Xenalyze
Finding meaning in the chaos

George Dunlap

george.dunlap@eu.citrix.com

Citrix Systems, UK Ltd

Introduction

- Modern operating systems are complex
- Xentrace for gathering in-depth information
- Too much information
- Xenalyze

Talk goals

- Those for whom xenalyze is useful will use it
- Basic understanding of what xenalyze does, and what it’s useful for

Outline

- Overview of Xen tracing
- When xentrace is useful
- Core functionality of xenalyze
- Xenalyze as a platform
- Case studies

Xen tracing

- Trace records
  - Single 4-byte event number
  - Optional TSC timestamp
  - Optional trace-specific data, up to 28 bytes
- Event mask to control which events are logged
- Per-cpu trace buffers
- Buffers read by process in dom0, copied to disk

Xen tracing: What it's good for

- Key attributes
  - Lots of detailed data
  - Moderate cpu, disk overhead
  - Not persistent on host crash
- Understand both macro and micro effects
  - Performance analysis
  - Debugging
  - Understanding guest behavior
- Comparing to other techniques
  - printk
  - Xenoprof
  - Xen performance counters

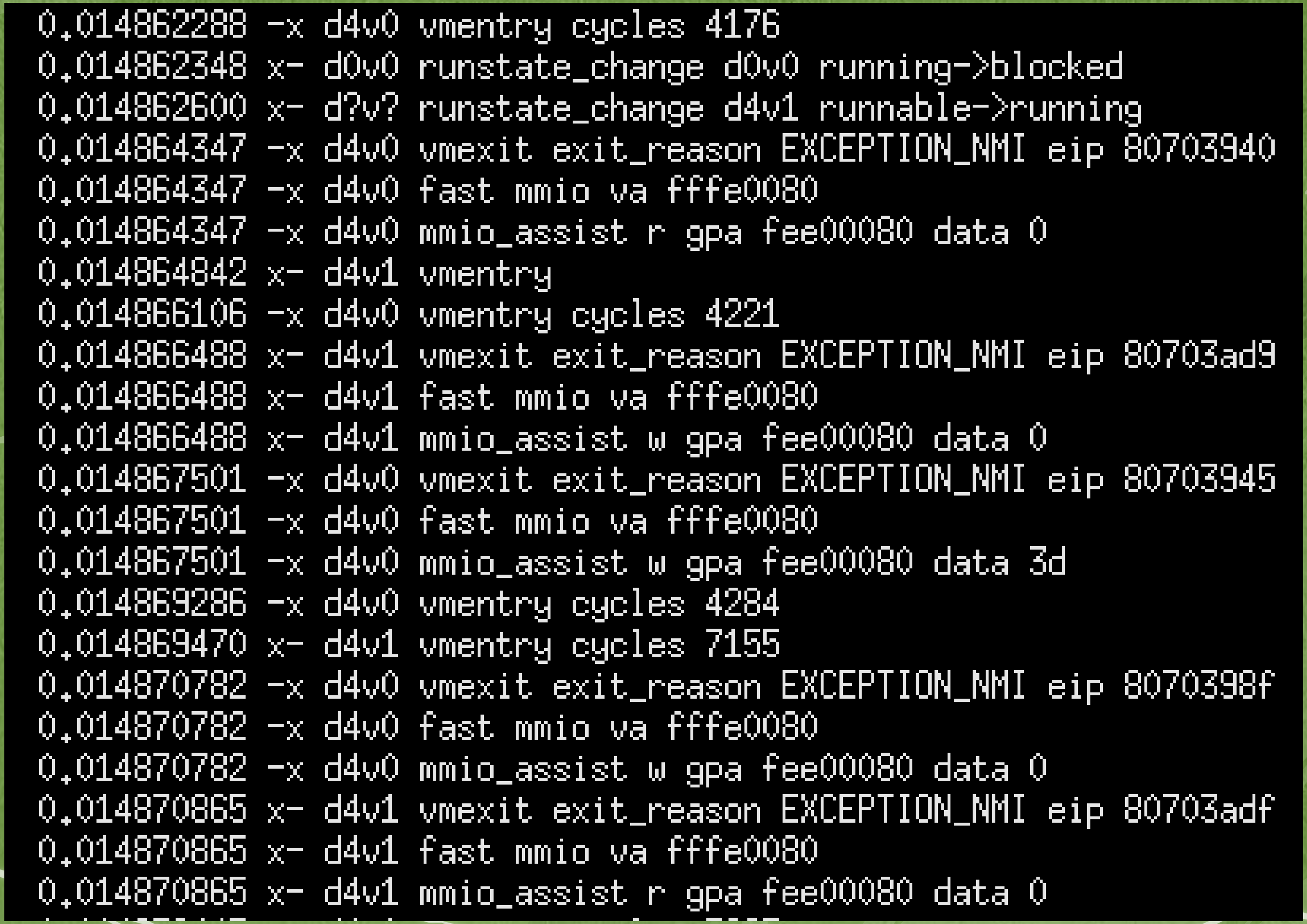
Key trace events

- Runstate change
  - Figure out who’s running where
  - Analyze how much time is spent blocked, preempted, waiting after wake, &c
- VMEXIT / VMENTER
  - How much time, and for what reason, we’re spending time in Xen

