
Meta

Adventures of Linux Userspace at Meta

Anita Zhang
engineerd managerd
“Linux Umbrella” family of teams

Meta

Agenda

01 What is Linux Userspace at Meta?

02 Look Back: systemd

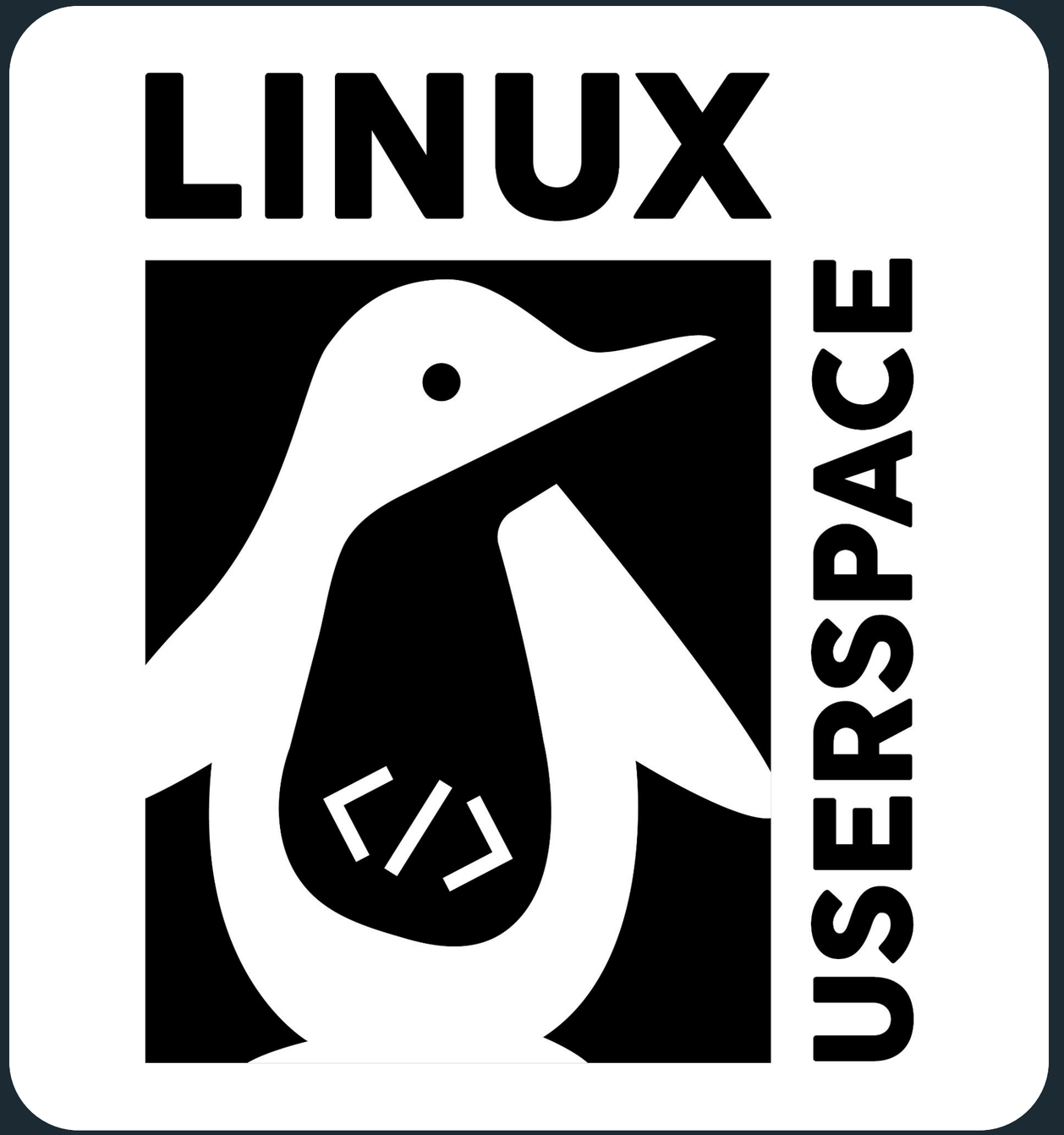
03 Look Back: Hyperscale SIG

04 Look Back: Fedora Asahi Remix

05 Look Back: Frame Pointers

06 New Initiatives: Applied BPF

What is Linux Userspace at Meta?



Upstream First

- Tackling big problems together with the open source community.
- Meta’s teams included Kernel and Operating Systems.
- Contributing to existing community projects benefits everyone.

A Brief History

- Operating Systems team’s charter was contributing upstream and community building.
  - Chef, CentOS, systemd, upstream packaging, etc.
- Responsibilities eventually split to different teams with narrower focus.
  - The open source parts were the primary inspiration for the Linux Userspace team.

