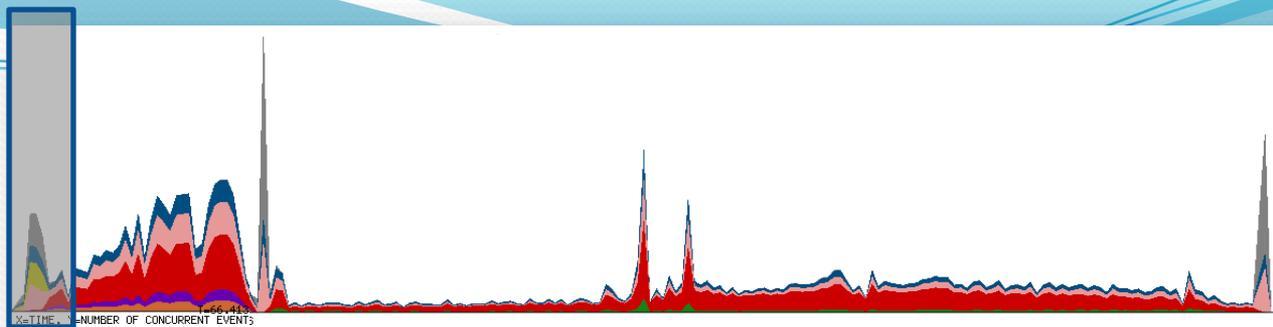
- Treemap view
 - Hierarchy of properties
 - E.g. Server->Rank->Call->...
- Way to see whole system at a single time.
- Links to graph view

Detail Views

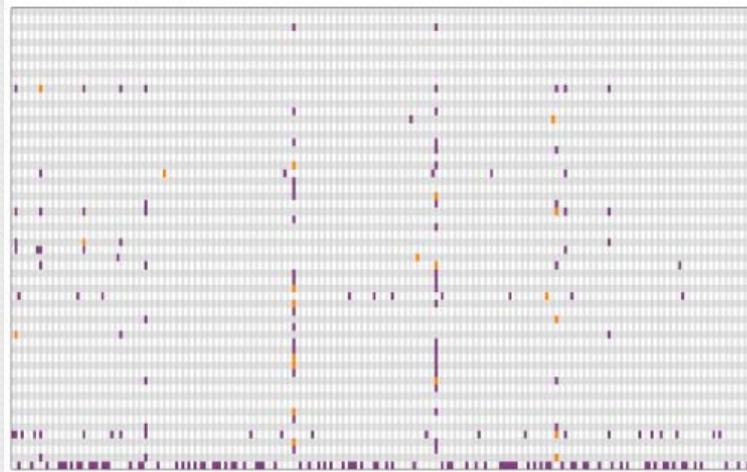


- Sigmoidal View
 - Non-linear time axis
 - Zoom level and linearity user adjustable
- Whole system around one or two points in time.
 - See a range of time
 - Compare two times
- For now, a Gantt chart
 - (Alternate representation?)

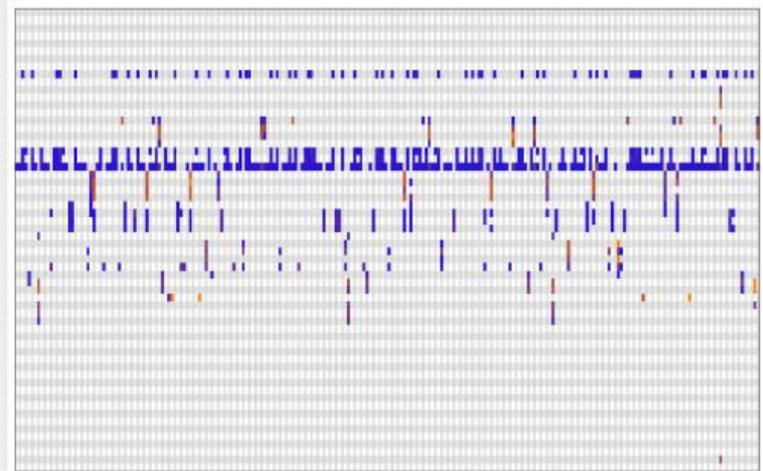
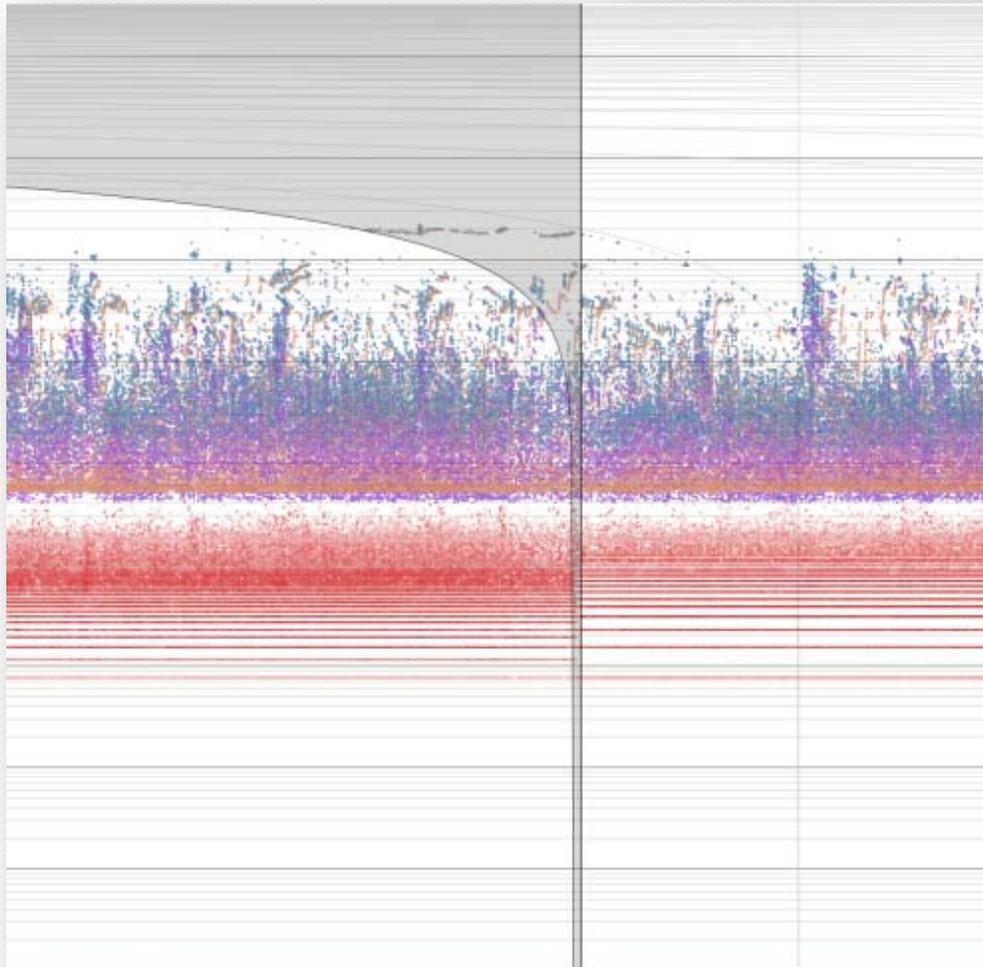
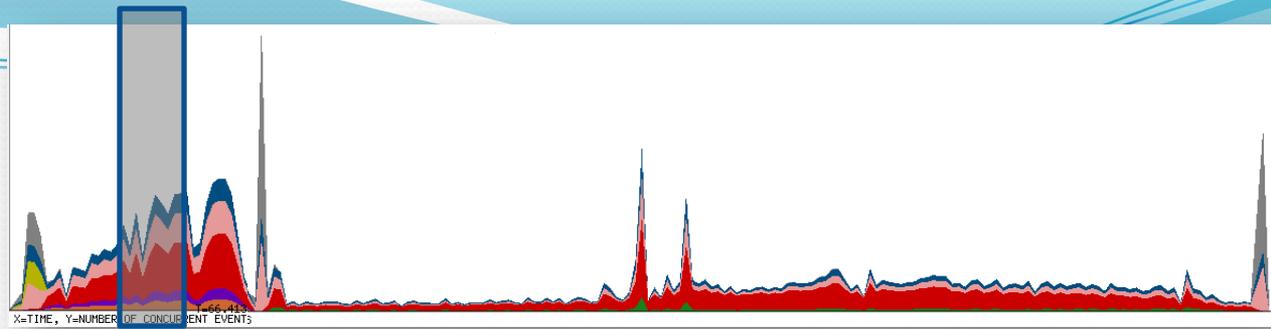
Examples



