Tracing ABI for efficient kernel-userspace tracing.

Linux Plumbers Conference 2008

Mathieu Desnoyers
École Polytechnique de Montréal

> Mathieu Desnoyers

- Author/Maintainer of :
  - Linux Trace Toolkit Next Generation
  - Linux Kernel Markers
  - Tracepoints
  - Linux Trace Toolkit Viewer

> Summary

- Userspace instrumentation
  - Requirements
  - Proposal
- Userspace data extraction
  - Requirements
  - Proposal

> Instrumentation Requirements

- Statically declared, enabled dynamically
- Activated across all or specific processes
  - e.g. instrument all pthread mutexes
- Early boot support
- Instrumentation enabled/disabled asynchronously wrt userspace execution
- Cross-layer instrumentation, including Java Virtual Machine.

> Instrumentation Proposal

- Tracepoints
  - Headers manage name-spacing
    - Packaging
  - Linker script modification
    - New tracepoint section

> Instrumentation Proposal

- Instrumentation updates done by statically linked object
  - Synchronous
    - Executables and shared libraries
      - Enable/disable instrumentation by querying the OS for instrumentation status through a system call in constructor (tp-sync.o)
  - Asynchronous
    - Executables handle an “update instrumentation” signal (tp-async.o). Use query system call.
    - Shared libraries register their callback to the executable (tp-async-lib.o). Use query system call.

> Instrumentation Proposal

- Quiescent state
  - Knowing if instrumentation has been activated or deactivated after performing the status change
  - Depends on signal delivery
    - If receiver thread is not running
    - If receiver thread is running
  - Use synchronize_sched() to insure all threads running when the signal has been sent have scheduled out, thus will run the signal handler before any other userspace code.

> Data Extraction Requirements

- Export data to
  - Disk, network, serial port
  - Flight-recorder mode to memory buffers
    - Killed processes
    - Part of kernel crash dump
- Early boot tracing (e.g. init process)
- Security/isolation
- Multiple active traces (nice-to-have)
  - Different filters/scripts

> Data Extraction Proposal

- Export data through system call or shared memory buffer ?
  - Speed/complexity trade-off
  - Time-stamping (vDSO)
  - Locking in userspace
    - no RCU, seqlock only suitable to protect reading
  - Global/per-cpu/per-thread buffers
  - Multiple trace handling
  - Filtering/scripting