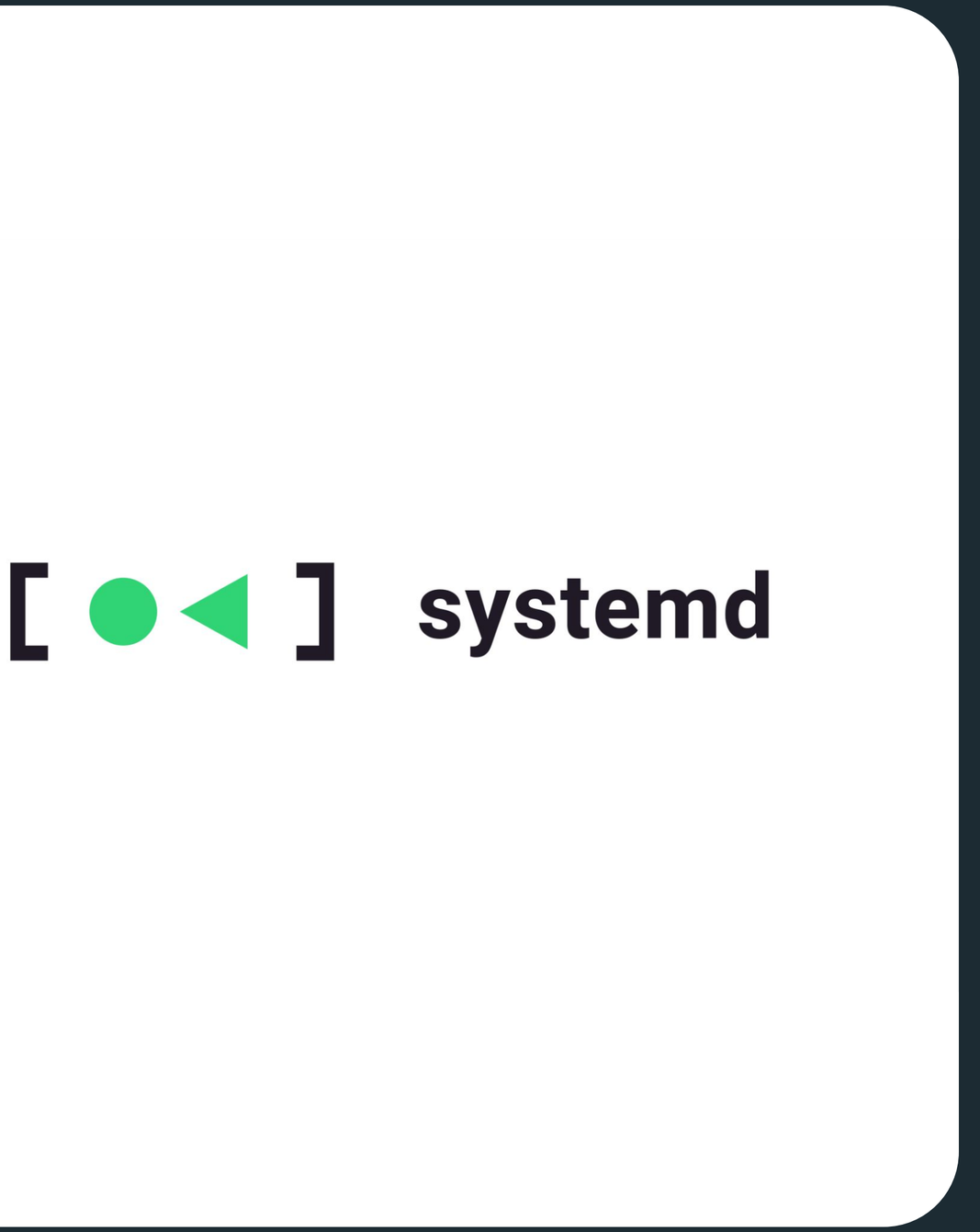
Today

- Linux Distributions
- Community Building
- Packaging
- systemd
- BPF

The Fleet We Support

- Millions of hosts!
  - Largely homogenous, increasing hardware diversity.
- Teams to maintain their own bare metal hosts:
  - Large collection of containerized services.
  - Many services also supported on bare metal.
- >98% upgraded from CentOS Stream 8 → CentOS Stream 9.
  - systemd we deploy as it is released.
    - "Adventures with systemd in Hyperscale" at CentOS Dojo, August 2022

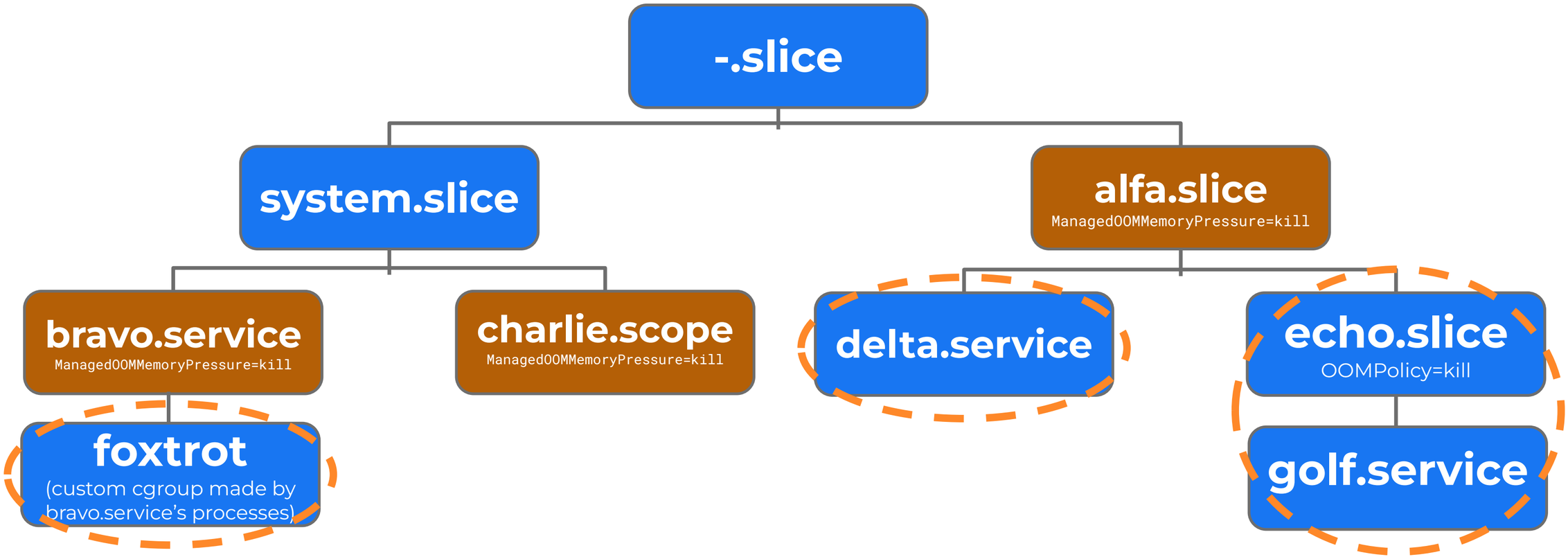
Look Back: systemd



systemd-oomd

- Userspace out of memory (OOM) killer.
- Uses cgroup2, pressure stall information (PSI), etc. to make kill decisions.
- Spun out of github.com/facebookincubator/oomd.
- "systemd-oomd: PSI-based OOM kills in systemd" at Linux Plumbers 2021

systemd-oomd



systemd-oomd

- fedoraproject.org/wiki/Changes/EnableSystemdOomd (Fedora 34)
- Multiple configuration tweaks to make a reasonable default.
  - Relies on desktop environments and browsers to split by cgroup.
- More PSI knobs in systemd.

