



Admin's Workshop: Eine Reise durch das system(d)



Hinter den Kulissen



systemd - Units

- Units sind die Building-Blocks des systemd Universums
- Deklarative Servicebeschreibung statt ausführbarer Dateien

systemd.unit(5)

- Sauberes Environment
- Socket-Aktivierung (aber auch inotify, dbus)
- Conditions

```
# /lib/systemd/system/cups.socket
[Unit]
Description=CUPS Printing Service Sockets

[Socket]
ListenStream=/var/run/cups/cups.sock
BindIPv6only=ipv6-only
PassCredentials=true

[Install]
WantedBy=sockets.target
```

```
# /lib/systemd/system/cups.service
[Unit]
Description=CUPS Printing Service
Documentation=man:cupsd(8) man:cupsd.conf(5)
Requires=cups.socket

[Service]
ExecStart=/usr/sbin/cupsd -f
PrivateTmp=true

[Install]
Also=cups.socket cups.path
WantedBy=printer.target
```



systemd systemd – Units

.service

.socket

.device

/etc/systemd/system

/run/systemd/system

/lib/systemd/system

.mount

.automount

.swap

.path

.timer

.target

.slice

.scope

\$XDG_CONFIG_HOME/systemd/user

\$HOME/.config/systemd/user

/etc/systemd/user

\$XDG_RUNTIME_DIR/systemd/user

/run/systemd/user

\$XDG_DATA_HOME/systemd/user

\$HOME/.local/share/systemd/user

/usr/lib/systemd/user

<https://www.freedesktop.org/software/systemd/man/systemd.unit.html>



systemd – was ist los auf dem System?

```
root@machine:~ systemctl status
```

- machine

```
State: running
Jobs: 0 queued
Failed: 0 units
Since: Mon 2015-05-07 05:35:11 CET; 2min 27s ago
CGroup: /
```

```
├─1 /sbin/init
├─system.slice
│   ├─dbus.service
│   │   └─112 /usr/bin/dbus-daemon --system --address=systemd: [...]
│   ├─system-serial\x2dgetty.slice
│   │   └─serial-getty@ttyLP0.service
│   │       ├──117 -sh
│   │       ├──134 systemctl status
│   │       └─135 less
│   ├─systemd-journald.service
│   │   └─59 /lib/systemd/systemd-journald
│   ├─systemd-logind.service
│   │   └─114 /lib/systemd/systemd-logind
│   ├─systemd-udevd.service
│   │   └─82 /lib/systemd/systemd-udevd
│   └─NetworkManager.service
│       └─125 /usr/sbin/NetworkManager --no-daemon [...]
└─
```



Targets

- Synchronisationspunkte im System
- `systemctl list-units -type=target`
- `systemctl isolate <target>`