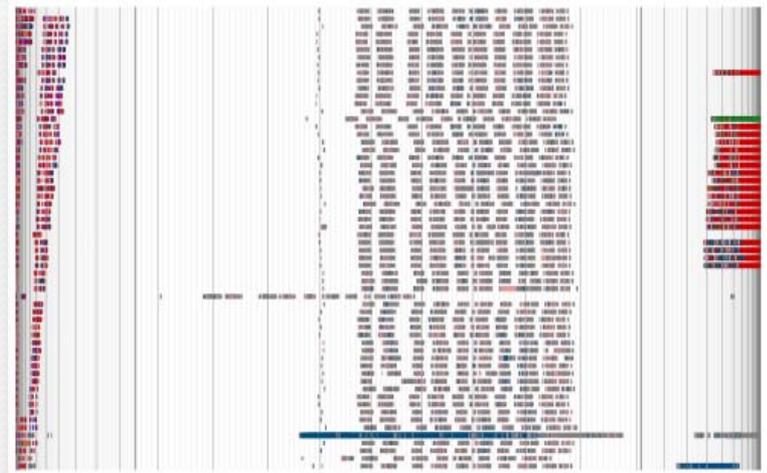
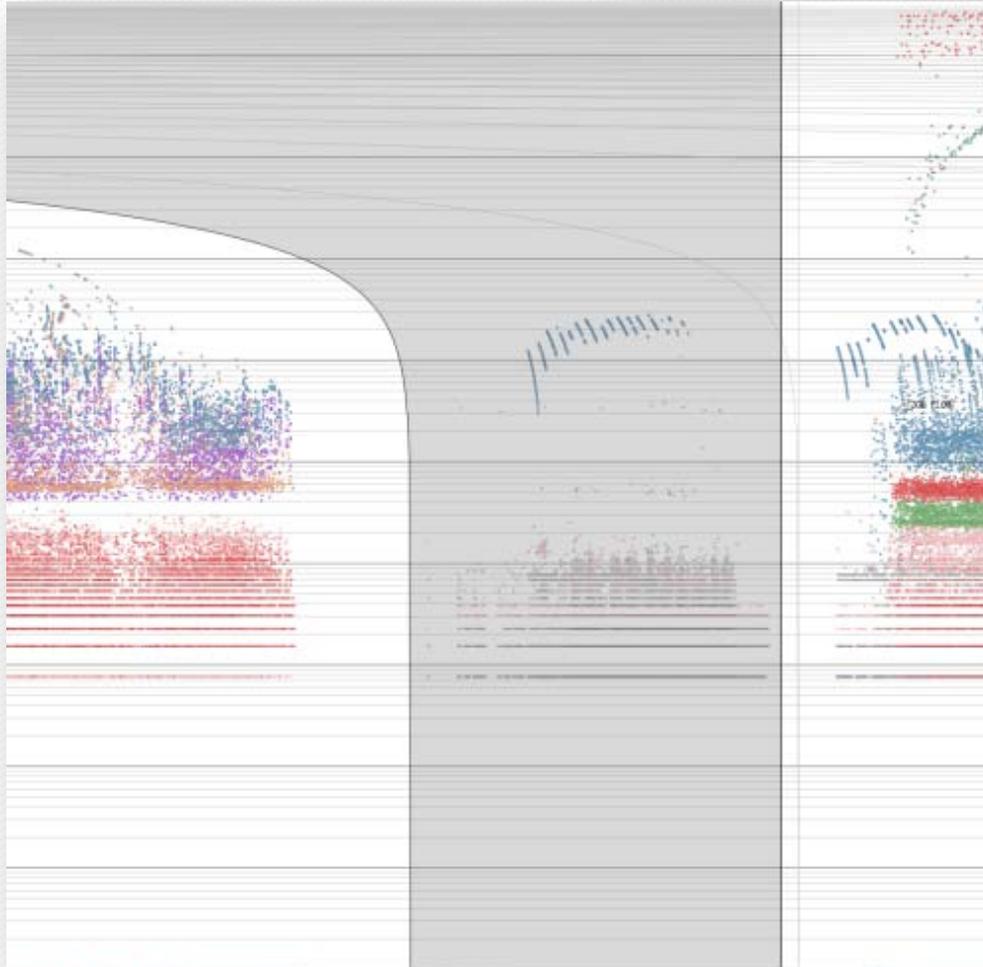
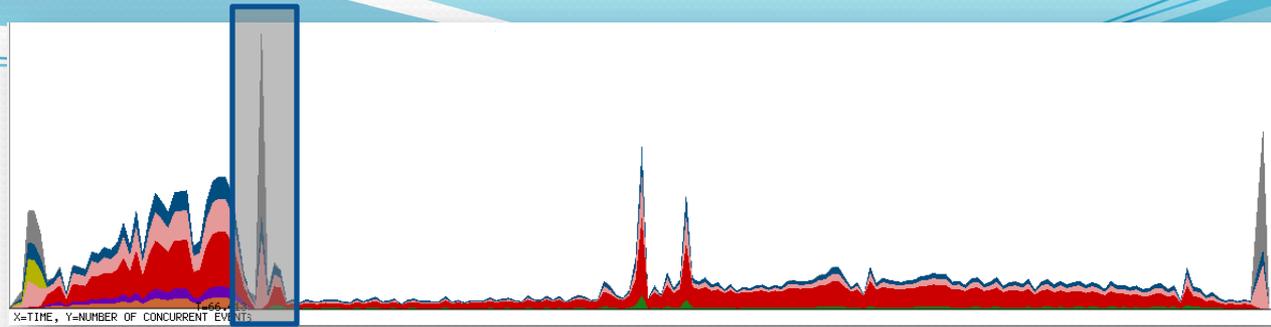
Time	Event ID	Source	Destination	Priority	Status
00:00:00	1
00:00:01	2
00:00:02	3
00:00:03	4
00:00:04	5
00:00:05	6
00:00:06	7
00:00:07	8
00:00:08	9
00:00:09	10
00:00:10	11
00:00:11	12
00:00:12	13
00:00:13	14
00:00:14	15
00:00:15	16
00:00:16	17
00:00:17	18
00:00:18	19
00:00:19	20
00:00:20	21
00:00:21	22
00:00:22	23
00:00:23	24
00:00:24	25
00:00:25	26
00:00:26	27
00:00:27	28
00:00:28	29
00:00:29	30
00:00:30	31
00:00:31	32
00:00:32	33
00:00:33	34
00:00:34	35
00:00:35	36
00:00:36	37
00:00:37	38
00:00:38	39
00:00:39	40
00:00:40	41
00:00:41	42
00:00:42	43
00:00:43	44
00:00:44	45
00:00:45	46
00:00:46	47
00:00:47	48
00:00:48	49
00:00:49	50
00:00:50	51
00:00:51	52
00:00:52	53
00:00:53	54
00:00:54	55
00:00:55	56
00:00:56	57
00:00:57	58
00:00:58	59
00:00:59	60
00:01:00	61
00:01:01	62
00:01:02	63
00:01:03	64
00:01:04	65
00:01:05	66
00:01:06	67
00:01:07	68
00:01:08	69
00:01:09	70
00:01:10	71
00:01:11	72
00:01:12	73
00:01:13	74
00:01:14	75
00:01:15	76
00:01:16	77
00:01:17	78
00:01:18	79
00:01:19	80
00:01:20	81
00:01:21	82
00:01:22	83
00:01:23	84
00:01:24	85
00:01:25	86
00:01:26	87
00:01:27	88
00:01:28	89
00:01:29	90
00:01:30	91
00:01:31	92
00:01:32	93
00:01:33	94
00:01:34	95
00:01:35	96
00:01:36	97
00:01:37	98
00:01:38	99
00:01:39	100