systemd-oomd... at Meta

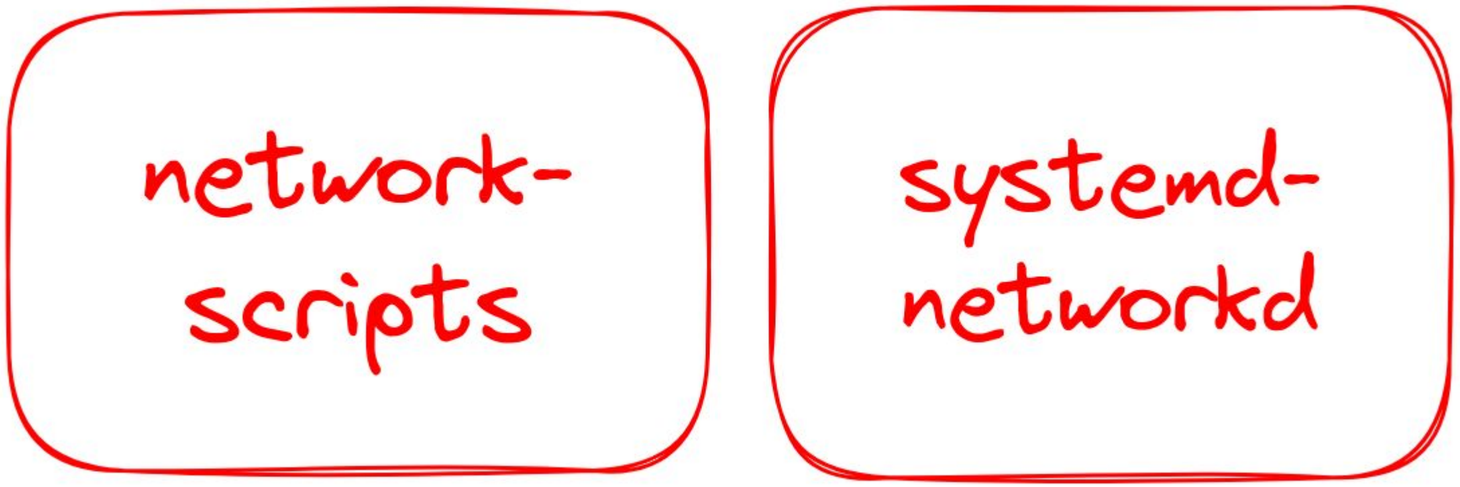
- Rolled out alongside the original FB OOMD!
- FB OOMD still used for experimentation:
  - Automated memory sizing tool: github.com/facebookincubator/senpai.
  - Memory profiling before an OOMD kill.

systemd-networkd

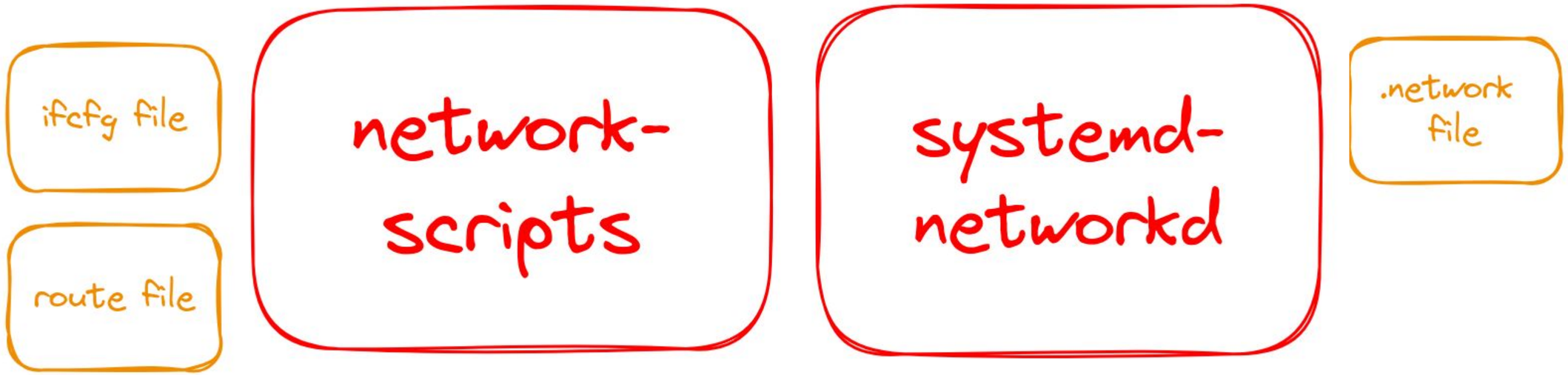
- Daemon shipped with systemd to manage network configurations.
- Configure your network with systemd-style configuration files.