Xenalyze: Core functionality

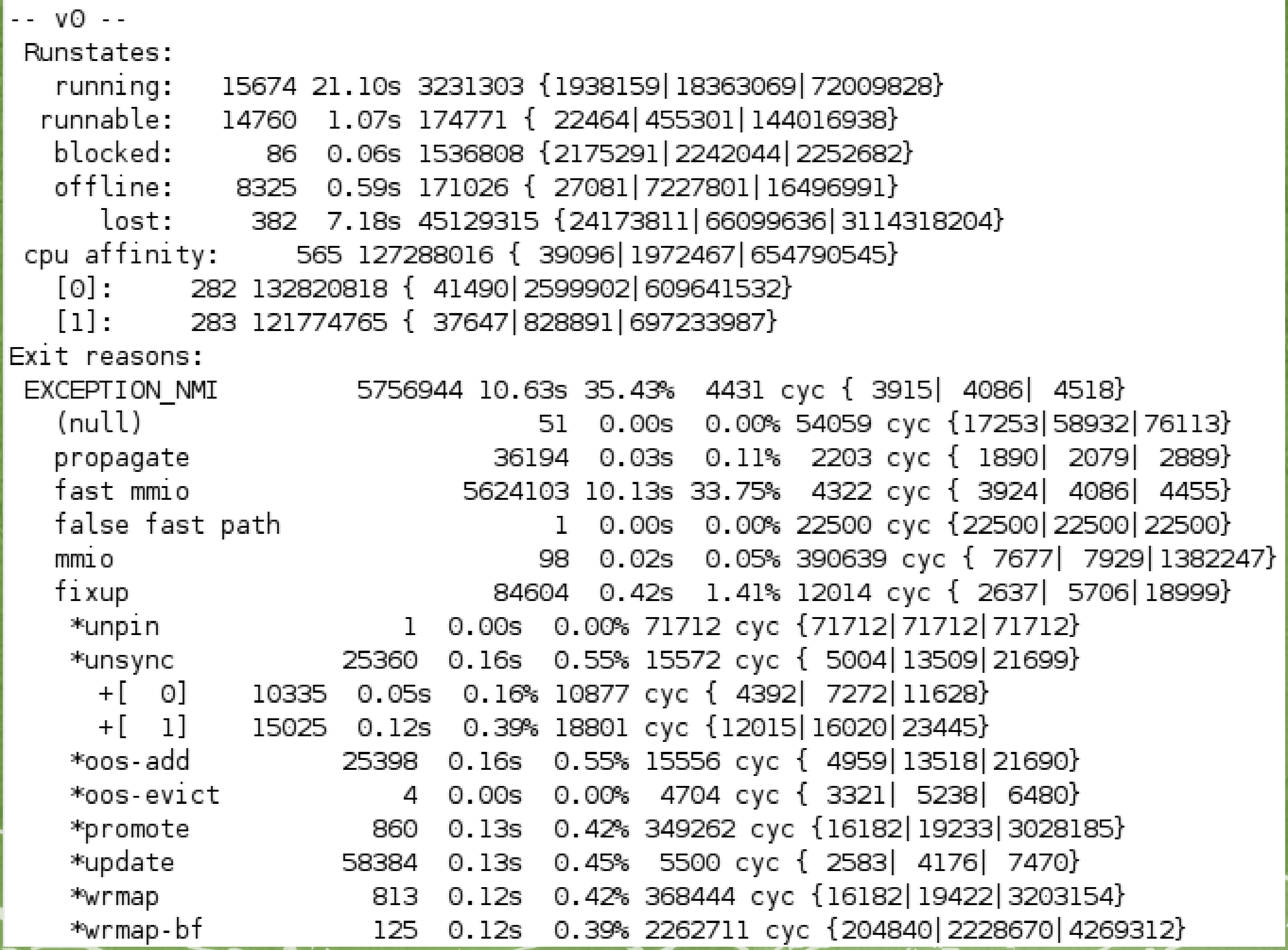
- Problem: xentrace file not in order
  - Attempt to process records in order
- Mapping small to large
  - Aggregate information to see larger trends
- Data is per-cpu, but we want per-vcpu
  - Track vcpus across physical cpus