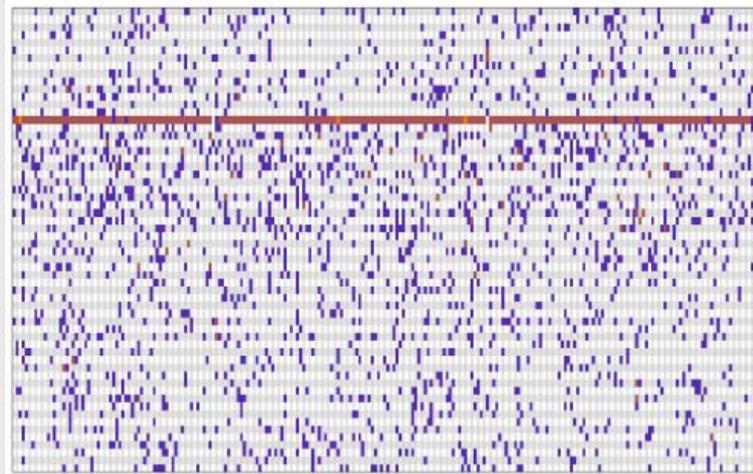
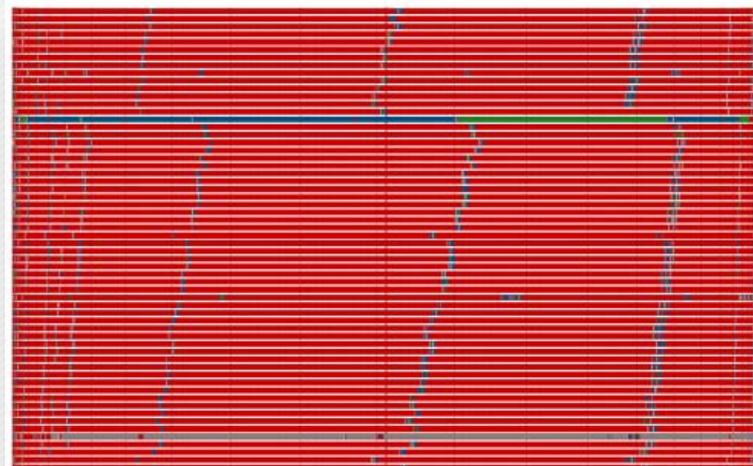
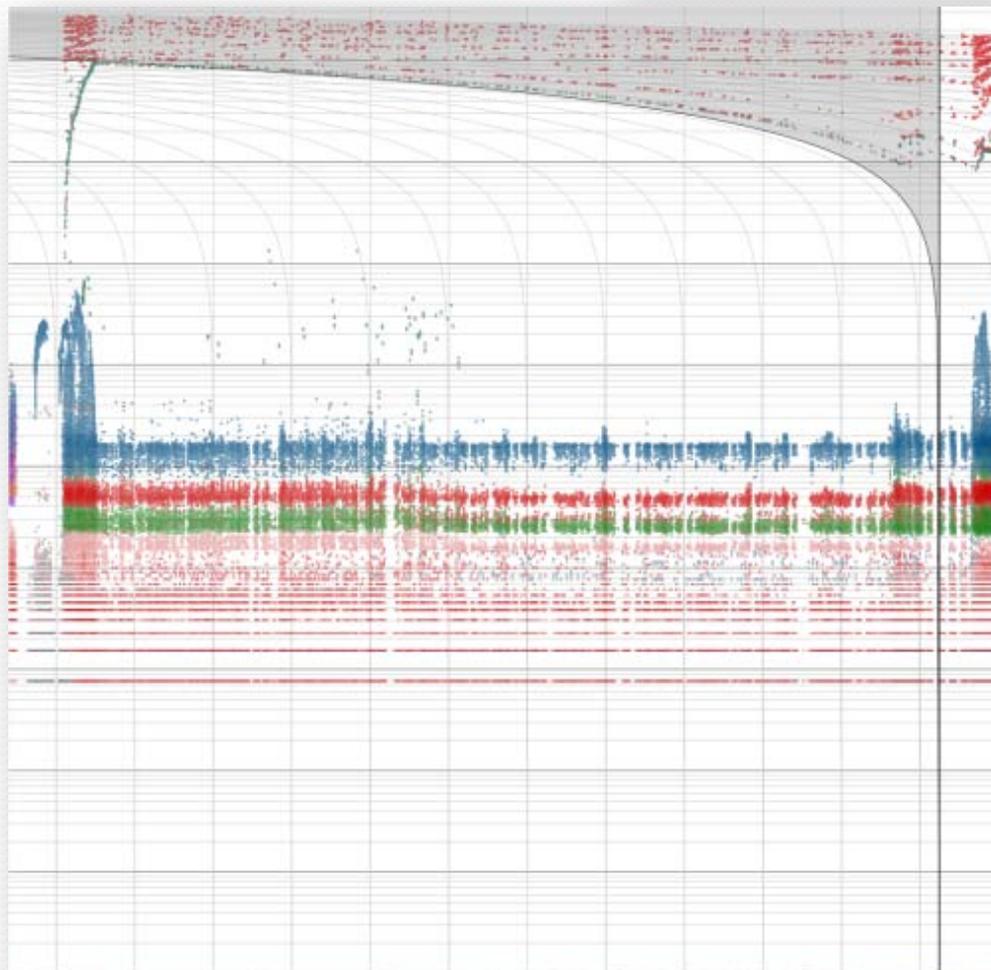
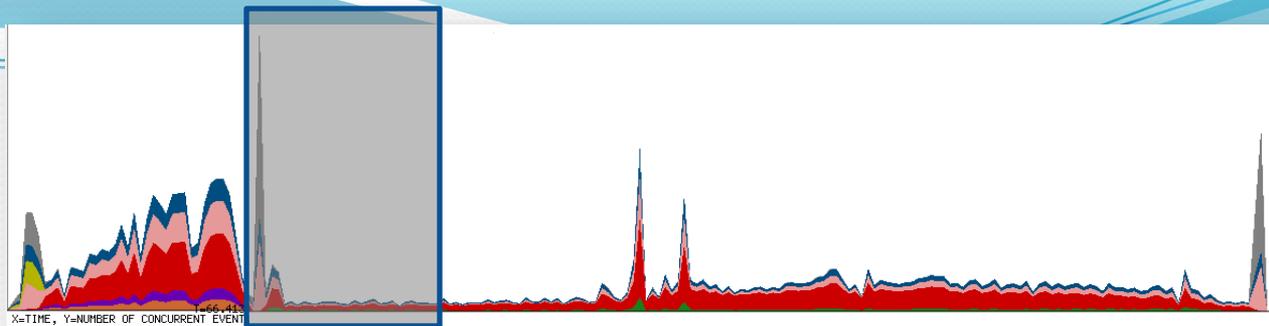
Examples