Network Configuration at Meta until ~2022

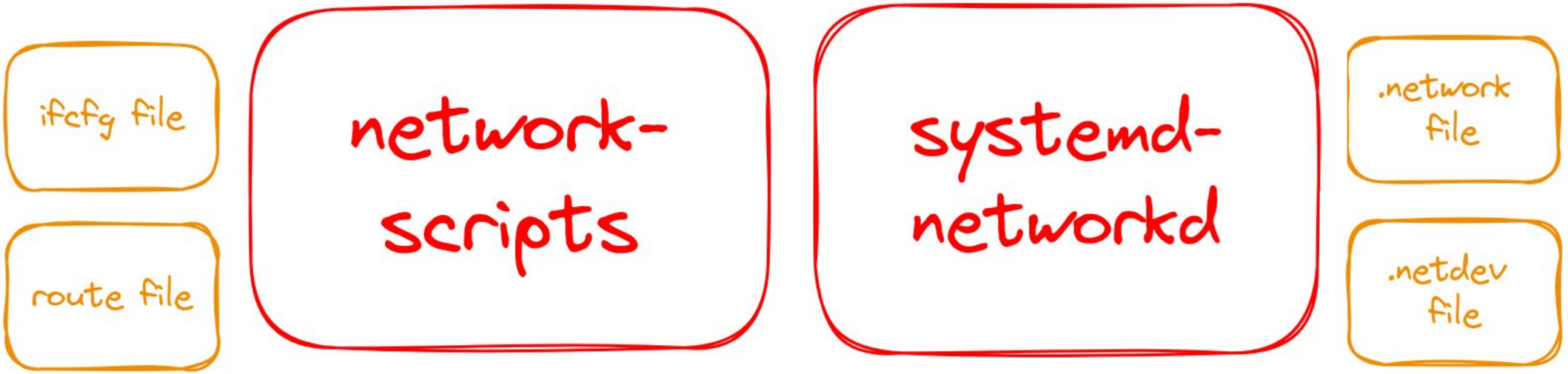
- Network Scripts were deprecated in RHEL 8.
  - Still used it in our fleet for CentOS Stream 8.
- Decided to go all in on systemd-networkd for CentOS Stream 9!


