BloodTest
快速检测 Linux 性能问题思路以及简单实现

朱辉
teawater@126.com
teawater.github.io

现有工具和问题

- Perf, Ftrace, Systemtap, KGTP， proc 目录导出数据， sysfs 导出数据。能力强，功能丰富，轻巧灵活。

- 面对不好复现问题，线上的问题，尤其是性能问题。需要长时间观测，需要对系统性能影响小，同时需要收集足够的数据。
  - 无法在对系统影响小和收集全面数据用来解决问题上无法找到平衡点。
- 接口各异，在收集数据的时候需要使用不同接口。
- 工具使用本身会耗费系统资源（尤其是边收集边分析的工具），影响测试结果准确度。

解决思路和问题 1: 努力复现

- 一些问题很难复现。
- 一些问题复现流程复杂，仍然面临影响系统性能小和收集足够的问题数据的矛盾。

解决问题思路 2 ：真人监控，程序监控

- 监控系统，一旦触发相关条件开始收集系统信息。

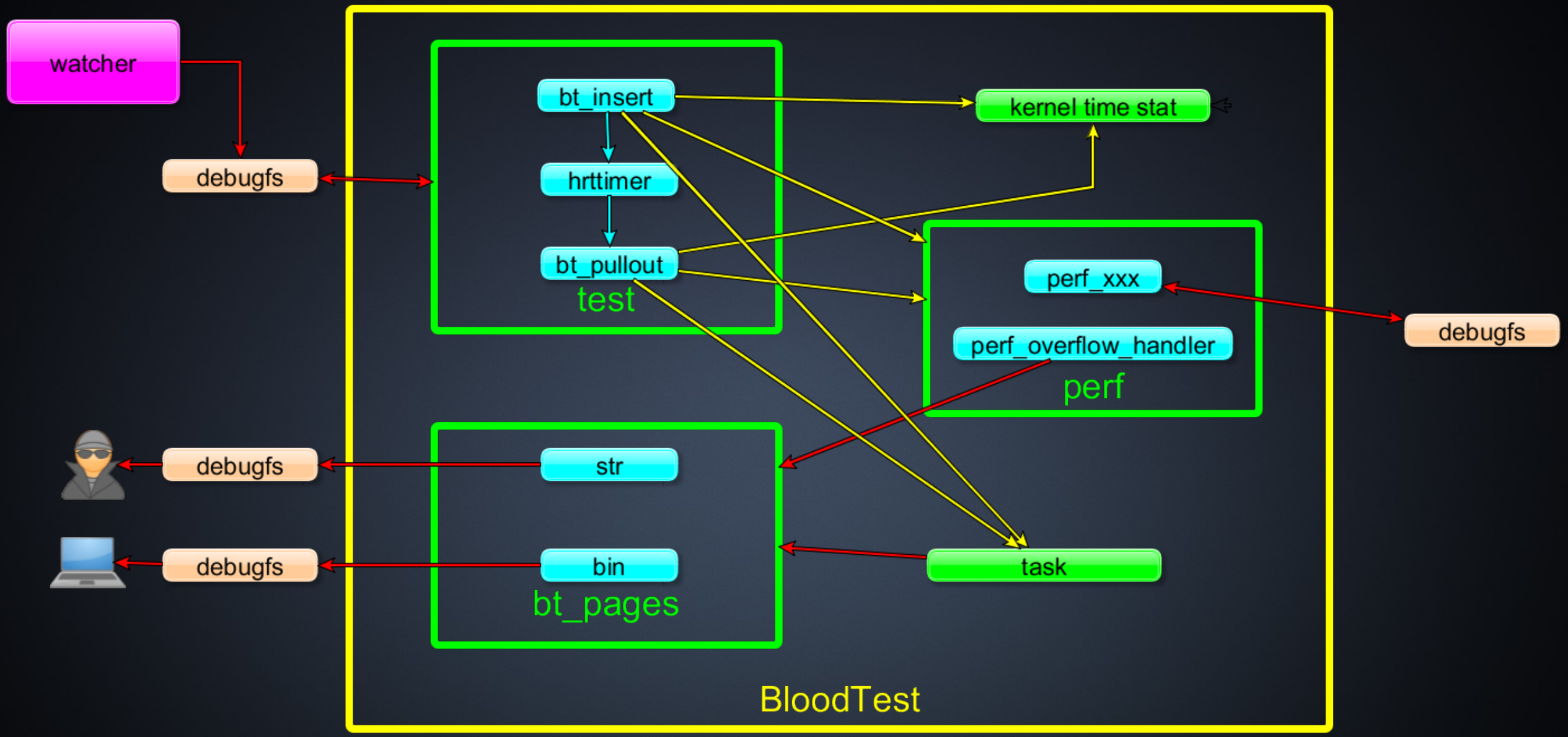
  在一定情况能收集到系统问题的蛛丝马迹。

  例如安卓在遇到 ANR 收集系统数据，用来分析原因。

- 因为是触发条件后开始收集数据，所以开始记录速度越快越好。
  - 但是现有接口不统一，不同工具可能需要分别调用，影响开始收集速度。
- 有些工具本身执行会耗费系统资源（比如一些工具会分配内存做 buffer），影响数据准确度。
- 有些工具是应用层工具，如果是在内核中发现需要开始监控，调用相关应用层工具不方便。

BloodTest—— 抽血检测

- 为思路 2 提供支持。
- 内核功能。
- 基本定位——监控小助手。
  - 不负责侦测问题。
  - 不负责分析问题。
  - 只做收集数据的第一环节。
- 不生产数据，只做数据的搬运工。


watcher
debugfs
bt_insert
hrttimer
bt_pullout
test
kernel time stat
perf_xxx
perf_overflow_handler
perf
debugfs
debugfs
str
debugfs
bin
bt_pages
task
BloodTest

BloodTest str 格式

- perf

cpu0

k swapper/0 0 native_safe_halt+0x6/0x10[0xffffffff92857496]
k swapper/0 0 native_safe_halt+0x6/0x10[0xffffffff92857496]
k cc1 0 __lru_cache_add+0x10/0x70[0xffffffff921a0c70]
u cc1 0 [0x555b7f]
u cc1 0 [0x15305adea1c0]
k cc1 0 clear_page_rep+0x7/0x10[0xffffffff92847e47]

cpu1

u cc1 0 [0xc56706]
u cc1 0 [0x634690]
u cc1 0 [0xc50de1]
k cc1 0 clear_page_rep+0x7/0x10[0xffffffff92847e47]