Example output, dump mode



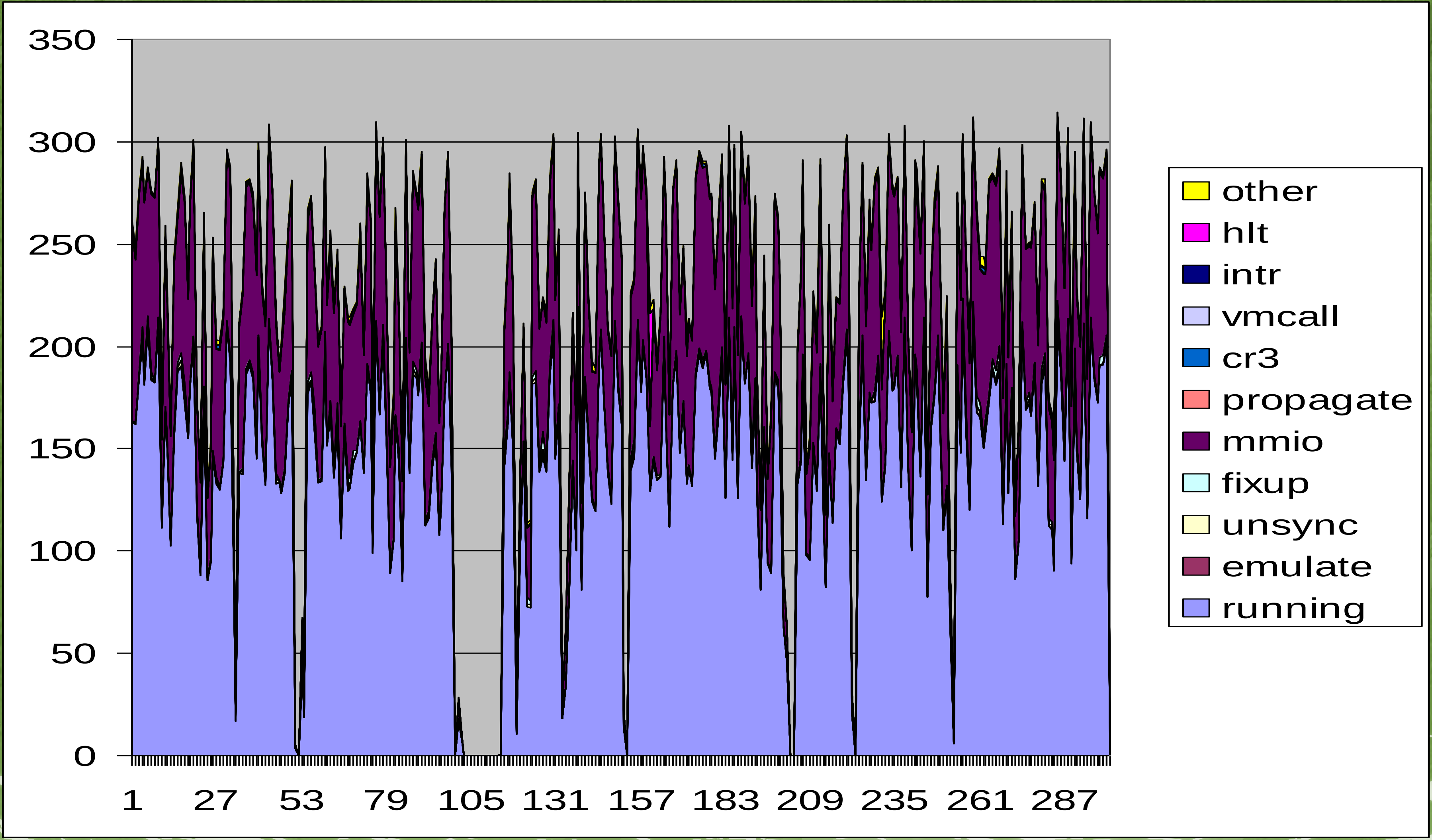
```
0.014862288 -x d4v0 vmentry cycles 4176
0.014862348 x- d0v0 runstate_change d0v0 running->blocked
0.014862600 x- d?v? runstate_change d4v1 runnable->running
0.014864347 -x d4v0 vmexit exit_reason EXCEPTION_NMI eip 80703940
0.014864347 -x d4v0 fast mmio va fffe0080
0.014864347 -x d4v0 mmio_assist r gpa fee00080 data 0
0.014864842 x- d4v1 vmentry
0.014866106 -x d4v0 vmentry cycles 4221
0.014866488 x- d4v1 vmexit exit_reason EXCEPTION_NMI eip 80703ad9
0.014866488 x- d4v1 fast mmio va fffe0080
0.014866488 x- d4v1 mmio_assist w gpa fee00080 data 0
0.014867501 -x d4v0 vmexit exit_reason EXCEPTION_NMI eip 80703945
0.014867501 -x d4v0 fast mmio va fffe0080
0.014867501 -x d4v0 mmio_assist w gpa fee00080 data 3d
0.014869286 -x d4v0 vmentry cycles 4284
0.014869470 x- d4v1 vmentry cycles 7155
0.014870782 -x d4v0 vmexit exit_reason EXCEPTION_NMI eip 8070398f
0.014870782 -x d4v0 fast mmio va fffe0080
0.014870782 -x d4v0 mmio_assist w gpa fee00080 data 0
0.014870865 x- d4v1 vmexit exit_reason EXCEPTION_NMI eip 80703adf
0.014870865 x- d4v1 fast mmio va fffe0080
0.014870865 x- d4v1 mmio_assist r gpa fee00080 data 0
```

