

Linux Perf Tools

Overview and Current Developments

Arnaldo Carvalho de Melo, Jiri Olsa

Red Hat Inc.

April 11, 2013

Overview

- Multiple events view
- Annotate GTK UI
- New 'perf mem' tool
- Per socket/core aggregation
- Diff enhancements
- Group leader sampling
- DWARF unwind
- Default precise
- Interval stat
- Toggling events
- Non architectural events
- The traceevent library
- The Queue

Multiple events without grouping

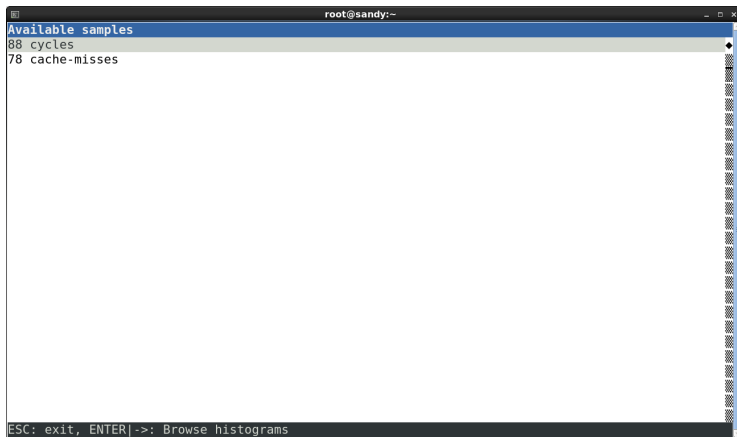
```
# perf record -e cycles,cache-misses -a usleep 1
[ perf record: Woken up 1 times to write data ]
[ perf record: Captured and wrote 0.616 MB perf.data (~26891 samples) ]
# perf evlist
cycles
cache-misses
#
```

Multiple events grouping

```
# perf record -e '{cycles,cache-misses}' -a usleep 1
[ perf record: Woken up 1 times to write data ]
[ perf record: Captured and wrote 0.621 MB perf.data (~27151 samples) ]
# perf evlist
cycles
cache-misses
# perf evlist --group
{cycles,cache-misses}
#
```

perf report - no grouping

perf report



The screenshot shows a terminal window titled 'root@sandy:~'. The output of the 'perf report' command is displayed. At the top, 'Available samples' is highlighted in blue. Below it, '88 cycles' and '78 cache-misses' are listed. The bottom of the window shows the prompt 'ESC: exit, ENTER|->: Browse histograms'.

```
root@sandy:~  
Available samples  
88 cycles  
78 cache-misses  
  
ESC: exit, ENTER|->: Browse histograms
```

perf report - single event

```
root@sandy:~  
Samples: 86 of event 'cycles', Event count (approx.): 4844145  
16.79%   swapper [kernel.kallsyms] [k] intel_idle  
7.01%     perf [kernel.kallsyms] [k] generic_exec_single  
5.57%   swapper [kernel.kallsyms] [k] try_to_wake_up  
3.95%   usleep [kernel.kallsyms] [k] avc_has_perm_noaudit  
3.75%   usleep [kernel.kallsyms] [k] __bitmap_weight  
3.70%   usleep libc-2.12.so      [.] __dl_addr  
3.65%   usleep ld-2.12.so       [.] __dl_check_all_versions  
3.59%   usleep ld-2.12.so       [.] __dl_map_object  
3.58%   qemu-kvm [kernel.kallsyms] [k] sys_timer_settime  
3.54%   qemu-kvm [kernel.kallsyms] [k] fget_light  
3.45%   usleep [kernel.kallsyms] [k] __do_page_fault  
3.44%   qemu-kvm [kernel.kallsyms] [k] user_exit  
3.36%   usleep [kernel.kallsyms] [k] flush_tlb_mm_range  
3.36%   usleep [kernel.kallsyms] [k] do_generic_file_read.clone.0  
3.28%   usleep [kernel.kallsyms] [k] __raw_spin_lock  
3.26%   usleep [kernel.kallsyms] [k] unlink_anon_vmas  
3.20%   usleep [kernel.kallsyms] [k] unmap_vmas  
3.05%   qemu-kvm libpthread-2.12.so [.] __sigaction  
2.80%   qemu-kvm [kernel.kallsyms] [k] __raw_spin_lock_irqsave  
2.47%   swapper [kernel.kallsyms] [k] call_function_single_interrupt  
2.26% plugin-containe libpthread-2.12.so [.] __pthread_enable_asynccancel  
2.21% plugin-containe [kernel.kallsyms] [k] cpuacct_charge  
2.05% plugin-containe [kernel.kallsyms] [k] do_prlimit  
1.78%   swapper [kernel.kallsyms] [k] tick_nohz_idle_exit  
0.89%   swapper [kernel.kallsyms] [k] menu_select  
0.83%   swapper [kernel.kallsyms] [k] cpuidle_idle_call  
Press '?' for help on key bindings
```