k swapper/1 0 native_safe_halt+0x6/0x10[0xffffffff92857496]
k swapper/1 0 native_safe_halt+0x6/0x10[0xffffffff92857496]

- task

comm:cc1 pid:9919

status:PROCESS_INSERT

utime:2126895 stime:2126894

read_bytes:28672 write_bytes:0 cancelled_write_bytes:0

comm:as pid:9920

status:PROCESS_INSERT

utime:0 stime:3119211

read_bytes:0 write_bytes:0 cancelled_write_bytes:0

BloodTest bin 格式

- Perf

little-endian

page_size:4096

size:32

pc format:u64 unsigned offset:0 size:8

is_user format:u8 unsigned offset:8 size:1

oom_score_adj format:s16 signed offset:10 size:2

comm format:char[] signed offset:12 size:16

- task

little-endian

page_size:4096

size:64

status format:u8 unsigned offset:0 size:1

pid format:pid_t signed offset:4 size:4

comm format:char[] signed offset:8 size:16

utime format:u64 unsigned offset:24 size:8

stime format:u64 unsigned offset:32 size:8

read_bytes format:u64 unsigned offset:40 size:8

write_bytes format:u64 unsigned offset:48 size:8

cancelled_write_bytes format:u64 unsigned offset:56 size:8

BloodTest 特点

- 需要的时候调用一个接口全部数据收集功能开启。
- 同时每个数据收集功能可事先配置为打开或者关闭。
- bt_pages 为事先分配好的内存，使用的时候不再分配。
- 数据读取接口：
  - str 可以直接读
  - bin 二进制，数据直接 mmap，分析程序快速访问。

BloodTest 还需要实现的东西

- 现有代码： https://lkml.org/lkml/2017/10/13/140
- 支持更多工具
- 支持从内核调用 BloodTest
- 需要时记录结束后唤醒应用层的后续处理程序

谢谢！ 问题？

- weibo: @teawater_z 欢迎在线吐槽
- 微信公众号：茶水倪山 (id: cschatcs)