Units müssen das explizit erlauben:
`AllowIsolate=yes`

- `/etc/systemd/system/default.target`

`/lib/systemd/system/runlevel0.target -> poweroff.target`

`/lib/systemd/system/runlevel1.target -> rescue.target`

`/lib/systemd/system/runlevel2.target -> multi-user.target`

`/lib/systemd/system/runlevel3.target -> multi-user.target`

`/lib/systemd/system/runlevel4.target -> multi-user.target`

`/lib/systemd/system/runlevel5.target -> graphical.target`

`/lib/systemd/system/runlevel6.target -> reboot.target`



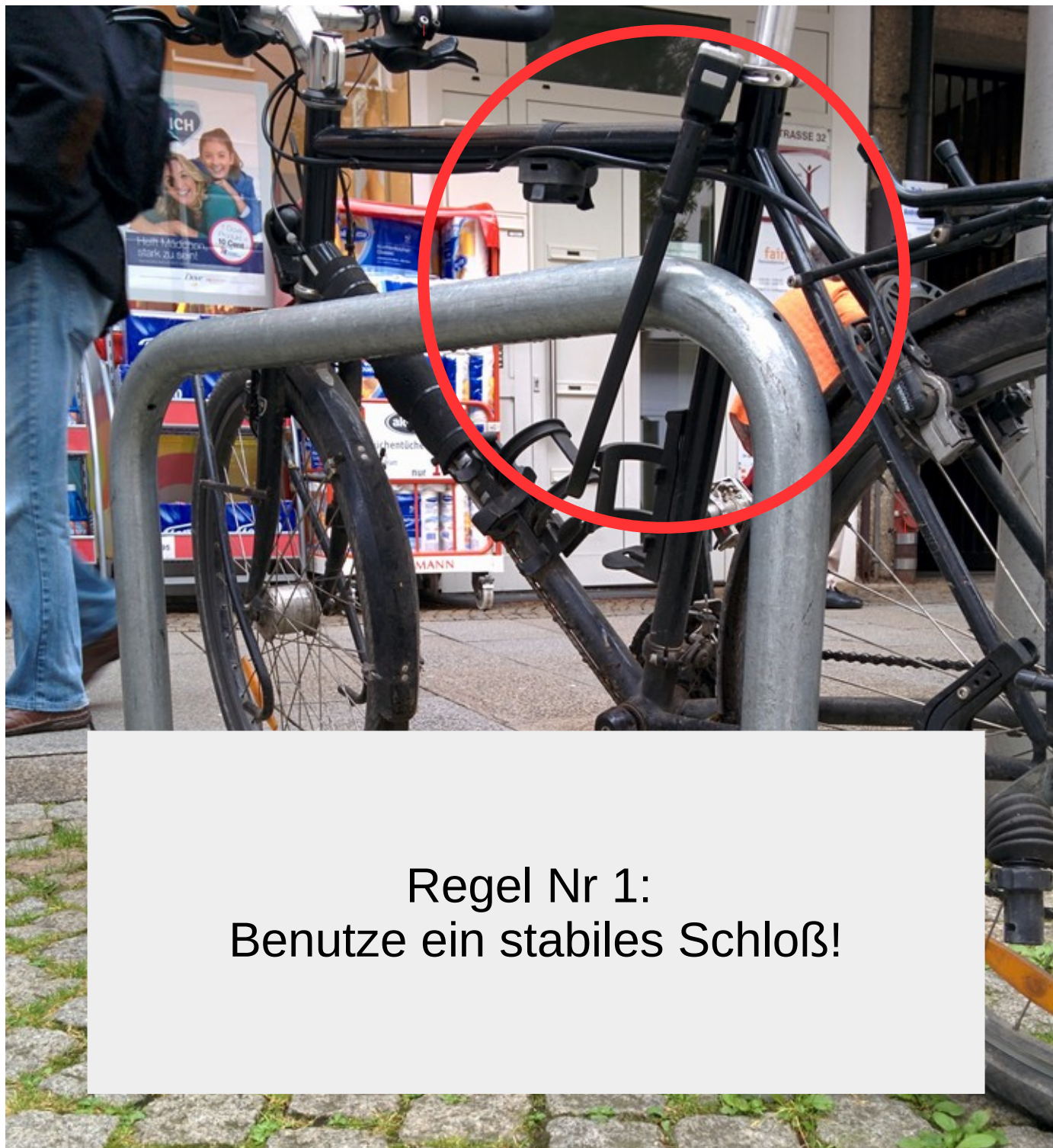
Nutzt Namespaces!

- ReadWriteDirectories=
- ReadOnlyDirectories=
- InaccessibleDirectories=
- PrivateTmp=Yes
- PrivateNetwork=Yes

```
[Service]
ExecStart=/lib/systemd/systemd-machined
BusName=org.freedesktop.machine1
CapabilityBoundingSet=CAP_KILL CAP_SYS_PTRACE [...]
WatchdogSec=1min
PrivateTmp=yes
PrivateDevices=yes
PrivateNetwork=yes
ProtectSystem=full
ProtectHome=yes
```

```
drwx----- 3 root root 4096 Mar  1 18:47
/var/tmp/systemd-private-ab..3bd13-systemd-machined.service-TSEHHU/
```

<http://www.freedesktop.org/software/systemd/man/systemd.directives.html>



Regel Nr 1:
Benutze ein stabiles Schloß!