network-
scripts
systemd-
networkd

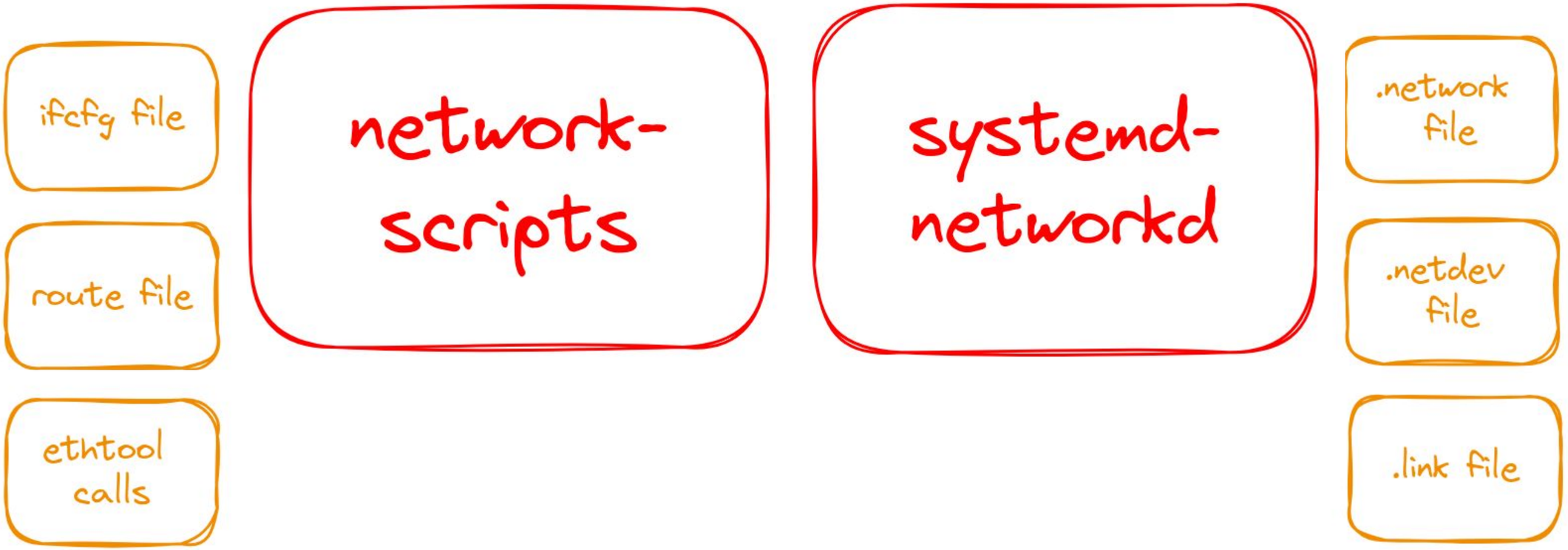
Look Back: systemd



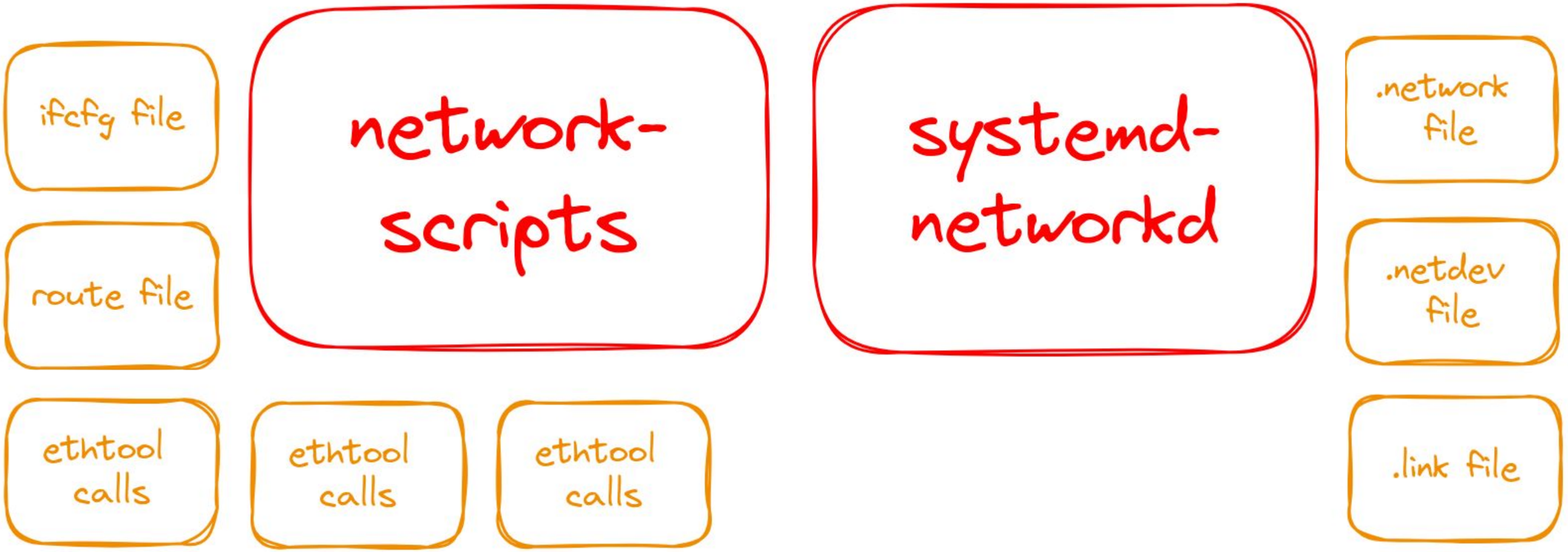
Look Back: systemd



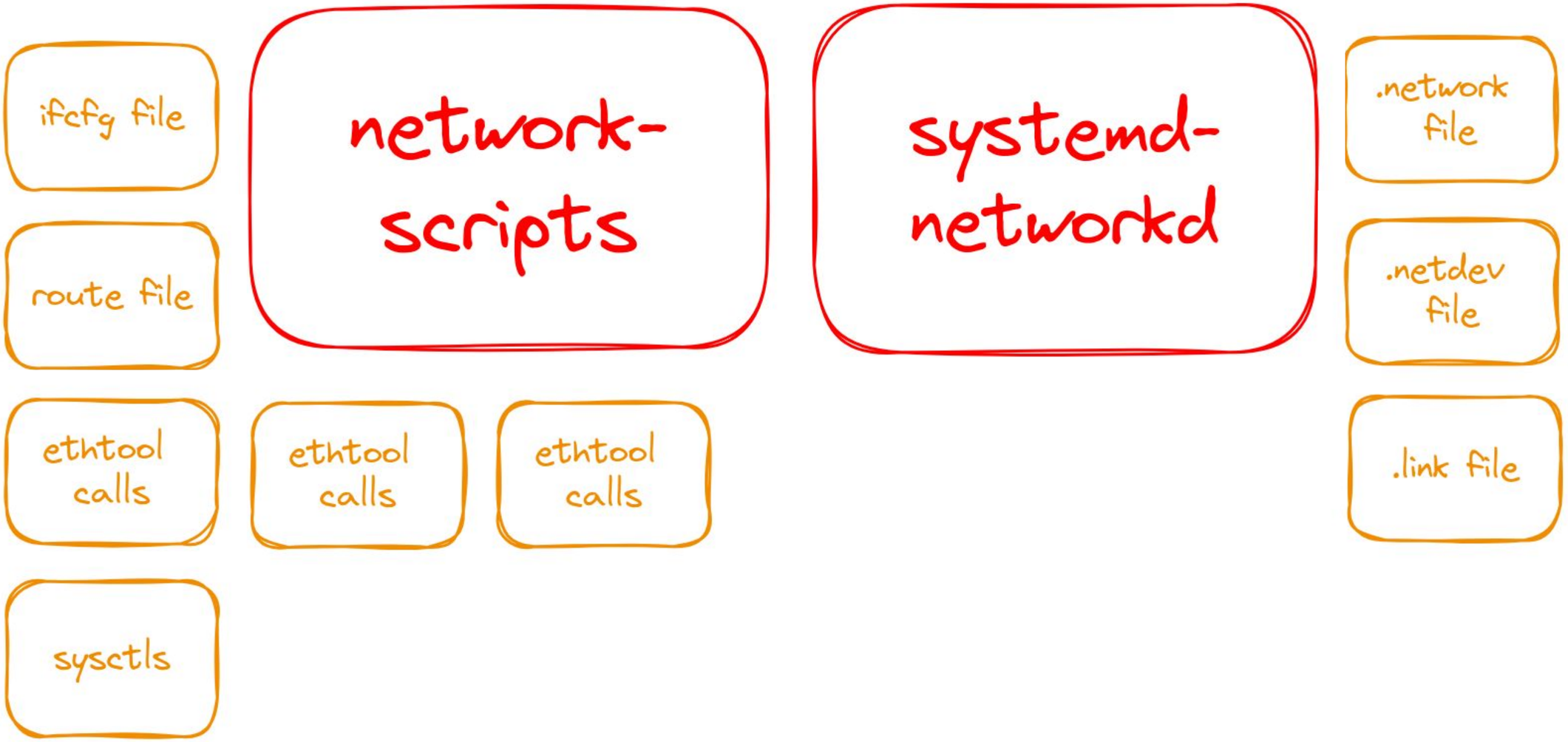
Look Back: systemd



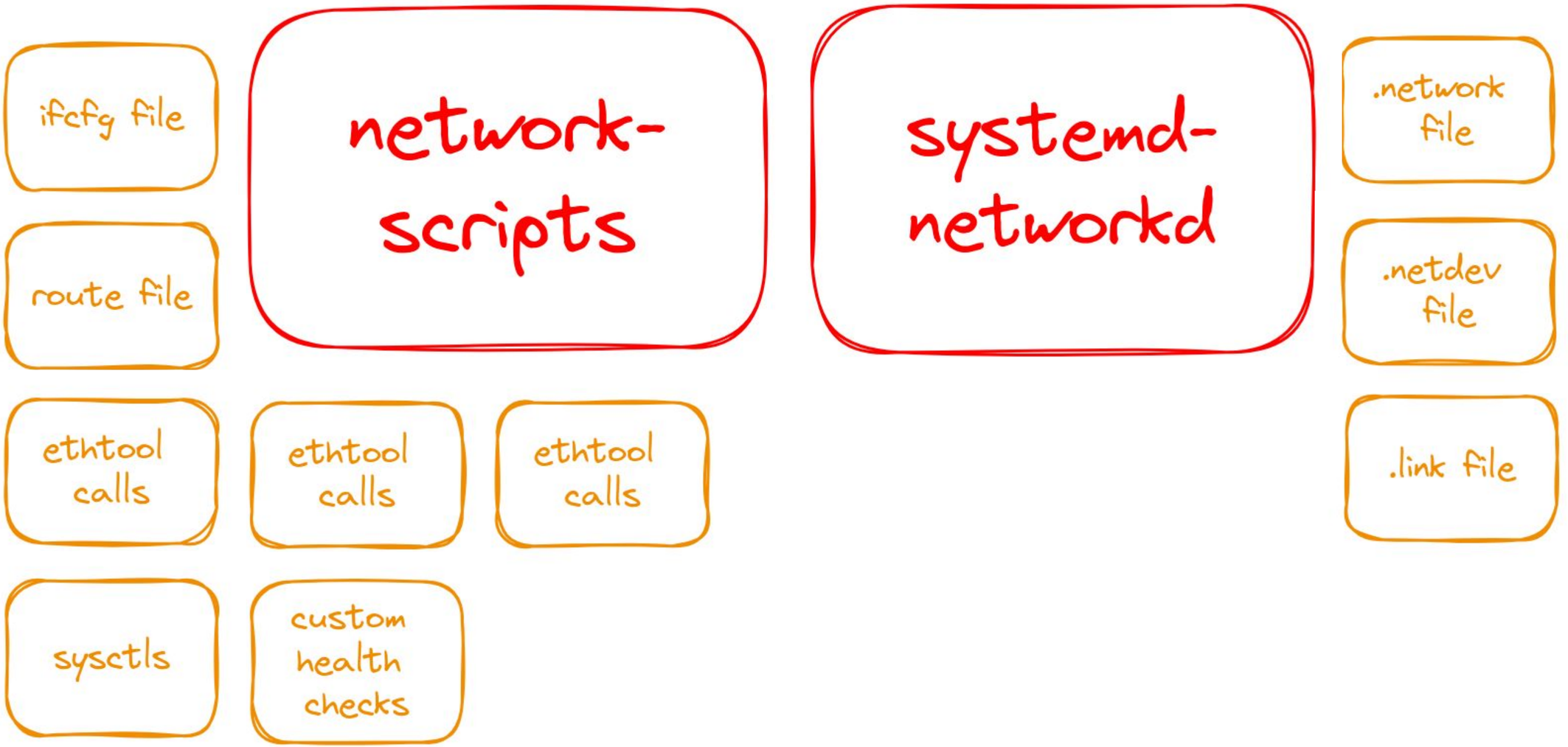
Look Back: systemd



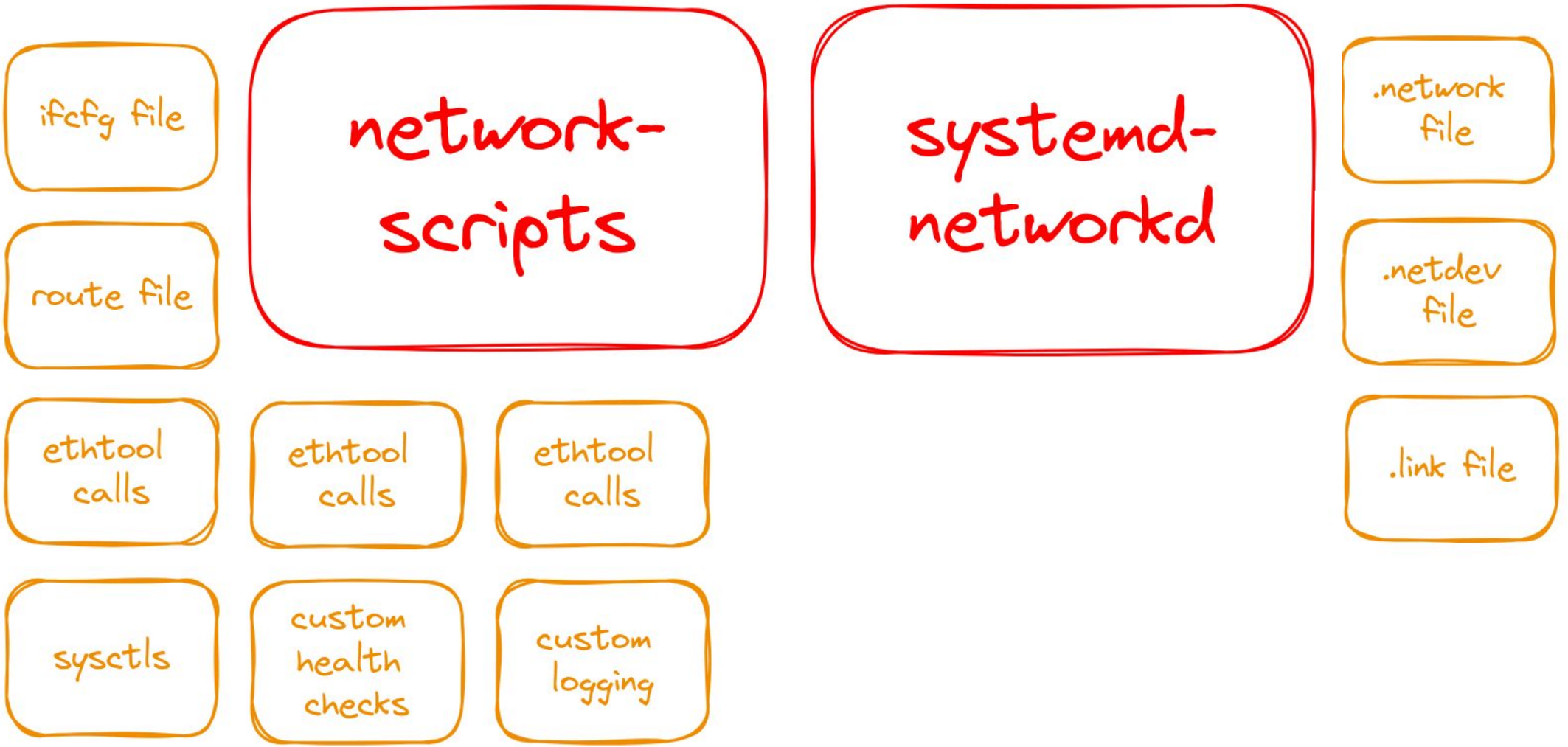
Look Back: systemd



Look Back: systemd



Look Back: systemd



systemd-networkd… at Meta

- Rolled out as part of CentOS Stream 9.