Example output, Summary mode

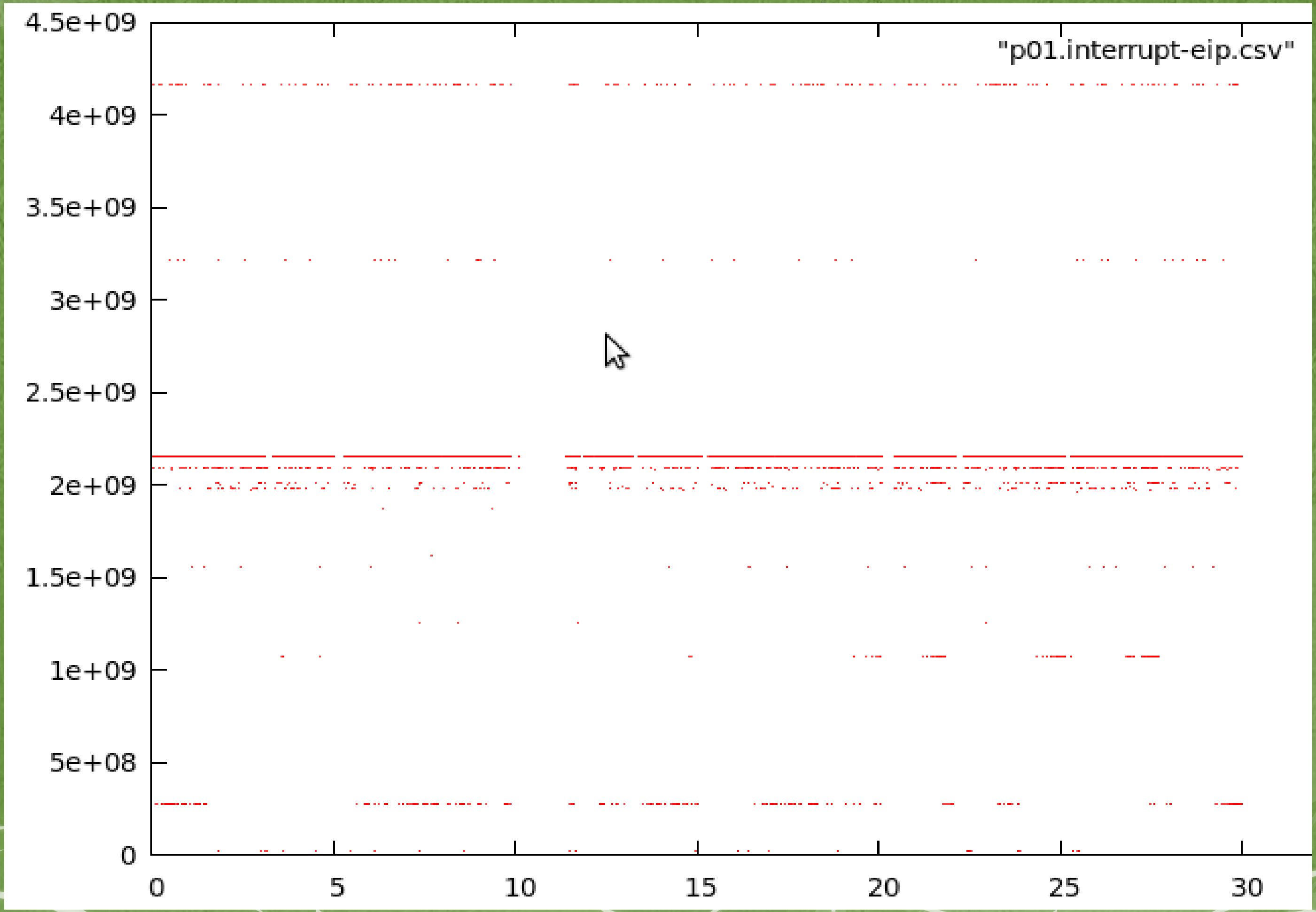


```
-- v0 --
 Runstates:
   running:   15674 21.10s 3231303 {1938159|18363069|72009828}
  runnable:   14760  1.07s 174771 { 22464|455301|144016938}
   blocked:      86  0.06s 1536808 {2175291|2242044|2252682}
   offline:    8325  0.59s 171026 { 27081|7227801|16496991}
      lost:     382  7.18s 45129315 {24173811|66099636|3114318204}
 cpu affinity:     565 127288016 { 39096|1972467|654790545}
   [0]:     282 132820818 { 41490|2599902|609641532}
   [1]:     283 121774765 { 37647|828891|697233987}
Exit reasons:
 EXCEPTION_NMI         5756944 10.63s 35.43%  4431 cyc { 3915|  4086|  4518}
   (null)                          51  0.00s  0.00% 54059 cyc {17253|58932|76113}
   propagate                    36194  0.03s  0.11%  2203 cyc { 1890|  2079|  2889}
   fast mmio                  5624103 10.13s 33.75%  4322 cyc { 3924|  4086|  4455}
   false fast path                  1  0.00s  0.00% 22500 cyc {22500|22500|22500}
   mmio                            98  0.02s  0.05% 390639 cyc { 7677|  7929|1382247}
   fixup                        84604  0.42s  1.41% 12014 cyc { 2637|  5706|18999}
    *unpin                1  0.00s  0.00% 71712 cyc {71712|71712|71712}
    *unsync           25360  0.16s  0.55% 15572 cyc { 5004|13509|21699}
      +[  0]    10335  0.05s  0.16% 10877 cyc { 4392|  7272|11628}
      +[  1]    15025  0.12s  0.39% 18801 cyc {12015|16020|23445}
    *oos-add          25398  0.16s  0.55% 15556 cyc { 4959|13518|21690}
    *oos-evict            4  0.00s  0.00%  4704 cyc { 3321|  5238|  6480}
    *promote            860  0.13s  0.42% 349262 cyc {16182|19233|3028185}
    *update           58384  0.13s  0.45%  5500 cyc { 2583|  4176|  7470}
    *wrmap              813  0.12s  0.42% 368444 cyc {16182|19422|3203154}
    *wrmap-bf           125  0.12s  0.39% 2262711 cyc {204840|2228670|4269312}
```

