BPF User Tracepoints

Enhancing User Defined Tracepoints

Mary Marchini

@mmarkini

sthima

Goals

- Instrument user-space applications
  - Trace with BPF tools (e.g., bpftrace)
- Sophisticated Argument Types (structures)
  - Possibly integrated with BTF
- Static and dynamic instrumentation

Current Approach

Systemtap SDT [1]

- Only static instrumentation
  - Workaround: libstapsdt [2]
- Simple (numeric) argument type
  - Workaround: casting arguments

Extending Systemtap SDT

(or: creating a new user-space tracepoint format)

- ELF Note Section → New ELF Section
- Revamped Arguments
    - Linux Tracepoint-inspired format
        - **`field:`** *`BTF_type_id name`*`;` **`offset:`** *`arg_offset`*`;`
        - Arguments located with offsets instead of strings
            - (might have perf impact, but increases portability)
- Dynamic tracepoints installed in a separate file
    - /proc/PID/events/*provider*/*probe*/[address,format]
    - or inspired in *perf* JIT support (on /tmp/ or $PWD/)

https://sourceware.org/systemtap/wiki/UserSpaceProbeImplementation

https://docs.oracle.com/cd/E23824_01/html/819-0690/chapter6-18048.html

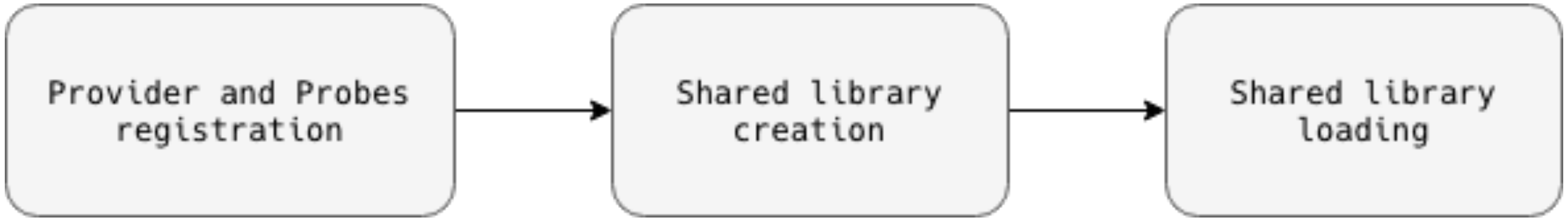
<table>
<tr><th></th><th>+0</th><th>+1</th><th>+2</th><th>+3</th><th></th></tr>
<tr><td><b>namesz</b></td><td colspan="4">8</td><td></td></tr>
<tr><td><b>descsz</b></td><td colspan="4">24 + len(provider) + len(probe) + len(arguments)</td><td></td></tr>
<tr><td><b>type</b></td><td colspan="4">3</td><td></td></tr>
<tr><td><b>name</b></td><td>s</td><td>t</td><td>a</td><td>p</td><td></td></tr>
<tr><td></td><td>s</td><td>d</td><td>t</td><td>\0</td><td></td></tr>
<tr><td><b>desc</b></td><td colspan="4">0x…</td><td><b>Probe PC</b></td></tr>
<tr><td></td><td colspan="4">0x…</td><td><b>Link-time sh_addr</b></td></tr>
<tr><td></td><td colspan="4">0x…</td><td><b>Link-time semaphore variable address</b></td></tr>
<tr><td></td><td></td><td></td><td>…</td><td>\0</td><td><b>Provider name (null-terminated)</b></td></tr>
<tr><td></td><td></td><td></td><td>…</td><td>\0</td><td><b>Probe name (null-terminated)</b></td></tr>
<tr><td></td><td></td><td></td><td>…</td><td>\0</td><td><b>Arguments (space-separated, null-terminated)</b></td></tr>
</table>

```
$ readelf -n ./example

Displaying notes found in: .note.stapsdt
  Owner                 Data size       Description
  stapsdt              0x00000047       NT_STAPSDT (SystemTap probe descriptors)
    Provider: example
    Name: second_probe
    Location: 0x00000000000006a8, Base: 0x0000000000000754, Semaphore: 
0x0000000000000000
    Arguments: 8@%rax -4@$1 16@-48(%rbp)
```

libstapsdt

- Creates shared libraries on runtime
  - With empty functions
  - Probes instrument those functions
- *dlopen* the shared library



https://github.com/sthima/libstapsdt