- Big (scary) migrations bring people together!
  - Everyone motivated to do this safely.
    - And make things more maintenance friendly.
  - systemd maintainers were really supportive.

systemd-journald

- Collects and stores logging data.
- Lots of metadata!

Meta: Can we stop using rsyslog?

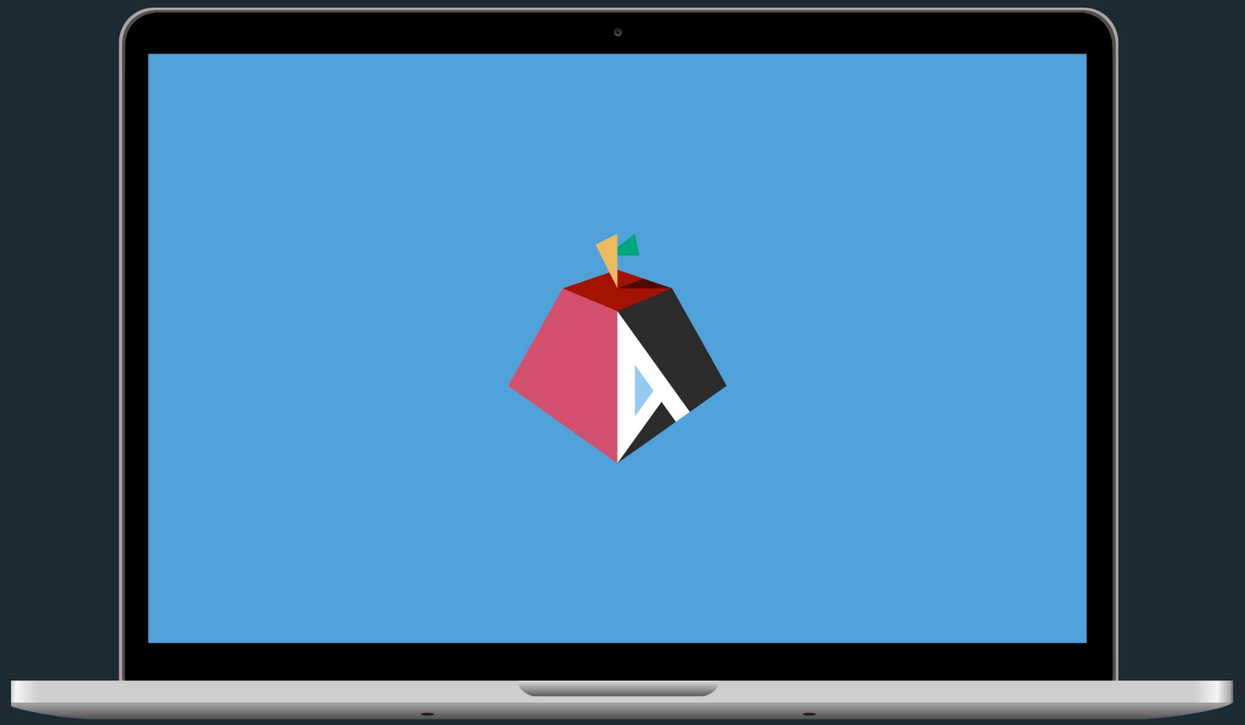
- "Slimming down the journal" at Linux Plumbers 2022
  - Journal compact mode
    - Less flash!
  - Compression fixes (with BTRFS)
  - Corruption improvements

Look Back:
Hyperscale SIG

CentOS Hyperscale SIG

- sigs.centos.org/hyperscale
- Special Interest Group (SIG) formed in 2021 focused on CentOS Stream on large-scale infrastructures.
  - Allows alternative policies.
  - Latest systemd based on Fedora Rawhide.
  - “hyperscale-intel” repository with optimized packages.
  - RPM copy-on-write (CoW).
- Plus automation to release Hyperscale images.