Example output, Interval



Example output, Scatterplot



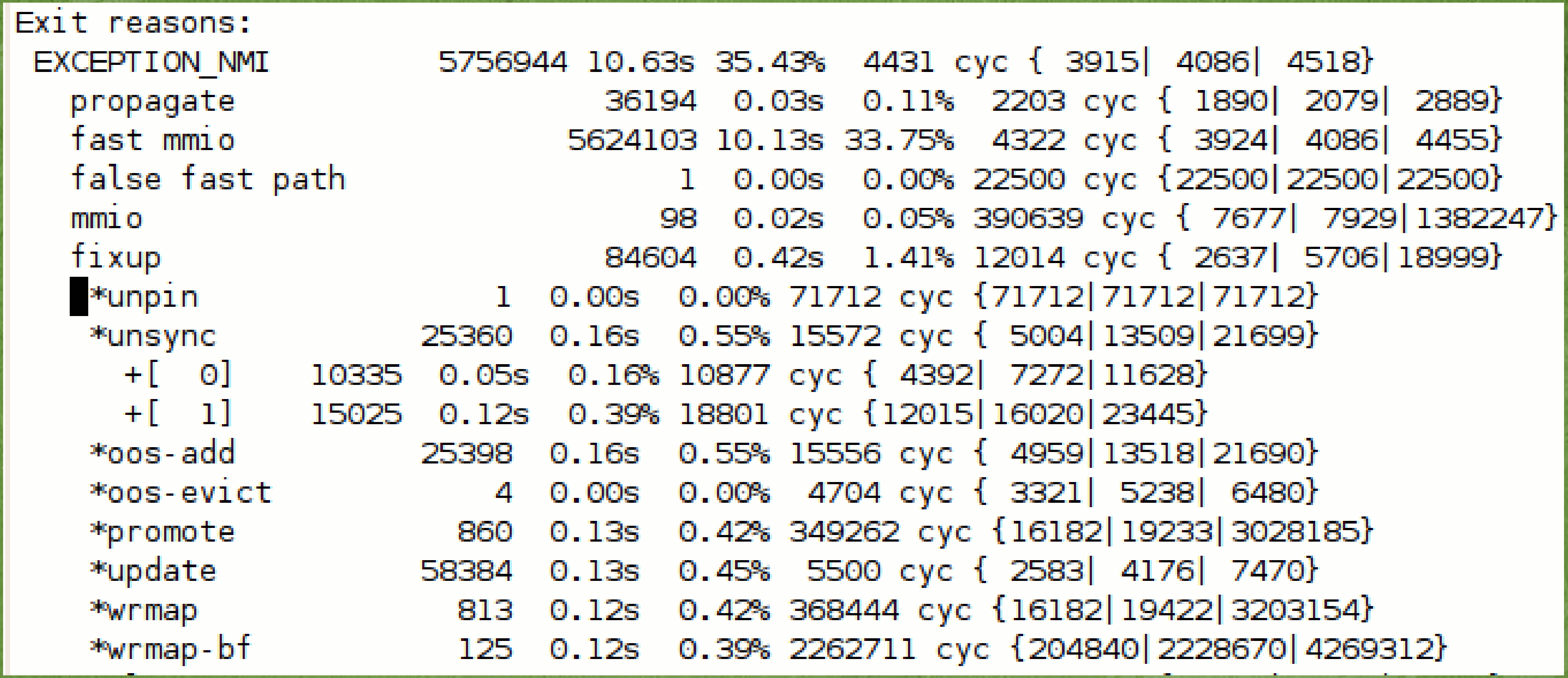
Advanced features

- “Enumeration” of MMIO, IO, addresses, and so on
- Symbol file translation
- Linear pagetable back-calculation
- Wake-to-halt, by interrupt