perf report - multiple events

perf report --group

```
root@sandy:~  
Samples: 155 of event 'anon group { cycles, cache-misses }', Event count (approx.): 4858423  
16.79% 19.71% swapper [kernel.kallsyms] [k] intel_idle  
7.01% 0.00% perf [kernel.kallsyms] [k] generic_exec_single  
5.57% 0.00% swapper [kernel.kallsyms] [k] try_to_wake_up  
3.95% 0.00% usleep [kernel.kallsyms] [k] avc_has_perm_noaudit  
3.75% 0.00% usleep [kernel.kallsyms] [k] __bitmap_weight  
3.70% 0.00% usleep libc-2.12.so [.] __dl_addr  
3.65% 0.00% usleep ld-2.12.so [.] __dl_check_all_versions  
3.59% 0.00% usleep ld-2.12.so [.] __dl_map_object  
3.58% 0.00% qemu-kvm [kernel.kallsyms] [k] sys_timer_settime  
3.54% 0.00% qemu-kvm [kernel.kallsyms] [k] fget_light  
3.45% 0.00% usleep [kernel.kallsyms] [k] __do_page_fault  
3.44% 0.00% qemu-kvm [kernel.kallsyms] [k] user_exit  
3.36% 0.00% usleep [kernel.kallsyms] [k] flush_tlb_mm_range  
3.36% 0.00% usleep [kernel.kallsyms] [k] do_generic_file_read.clone.0  
3.28% 2.92% usleep [kernel.kallsyms] [k] __raw_spin_lock  
3.26% 0.00% usleep [kernel.kallsyms] [k] unlink_anon_vmas  
3.20% 0.00% usleep [kernel.kallsyms] [k] unmap_vmas  
3.05% 0.00% qemu-kvm libpthread-2.12.so [.] __sigaction  
2.80% 0.00% qemu-kvm [kernel.kallsyms] [k] __raw_spin_lock_irqsave  
2.47% 0.00% swapper [kernel.kallsyms] [k] call_function_single_interrupt  
2.26% 0.00% plugin-containe libpthread-2.12.so [.] __pthread_enable_asynccancel  
2.21% 0.00% plugin-containe [kernel.kallsyms] [k] cpuacct_charge  
2.05% 0.00% plugin-containe [kernel.kallsyms] [k] do_prlimit  
1.78% 0.00% swapper [kernel.kallsyms] [k] tick_nohz_idle_exit  
0.89% 0.00% swapper [kernel.kallsyms] [k] menu_select  
0.83% 0.00% swapper [kernel.kallsyms] [k] cpuidle_idle_call  
Press '?' for help on key bindings
```

Diff enhancements

- compare methods: delta, weighted diff, ratio
- multiple data files

Diff enhancements - basics

```
12.62% _raw_spin_lock_irqsave
3.44%  mutex_unlock
2.28%  __wake_up
2.09%  fget_light
2.06%  n_tty_write
1.74%  system_call
1.71%  pty_write
1.39%  enqueue_entity
1.38%  vfs_write
1.37%  __srcu_read_lock
```

DATA 1

DATA 1/2 INTERSECTION

```
_raw_spin_lock_irqsave
mutex_unlock
n_tty_write
pty_write
__srcu_read_lock
```

PAIRS

```
10.30% _raw_spin_lock_irqsave
2.83%  native_write_msr_safe
2.70%  n_tty_write
2.49%  tty_write
2.31%  update_curr
2.06%  __schedule
2.00%  mutex_unlock
1.61%  try_to_wake_up
1.54%  __srcu_read_lock
1.30%  pty_write
```

DATA 2

PAIR DATA:

COUNT 1 SYMBOL COUNT FOR DATA 1
COUNT 1 TOTAL TOTAL COUNT FOR DATA 1

COUNT 2 SYMBOL COUNT FOR DATA 2
COUNT 2 TOTAL TOTAL COUNT FOR DATA 2

COMPUTE

**DELTA
WEIGHTED DIFF
RATIO**

Diff enhancements - delta

%1 = (COUNT1 * 100) / COUNT1 TOTAL

%2 = (COUNT2 * 100) / COUNT2 TOTAL

DELTA = %2 - %1

```
$ perf diff -c delta
# Event 'cycles'
#
# Baseline      Delta      Shared Object      Symbol
# .....
#
# 12.62%    -2.32% [kernel.kallsyms] [k] _raw_spin_lock_irqsave
# 3.44%     -1.44% [kernel.kallsyms] [k] mutex_unlock
# 2.06%     +0.64% [kernel.kallsyms] [k] n_tty_write
# 1.71%     -0.42% [kernel.kallsyms] [k] pty_write
# 1.37%     +0.17% [kernel.kallsyms] [k] __srcu_read_lock
...
```

Diff enhancements - weighted diff

WEIGHT1 = USER DEFINED

WEIGHT2 = USER DEFINED

WEIGHTED DIFF = COUNT2 * WEIGHT1 – COUNT1 * WEIGHT2

```
$ perf diff -c wdiff:1,2
# Event 'cycles'
#
# Baseline   Weighted diff   Shared Object   Symbol
# .....
#
12.62%      100376692 [kernel.kallsyms] [k] _raw_spin_lock_irqsave
3.44%       17128216 [kernel.kallsyms] [k] mutex_unlock
2.06%       29267199 [kernel.kallsyms] [k] n_tty_write
1.71%       12346582 [kernel.kallsyms] [k] pty_write
1.37%       16196601 [kernel.kallsyms] [k] __srcu_read_lock
...
```

Diff enhancements - ratio

RATIO = COUNT2 / COUNT1

```
$ perf diff -c ratio
# Event 'cycles'
#
# Baseline      Ratio      Shared Object      Symbol
# .....
#
12.62%      2.168020 [kernel.kallsyms] [k] _raw_spin_lock_irqsave
3.44%      1.542882 [kernel.kallsyms] [k] mutex_unlock
2.06%      3.477702 [kernel.kallsyms] [k] n_tty_write
1.71%      2.010982 [kernel.kallsyms] [k] pty_write
1.37%      2.986062 [kernel.kallsyms] [k] __srcu_read_lock
...
```

Diff enhancements - multiple data files

```
12.62% _raw_spin_lock_irqsave
3.44% mutex_unlock
2.28% wake_up_tty_insert_flip_string_fixed_flag
2.09% fget_light_queue
2.06% n_tty_write
1.74% system_call_strlen
1.71% pty_write_tty_write
1.39% enqueue_entity
1.38% vfs_write_process
1.37% _srcu_read_lock
0.05% s 0.89% cuTrent
0.05% s 0.87% __schedule
0.05% i 0.86% put_page
0.05% v 0.86% update_curr
0.05% v 0.85% _IO_file_write@@GLIBC.2.2.5
0.75% ttwu_stat
0.71% _srcu_read_unlock
0.70% _IO_file_write@@GLIBC.2.2.5
0.69% _tty_buffer_request_room
0.67% native_read_tsc
0.66% set_next_entity
0.65% memset
0.65% _IO_file_overflow@@GLIBC.2.2.5
0.64% enqueue_task
0.64% sched_clock_cpu
0.63% wake_up_common
0.63% tty_flip_buffer_push
0.63% native_sched_clock
0.62% do_output
0.61% ttwu_do_wait
10.30% _raw_spin_lock_irqsave
2.83% native_write_msr_safe
2.70% n_tty_write
2.49% tty_write
2.31% update_curr
2.06% schedule
2.00% mutex_unlock
1.61% try_to_wake_up
1.54% _srcu_read_lock
1.30% pty_write
1.30% _IO_file_write@@GLIBC.2.2.5
1.30% tty_buffer_request_room
1.30% native_write_msr_safe
2.70% n_tty_write
2.49% tty_write
2.06% schedule
2.00% mutex_unlock
0.61% try_to_wake_up
0.61% _srcu_read_lock
0.61% pty_write
0.09% task_waking_fair
0.09% ctx_sched_out
```