Regel Nr 2:
Überblick nicht verlieren!



Nutzt die Sicherheits-Funktionen!

- Brot und Butter Technik ...
- Vieles geht auch mit sysvinit ... aber kaum jemand macht es!
- User, Group (su)
- RootDirectory= (chroot)
- Classic Unix Ressource Limits (ulimit)
- CapabilityBoundingSet= CAPABILITIES(7)

CapabilityBoundingSet=~CAP_SYS_PTRACE

- SELinuxContext=
- AppArmorProfile=
- SystemCallFilter=
- SystemCallArchitecture=
- RestrictAddressFamilies=
- CPUQuota=25%
- [...]



systemd – Netzwerkservices ohne Netz

```
[Unit]
Description=Journal Gateway Service
Documentation=man:systemd-journal-gatewayd(8)
Requires=systemd-journal-gatewayd.socket

[Service]
ExecStart=/lib/systemd/systemd-journal-gatewayd
User=systemd-journal-gateway
Group=systemd-journal-gateway
SupplementaryGroups=systemd-journal
PrivateTmp=yes
PrivateDevices=yes
PrivateNetwork=yes
ProtectSystem=full
ProtectHome=yes
[...]
```

```
[Unit]
Description=Journal Gateway Service Socket
Documentation=man:systemd-journal-gatewayd(8)

[Socket]
ListenStream=19531
[...]
```



systemd – resource limits

- systemd.resource-control (5)
- systemd-cgtop

Control Group	Tasks	%CPU	Memory	Input/s	Output/s
/	–	9.7	–	–	–
/machine.slice	5	1.2	–	–	–
/machine.slice/[...]mo.scope	5	1.2	–	–	–
/machine.slice/[...]mo.scope/emulator	–	0.7	–	–	–
/machine.slice/[...]mo.scope/vcpu0	–	0.2	–	–	–
/machine.slice/[...]mo.scope/vcpu1	–	0.2	–	–	–
/system.slice	41	0.0	–	–	–
/init.scope	1	–	–	–	–
/user.slice	290	–	–	–	–
/user.slice/user-1000.slice	224	–	–	–	–
/user.slice/user-1003.slice	24	–	–	–	–
/user.slice/user-117.slice	42	–	–	–	–



systemd - Das Journal

- journald ist ... system log ... „neu gedacht“
- speichert strukturierte Daten
- nützlich Clients: Cursor Support
- deckt auch die early-boot Phase ab
- rechnet Zeitangaben immer erst beim Anzeigen um
 - Ausgabe konfigurierbar
- **alles**, was ein Daemon rausschreibt, wird ins Log übernommen und passend indiziert
- ein Großteil wird implizit vom System mitgespeichert, kann also vom Dienst nicht gefälscht werden



systemd - Das Journal

Beispiel:

```
Di 2014-10-07 22:30:43.170745 CEST [s=95729ff918ce4...
  _TRANSPORT=syslog
  _PRIORITY=6
  _SYSLOG_FACILITY=4
  _SYSLOG_IDENTIFIER=sshd
  _MESSAGE=Received disconnect from
           2001:67c:670:100:1d::6: 11: disconnected by user
  _UID=999
  _GID=4001
  _COMM=sshd
  _CAP_EFFECTIVE=0
  _AUDIT_LOGINUID=999
  _SYSTEMD_OWNER_UID=999
  _SYSTEMD_SLICE=user-999.slice
  _BOOT_ID=f37591b9a4fd4418b18a27afaad0b0dd
  _MACHINE_ID=1f9112a51b3edc4d8ad934c800000007
  _HOSTNAME=dude
  _SYSLOG_PID=24591
  _PID=24591
  _AUDIT_SESSION=38504
  _SYSTEMD_CGROUP=/user.slice/user-999.slice/session-38504.scope
  _SYSTEMD_SESSION=38504
  _SYSTEMD_UNIT=session-38504.scope
  _SOURCE_REALTIME_TIMESTAMP=1412713843170745
```