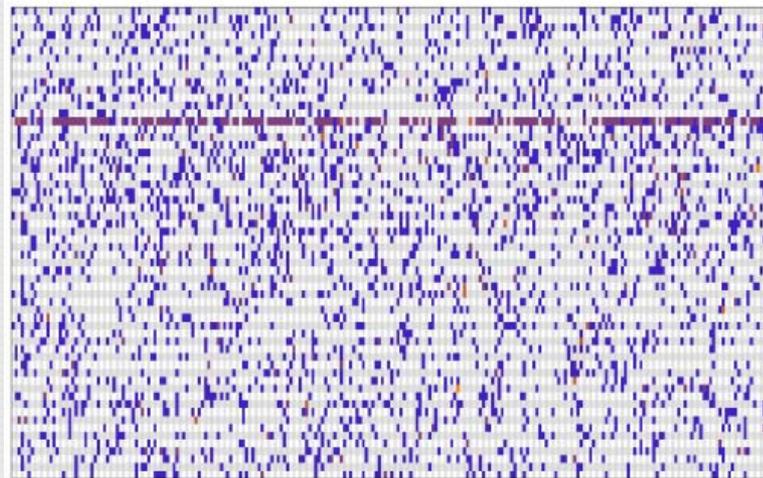
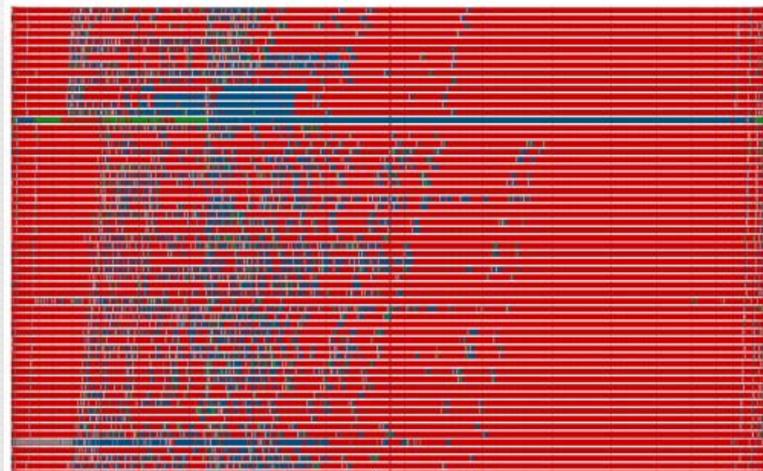
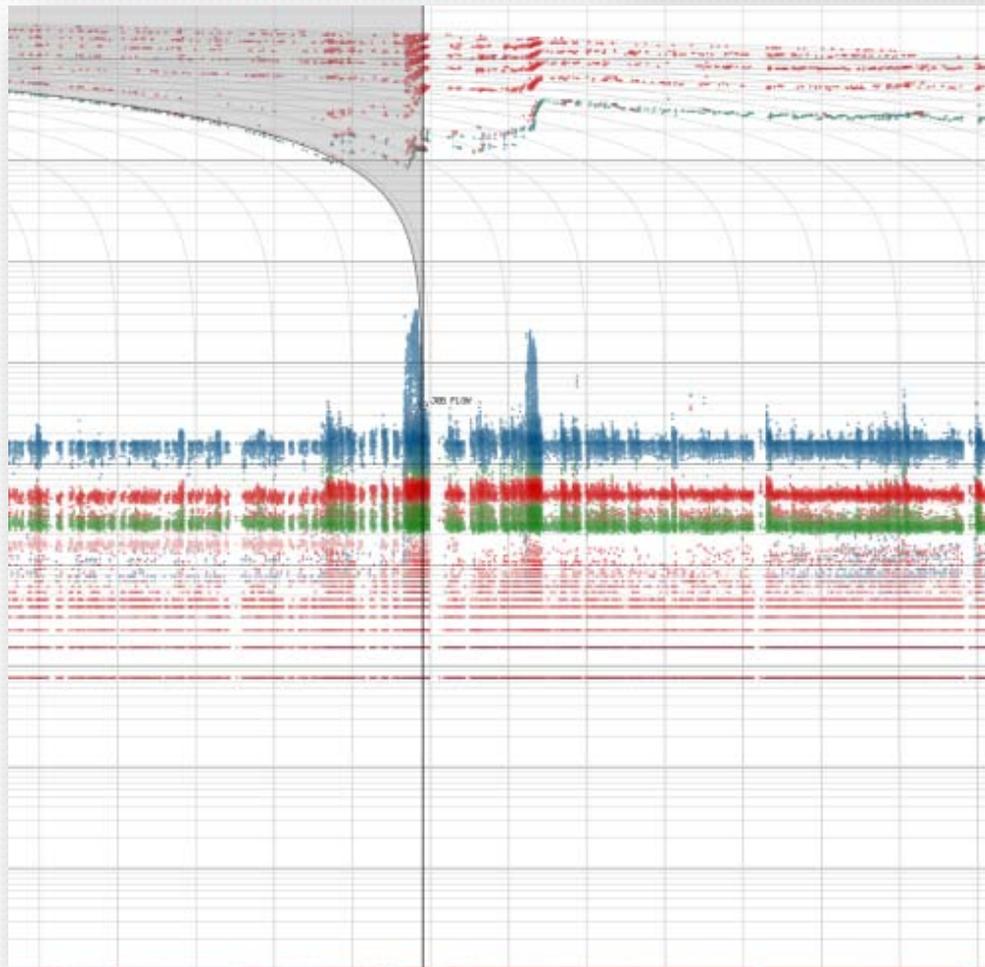
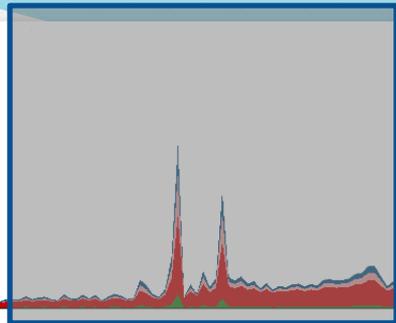
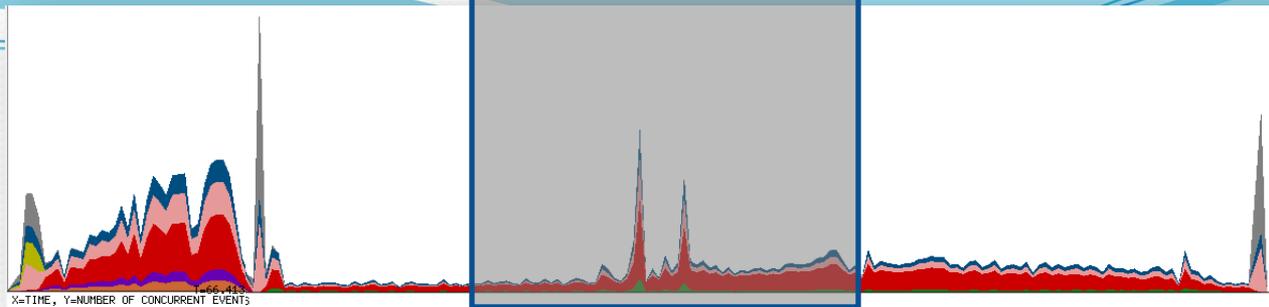
Examples