PAIRS

**DELTA
WEIGHTED DIFF
RATIO**

Diff enhancements - multiple data files - example

```
$ perf diff -b ./perf.data.[123456]
# Event 'cycles'
#
# Data files:
# [0] ./perf.data.1 (Baseline)
# [1] ./perf.data.2
# [2] ./perf.data.3
# [3] ./perf.data.4
# [4] ./perf.data.5
# [5] ./perf.data.6
#
# Baseline/0  Delta/1  Delta/2  Delta/3  Delta/4  Delta/5  Shared Object  Symbol
# .....
#
36.44% +0.27% +7.81% +1.18% +0.72% +0.74% libc-2.15.so  [.] _IO_file_xsputn@@GLIBC_2.2.5
32.70% -2.74% -12.76% -0.90% -2.16% -1.11% yes        [.] 0x0000000000000140b
15.01% +1.75% +0.50% +1.03% +1.80% +0.13% libc-2.15.so  [.] __strlen_sse2
14.88% +0.45% +4.45% -1.38% -0.64% +0.11% libc-2.15.so  [.] fputc_unlocked
0.25% +0.31% -0.08% +0.04% +0.33% +0.03% yes        [.] fputc_unlocked@plt
0.11% -0.05% -0.02% -0.05% -0.06% -0.06% [kernel.kallsyms] [k] _sru_read_lock
0.06% -0.03% -0.05% -0.05% -0.04% -0.02% [kernel.kallsyms] [k] fget_light
0.05% -0.03% -0.02% -0.02% -0.02% -0.02% [kernel.kallsyms] [k] native_write_msr_safe
0.05% -0.01% -0.01% +0.01% +0.06% -0.02% [kernel.kallsyms] [k] system_call
0.05% +0.02% -0.02% -0.03% -0.03% -0.03% [kernel.kallsyms] [k] __audit_syscall_exit
0.05% -0.04% -0.03% -0.02% -0.02% -0.02% [kernel.kallsyms] [k] sysret_check
0.03% -0.02% -0.02% -0.02% -0.02% -0.02% libc-2.15.so  [.] _IO_file_overflow@@GLIBC_2.2.5
0.03% -0.02% -0.02% -0.02% -0.02% -0.02% [kernel.kallsyms] [k] security_file_permission
0.03% -0.02% -0.02% -0.02% -0.02% -0.02% [kernel.kallsyms] [k] fsnotify
0.03% -0.02% -0.02% -0.02% -0.02% -0.02% [kernel.kallsyms] [k] __audit_syscall_entry
...
```

Diff enhancements - multiple data files - example

```
$ perf diff -b -o 1 -c ratio /perf.data.{1234}
# Event 'cycles'
#
# Data files:
# [0] ./perf.data.1 (Baseline)
# [1] ./perf.data.2
# [2] ./perf.data.3
# [3] ./perf.data.4
#
# Baseline/0      Ratio/1      Ratio/2      Ratio/3      Shared Object      Symbol
# .....
#
0.25%      2.911074      0.996950      2.366541      yes                [.] fputs_unlocked@plt
0.03%      2.033060      3.077690      2.589222      libc-2.15.so       [.] __GI__libc_write
0.02%      2.024182      2.978838      1.005138      libc-2.15.so       [.] new_do_write
15.02%     1.447379      1.487681      2.198029      libc-2.15.so       [.] __strlen_sse2
14.88%     1.335010      1.870979      1.865701      libc-2.15.so       [.] fputs_unlocked
36.44%     1.305951      1.748771      2.123142      libc-2.15.so       [.] _IO_file_xsputn@@GLIBC_2.2.5
32.71%     1.187609      0.878003      1.999755      yes                [.] 0x0000000000000140b
0.01%      1.057943      1.055648      [kernel.kallsyms] [k] unroll_tree_refs
0.05%      1.023360      1.992786      2.052029      [kernel.kallsyms] [k] __audit_syscall_exit
0.03%      1.019583      1.006175      1.547110      [kernel.kallsyms] [k] __audit_syscall_entry
0.05%      0.992093      0.995756      2.688763      [kernel.kallsyms] [k] system_call
0.03%      0.986661      0.502363      2.520055      [kernel.kallsyms] [k] fsnotify
0.06%      0.750137      0.256142      0.257698      [kernel.kallsyms] [k] fget_light
0.11%      0.722499      1.139460      1.172109      [kernel.kallsyms] [k] srcu_read_lock
0.03%      0.507413      1.008720      0.507093      libc-2.15.so       [.] _IO_file_overflow@@GLIBC_2.2.5
...

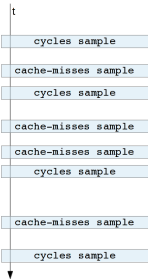
```