journalctl – Einige nützliche Optionen

- Vollständiger Kernel Ringbuffer seit Boot:
--dmesg / -k
- tail / follow mode
-n
-f
- filter
-t <syslog identifier>
-p <priority>
-u <unit>
- output optionen
--utc
-x
-o short, short-iso, short-precise, short-monotonic,
verbose, export, json, json-pretty, json-sse, cat



systemd - Das Journal

- Es wird immer das gesamte verfügbare Journal durchsucht
--system
--user
--list-boots
--boot=<ID>
--since="2015-05-23 01:23:42" --until="2015-05-24"
- Bash: Autocompletion ist hilfreich, kann sich zur Bremse entwickeln
- Zeige alle Werte von <FIELD>
journalctl -F _COMM
journalctl _PID=20400

Container Support: --machine / --merge

```
bbu@pizza:~# journalctl -M foo -n 3
-- Logs begin at Wed 2015-03-18 13:12:27 CET, end at Thu 2015-03-19 15:42:26 CET.
Mar 18 15:42:06 foo systemd[1]: Started Cleanup of Temporary Directories.
Mar 19 15:42:26 foo systemd[1]: Starting Cleanup of Temporary Directories...
Mar 19 15:42:26 foo systemd[1]: Started Cleanup of Temporary Directories.
```



systemd - Das Journal - Neuigkeiten

- Journal Management
 - disk-usage
 - vacuum-size --vacuum-time --vacuum-size --vacuum-files
 - verify
 - sync --flush --rotate
- seamless restart
 - journald übergibt seine offenen FD nun an PID 1
 - ... und bekommt sie bei restart wieder zurück
- systemd-journal-gatewayd
 - http GET / POST (push model)
 - Cursors → remember, continue
- coredump support
- Einige Drittprojekte unterstützen nun auch systemd-journal
 - z.B. fail2ban



coredumpctl

```
bbu@calvin:~# coredumpctl
```

TIME	PID	UID	GID	SIG	PRESENT	EXE
Do 2016-06-09 23:26:59 CEST	27329	1000	1000	11		/usr/lib/icedove/icedove
Fr 2016-06-10 13:28:37 CEST	29295	1000	1000	11		/usr/lib/icedove/icedove

```
bbu@calvin:~# coredumpctl info 29295
```

PID: 29295 (icedove)

UID: 1000 (bbu)

GID: 1000 (bbu)

Signal: 11 (SEGV)

Timestamp: Fr 2016-06-10 13:26:19 CEST (5 days ago)

Command Line: icedove

Executable: /usr/lib/icedove/icedove

Control Group: /user.slice/user-1000.slice/session-11.scope

Unit: session-11.scope

Slice: user-1000.slice

Session: 11

Owner UID: 1000 (bbu)

Boot ID: e1c4dbaa4de445158a4cd072bfd6ced0

Machine ID: 8420a88fba18eab860086f8851feb560

Hostname: calvin

Message: Process 29295 (icedove) of user 1000 dumped core.

Stack trace of thread 29295:

```
#0  0x00007f0145584c09 raise (libpthread.so.0)
```

```
#1 0x00007f014148e7e1 n/a (libxul.so)
```

```
#2 0x00007f0141c4ffb1 n/a (libxul.so)
```

```
#3  0x00007f0145584d30  restore rt (libpthread.so.0)
```