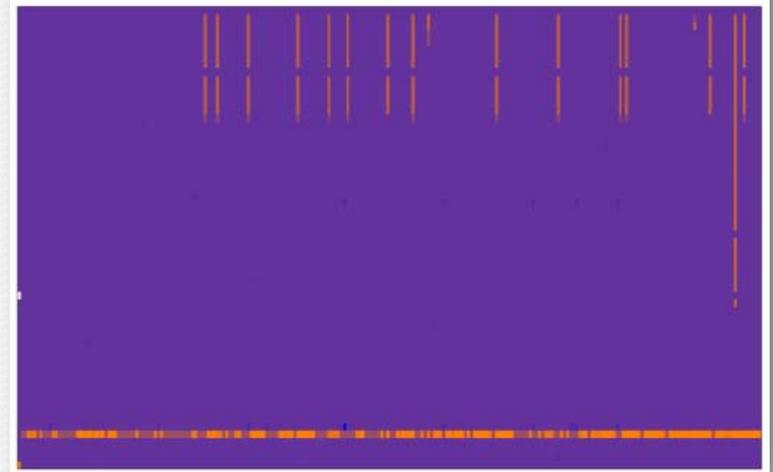
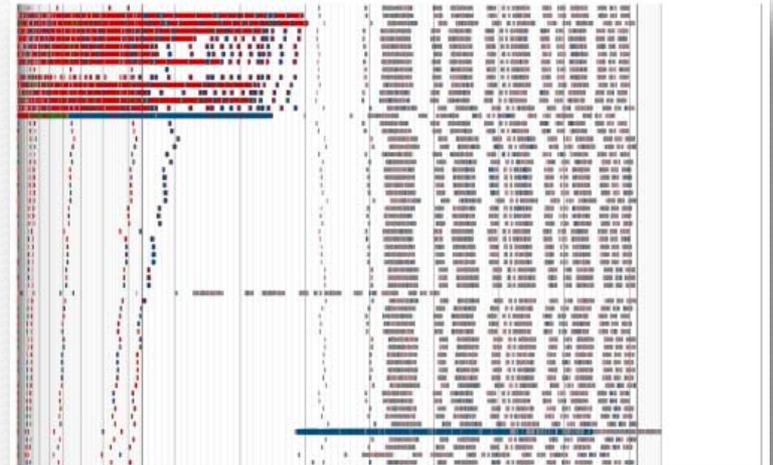
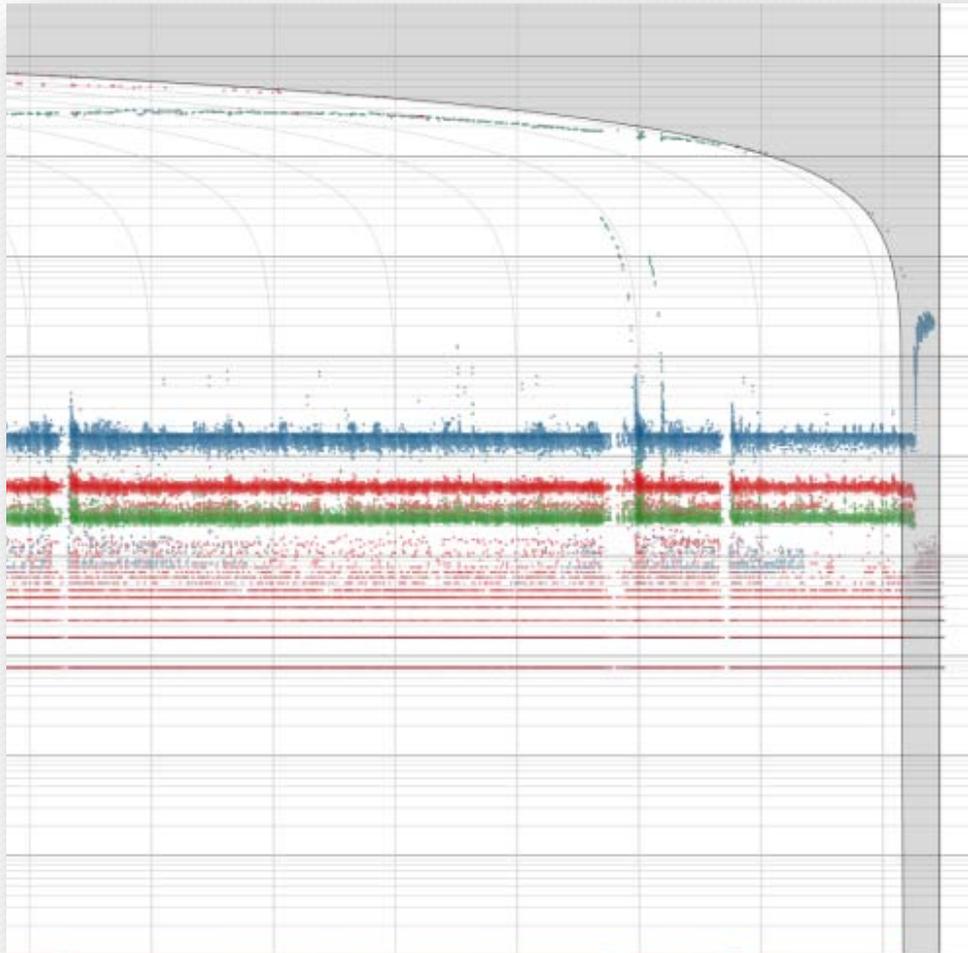
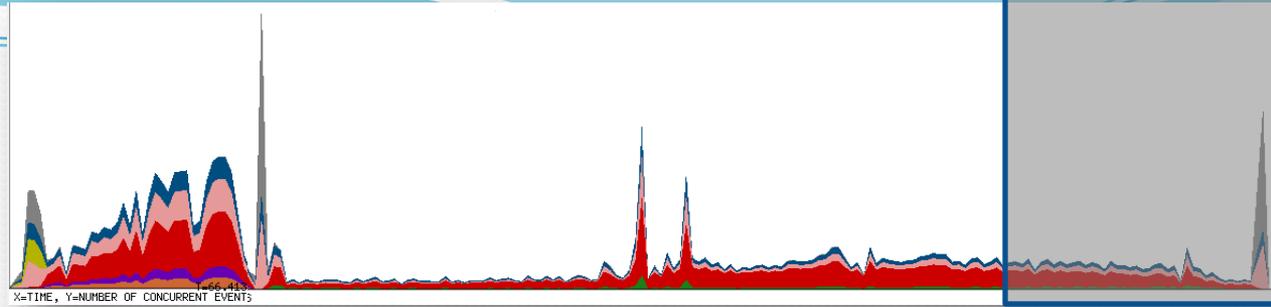
Examples