Look Back: Fedora Asahi Remix

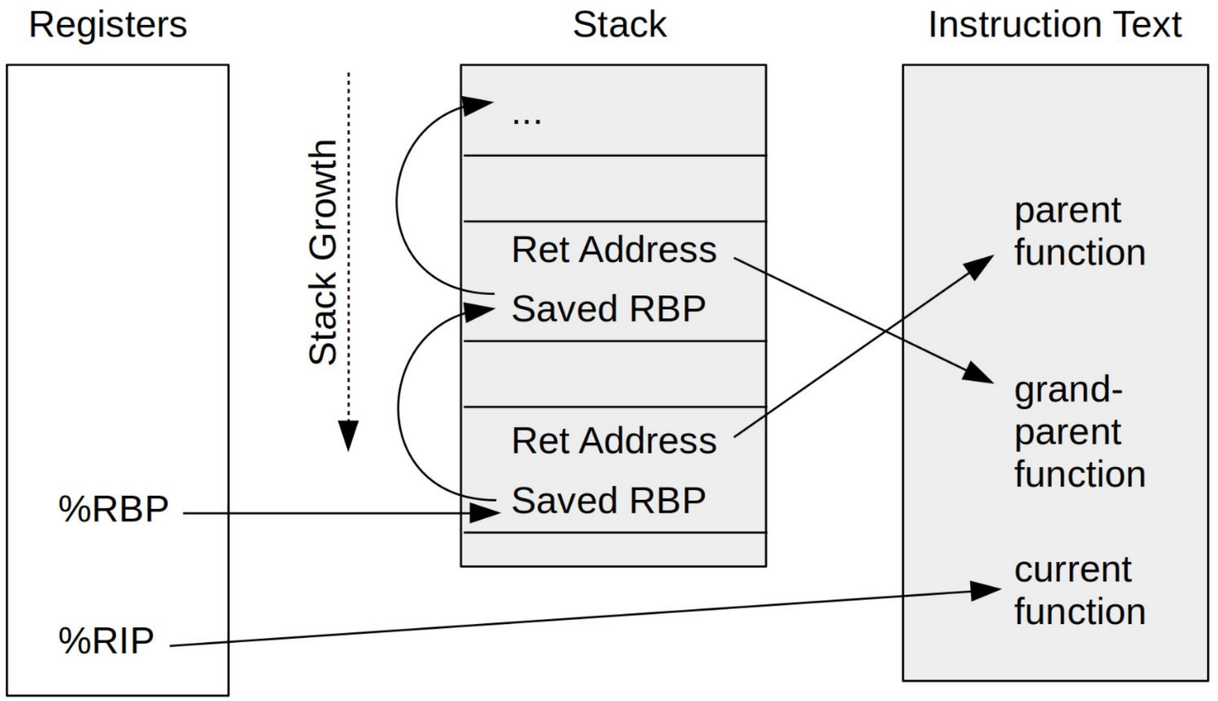


Fedora Asahi Remix

- fedora-asahi-remix.org
- The Asahi Linux flagship distribution.
  - “Asahi Linux aims to bring you a polished Linux® experience on Apple Silicon Macs.”
    - High performance, reliable, and readily available native aarch64 platform.
  - Over a year of collaboration to port Fedora to Asahi Linux.
    - Bringing together reverse engineers, kernel developers, distribution integration experts, and other motivated community members.
- Actively being used as daily and testing machines for kernel development!
  - Setting up GitHub Actions to run fstests for Btrfs.

Look Back:
Frame Pointers

What Are Frame Pointers?



Frame Pointer-Based Stack Walking from [BPF Performance Tools: Linux System and Application Observability] book.

Frame Pointers

- Since ~2004 frame pointers were not compiled in by default.
  - Pro: Performance wins!
  - Con: Makes stack unwinding (needed for debugging, profiling, or tracing) very hard/inconvenient.
- Brendan Gregg's The Return of the Frame Pointers is a great overview of frame pointers until 2024 and beyond.

Frame Pointers by Default in Fedora

- fedoraproject.org/wiki/Changes/fno-omit-frame-pointer (Fedora 38)
  - Many benchmarks and alternatives were explored.
- Reasonable enough overhead to (conditionally) enable by default starting in Fedora 38.
  - Aarch64 and ppc64le already had frame pointers.
  - i686 and s390x were omitted for more testing and benchmarks.
  - Python 3.11 opted out due to 1-10% benchmark regressions.
    - Python 3.12 recommends enabling frame pointers due to the Linux perf profiler.
      - Will also have speedups that could offset/fix the noted regression.
- Reevaluated in Fedora 40. Frame pointers are here to stay!

Frame Pointers... Everywhere?

- Frame pointers also enabled by default in:
  - Ubuntu 24.04 LTS.
  - Arch Linux.
- What benefits have we gained from frame pointers?
  - Fedora 39 redesigned Sysprof (system wide profiling tool). Together with frame pointers this lead to several performance improvements in Fedora Linux.

New Initiatives: Applied BPF

Ongoing/Future

- systemd-bpfd
- BPF token
- sched_ext
- bpftrace
- bpfilter

systemd-bpfd

- github.com/systemd/systemd/pull/28268
  - In progress.
- Monitoring and auditing BPF programs. And more?

BPF Token

- https://lore.kernel.org/all/20240124022127.2379740-1-andrii@kernel.org/
  - Merged!
- Allow BPF to be used inside user namespaces.
  - Use “tokens” to delegate BPF permission from a privileged process to a trusted, unprivileged process.
- Will be adding this functionality to userspace soon!

Questions?

THANK YOU FOR YOUR TIME

github.com/anitazha
twitter.com/the_anitazha
@anitazha@fosstodon.org
@anitazha:matrix.org

Bonus Slides

Digression on systemd and GPUs

- Audience question from All Systems Go 2023: Any interesting systemd work come out of AI making the fleet increasingly more heterogeneous?
  - More than 25 seconds for netlink calls!
  - Default netlink timeout in systemd is configurable starting in systemd 254.

“Could it be that we have systemd build with -fomit-frame-pointer compiler flag? I vaguely remember someone complaining that we are not getting good stack traces in Strobelight for systemd.”

- **Andrii Nakryiko on Daan De Meyer’s status report**