Group leader sampling

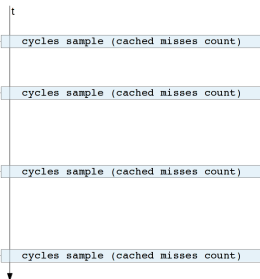
- ① leader sampling
- ② :S modifier
- ③ -e cycles:S
- ④ -e {cycles,cache-misses}:S
- ⑤ attach rest of the group data to sample
- ⑥ report group view by Namhyung Kim

Group leader sampling

```
-e '{cycles,caches-misses}'
```



```
-e '{cycles,caches-misses}:S'
```



Group leader sampling - example

```
$ perf record -e '{cycles,cache-misses}:S' yes > /dev/null
^C( perf record: Woken up 3 times to write data )
[ perf record: Captured and wrote 0.692 MB perf.data (-30242 samples) ]
yes: Interrupt
$ perf report --group --show-total-period --stdio
# group: {cycles,cache-misses}
# =====
#
# Samples: 23K of event 'anon group ( cycles, cache-misses )'
# Event count (approx.): 15458572434
#
#
```

Overhead			Period	Command	Shared Object	Symbol
32.52%	10.06%	5027754214	500	yes	libc-2.14.90.so	[.] _IO_file_xsputn@@GLIBC_2.2.5
18.75%	5.84%	2898593210	290	yes	yes	[.] main
16.84%	0.60%	2603347064	30	yes	libc-2.14.90.so	[.] __strlen_sse2
15.39%	8.90%	2378987098	442	yes	libc-2.14.90.so	[.] fputs_unlocked
3.50%	3.50%	540291244	174	yes	yes	[.] fputs_unlocked@plt
2.14%	0.02%	331186974	1	yes	[kernel.kallsyms]	[k] __lock_acquire
0.98%	0.18%	150828592	9	yes	[kernel.kallsyms]	[k] sched_clock_local
0.96%	0.32%	148200260	16	yes	[kernel.kallsyms]	[k] debug_smp_processor_id
0.81%	23.27%	125803028	1156	yes	[kernel.kallsyms]	[k] lock_release
0.76%	0.00%	117272260	0	yes	[kernel.kallsyms]	[k] native_sched_clock
0.55%	0.00%	84947064	0	yes	[kernel.kallsyms]	[k] intel_pmu_disable_all
0.49%	0.02%	76024090	1	yes	[kernel.kallsyms]	[k] perf_event_task_tick
0.48%	0.00%	73442096	0	yes	[kernel.kallsyms]	[k] local_clock
0.38%	0.00%	57982898	0	yes	[kernel.kallsyms]	[k] lock_acquired
0.33%	0.56%	51541448	28	yes	[kernel.kallsyms]	[k] lock_acquire
0.33%	14.81%	51096950	736	yes	[kernel.kallsyms]	[k] lock_release_holdtime.part.20
0.31%	0.00%	47667534	0	yes	[kernel.kallsyms]	[k] mark_lock
...						

- ① libunwind support
- ② not ditros favorite lib (Fedora/RHEL)
- ③ elfutils remote DWARF unwind support by Jan Kratochvil, pending review
- ④ testable perf support ready

Default precise

- ① different level of precise in CPUs
- ② sysfs precise level export
- ③ enable the precise event by default

- 1 -l 'interval' option

Toggling events

- ❶ want to have feature
- ❷ configure event to trigger another event
- ❸ initial patchset sent by Frederic Weisbecker

Non architectural events

- ① specify non-architectural events
- ② libpfm support
- ③ alias support

The traceevent library

1

The queue

- Kconfig
- Others - fill in till presentation day!

That is all folks!

Thanks!

Arnaldo Carvalho de Melo

acme@infradead.org

acme@redhat.com

Jiri Olsa - jolsa@redhat.com

linux-perf-users@vger.kernel.org