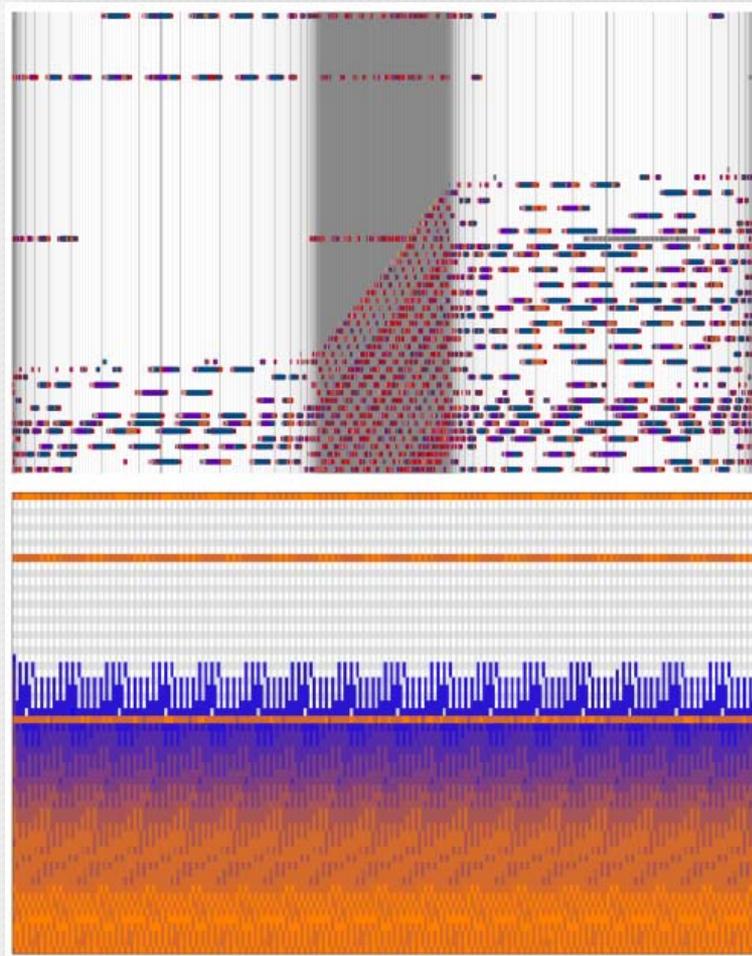
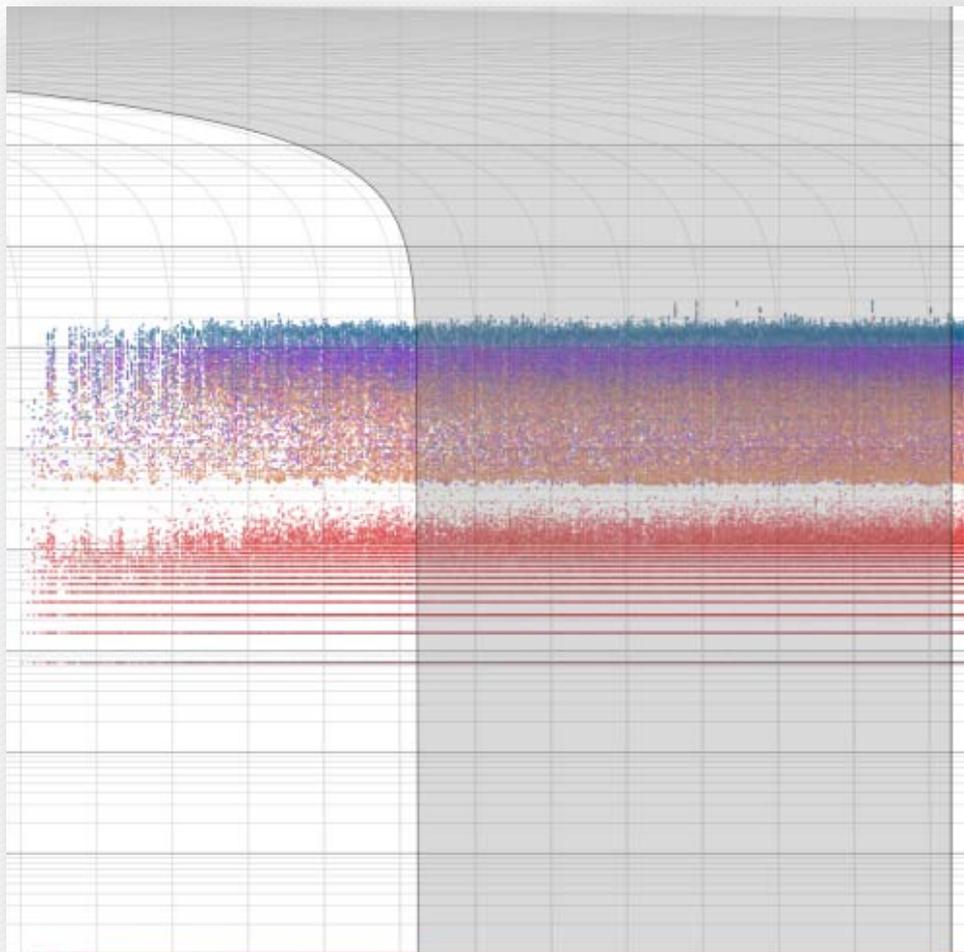
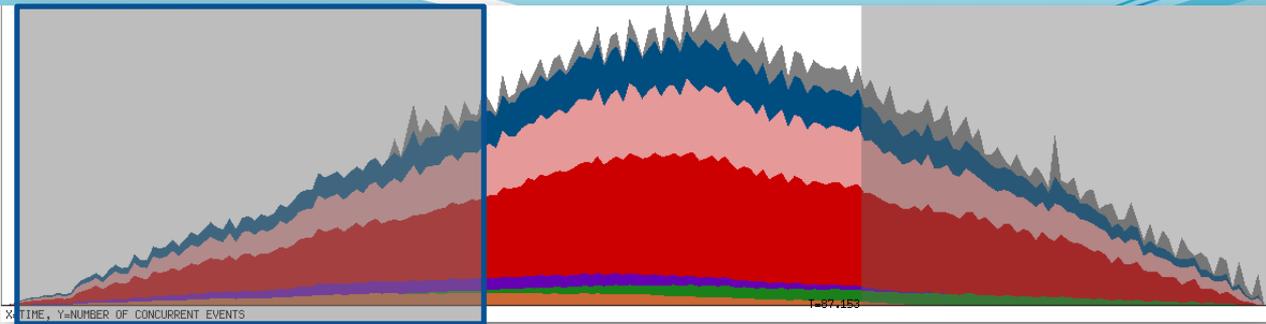
Examples



