systemd-oomd

PSI-based OOM kills in systemd

Anita Zhang
the.anitazha@gmail.com
Software Engineer, Containers, Facebook

22 September 2021

Agenda

1) Overview of oomd
2) Integrating into systemd
3) Outcomes from Fedora
4) Future Plans
5) Discussion

Overview of oomd

oomd

https://github.com/facebookincubator/oomd

Daniel Xu's 2019 LPC talk on oomd

Userspace out of memory (OOM) killer

Advantages over the kernel OOM killer

- Flexible configuration
- Deterministic kills

Uses cgroup2, pressure stall information (PSI), etc. to make decisions

cgroup2 and PSI

cgroup2

- Allows grouping processes together to control/measure resources (CPU, IO, memory)

Pressure Stall Information (PSI)

- https://facebookmicrosites.github.io/psi/docs/overview
- Measures percentage of time tasks were delayed due to lack of resources

oomd Configuration (Snippet)

```
"name": "protection against heavy workload thrashing",
  "detectors": [
    [
      "sustained high workload memory pressure",
      {
        "name": "pressure_above",
        "args": {
            "cgroup": "workload.slice/workload-tw.slice",
            "resource": "memory",
            "threshold": "80",
            "duration": "180"
        }
      }
    ]
  ]
```

oomd Configuration (Snippet Cont.)

```
"actions": [
  {
    "name": "kill_by_pg_scan",
    "args": {
      "cgroup": "workload.slice/workload-tw.slice/*",
      "recursive": "true"
    }
  }
]
```

Integrating into systemd

Why systemd-oomd

oomd expects you to use systemd

systemd is well positioned between kernel and applications

- Open to novel uses of resource control

Make it easier to adopt userspace OOM killing

- systemd is widely used
- No additional packaging dependencies
- Familiar configuration syntax

Simplifying oomd for systemd

oomd is C++; systemd is C

systemd’s configuration interface is limited

- INI files

Needed to balance ease/flexibility of configuration with interface constraints

Simplifying oomd for systemd

Only integrate the key features/plugins of oomd

Detect on **memory pressure** and reclaim activity

- Kill based on pgscan rate

Detect on **swap**

- Kill based on the largest consumer

Simplifying oomd for systemd

/etc/systemd/oomd.conf

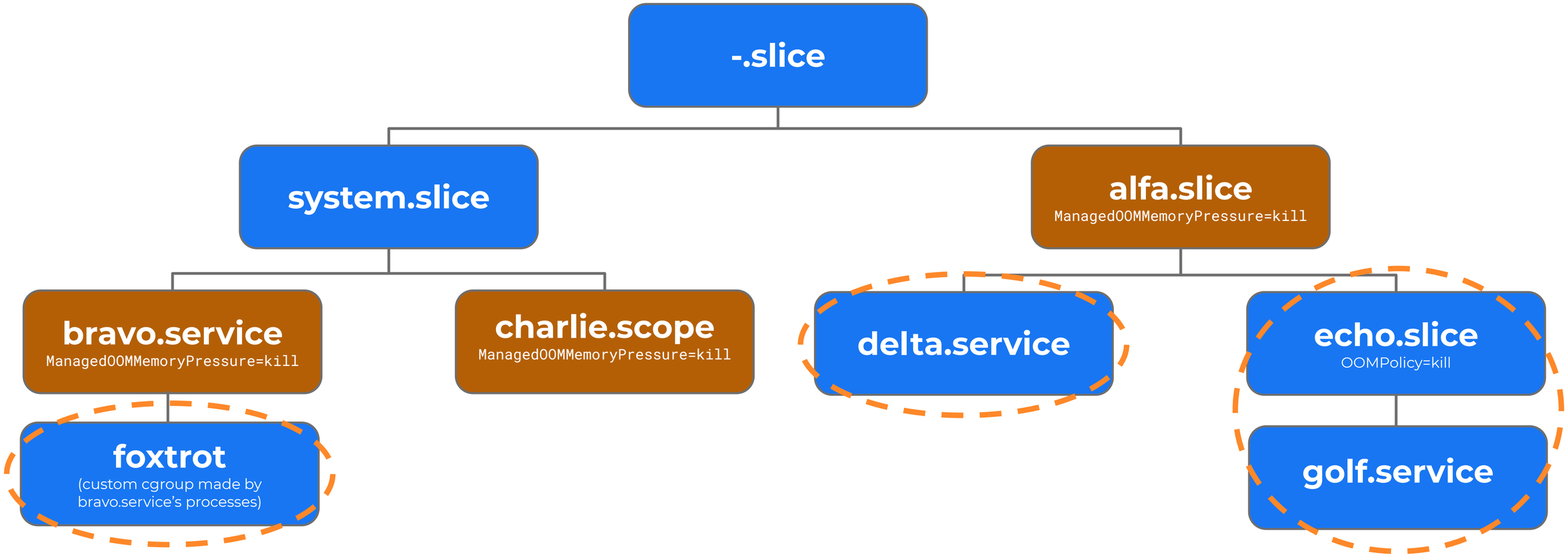
```
[OOM]
SwapUsedLimit=90%
DefaultMemoryPressureLimit=60%
DefaultMemoryPressureDurationSec=30s
```

Simplifying oomd for systemd

/etc/systemd/systemd/birb.slice

```
[Slice]
ManagedOOMSwap=auto|kill
ManagedOOMMemoryPressure=auto|kill
ManagedOOMMemoryPressureLimit=0%
ManagedOOMPreference=none|avoid|omit
```

Candidate Selection for Kills



Outcomes from Fedora

systemd-oomd by default in Fedora 34

user@.service (all user services) with memory pressure above 50% for 20 seconds

- All user unit leaf nodes are candidates

-.slice (root slice) with swap used limit 90%

- All leaf nodes in the hierarchy are candidates

Works best in environments that support splitting applications into cgroups

- GNOME is one of the best examples of this

Resolved Items

Initial limits too low

Swap killing too aggressive

High CPU

Future Plans

Future Plans

Enabling systemd-oomd settings for user units

- https://github.com/systemd/systemd/pull/20690

Improvements for systemd-oomd kill insight

- https://github.com/systemd/systemd/issues/20649

Thanks!

Facebook

Davide Cavalca

Daan De Meyer

Tejun Heo

Jared Pochtar

Michel Salim

Dan Schatzberg

Johannes Weiner

Daniel Xu

GNOME

Benjamin Berg

Fedora

Neal Gompa

Chris Murphy

Systemd

Lennart Poettering

Zbigniew Jędrzejewski-Szmek

For further questions: Anita Zhang <the.anitazha@gmail.com>