kexec: rychlý restart bez restartu

Ondřej Caletka

3. října 2020

Jak vzniká proces

- volání `fork(2)` naklonuje běžící proces
- původní proces pokračuje v činnosti
- klon původního procesu voláním `exec(2)` nahradí svůj kód novým
- lze volat i pouze `exec(2)`, čímž původní proces přestane existovat

RIPE NCC cesnet

Volání kexec

- obdoba `exec(2)` pro linuxové jádro
- *zavádějící:* výměna jádra za běhu
  - jádro během výměny ztratí vnitřní stav
  - *všechny procesy* jsou vnitřním stavem jádra
  - fakticky jde tedy o restart počítače
- *přesnější:* linux v roli bootloaderu
  - zavede do paměti obraz jádra, popř. initramfs
  - předá jádru parametry příkazového řádku
  - předá řízení novému jádru

RIPE NCC cesnet

K čemu to může být dobré

- přeskočení extrémně pomalého POST u profesionálních serverů
  - úspora i více než pěti minut
- restart bez nutnosti zadávat heslo k rozšifrování disku
  - v případě, kdy je šifrovaný i oddíl `/boot`
  - heslo je součástí initramfs, který leží na šifrovaném disku

RIPE NCC cesnet

Dva kroky k restartu

1. `kexec -l <vmlinuz>`
   - nahraje jádro do paměti
   - lze přidat initramfs (`--initrd`) a volby (`--append`)
2. `kexec -e`
   - předá řízení novému jádru
   - **nastane okamžitě**, stav aktuálního jádra je ztracen
   - je integrován ve vypínacích skriptech současných distribucí

RIPE NCC cesnet

Praktické použití s balíčkem `kexec-tools`

- `systemctl kexec`
  - vyžaduje UEFI a `systemd-boot`
  - netestoval jsem
- `kexec -l /vmlinuz --initrd=/initrd.img --reuse-cmdline`
  - funguje na Debianu a podobných
  - na jiných distribucích stačí upravit cesty k obrazům jádra a initramfs
  - k výměně jádra dojde při nejbližším restartu

RIPE NCC cesnet

Demo

```
[  OK  ] Reached target Unmount All Filesystems.
[  OK  ] Stopped File System Check on /dev/disk/by-uuid/3143-1D20.
[  OK  ] Removed slice system-systemd\x2dfsck.slice.
[  OK  ] Stopped target Local File Systems (Pre).
[  OK  ] Stopped Create Static Device Nodes in /dev.
[  OK  ] Stopped Create System Users.
[  OK  ] Stopped Remount Root and Kernel File Systems.
[  OK  ] Reached target Shutdown.
[  OK  ] Reached target Final Step.
         Starting Reboot via kexec...
         Stopping Monitoring of LVM2 mirrors, snapshots etc. using dmeventd or progress polling...
[  167.997222] systemd-shutdow: 58 output lines suppressed due to ratelimiting
[  168.022512] systemd-shutdown[1]: Syncing filesystems and block devices.
[  168.031177] systemd-shutdown[1]: Sending SIGTERM to remaining processes...
[  168.038073] systemd-journald[261]: Received SIGTERM from PID 1 (systemd-shutdow).
[  168.049414] systemd-shutdown[1]: Sending SIGKILL to remaining processes...
[  168.056850] systemd-shutdown[1]: Unmounting file systems.
[  168.060888] [1093]: Remounting '/' read-only in with options 'errors=remount-ro'.
[  168.068814] EXT4-fs (sda1): re-mounted. Opts: errors=remount-ro
[  168.077880] systemd-shutdown[1]: All filesystems unmounted.
[  168.080692] systemd-shutdown[1]: Deactivating swaps.
[  168.083568] systemd-shutdown[1]: All swaps deactivated.
[  168.085821] systemd-shutdown[1]: Detaching loop devices.
[  168.088071] systemd-shutdown[1]: All loop devices detached.
[  168.089524] systemd-shutdown[1]: Detaching DM devices.
[  168.112055] Unregister pv shared memory for cpu 0
[  168.116687] sd 2:0:0:0: [sda] Synchronizing SCSI cache
[  168.178623] kexec_core: Starting new kernel
```

RIPE NCC cesnet

Děkuji za pozornost

Ondřej Caletka
Ondrej.Caletka@ripe.net
https://**Ondřej.Caletka.cz**

RIPE NCC
RIPE NETWORK COORDINATION CENTRE

RIPE NCC cesnet

Prezentace je již nyní k dispozici ke stažení.