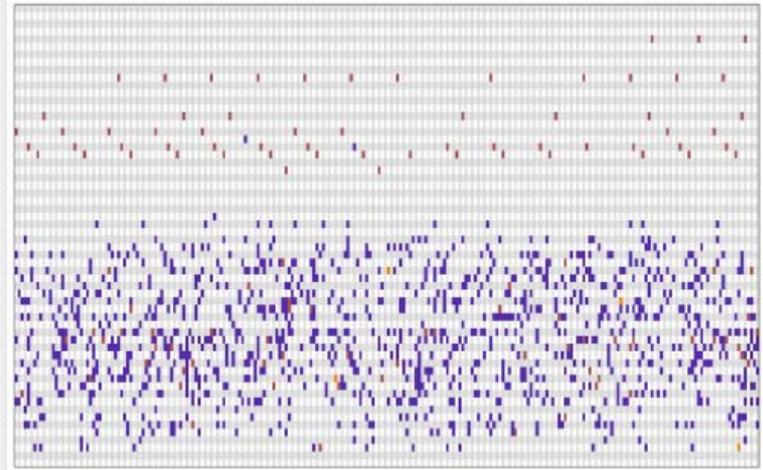
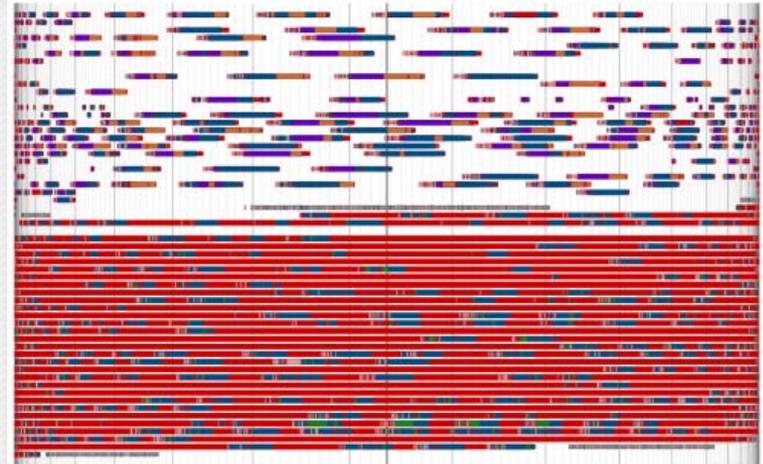
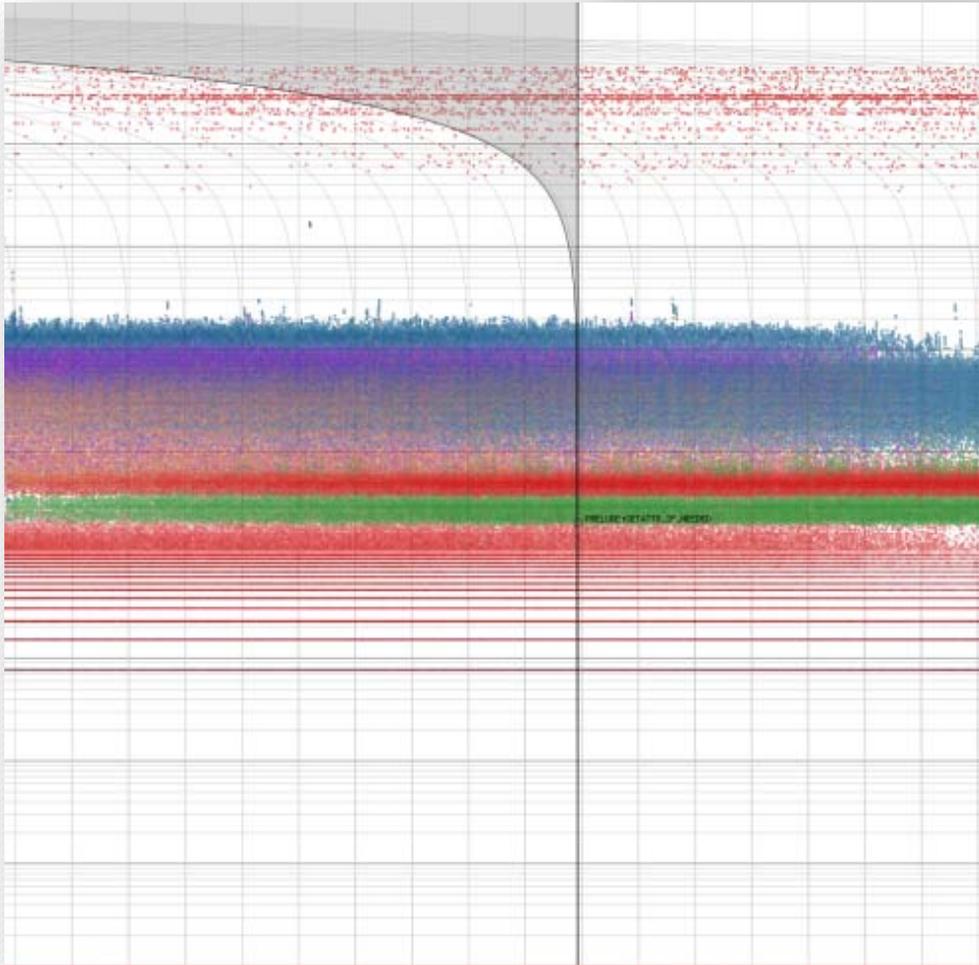
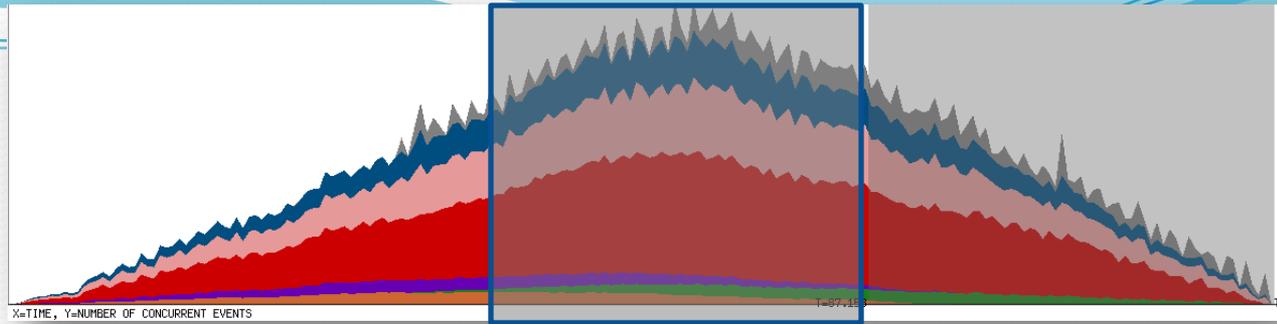
Examples