- …and many more

Platform for new analysis

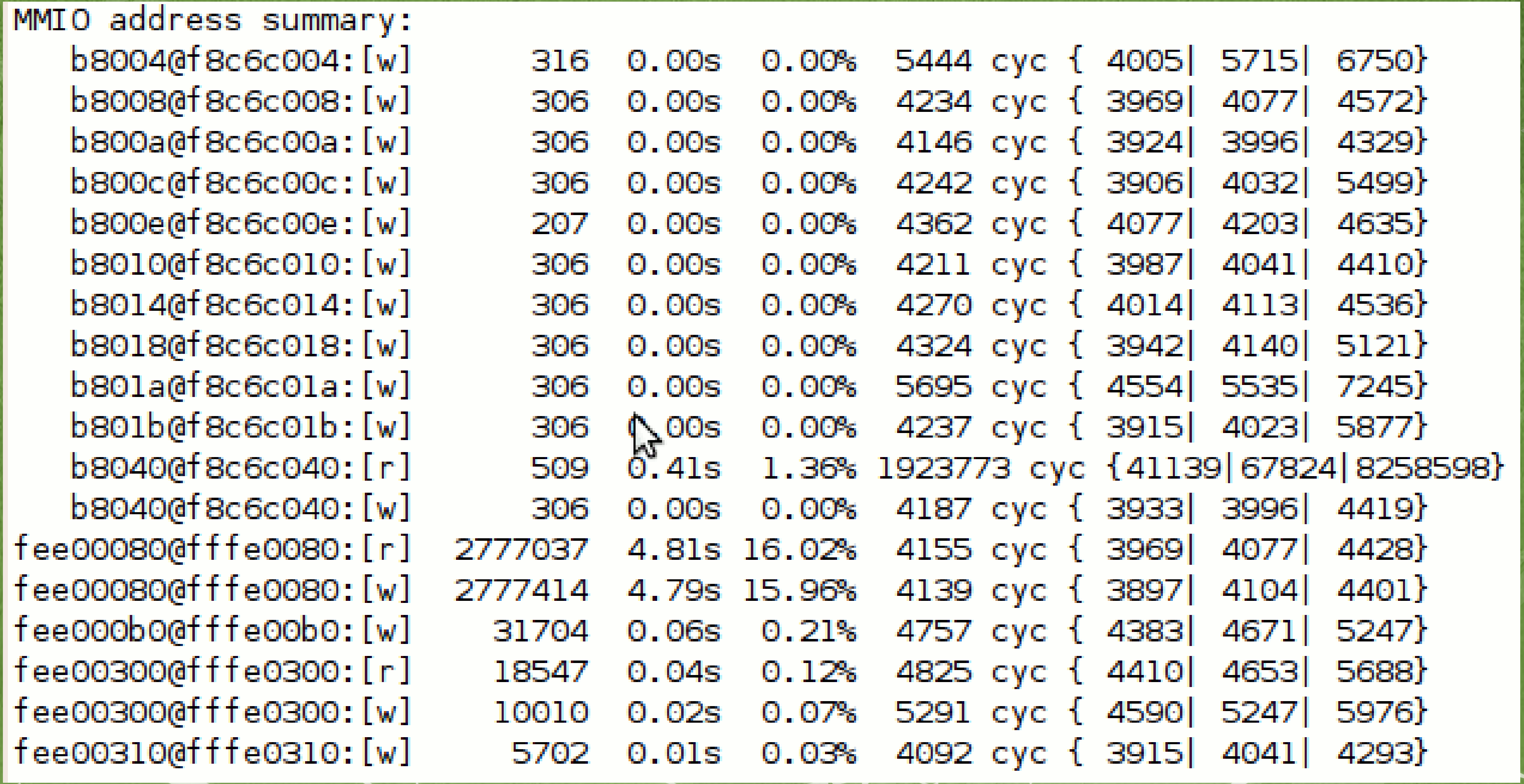
- Xenalyze may not be able to answer the questions you have
- But it’s a great platform to modify, because it’s already done a lot of the hard work for you