#4 0x00007f00d2ba5578 n/a (n/a)



systemd - nspawn

- Eigentlich nur ein simpler Test-Container
- systemctl / machinectl / journalctl
- Guter Ersatz für chroot, ~500 Zeilen

```
bbu@pizza:~# systemctl status systemd-nspawn@fedora21
● systemd-nspawn@fedora21.service - Container fedora21
   Loaded: loaded (/usr/lib/systemd/system/systemd-nspawn@.service; enabled)
   Active: active (running) since Wed 2015-03-18 15:27:32 CET; 1 day 10h ago
     Docs: man:systemd-nspawn(1)
 Main PID: 702 (systemd-nspawn)
   Status: "Container running."
    CGroup: /system.slice/system-systemd\x2dnspawn.slice/systemd-nspawn@fedora21.service
            └─702 /usr/bin/systemd-nspawn --quiet --keep-unit --boot --link-journal=gu...
               └─776 /usr/lib/systemd/systemd
                  └─system.slice
                     └─dbus.service
                        └─1771 /bin/dbus-daemon --system --address=systemd: --nofork --nopid...
                     └─systemd-journald.service
                        └─888 /usr/lib/systemd/systemd-journald
                     └─systemd-logind.service
                        └─1766 /usr/lib/systemd/systemd-logind
                     └─console-getty.service
                        └─2121 /sbin/agetty --noclear --keep-baud console 115200 38400 9600 vt102

Mar 18 15:28:12 pizza systemd-nspawn[702]: Starting Update UTMP about System Runlevel ...
Mar 18 15:28:12 pizza systemd-nspawn[702]: [ OK ] Started Update UTMP about System ...
Mar 18 15:28:15 pizza systemd-nspawn[702]: Fedora release 21 (Twenty One)
Mar 18 15:28:15 pizza systemd-nspawn[702]: Kernel 3.18.9-200.fc21.x86_64 on an x86_64 ...
```



systemd - nspawn

- Interessante Neuigkeiten bei nspawn
 - Shares per overlayfs im Container
 - btrfs Integration: Snapshots
 - Templates
 - Flüchtige VM Images: --ephemeral
 - foo | systemd-nspawn | bar
- systemd-importd
 - gpg1 / gpg2 support
 - Import: raw, tar, qcow2
 - Export: tar, qcow2
 - Auch hier: btrfs Integration
- Achtung! Nach wie vor Baustelle!
- Docker Integration in Arbeit



systemd - autofs integration

Quelle (/etc/fstab):

```
nfssrv:/path/ /target nfs4 noauto,x-systemd.automount,defaults
```

```
root@dude:~# systemctl status ptx-work-himalia.automount
```

```
● ptx-work-himalia.automount
  Loaded: loaded (/etc/fstab)
  Active: active (running) since Mi 2015-01-21 10:53:26 CET;
          1 months 23 days ago
    Where: /ptx/work/himalia
     Docs: man:fstab(5)
           man:systemd-fstab-generator(8)
```

```
Mär 15 21:50:38 dude systemd[1]:
```

```
Got automount request for /ptx/work/himalia, triggered by 31789 (ls)
```

ab systemd-220:

- TimeoutIdleSec=60
- x-systemd.idle-timeout=60
- x-systemd.requires=



systemd - Zeitmanagement

- systemd-timesyncd Dump / Restore für Systeme ohne RTC
- timedatectl praktisches Helferlein

```
events@calvin:~$ timedatectl
Local time: Do 2015-04-02 12:56:51 CEST
    Universal time: Do 2015-04-02 10:56:51 UTC
        RTC time: Do 2015-04-02 10:56:51
            Time zone: Europe/Berlin (CEST, +0200)
                NTP enabled: no
NTP synchronized: yes
    RTC in local TZ: no
        DST active: yes
Last DST change: DST began at
                  So 2015-03-29 01:59:59 CET
                  So 2015-03-29 03:00:00 CEST
Next DST change: DST ends (the clock jumps one hour backwards) at
                  So 2015-10-25 02:59:59 CEST
                  So 2015-10-25 02:00:00 CET
```



systemd – DNS / Netzwerk

- systemd-hostnamed Einheitliche Repräsentation der Host-Details
- systemd-networkd Container Support, dhcp / dhcpv6 Server
- systemd-resolved Caches per interface, LLMNR, dnssec
- nssmyhostname /etc/hosts wird ggf. nicht mehr gebraucht
- systemd-netdev Virtual Device Management

```
bbu@pizza:~$ cat /usr/lib/systemd/network/80-container-ve.network
[Match]
Name=ve-*
Driver=veth

[Network]
Address=0.0.0.0/28
IPv4LL=yes
DHCPServer=yes
```