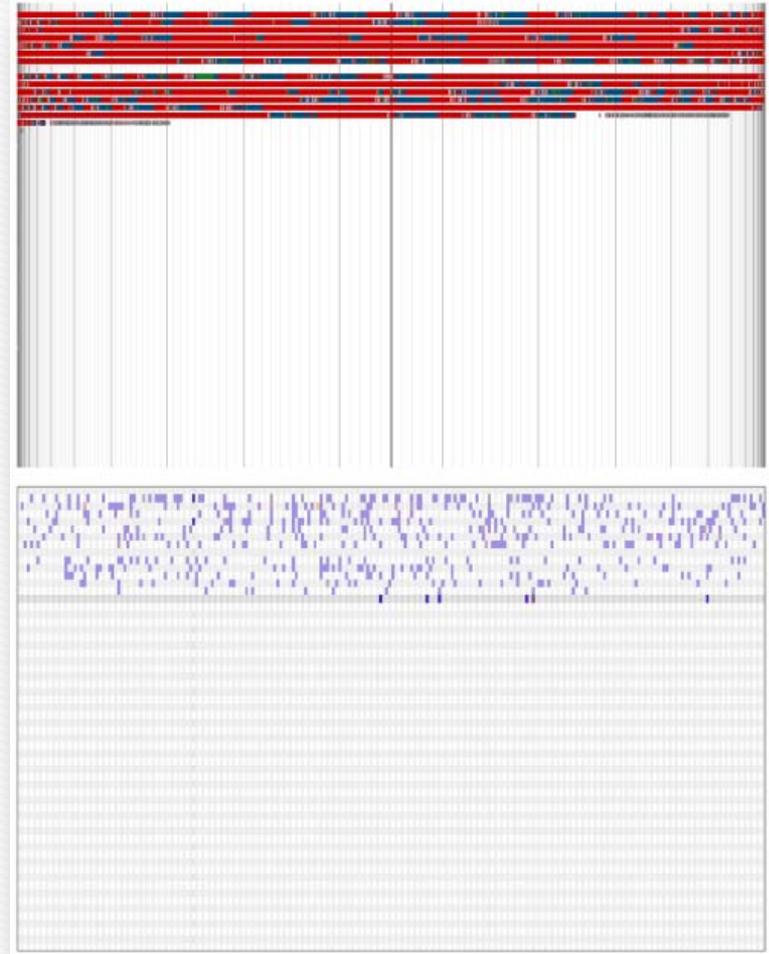
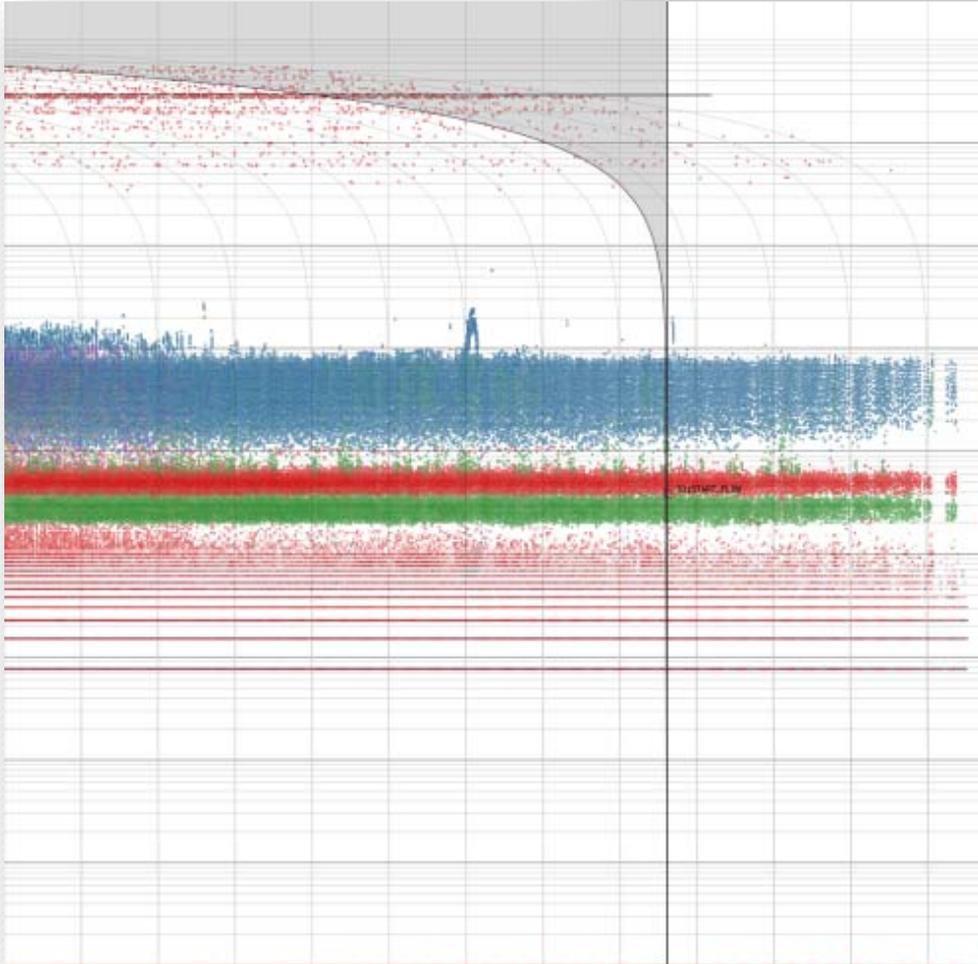
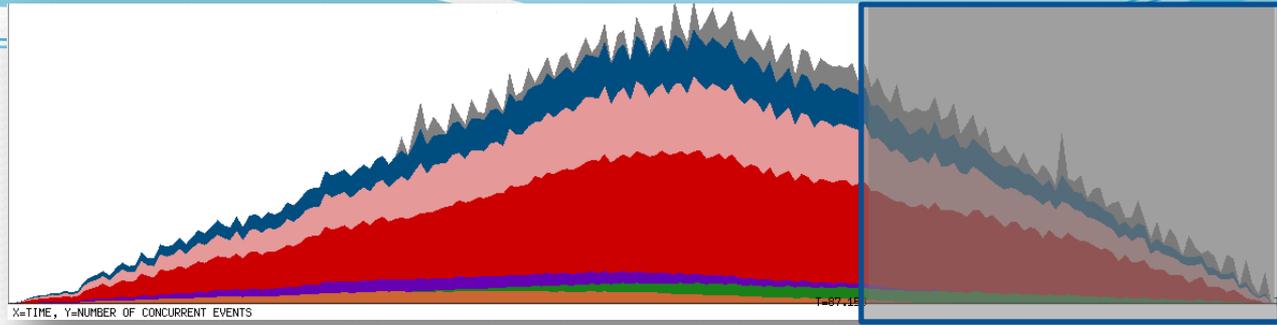
Examples



Examples



Examples



Observations

- ‘In order’ ramp up/down
 - More so on single file than FPP
- Centralized metadata
 - FS dependent
- Fairly consistent disc rate and network functions, but variable length enqueue
 - Limiting factor disc rate not network?

Next steps

- Out-of-core data management
 - Larger data sets
- Sigmoidal view variants
 - Gantt chart not scalable
- Other view variants
- MPI calls on compute nodes
 - Show end-to-end communication better
- Intermediate data servers
 - SSDs
 - Network and data flow



?