Case study: WinXP and TPR



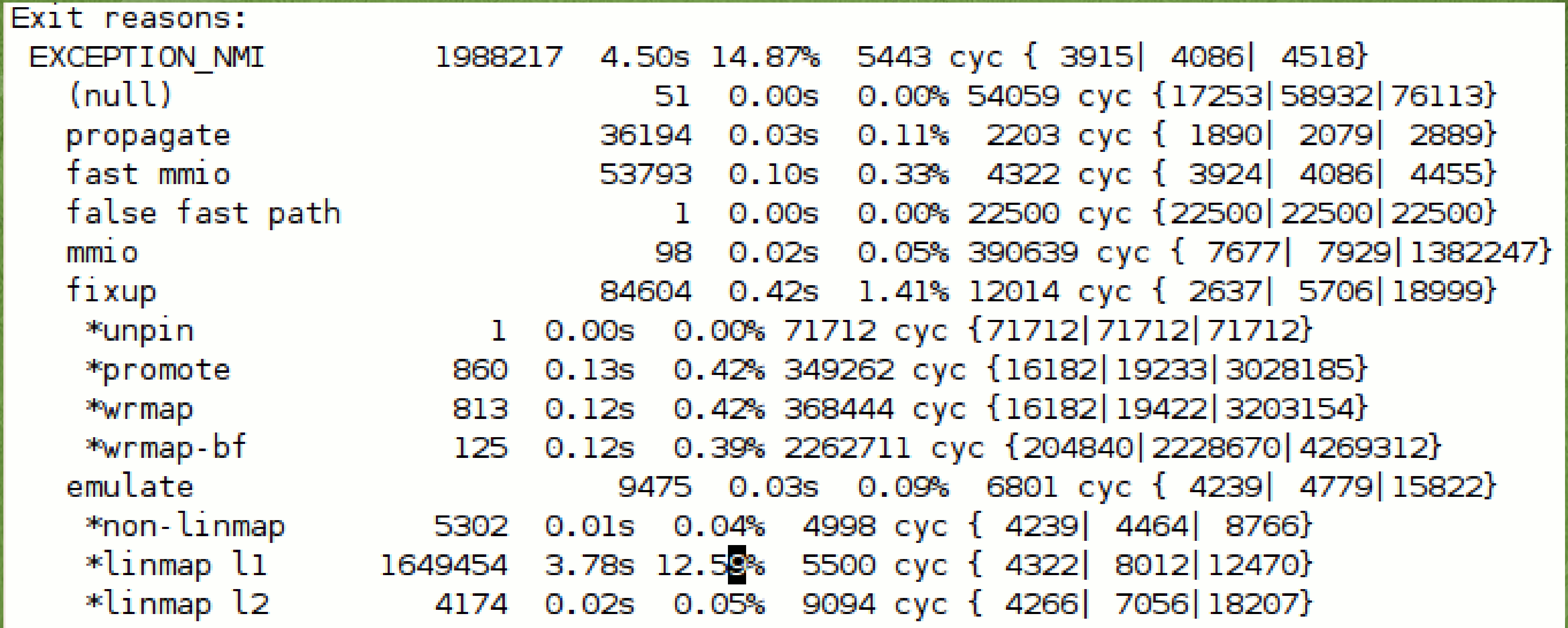
```
Exit reasons:
 EXCEPTION_NMI          5756944 10.63s 35.43%  4431 cyc { 3915| 4086| 4518}
   propagate                      36194  0.03s  0.11%  2203 cyc { 1890| 2079| 2889}
   fast mmio                    5624103 10.13s 33.75%  4322 cyc { 3924| 4086| 4455}
   false fast path                    1  0.00s  0.00% 22500 cyc {22500|22500|22500}
   mmio                              98  0.02s  0.05% 390639 cyc { 7677| 7929|1382247}
   fixup                          84604  0.42s  1.41% 12014 cyc { 2637| 5706|18999}
    *unpin                  1  0.00s  0.00% 71712 cyc {71712|71712|71712}
    *unsync             25360  0.16s  0.55% 15572 cyc { 5004|13509|21699}
      +[  0]    10335  0.05s  0.16% 10877 cyc { 4392| 7272|11628}
      +[  1]    15025  0.12s  0.39% 18801 cyc {12015|16020|23445}
    *oos-add            25398  0.16s  0.55% 15556 cyc { 4959|13518|21690}
    *oos-evict              4  0.00s  0.00%  4704 cyc { 3321| 5238| 6480}
    *promote              860  0.13s  0.42% 349262 cyc {16182|19233|3028185}
    *update             58384  0.13s  0.45%  5500 cyc { 2583| 4176| 7470}
    *wrmap                813  0.12s  0.42% 368444 cyc {16182|19422|3203154}
    *wrmap-bf             125  0.12s  0.39% 2262711 cyc {204840|2228670|4269312}
```