- IP Forwarding / Masquerading
- Application Firewall Support (via cgroup netclass)
- DHCPd
- PPPoE



systemd – link, network, netdev

- Default Network Policies machbar
- Resource Limits
- Dependencies
- Rücksicht auf externe Abhängigkeiten
z.B: DHCP -> CriticalConnection=yes

```
[Match]
MACAddress=12:34:56:78:9a:bc
Driver=brcmsmac
Path=pci-0000:02:00.0-*
Type=wlan
Virtualization=no
Host=my-laptop
Architecture=x86-64
```

```
[Link]
Name=wireless0
MTUBytes=1450
BitsPerSecond=10M
WakeOnLan=magic
MACAddress=cb:a9:87:65:43:21
```

Praxisbeispiel:
Defaults für Entwickler-WS

```
bbu@ptx:network (master=)# ls
```

```
00_ptxnet.network
01_labnet.network
02_lo.network
99_other.network
labnet.netdev
ptxnet.link
v_lab.netdev
```



Inhibitors

Unterbinden von shutdown / sleep / idle

```
root@calvin:~# systemd-inhibit --list
```

```
Who: NetworkManager (UID 0/root, PID 1142/NetworkManager)
What: sleep
Why: NetworkManager muss Netzwerke abschalten
Mode: delay
```

```
Who: Debian-gdm (UID 117/Debian-gdm, PID 1670/gnome-settings-)
What: handle-power-key:handle-suspend-key:handle-hibernate-key
Why: GNOME handling keypresses
Mode: block
```

```
Who: Debian-gdm (UID 117/Debian-gdm, PID 1670/gnome-settings-)
What: sleep
Why: GNOME needs to lock the screen
Mode: delay
```

```
3 inhibitors listed.
```



systemd – resolved / resolve

- optional
- dnssec support
- vpn support
- cache

```
bbu@ptx:~# systemd-resolve www.mailbox.org
www.mailbox.org: 80.241.60.194
```

```
-- Information acquired via protocol DNS in 3.1ms.
-- Data is authenticated: yes
```

```
bbu@ptx:~# systemd-resolve www.mailbox.org --tlsa
_443._tcp.www.mailbox.org IN TLSA 3 1 1 \
4758af6f02dfb5dc8795fa402e77a8a0486af5e85d2ca60c294476aad40b220
-- Cert. usage: Domain-issued certificate
-- Selector: SubjectPublicKeyInfo
-- Matching type: SHA-256
```

```
-- Information acquired via protocol DNS in 3.6ms.
-- Data is authenticated: yes
```



systemd - timers

- Soll kein vollständiger Cron Ersatz sein
- Basis für andere Projekte
- Praktisch für Distributionen / Upstream Packages

```
bbu@ptx:~# systemctl cat systemd-tmpfiles-clean.timer
# /lib/systemd/system/systemd-tmpfiles-clean.timer

[Unit]
Description=Daily Cleanup of Temporary Directories
Documentation=man:tmpfiles.d(5) man:systemd-tmpfiles(8)

[Timer]
OnBootSec=15min
OnUnitActiveSec=1d
```



systemd - timers

```
bbu@calvin:~# systemctl cat systemd-tmpfiles-clean.service  
# /lib/systemd/system/systemd-tmpfiles-clean.service
```

```
[Unit]  
Description=Cleanup of Temporary Directories  
Documentation=man:tmpfiles.d(5) man:systemd-tmpfiles(8)  
DefaultDependencies=no  
Conflicts=shutdown.target  
After=local-fs.target time-sync.target  
Before=shutdown.target
```

```
[Service]  
Type=oneshot  
ExecStart=/bin/systemd-tmpfiles --clean  
IOSchedulingClass=idle
```