Case study: WinXP and TPR, cont



```
MMIO address summary:
   b8004@f8c6c004:[w]       316  0.00s   0.00%  5444 cyc { 4005| 5715|  6750}
   b8008@f8c6c008:[w]       306  0.00s   0.00%  4234 cyc { 3969| 4077|  4572}
   b800a@f8c6c00a:[w]       306  0.00s   0.00%  4146 cyc { 3924| 3996|  4329}
   b800c@f8c6c00c:[w]       306  0.00s   0.00%  4242 cyc { 3906| 4032|  5499}
   b800e@f8c6c00e:[w]       207  0.00s   0.00%  4362 cyc { 4077| 4203|  4635}
   b8010@f8c6c010:[w]       306  0.00s   0.00%  4211 cyc { 3987| 4041|  4410}
   b8014@f8c6c014:[w]       306  0.00s   0.00%  4270 cyc { 4014| 4113|  4536}
   b8018@f8c6c018:[w]       306  0.00s   0.00%  4324 cyc { 3942| 4140|  5121}
   b801a@f8c6c01a:[w]       306  0.00s   0.00%  5695 cyc { 4554| 5535|  7245}
   b801b@f8c6c01b:[w]       306  0.00s   0.00%  4237 cyc { 3915| 4023|  5877}
   b8040@f8c6c040:[r]       509  0.41s   1.36% 1923773 cyc {41139|67824|8258598}
   b8040@f8c6c040:[w]       306  0.00s   0.00%  4187 cyc { 3933| 3996|  4419}
fee00080@fffe0080:[r]   2777037  4.81s  16.02%  4155 cyc { 3969| 4077|  4428}
fee00080@fffe0080:[w]   2777414  4.79s  15.96%  4139 cyc { 3897| 4104|  4401}
fee000b0@fffe00b0:[w]     31704  0.06s   0.21%  4757 cyc { 4383| 4671|  5247}
fee00300@fffe0300:[r]     18547  0.04s   0.12%  4825 cyc { 4410| 4653|  5688}
fee00300@fffe0300:[w]     10010  0.02s   0.07%  5291 cyc { 4590| 5247|  5976}
fee00310@fffe0310:[w]      5702  0.01s   0.03%  4092 cyc { 3915| 4041|  4293}
```

Case study: Shadow Performance



```
Exit reasons:
 EXCEPTION_NMI         1988217  4.50s 14.87%  5443 cyc { 3915|  4086|  4518}
   (null)                          51  0.00s  0.00% 54059 cyc {17253|58932|76113}
   propagate                    36194  0.03s  0.11%  2203 cyc { 1890|  2079|  2889}
   fast mmio                    53793  0.10s  0.33%  4322 cyc { 3924|  4086|  4455}
   false fast path                  1  0.00s  0.00% 22500 cyc {22500|22500|22500}
   mmio                            98  0.02s  0.05% 390639 cyc { 7677|  7929|1382247}
   fixup                        84604  0.42s  1.41% 12014 cyc { 2637|  5706|18999}
    *unpin                1  0.00s  0.00% 71712 cyc {71712|71712|71712}
    *promote            860  0.13s  0.42% 349262 cyc {16182|19233|3028185}
    *wrmap              813  0.12s  0.42% 368444 cyc {16182|19422|3203154}
    *wrmap-bf           125  0.12s  0.39% 2262711 cyc {204840|2228670|4269312}
   emulate                       9475  0.03s  0.09%  6801 cyc { 4239|  4779|15822}
    *non-linmap        5302  0.01s  0.04%  4998 cyc { 4239|  4464|  8766}
    *linmap l1      1649454  3.78s 12.59%  5500 cyc { 4322|  8012|12470}
    *linmap l2         4174  0.02s  0.05%  9094 cyc { 4266|  7056|18207}
```

Case study: Shadow perf, con't

OS action	Sync	Out-of-sync
Page fault	Propagate	Propagate
Transition PTE	Emulate	
Real PTE	Emulate	
Access	(TLB miss)	Fix-up fault

Case study: Shadow perf, con't

OS action	Sync	Out-of-sync
Map PTE	Emulate	
Access	(TLB miss)	Fix-up fault
Unmap PTE	Emulate	

Outline

- Overview of Xen tracing
- When xentrace is useful
- Core functionality of xenalyze
- Xenalyze as a platform
- Case studies

Talk goals

- Those for whom xenalyze is useful will use it
- Basic understanding of what xenalyze does, and what it’s useful for

Questions

- Download now:

http://xenbits.xensource.com/ext/xenalyze